St. Petersburg University
Graduate School of Management

Master in Business Analytics and Big Data - MiBA

STUDENT PROFILING IN ONLINE SOCIAL NETWORKS

Master's Thesis by the 2nd year students:
Marina Talianskaia
Sergei Babushkin

Research advisor:
Elvira V. Strakhovich, Associate Professor

St. Petersburg
2020

ЗАЯВЛЕНИЕ О САМОСТОЯТЕЛЬНОМ ХАРАКТЕРЕ ВЫПОЛНЕНИЯ ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЫ

Мы, Талянская Марина Андреевна и Бабушкин Сергей Сергеевич, студенты второго курса магистратуры направления «Менеджмент», заявляем, что в нашей магистерской диссертации на тему «Построение профиля студента на базе данных социальных сетей», представленной в службу обеспечения программ магистратуры для последующей передачи в государственную аттестационную комиссию для публичной защиты, не содержится элементов плагиата.

Все прямые заимствования из печатных и электронных источников, а также из защищенных ранее выпускных квалификационных работ, кандидатских и докторских диссертаций имеют соответствующие ссылки.

Нам известно содержание п. 9.7.1 Правил обучения по основным образовательным программам высшего и среднего профессионального образования в СПбГУ о том, что «ВКР выполняется индивидуально каждым студентом под руководством назначенного ему научного руководителя», и п. 51 Устава федерального государственного бюджетного образовательного учреждения высшего профессионального образования «Санкт-Петербургский государственный университет» о том, что «студент подлежит отчислению из Санкт-Петербургского университета за представление курсовой или выпускной квалификационной работы, выполненной другим лицом (лицами)».

_____ _____ (Бабушкин С.С.)

_____ _____ (Талянская М.А.)

STATEMENT ABOUT THE INDEPENDENT CHARACTER OF THE MASTER THESIS

We, Marina Talianskaia and Sergei Babushkin, second year master students, program «Management», state that our master thesis on the topic «Student Profiling in Online Social Networks», which is presented to the Master Office to be submitted to the Official Defense Committee for the public defense, does not contain any elements of plagiarism.

All direct borrowings from printed and electronic sources, as well as from master theses, PhD and doctorate theses which were defended earlier, have appropriate references.

We are aware that according to paragraph 9.7.1. of Guidelines for instruction in major curriculum programs of higher and secondary professional education at St.Petersburg University «A master thesis must be completed by each of the degree candidates individually under the supervision of his or her advisor», and according to paragraph 51 of Charter of the Federal State Institution of Higher Professional Education Saint-Petersburg State University «a student can be expelled from St. Petersburg University for submitting of the course or graduation qualification work developed by other person (persons)».

_____ _____ (S. Babushkin)

_____ _____ (M. Talianskaia)

АННОТАЦИЯ

Авторы	Талянская Марина Андреевна, Бабушкин Сергей Сергеевич
Название магистерской диссертации	«Построение профиля студента на базе данных социальных сетей»
Факультет	Высшая школа менеджмента
Направление подготовки	Management
Год	2020
Научный руководитель	Страхович Эльвира Витаутасовна
Описание цели, задач и основных результатов	Целью данной статьи является определение специфических особенностей студентов бакалавриата ВШМ СПбГУ на основе анализа данных аккаунтов студентов в социальной сети ВКонтакте для предоставления администрации Школы инсайтов на основе полученных результатов. Основными задачами исследования являются сбор данных, кластеризация студентов по методу k-means с учетом их интересов, определение специфических особенностей студентов с высокой успеваемостью и выявление особенностей интересов студентов и коммуникативных паттернов в контексте их специализации, успеваемости и курса обучения. В работе также приведен подробный анализ существующих подходов к анализу данных социальных сетей и построению профилей студентов. Методологической основой исследования являются такие методы профилирования пользователей и интеллектуального анализа данных, как описательная статистика, статистический анализ, ft-idf, кластеризация k-средних. Основными выводами являются различия в коммуникативном поведении студентов HR и маркетинга, демонстрирующих совершенно разный уровень экстраверсии; перечень специфических различий между интересами

	первокурсников и студентов последнего курса, подтверждающих выдвинутую гипотезу о влиянии обучения в ВШМ на интересы студентов. Также были определены ключевые особенности студентов с высокой успеваемостью, такие как интерес к профессиональной академической тематике и разница в количестве друзей и подписчиков в социальной сети. На основании результатов даны управленческие рекомендации, такие как внедрение дополнительных курсов, взаимодействие со студентами с помощью коротких видеороликов и конкурсы пользовательского контента для продвижения Школы, а также освещены перспективы дальнейших исследований.
Ключевые слова	построение профиля студента, анализ социальный сетей, tf-idf, кластеризация методов k-средних, статистические методы, бизнес образование

ABSTRACT

Master Students' Names	Marina Talianskaia, Sergei Babushkin
Master Thesis Title	«Student Profiling in Online Social Networks»
Faculty	Graduate School of Management
Main field of study	Management
Year	2020
Academic Advisor's Name	Elvira V. Strakhovich
Description of the goal, tasks and main results	The purpose of the paper is to define specific features of GSOM undergraduate students based on analysis of data from students' profiles from online social network Vkontakte in order to provide the GSOM administration with insights based on analysis of these features. The main objectives of the study include data collection, student clustering with k-

	means method based on their interests, defining specific features of high-performing students and revealing peculiarities in students interests and communication patterns in the context of students' concentration, academic progress and year of study. Paper also contents theoretical overview of issues concerning student profiling in social networks. The methodological basis of research is such user profiling and data mining techniques as descriptive statistics, statistical analysis, tf-idf, k-means clustering. The main findings regard the differences in communication behavior of HR and marketing students as they tend to show completely different level of extraversion; range of dissimilarities between interests of freshmen and last-year students proving the hypothesis that GSOM has an impact on student interest; the key features of well performing students such as interests to professional topics, job and academic issues and tendency to have less friends and more followers are also revealed. Managerial implications and recommendations like additional courses implementation, interaction with students via short videos and school promotion contests, and further research prospects are given.
Keywords	student profiling, social network analysis, tf-idf, k-means clustering, statistical methods, business education

Table of contents

Introduction

Nowadays the vast majority of Planet Earth's population is using information and communications technologies during their day to day activity. People are leaving digital footprints, which are then collected and analyzed. Companies are spending enormous sums of money on digital marketing in order to offer the best possible deals to their audience. In the sphere of higher education universities are on the business side, whereas enrollees and students are on the customers' side. It is no secret that universities aim to attract not only the most gifted students, but also those who would be a perfect fit for their programs. This requires an understanding of which marketing strategies to use, which, in turn, raises questions about the unique characteristics of the audience. And as youngsters (the so called Generation Z), including prospective and current university students, are even more likely to be engaged in such type of technological communication as online social networks, the data that could be extracted from their profile pages could be of a great help to complete this task.

Graduate School of Management of St. Petersburg State University (GSOM SPbU), the only Russian business school that is in the top 95 of the best European schools in the Financial Times ranking and has prestigious international accreditation AMBA and EQUIS, has reached out to us with a problem to solve. There are over 700 students enrolled in the undergraduate programs of GSOM. To no surprise, most of them are registered in various online social networks, including Russian social network Vkontakte, which is used by the School administration as one of the means of feedbacks and for urgent announcements. According to GSOM's hypothesis, an analysis of the data from students' profiles can become a source of valuable insights about their interests and preferences. Moreover, some dependencies could be traced by combining this data with the information about students' academic progress. This may help to come up with some noteworthy conclusions on which categories of students are better prepared to study at GSOM and how to allure them.

On the meeting with the company representative (Vitaly V. Mishuchkov, Director of Bachelor and Master Programmes Office) were discussed the key questions that were expected to be answered:

1. What are the typical interests of current GSOM bachelor students?
2. What is their common behavior in online social networks (how many friends and subscribers they have, what their interests are, including what communities they are following)?
3. Do they tend to take part in related to GSOM activities in online social networks (communities, conversations, etc.)?

4. Are there any features that distinguish students with better academic progress from those with lower results?

The emphasis on undergraduate students is made due to their very nature - most students go to universities directly from school, and not all of them are confident in their choice. Master students, in contrast, are older and know better what they want. Therefore, for GSOM it is of a greater importance to understand the peculiarities of bachelor students. Moreover, the sample of latter is 3-4 times higher than the one of the graduate students (whose enrollment is less and studies last 2 years instead of four), which will make the results less biased and more reliable.

The project goal, therefore, was set as following: to define specific features of GSOM undergraduate students based on analysis of data from students' profiles from online social network Vkontakte in order to provide the School's administration with insights based on analysis of these features.

In order to reach the above-mentioned project goal we have set up the following project objectives:

1. Develop a project plan.
2. Receive the information on the academic progress of students from the Client company coordinator (CC).
3. Obtain students' profiles data from the online social network Vkontakte.
4. Build an image of an undergraduate student based on available data.
5. Classify bachelor students based on the data from Vkontakte.
6. Investigate students' specific features based on different dimensions.
7. Develop proposals to GSOM administration based on analysis of these features.

The relevance of this project is based on two pillars. Firstly, on the impact of the rapid development of information technologies in terms of transparency and accessibility of data posted by people on the Internet. Secondly, on no less rapidly changing approaches in modern education, including the concept of Education as a Service, and the strive of educational organizations not to lose in this race.

The object of the study is an image of a student, based on information from online social networks, which allows classifying students according to the collected data. The subject of the study is the data of undergraduate students of the Graduate School of Management at St. Petersburg State University, which they post on online social networks, as well as information about their academic performance.

Theoretical significance of the thesis lies in the answer to the question: is it possible build a picture describing current students' profiling in online social networks? Under the practical

significance of the project is understood the opportunity to implement the model by any educational organization in Russia and even abroad (with adequate adjustments, of course).

The thesis will consist of a title page, statement on the independent nature of the work, annotation, table of contents, introduction, main part, conclusion, reference list and applications. The study itself consisted of the two big parts that are reflected in the organization of chapters and contents of the paper. The first stage included the theoretical overview of all relating topics such as the modern trends in education, approaches to social networks analysis and student profiling and data mining tools used for these purposes that are resulted in chapter one. Here the broad justification on the choice of approach and methodology is given. The second part of the study is practical implementations of the methodological issues from the first chapter. The second chapter represents two stage of the practical part of the study: primary data analysis including data collections, exploratory data analysis and attempt to build k-means based classifier on the whole set of students, the second part consists of testing the set of hypothesis within the frameworks of certain dimensions describing students (year of study, academic progress and concentration).

As the group of researchers consisted of two people, the responsibilities were distributed as follows:

- Sergey Babushkin was responsible for global and managerial aspects including communication with customer and academic adviser, setting goals, frameworks; theoretical parts regarding overview of modern educational trends, setting research in the framework of GSOM strategy and other related theoretical issues, in practical part he was responsible for clustering students based their academic progress, labeling students according to their concentrations, formulation and testing of some hypotheses regarding field of study and interests, etc. and elaboration of final recommendations for GSOM administration.
- Marina Talianskaia was responsible mainly for more specific and technical issues, thus, theoretical overview of existing approaches to profiling, social networks analysis and data mining techniques; practical issues such as data collection from VK, NLP, exploratory data analysis and testing and formulating some hypothesis mainly regarding communicative patterns and defining specific features of well-performing students; final findings and further research prospects formulation.

CHAPTER 1. THEORETICAL BACKGROUND OF STUDENT PROFILING

1.1. Prerequisites for student profiling in terms of trends in modern education

As any other industry, education faces challenges related to constantly changing environment. An educational organization is hard to imagine without its students - they are the immediate beneficiaries of the services it offers. Thus, such organizations have to compete for students by offering best possible educational services. Universities, and especially top business schools, are at the forefront of the competition for students, as they do their best to attract the brightest minds. According to the Undergraduate Education Value Chain (see Figure 1) undergraduate education is one of the pillars that create value of the business school. It is an iterative, continuous process from finding students to graduating them (and even further).

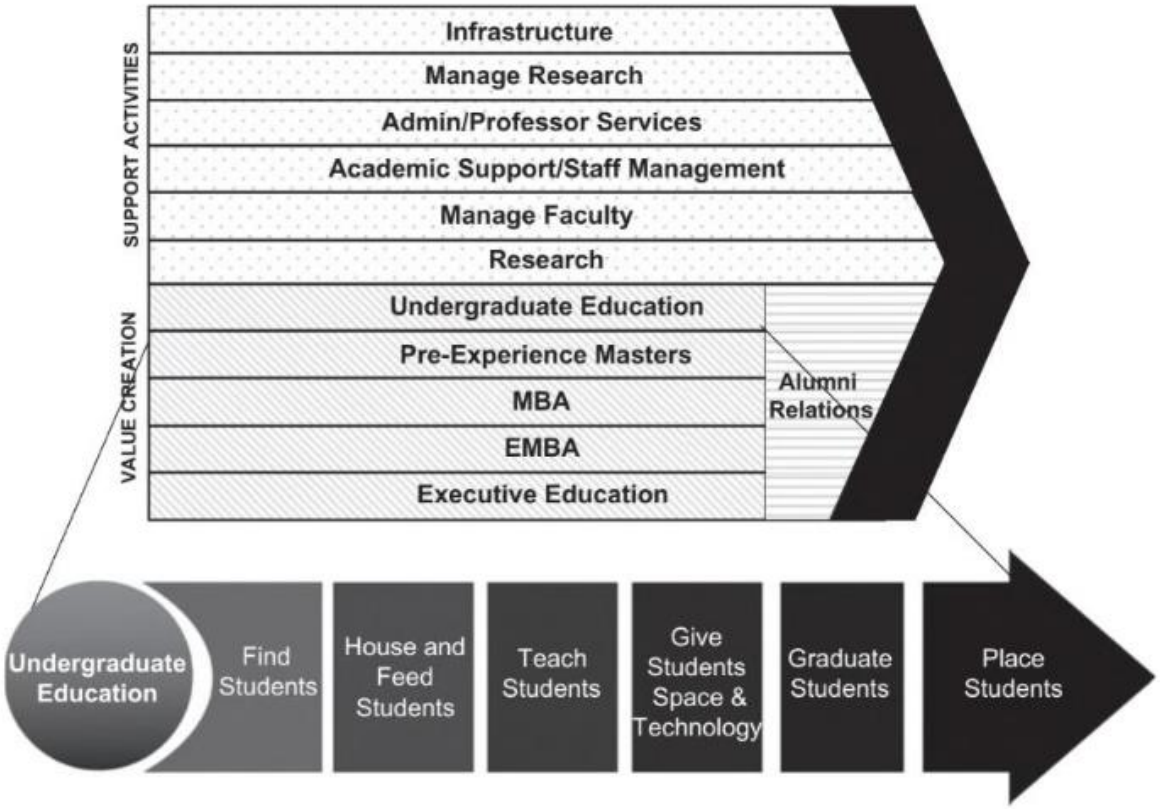


Figure 1. Undergraduate Education Value Chain
Source: (Peters, Smith, & Thomas, 2018)

Marketing is considered effective when the needs of customers are satisfied (Kotler & Keller, 2006). If we consider students to be an educational organization's customers, the conclusion suggests itself - to attract students, such organization is supposed to develop its environment so that it serves students' needs, although it is not always an easy task (Ng & Forbes, 2009). The importance of this issue is difficult to underestimate, as it is well known, the costs of attracting a new student significantly exceed the costs of retaining a current one (Howard, Boettcher, Justice, Schenk, & Rogers, 2005). In other words, retaining current students is at least as important as attracting new ones, and it is hard to imagine it without having a clear idea on

students' interests and expectations. For the retention of students are responsible three other significant stages of the Undergraduate Education Value Chain: Housing and Feeding Students, Teaching them and Giving them Space and Technology. Obviously, value creation is hard to imagine without answering the main question that arises while trying to successfully attract and retain students: how much do universities know about them, their interests and needs?

The research conducted in 2017 shows that quality of university education issues are primarily considered in terms of compliance with the requirements and standards of accrediting organizations, while the satisfaction of students themselves in the vast majority of cases is left overboard (Aikins, Adu-Oppong, & Darko, 2018). Paul Woodgates, an education expert at PA Consulting, names 3 crucial features of a university that strives to be successful and adaptive: Relevance, presuming that universities need to be relevant to the characteristics of the environment in which they operate, Excellence, which means that will succeed only those universities that are really good at what they do, and Agility, which, as expected, implies the ability not only to respond quickly to changes in context, but rather to anticipate them (Woodgates, 2018). Indeed, these features are very significant for sustaining the prosperity of a university, but it is of a great importance to offer support and flexible services to the student community to meet their rapidly changing and sometimes multidirectional needs (Fishman, Ludgate, & Tutak, 2017).

Due to increasing competition business schools face the importance of having a development strategy. In order to stay on the crest of a wave, they must not only set strategic goals and KPIs to stick to these goals, but also constantly update them based on the current situation (Thomas, 2007).

At the beginning of 2020, GSOM SPbU launched the process of updating its development strategy until 2025. According to the Strategy Commision under the GSOM Advisory Board opinion, the next generations of graduates of the School must possess such qualities as: being open-minded, ready to work in diverse and multicultural teams, able to implement brand new practices of Digital Transformation and Industry 4.0 (Graduate School of Management, 2020). As per School's Director, Olga Dergunova (Graduate School of Management, 2020), managers of the future will have to possess the following skills: manage projects, working in flexible teams and make data-driven decisions. This is what GSOM is going to teach its future students. Dergunova also points out that current undergraduate alumni are lacking practical skills which cannot be acquired through internships, so the School's administration and its corporate partners' aim is is to make internships more entertaining and to show students the true beauty of a modern industrial world, which is just as data-saturated as, for example, banking industry.

These ideas are confirmed by Andrew Crisp, founder of education research company CarringtonCrisp: "The potential market for business schools is vast, but the market for learning

and development is evolving rapidly. No longer is it dominated by classroom learning or executive retreats, but instead digital is to the fore with everything from microcredentials to digital badges to stackable certificates" (Moules, 2019).

GSOM's focus on digitalization is understandable. According to the results of monitoring of the quality of admission to undergraduate programs in Russia in 2019, traditionally popular areas of admission to fee-paying education, such as management, public administration and economics, - are gradually losing their ground. On the contrary, areas of study related to IT technologies are breaking all records of popularity (Agranovich, 2019). Business schools also have to adapt to the changing demand by increasing the number and quality of IT-related courses, so as not to lose promising students. This is the case not only for home students, but also for those who come from abroad. In 2019 the number of international applicants has grown by 20%. It is worth mentioning that St. Petersburg State University remains one of the leaders in attracting students from abroad, and management is one of the most popular "destinations" among them (Mironova, 2019).

A sound volume of research is devoted to finding features that correlate to students' academic progress (Abedi, 1991; McManus et al., 1998; Harackiewicz et al., 2002; Moruzi & Norman, 2002). Among those, for instance, are considered: students' personality factors, self-confidence, motivation levels, learning styles, prior academic results and even spatial abilities (O'Connor & Paunonen 2007). In relation to social media, research is, again, mainly focused on its influence on academic progress (Kirschner & Karpinski, 2010; Paul, J., Baker, H., & Cochran, J., 2012) or research skills of undergraduates (Nwangwa, Yonlonfoun, & Omotere, 2014).

1.2. Student profiling and social network research approaches

Social networks research approaches

As far as social networks analysis implies research over social media data produced by people in digital space, it can be regarded as interdisciplinary field of study posed somewhere in between social studies and computational sciences that entail usage of methodological and theoretical concepts of social sciences such as sociology and psychology and exact sciences (mostly mathematics and computer science).

There are two types of online profiling: reactive and non-reactive data collection (Wiedmann, Buxel, & Walsh, 2002). Non-reactive data collection implies gathering information on the users' behavior on the Internet, i.e. all the data that is collected automatically: IP address, device type, clicks and the amount of time spent on certain webpages, while by reactive data collection the input of information by a user is understood.

Social networks research is, in general, conducted within three different approaches: qualitative, quantitative, and mixed studies. The first one consists of qualitative analysis of relatively small samples of datasets and manual or poorly automated exploratory research social media data. This approach was inherited by SNA from its social sciences field and approach is mainly focused on content analysis of visual and textual materials and represents the micro-level of social interactions.

The example of such research is research conducted by a group of Brazilian social scientists that conducted qualitative study (Burns, Blumenthal, & Sitter, 2018) of Twitter posts of Brazilian college students and revealed that Twitter messages' content and comments fully represent the existing tensions and relationship within college groups in their everyday offline behavior. Another example of qualitative approach to social network analysis (SNA) is represented by content analysis of Vkontakte profiles (Tcheremisova, 2016) that implies careful investigation of 100 user profiles within 12 different dimensions characterizing each profile in terms of interests and communication patterns. Both these papers represent the classical qualitative methods towards digital data of social networks and make certain conclusion of specific people and features within the social network context.

Another approach to SNA is a set of quantitative methods that is commonly used for the other types of numerical data such as financial market data or other numerical features. The typical implication of such research is revealing general trends and discovering features at the macro-level not digging into specific characteristics of each profile. As a part of this approach, some other types of data such as texts, images and rank values are converted into their numerical representation in order to further apply statistical analysis and other mathematical analytical tools.

The example of numerical research in social networks is analysis of information spreading within Facebook (Khrapov & Stolbova, 2019) where authors performed a simulation of histograms of the number of comments to popular posts within a certain time interval, using the extended epidemiological model SIR and systems of differential equations. The proposed model was used to track the behavioral dynamics of active users regarding their commenting patterns and social-demographic background.

A particular application of numerical approach to SNA is social graph analysis. As any social network represents an instance of a network of interconnected object, classical tools of graph analysis find an interesting application within social network context. Traditional social networks graph analysis includes the following features: the depth, the width, the level of graph clusterization, graph density, etc. The application of graph analysis to SNA is, for example, the work of group of Canadian researchers who created a model being able to identify the characteristic of elements of graph and their interrelations basing on only a part of certain social

graph solving the problem of computational difficulty of analyzing the whole social network as a single graph (Dougnon, Fournier-Viger, & Nkambou, 2015). In general, social graph analysis is used to identify the key nodes that are actually the most influential people within certain social groups within networks or to investigate interconnections between different groups.

Mixed approach to social networks analysis implies the combination of qualitative and quantitative methods. Social networks tend to be suitable for both qualitative and quantitative tools as they represent since they embody both the structure as well as content of social relations (Yousefi Nooraie & al., 2020). Classical approach proposes sequential or parallel implementation of qualitative and quantitative research methods for SNA. For example, the preliminary quantitative research can be conducted in order to identify the suitable profiles for further qualitative research as deep interviews and focus groups such for deeper investigation of revealed features of for 'tuning' quantitative methods in case of lack of reliable data or its poor structure.

Approaches to student profiling

As Cambridge Business Dictionary defines it, profiling means 'the activity of collecting important and useful details about someone or something' (Cambridge University Press, n.d.). In business context, it is mainly related to 'customer profiling', that includes collection and systematisation of set of features describing a real or potential consumer of certain product. In information science, this term refers to the process of building and implementation of user profiles created by computer-based data analysis tools. In general, the profiling process towards people includes the following steps (Hawtin & Percy-Smith, 2007):

1) Preliminary grounding during which the researcher defines the main sources and purposes of profiling process.
2) Data collection that implies understanding of what data is necessary, how and where can we get it and what is desired dataset structure.
3) Data preparation including eliminating noise, sampling, data structure changes and data type conversion.
4) Data analysis with implication of certain statistical methods and computational algorithms in compliance with goals defined on the first stage.
5) Interpretation as making conclusions on the obtained results and estimation of their relevance within the context of research.
6) Results application that means incorporating the findings into business processes or research results definition.

This algorithm can be applied to student profiling as well but for deeper and more careful analysis we should define the specific features of student profiling. Student profiling is divided in

two large research areas: learning-style analytics and profiling of students based on their major discipline or another common feature.

The first group of studies covers the wide range of questions regarding particular behavioral patterns in learning process. Most of papers in this field are devoted to behaviour of online courses student. The majority of studies here are based on analysis of logs of different educational systems allowing to know who, how, when and how long have been learning certain content or solving certain task. After analysis of these logs with statistical methods and data mining tools. All these papers are based on the assumption that there are a set of so-called 'learning styles' describing student's ability to perceive, process and remember learning materials and these styles can be traced over logging on students' actions and attributed to certain demographic features such as age, gender, cultural experience, educational background, etc. One of the earliest examples of such paper is research of group of Asian mathematicians who proposed a 'fuzzy model', a model of adaptive learning system that logs data on online course performance, computes certain statistical patterns and then attributes student to a certain learning track (Xu, Wang, & Su, 2002). Another interesting model was proposed by S. Shorko and S.K. Jha: they developed a calculative associative logical memorable (CALM) algorithm is proposed which integrates calculative thinking, association, logical perception, and memorizing features (Jha & Shorko, 2018).

Recent studies in this field are focused on student profiling with machine learning methods to optimize the learning experience of a student. For example, in the series of papers by Tempelaar and his colleagues (Tempelaar, Rienties, Mittelmeier, & Nguyen, 2018) student clustering based on dispositioning and time logs data is used to define the profile of at-risk student. However, the most popular approach to learning profiling is Decision tree algorithm. It is proposed by several studies that aim to develop the optimal adaptive e-learning algorithm for personalized learning (Strang, 2008; Fok et al., 2014; Desai, Shah, & Dhodi, 2016).

The second large groups of studies is dedicated to gaining special features of students that share similar features, the most common features are high level of academic progress and/or certain field of study. The main approach in this studies is to determine a picture of successful student based on statistical analysis of demographic features. A good example of such approach is illustrated in paper devoted to defining specific features of well-performing students from business programs at an Australian university. The study revealed that typical successful profile is determined by young age of the student and high level of English proficiency and does not depend on the nature of previous degree or work experience (Eddey & Baumann, 2009).

Similar study was conducted over English students majoring in telecommunications, however, the research was based not on demographic features but on psychological characteristics

obtained from specially designed questionnaire. The study revealed that well-performing students are characterized self-sufficient, with low level of compulsivity and having high level of expediency in their daily lives (Biner & Dean, 1997). This paper illustrated the psychological and interview/survey based approach to student profiling that is more applicable for psychological and sociological studies that are more focused on qualitative research.

Another approach to student profiling is proposed in the paper devoted to students future employment. Here the authors suggest to elaborate a skills map for each student and classify them basing on the level of development of certain skills and competences in order to compensate weak parts of their personality and ensure future employment of the student. An interesting feature here is that education process and obtained competences are perceived as saleable product, thus, some customer profiling tools are implemented namely estimation of education background, 'customer journey map' construction and estimation of current social status (Dumbre & Ingawale, 2016). Here, as an instance of non traditional approach, we should also mention the approach that embeds extracurricular activities of students and combines qualitative and quantitative tools that can be illustrated on the example of the paper that main objective was to develop a profile of student co-curricular activities in uniform units and test several variables describing characteristics for student co-curricular profile (Jamil & Shaharanee, 2017). As the processing techniques, clustering and data visualizations were applied.

Finally, very classic and the most common approach to student profiling in recent years is based on academic progress and academic background analysis but with usage of machine-learning techniques. The distinctive feature here that the classes or profiles are not predetermined as they do not build an image of successful student but apply it only as one of dimensions. An example of this approach is the research on student profiling on academic performance using cluster analysis provided by a group of Turkish researchers (Darcan & Badur, 2012).

To sum up, there are three main approaches to social network analysis that are differed by the methods applied and the choice of approach is determined by the nature of research and the desired level of abstraction of final results. For micro-level interaction research, qualitative methods is better choice, whereas the general and macro-level trends are discovered with quantitative tools and the desire to get the middle-level picture or test the overall trends with some examples impels the researcher to apply mixed approach.

As for the approaches for user profiling, they are divided into two big groups: learning profile analytics and personal features profiling in the context of academic major or progress. In general, overall algorithm to people profiling might be applied to student profiling but at every stage the learning and educational context of the study should be kept in mind. For example, for

learning analytics approach, the data collection and data mining stages in most of cases should include data regarding the performance of student and the goal should be set as enabling better performance of learning system. For ‘common feature’ profiling approaches, such issue as the concentration or field of study might be a more important issue for building an image of a student.

1.3. Data mining techniques in social networks analysis

Machine learning (ML) in its classic definition is regarded as a set of computer algorithms that improve automatically through experience building a mathematical model based on sample data, known as “training data”, in order to make predictions or decisions without being explicitly programmed to do so (Mitchell, 1997; Bishop, 2006). The very close concept to ML is ‘Data mining’ that is perceived as a mix of machine learning, statistics and database analysis techniques used to discover hidden patterns in large datasets (Cabena, Hadjinian, Stadler, Verhees, & Zanasi, 1998). Machine learning and data mining techniques application to analysis of social media and social networks implies the various sets of methods. The most common ones were described in exploratory research by group of international academics from Robert Gorton’s University (Stahl, Gaber, & Adedoyin-Olowe, 2014) where all the tools are divided into three big groups: tools for data processing, data analysis and data interpretation. Basing on the nature of data in social networks that is mainly represented by images, texts, numerical and relational attributes of objects, we can also subdivide all data mining techniques towards social networks onto two even larger categories: those related to certain data processing and analysis of different types of data (that is actually the combination of the first two groups from previous classification) and those related to making certain classifications, grouping and predictions based on previously revealed features. The methods of both categories can be applied both simultaneously and separately and the most common applications imply sequent or iterative usage.

Natural language processing methods

As the data in social networks can be of various nature, it is reasonable to separately consider different applications of various techniques for various data types (Richthammer, Netter, Riesner, Saenger, & Pernul, 2014). Let us begin with text analysis tools. One of the first and the simplest natural language processing (NLP) tools is Bag of words that represents text as set of words (‘bag’) regardless of word order and punctuation issues. The frequency of each word within the document is counted and the ‘bag’ containing the words with their frequencies is regarded as features of certain document and when then used for attributing document to a specific category. The main disadvantage of this approach that it does not take into account the word order, so, it cannot help to properly distinguish the context in which specific term is used. Another

disadvantage is that it poorly deals with the most common words in general and within a certain topic and the 'stop words' list or any other additional preprocessing techniques are required. The most common application of Bag-of-words is spam detection and email filtering. In terms of social networks analysis, its implementation can be illustrated on the example of research Philippine academics who developed model that identifies local disasters analyzing 'hashtags' (analogue of key words for messages) in Twitter (Camama, 2019). Authors apply Bag-of-words to Twitter messages and hashtags and then use Naïve Bayes algorithm to classify if the message contain any report regarding local catastrophes or not. In this case, the Bag-of-word is appropriate decision as it is quick and effortless to implement and Twitter messages text specifics (short, commonly poorly structured pieces of text with little punctuation and other extra marks) make it possible to pay less attention to the text structure.

The further development of Bag-of-words approach to text mining is tf-idf method. Abbreviation 'tf-idf' stands for 'term frequency–inverse document frequency' that implies a numerical statistic representing the relative value of document within a collection of documents called 'corpus'. As it helps to define specific terms for each document and helps to distinct each document within the context of other documents, it is quite common tool in user modeling, information retrieval and other fields requiring text analysis for quick search and classification. In general form the tf-idf is calculated as follows (Robertson, 2004):

1) We calculate term frequency (tf) (1):

$$(1) \quad tf(t,d) = \frac{n_t}{\Sigma_k n_k}$$

Where n_t is a total count of a term (t, specific word) within a document (d, a collection of words) and the sum in denominator represents the total word count in a document.

2) Calculate inverse document frequency (2):

$$(2) \quad idf(t,D) = log\frac{|D|}{|\{d_i \in D | t \in d_i\}|}$$

Where:

|D| is a number of documents in a corpus (set of documents),

the number of documents in corpus D that contains a term (t) (when $n_t \neq 0$)

The logarithm is used to avoid extreme values of idf and normalize it.

3) Finally, we take a product and compute the final value of tf-idf (3):

$$(3) \quad tf\_idf(t,d,D) = tf(t,d) \times idf(t,D)$$

There are several modifications of this method that imply normalized values for both parts of the original formula. For example, tf is computed not only with the raw terms count, but also with 'boolean frequences', adjusted term frequencies in relation to document length, logarithmic scale, etc., idf uses smooth, max and probabilistic and other normalizations as well. The application of tf-idf to social media analysis can be illustrated on the example of hate speech detection model proposed by a group of Asian academics (Ruwandika & Weerasinghe, 2018). They used tf-idf and BoW approaches with several classification models to a set of English-language messages and publications in social media to identify documents containing any kind of insulting words, and, in comparison to several BoW models, tf-idf performed better in terms of accuracy and F1-score models.

Going further to more complicated tools, we should consider a set of so-called 'embedding' approaches in NLP. The key idea of embeddings is transforming terms of collocations into mathematical vectors of real numbers. There are several ways to map words to vector space including dimensionality reduction on word co-occurrence matrix, context-related representations, probabilistic models, etc. For example, context-related models include representing each word from a set of words (dictionary) to an Euclidean space attributing semantic features of different words as geometric distances into Euclidean space. Traditional approach includes creating a dictionary of words with one-hot encoding, where one word is set as it own dimension and then different dimensionality reduction techniques such as neural networks with hidden layers are applied, that enables the model to predict the next word basing on the set of previous ones. So, finally we get vectors representing each word within corpus and the vectors with close semantics or contexts are placed close to each other. The most common manifestation of embedding methods is word2vec algorithm, that can use two ways to construct a distributed representation of words: Continuous-bag-of-words (CBOW) and skip-grams. CBOW model predicts the upcoming word taking into account the set of close words within 'window' regardless of the distance between current term and the others. Skip-gram model works pretty the same, but it gives the closest words higher value than the furthest ones (Goldberg & Levy, 2014). The main advantage of this approach as it lets to define the semantic and context of words, but it poorly deals with polysemous words like 'fine' as a punishment or as positive adjective. The example of application of word2vec to social networks user profiling is given in the paper (Alekseev & S., 2017) of Russian researchers who created a model predicting profile attribution to certain demographic group basing on dataset of 868 thousands of user profiles from Russian social network Odnoklassniki and information on the interrelations, preferences ('likes') and publications.

Image processing approaches

Proceeding the overview of data mining tools within SNA context, we should address to image processing methodologies as the graphical data constitutes a large part of all social media data. In the context of social media and social networks, three main interconnected groups of tasks of image analysis can be defined: image segmentation, object detection and object recognition. It is important to mention that, in contrast to NLP-methods, in most cases all these three issues come together.

The main purpose of image segmentation is partitioning digital image onto a set of segments aiming to simplify further representation of image and making it more appropriate for further analysis. In general, it implies assigning label to a certain part of image (in most cases – pixel) in the way that pixels of the same group share similar characteristics. The typical result of this process is a defined set of contours that describe the image. The most common method is the thresholding that involves replacing each pixel in the image either by the black (if its value is less than predefined constant) or by a white one (if its value is greater) for grayscale images and to RGB-labels for colorful pictures (Mehmet & Bulent, 2004). In the context of social networks analysis, it is usually used as part of object detection process. For example, the group of Korean researches applied it to detect woman breast images in local social media to avoid private and porn images distribution (Joo, 2014). Another interesting research applying image segmentation is flood detection study in social media graphical content that detected changes in water level as changes in the edges between water and the other objects (Chaudhary, 2019).

Continuing the topic of image analysis, we should turn to object detection and object recognition issues that in framework of social media analysis is, first of all, connected with face detection into photos. The most popular tool is Viola–Jones algorithm that was invented namely for face detection purposes. The algorithm belongs to machine learning methods and consists of four stages. First, the Haar feature selection is used to detect specific black and white zones on the face, then it composes the integral image of the face evaluating a rectangular feature of face and learning algorithms that use a modified AdaBoost classifier for electing the best feature and training algorithm at the same time. Finally, the cascade architecture of classifiers is used and each stage of processing implies its own strong classifier with its own limited set of features (Viola & Jones, 2001). Scale-invariant feature transform (SIFT) is another common method for object detection. The main stages here are Scale-invariant feature detection that applies Lowe's method to image feature generation by transforming image into a set of image vectors invariant to any kind of picture changes, Feature matching and indexing applying best-bin-first search algorithms, Cluster identification by Hough transform voting, Model verification by linear least squares and

outliers detection and removal (Lowe, 1999). The third popular technique is histogram of oriented gradients (HOG) that is quite similar to scale-invariant feature transform and edge orientation histograms. This technique is based on counting the number of gradient directions in local areas of the image that is calculated on a dense grid of evenly distributed cells and uses normalization of overlapping local contrast to increase accuracy (Dalal & Triggs, 2005). The example of object detection within social media analysis is a paper of group of Indian researches who implemented their own modification of the mentioned above algorithms called UP3 for face detection and further user profiling based on their profiles photos. The key idea of the developed algorithm is to compute Logarithm of Determinant of Euclidean Distance Matrix (LDEDM) in Relative-Distance based on the Bounded Boxes of the face objects (Vasanthakumar, 2015).

Classification tools

The main aim of user social network profiling is attributing each user to a certain category or just 'labeling' them. ML-techniques that help to fulfill this task are in general divided into two large groups: supervised methods that use predefined classes and labels and unsupervised that perform the classification task without any predefined parameters.

The simplest manifestations of supervised learning models are linear and logistic regression models that apply a statistical regression method as a classifier. The linear regression is supposed as a weaker tool, it consists of calculating linear equation that better describes the difference between classes. It is a useful tool for classifier namely numeric values and quite handy and quick to implement, however, it cannot be used when the data dependencies are described with non-linear equations, moreover, it is prone to overfitting. The logistic regression is more powerful classifier and pretty similar to its linear counterpart, the only difference is that it uses logistic (sigmoid) function to describe feature dependencies and define the borders of two classes, the key estimators here are coefficients that are calculated with the help of maximum-likelihood estimator. The log-regression model is a simple and relatively powerful algorithm for binary classification that does not require any feature scaling and hyperparameters tuning, at the same time it is vulnerable to poor data representation and irrelevant and over-correlated data features. Due to its efficiency and ease of implementation, log-regression is commonly used in social network analysis. The examples are two papers that deal with the issue of false objects detecting in social media. The first research introduces the model classifying fake news or posts in Facebook basing on who had liked it and defining specific features of 'likers' profiles (Tacchini, Ballarin, Della Vedova, Moret, & de Alfaro, 2017). Another study proposes log-regression-based model to define fake user profiles in LinkedIn, the main features used to build model are user-generated text, such as name, email address, company or university; including both frequencies of patterns within the

cluster and comparison of text frequencies across the entire user base (Xiao, Freeman, & Hwa, 2015).

The next ML-method in terms of complexity is Naïve Bayes classifier. In general, it is conditional probability model that represents its instance as vector that is described as a number of its features and then assigns the probability of each instance to belong to a certain class and then applying Bayes' theorem to construct a whole probability model. The main advantages of this method are that it is scalable and can be applied to relatively large datasets, it is insensitive to irrelevant features and performs well on large set of dimensions and even for multiclass predictions. At the same time, it fails to deal with cases with huge differences in training and overall population data and quite commonly the condition of independence of features is not met. The implantation of Naïve Bayes algorithm for social media analysis is quite common, for example, it is used for spam/fake accounts detection in Twitter and Facebook, however, the model id based on 14 pre-defined features describing one of the classes (Ahmed & Abulaish, 2013). It also applied for such a wide topic as a sentiment analysis after converting words to vectors with different (mainly with uni and n-gram feature extraction models) tools as it is presented in the paper devoted to Indian parliamentary election topic analysis via Twitter messages (Anjaria & Guddeti, 2014).

It is also quite common to apply support-vector machines (SVMs) technique when it comes to extreme binary classification scenarios. SVM model represents each observation as a point into (p-dimensional vector) a space and the classes are divided by a wide range of empty space. This classifier performs very well when there are two easily separated classes and the dataset has many (almost infinite) dimensions. On the other hand, it is quite slow and is not an option if the classes are difficult to distinguish and require paying much attention to proper hyperparameters and kernel functions setting. The SVMs are quite popular for any kind of text mining tasks, for instance, as the second option in the above mentioned research regarding sentiment analysis based on Twitter publications.

More complex decision for SNA and users classification is proposed by k-nearest neighbors classifier (k-NN). The algorithm includes representing each observation as a set of vectors in multidimensional space and then attributing each unlabeled observation to certain class depending on labeling of its k neighbors. As a metric of 'closeness' the most common choice is Euclidean distance, however, for social networks classifications especially related with text mining, the Hamming distance is used. K-NN finds its application for relatively small datasets with not very high dimensionality and applicable for multiclass classification, however, it requires additional data pre-processing such as scaling, dimensionality reduction, balancing, etc. The example of its implementation is a paper where researchers applied this classifier in the model for detecting false

authorship of messages posted by compromised accounts on Twitter. In this research, namely overlap (Hamming) measure was chosen to deal with text classification task (Barbon, Igawa, & Bogaz Zarpela, 2017).

The last but not least, group of methods of supervised learning is artificial neural networks that is definitely the most popular in the last few years due to its high flexibility and variability. In general, it supposes the set of interconnected units (neurons) that can process and transmit information 'learning' as a real animal or human brain. There are a plenty of types of artificial neural networks that are used for profiling purposes, they vary from each other depending on the number of layers, nodes, structure of networks, sequence of layers, etc. For example, a group of Russian researchers applied convolutional neural networks (CNN) for user profiling based on their reviews on drugs in social media, and managed to predict demographic features of users (age, gender) with accuracy over 90% (Tutubalina & Nikolenko, 2018). Another study proposes a dynamic profile of the user (Cui, Agrawal, & Ramnath, 2020) based on the text and frequency of publications in Twitter using hybrid gated recurrent neural network (GRNN)-based model for rich contextual learning.

1.4. Summary of Chapter 1

Undergraduate education is one of the elements of the Undergraduate Education Value Chain. It involves certain iterative steps, that fall into two main categories: attracting and retaining students. To successfully cope with these tasks, an educational organization must have a clear image of its student.

At the beginning of 2020, GSOM SPbU launched the process of updating its development strategy. The School is going to focus on preparing data-driven managers, who will be able to manage projects, working in flexible teams. Such intentions are well-grounded, because, according to expert opinion and the actual results of admission to universities in 2019, the focus of applicants is also shifting to the IT-related side.

Regarding existing research devoted to the analysis of students in social media, it is mainly focused on finding features that influence students' academic progress or their research abilities.

Finally, after the careful consideration of the existing methods of data mining in social networks analysis and user profiling, we decided to apply some natural language processing tools to define specific textual features of students, some classic statistical methods for some numerical values namely the frequencies, central tendencies, etc. and k-means clustering for classification tasks. The main arguments for refusal to apply other methods are:

1) Refusal of image processing was explained by the privacy limitations of social network and personal user settings as, in most cases, they do not allow to obtain any image content

except the main profile image (avatar) and, after random check, we concluded that profile photos cannot give any valuable inputs for the tasks of the research.

2) We decided to abstain to apply supervised method of machine learning algorithms as they require predefined classes that we were not able to distinguish and indicate. The existing labels such as year of study or academic progress could be used as a labels for prediction tasks, but the purpose of the study was not to predict if a certain profile class but to define if there are any valuable insights concerning existing classes or if they are any talent groups of students that are not unified by any formal criteria. That is why the preference was given to unsupervised methods.
3) Decision to avoid using artificial neural networks classifiers is explained by the size of the dataset and the computational complexity in relation to expected results. The preliminary investigation showed that ANNs are good for large datasets (thousands of observations) and defining general features such as age, gender and level of education and is less applicable for more precise issues as our dataset is quite homogeneous in this terms. As for the computational complexity, the example of Facebook user preferences profiling study proves that not very complex ANN classifiers in general perform less successfully in comparison to more simple and less 'buzzy' methods (Jiamthapthaksin & Aung, 2017).
4) The preference of more 'statistical' tf-idf approach rather than 'semantic' word embeddings is explained by the size and nature of documents analyzed: each student is represented by the set of poorly connected short text describing his or her subscriptions and unified in one document. In this case, word2vec and the similar probabilistics methods based on the analysis of frequencies of neighbouring words are not very useful as there is very high probability that words from very different sources would stay close to each other and distort the final results. At the same time, tf-idf allows to distinct specific and unique words to each student and compare them with others in contradiction to Bag-of-Words in this case will give only the most frequent ones ignoring any peculiarities.

CHAPTER 2. APPLYING CHOSEN METHODS FOR STUDENT PROFILING IN ONLINE SOCIAL NETWORKS

2.1. Data collection and first steps of data processing

Data collection and dataset characteristics

The data for the project comes from two sources. The first of them was received directly from the School representative and consists of a table with all marks received by present 1st - 4th years students for passing courses during the last 4 years, totalling of over 25 thousands rows. Besides marks, the table contains information on the group and curriculum, according to which a student is studying.

Another dataset consists of the data obtained from personal profiles of 685 GSOM students on Vkontakte and represent 91,8% (varies from 90% to 96% depending on the year of studies with less profiles belonging to elder students) of bachelor's degree students of Graduate School of Management on March 2020. The profiles were attributed to certain students basing on the following criteria:

- First and/or last name coincide with the list given by GSOM representative.
- Students have direct linkage with GSOM and/or SPbU on their social network page.
- Students are members of official VK conversations with GSOM administration.

Any two of these criteria were considered enough to identify profile as a GSOM student profile, if profile with high degree of probability belonged to GSOM student but had a nickname instead of a real name (only 0,4%), it was considered as relevant one.

The main data collected from student profile pages included:

- ID in social network;
- First and last name as given on the page;
- Demographic information: age, gender, home city;
- Education information: secondary school, university, additional education;
- Career information: company, position;
- Personal preferences: favorite books, music, interests;
- Other social networks info: Twitter, Facebook, etc.;
- Most/last used device/browser;
- Communities and subscriptions: name, description, role in community;
- Personal views: life credo, values, attitude to harmful habits, etc.;
- Publications on the 'wall' (specific platform allowing owner of the profile and other people, if not forbidden by the owner, post text messages, music and graphical content);
- Some further information if available.

To extract such data from Vkontakte, the VK API was used.[1] API (or application programming interface) is a tool with the use of which an information from website (in our case - Vkontakte) databases could be extracted by sending http-requests to the server. For obtaining different kinds of information different methods are to be used. For instance, to get information from students' profiles, the method users.get was used[2], whereas the information on students' groups was extracted by the method groups.get[3]. The interaction with VK APIs was performed via methods available in open-sources libraries in Python programming language

It should be mentioned that due to the following reasons data collection from social networks has the certain difficulties and limitations:

1) Data privacy: in most cases, access to user data is allowed only for registered and authorized users network participants, which requires support for emulation of the user sessions using special tools (in our case, our personal accounts authorization and special build-in application were applied). Moreover, the privacy settings of each user can be different and quite strict that may cause limitations for overall data collection.

2) Technical limitations on quantity of collected data: social network programming interfaces (APIs) have limited functionality that causes constraints on the number of requests per second or a nature of data we can request.

For example, the information on what certain user 'liked' (tagged with 'I like it' sign) can be get only as a boolean answer if the certain object was liked by a certain user or not, that in case if we would like to get the list of 'liked' posts for 600 hundred of students with around 100+ subscriptions with 200+ posts on average within last 3 months and limitations of only three requests per second that resulted. in the most optimistic realizations, in 15+ days of constant work of algorithm.

3) Poor data structure: due to the nature of social network APIs, the requests can be send from and for different objects and the results can be got only in a certain form, that, together with privacy limitations, require further data pre-processing and structuring steps and create additional limitations to data obtained.

So, the main features of dataset that we used for further analysis are:

1) Multiple data sources that require further integration and association of data gathered from them, in some cases, in semi-manual mode.

2) Poor data structure due to multiple source of different nature (MS Excel tables, complex and mixed lists, arrays, etc. of data collected via APIs)

[1] Learning API. VK 2006-2020. URL: https://vk.com/dev/first_guide
[2] users.get. VK 2006-2020. URL: https://vk.com/dev/users.get
[3] groups.get. VK 2006-2020. URL: https://vk.com/dev/groups.get

3) Wide range of data types: numerical, ‘relational’, text, date-time data, etc.

4) Large differences in completeness of data (some accounts are almost empty whereas the others contain dozens of filled fields and hundreds of publications) due to different privacy settings and personal attitude to data completion of users.

Basic dataset insights

Collected data includes 260 male and 425 female profiles representing 90,91% and 93% of male and female population of students respectively, that indicates that female students slightly more tend to communicate via VKontakte under recognizable names. It was very delightful finding that only 4 of 685 profiles (less than 1%) belonged to obvious pseudonyms (like ‘Time Machine’) and the absolute majority of students do not hide their identity behind fake pages. This parameter is much more higher (around 10 times) the estimated quantity of fake users in social networks (Fire, Kagan, Elyashar, & Elovici, 2014).

107 of 685 student profiles (15,6%) were fully closed for public (have limited number of people having access to personal information). Notably that male students tend to hide personal information by 66% more in comparison with female students: only 12,9% of female profiles had some privacy limitations in contrast to 20% among male population.

14.3% of students do not give any information concerning their date of birth and 57,5% hide the year of their birth (and the age as well) or specify deliberately false information. The rest of the students are evenly distributed around 1998-2002 years of birth with median in 2001 and the average age 19,6 years. Notably that in contrast to private settings, male students do not tend to indicate their real age less than females, for both gender groups this value is around 72%.

The issue of correct or full date of birth indication lies close to the problem of completeness of other fields. 545 of 685 (79,5%) profiles contented any information regarding owner’s occupation. To no surprise, around 91,6% (499 profiles) of students specified any forms of mentioning of Saint Petersburg State University. The other references were to work (29 profiles or 5,1% of those who filled the field) and school (18 profiles or 3,3%). Most of those who stated work as the main occupation (24 of 28 students) are on the last year of their studies and this can easily be explained by fact that data was collected during their last semester and it is quite natural for future graduates to start their career slightly before graduation. All the companies stated as the place of work turned to be real and recognizable companies, but some students (3 profiles) tend to indicate as a job some temporary occupations like volunteering or certain event organization. All the students who stated schools as their main occupation were freshman and this can be explained by just lack of intention to update profile information.

486 of 515 (or 94,3%) of students who stated their hometown chose Saint Petersburg and the other cities indicated included in 93% of cases the previous location where students were born or got their secondary education. Here should be mentioned that the majority of those who stated other location (79,3%) were people from near abroad or ethnic regions within the country and that may point to their latent reluctance to perceive Saint Petersburg as their main location and disinclination to loose connection with their home regions.

There is a clear tendency to leave unfilled of fields related to non-institutional or materialized concepts such as personal views, languages, attitude to alcohol, smoking, etc. From 89,9% to 91,3% (depending on the field) of all students tend to leave such fields empty. This pattern seem does not depending on gender as 23 males and 38 females respectively who signed such kind of personal information reflect the initial sex distribution within the dataset (8,8-9,0% of students of both genders are ready to share such data). The revealed pattern could be explained via several hypothesis. In our opinion, the possible explanation is that the majority of these options were introduced by social network much later than the fields connected with locations and organizations when users had already formed a behavioral pattern to stick to more materialized issues and continued to update only respective fields of their profiles. Another possible explanation is that locations and institutions can be more efficient way to form community and to build connections based on these common features rather than based on the more abstract concept personal views. Unfortunately, all the suggestions and assumptions lie beyond the scope of research cannot be checked with the available data and may be a subject of other research of more psychological or sociological nature.

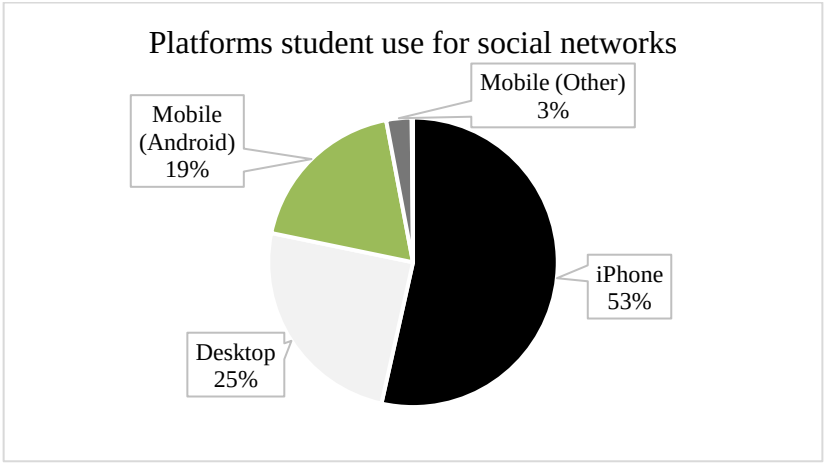
Analyzing those profiles, who indicated their personal information, we can state that, for example, the most common foreign language is English (100% of profiles) that is no surprise as this competence is required for the majority of economics and business faculties. The other common languages (except Russian) are German that was indicated in 22,5% of user profiles and French (14,5%), some people from ethnic regions or abroad stated their local languages such as Kazakh or Belarusian (27,3% in total). Note that people can indicate several languages, so, the total can not to sum up to 100%. As for the personal views, the most common life values were indicated as ‘self-development’ (43%) and ‘world improvement’ (38%) that pretty clear reflects the expected life orientation of business school students. The most frequent religious views indicated by the students were secularism (68%) and different branches of Christianity (23%). However, it should be outlined again that the sample of those who indicated certain preferences is extremely small and thus the sample is unrepresentative.

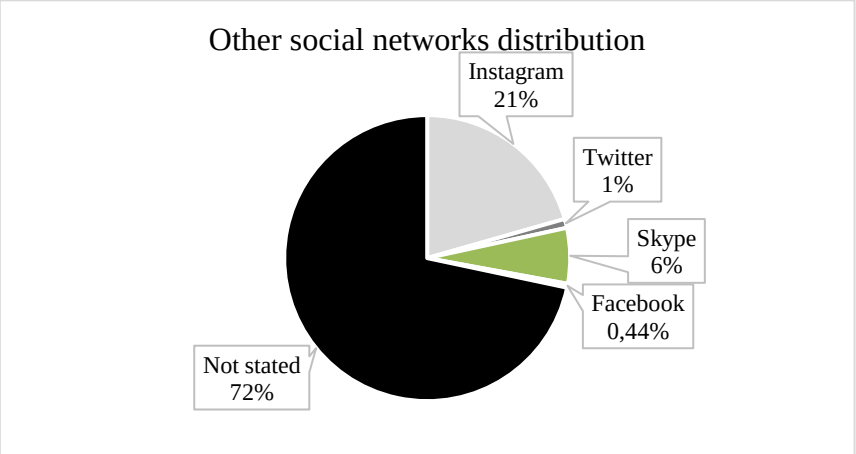
Another feature in the dataset that can shade a light on such characteristics as level of income and technological preferences is the last of frequently used device. On the pie chart below

(Figure 2) the distribution of devices and platforms students tend to use is represented. To no surprise, more than a half of students prefer using iPhone and 75% of all users are active users of smartphones. Only one quarter of GSOM students surf into social media via desktops that can be indirect evidence that the majority of students tend to consume graphical content and communicate using short text messages as traditional behavioral pattern for smartphone user, however, this is just an assumption that cannot be properly tested due to privacy issue. It is partly could be proven by distribution of other social networks accounts that students indicated as their contact (Figure 3). As we can see from the chart, students tend to use more smartphone-oriented social network (Instagram) rather than more neutral on these terms Facebook or Twitter.

To sum up, the overall image of GSOM student is quite obvious in the context of his or her age and socio-demographic background:

1) The distribution of males and females on social networks reflects the actual gender distribution (with small prevalence of females over males)
2) Student tend to hide their real age, but the actual age calculated from the rest of students is relevant to real life (19,6 years)
3) The vast majority of the students indicate their actual location (Saint Petersburg) and occupation (studying at St. Petersburg State University)
4) There is a tendency to indicate the real name and identity in social network (over 99% of students) but a large share of students tend to hide their private data from others and mal students tend to be more strict in terms of privacy.
5) The obsolete majority (90%) prefer to leave empty the fields with more abstract or personal information or concepts but ready to indicate geography or institution related data (50-70%).
6) GSOM students tend to use smartphones (mostly iPhones) for social media surfing and use Instagram rather more than other social networks that may serve as indirect evidence that they prefer to consume graphic content rather than text one.





Figures 2, 3.

K-means clustering on all students

The first attempt to get meaningful information concerning student preferences was made by getting the list of their communities and subscriptions and extracting the main words characterizing these students. The initial idea was to apply k-means-clustering on the vectorized words describing key interests of users and to get certain clusters of users for further data processing and description.

The algorithm was as follows:

1) Get lists of users' subscriptions
2) Get descriptions of these subscriptions
3) Represent these subscriptions as a document (set of words) for each of user
4) Preprocess text data for further analysis:
    a) Lowercase transformation
    b) URL and emails removal
    c) Removal of punctuation and digits
    d) Removal of emoji and other extra characters
    e) Words lemmatization (getting 'normalized' word forms)
    f) 'Stop-words' exclusion basing on the Russian and English 'stop-words' dictionaries
    g) Words tokenization
5) Applying tf-idf method to getting relative word frequencies and word 'weights'

As a result of tf-df method we got a set of words (a document) describing each student with a relative frequency of each word.

6) Defining the optimal number of clusters using Maximum-Likelihood Estimation method (MLE) and sensible limits
7) Getting key words describing the interests of each student cluster.

This algorithm was applied iteratively with exclusion of the most popular words, different lemmatization techniques and number of clusters starting with six and ending with two clusters. We get the best results for three and four number of clusters. The example describing the 15 most frequent words describing interest of each cluster are given in the following tables (Table 1, 2):

Table 1. Examples of words for clusters after applying k-means-clustering with k=3.

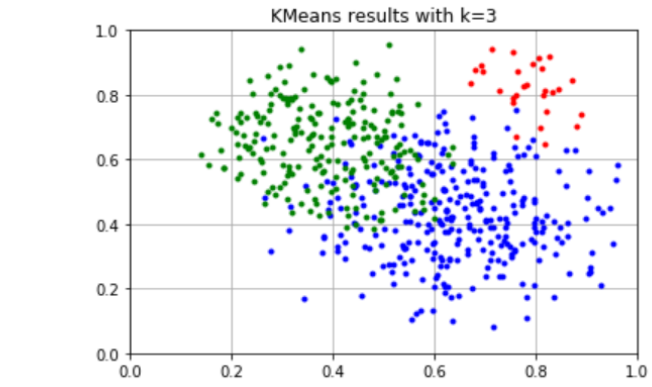
Cluster	Most popular words with k = 3	Number of students in cluster
Cluster 1	газета, мем, обязан, совет, год, все, рассказывать, председатель, второе, cats, совет, клуб, студенческий, мгу, пупа, правило, лупа	37
Cluster 2	пожаловать, memes, просто, истории, предложить, пост, комментарии, самые, актив, добро, другие, год, онлайн, петербург, школа, время, фото, быть, бан	376
Cluster 3	вопросы, наш, язык, просто, вакансии, проект, школа, видео, группа, менеджмента, найти, мероприятия, время, вшм, год, онлайн, работа, петербург, бизнес	276

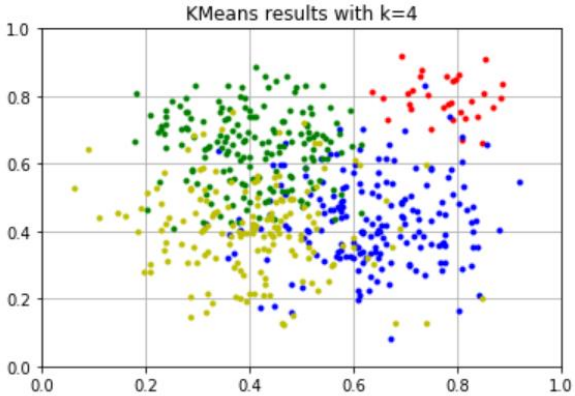
Table 2. Examples of words for clusters after applying k-means-clustering with k=4.

Cluster	Most popular words with k = 4	Number of students in cluster
Cluster 1	газета, мем, обязан, совет, год, всем, рассказывать, председатель, второе, cats, клуб, студенческий, мгу, пупа, правило, лупа	37

Cluster 2	новость, проект, петербург, онлайн, family, успешный, другой, case, предлагать, выпускник, memes, бизнес, gsom, мероприятия, актив, вшм, быть	216
Cluster 3	ты, пост, пожаловать, видео, год, лучший, пост, добро, предложить, петербург, комментарии, самый, просто, время, онлайн, истории, фото, егэ, бан	221
Cluster 4	компания, школа, видео, проект, найти, менеджмент, вакансии, группа, sf, время, петербург, вшм, онлайн, год, мероприятия, работа, бизнес	215

As we can see from these tables, the clusters are poorly separated and give as only a general picture on who GSOM students are with no specific features for any cluster except a small one (with 36-37 students) characterized by special words related to entertainment and memes whereas the other students could be characterized with general business and student life related terms. The level of interosculation the clusters proving their weak level of distinction is presented in the visualisations below (Figure 4):





Legend: Cluster 1 - red dots

Cluster 2 - blue dots

Cluster 3 (for k = 3) - green dots

Cluster 4 (for k = 4) - yellow dots

Figure 4 . Clusters visualization for k-means clustering (k=3, k=4).

We also tried to check if there is any dependency on clustering based on interests and average academic score but the observable difference did not give any statistical significance (p-value = 0,56 for three clusters and p-value = 0,42 for four clusters for ANOVA analysis).

Finally, after several iterations we concluded that this approach is not appropriate for getting student clusters based on the interests. Moreover, further attempts to develop unsupervised ML-model with k-means clustering by adding extra features such as year of study or average score during studies failed to give any fruitful results: clusters still were poorly described. However, basing on the information we obtained applying k-means clustering we can conclude that in general GSOM students are very homogeneous in terms of their interests and can be characterized as:

1) Interested in student life
2) Professionally/business oriented
3) Having a strong association with GSOM and SPbU
4) Geographically connected to St. Petersburg
5) Not very different from typical young people of their educational status and interested in general entertaining content as well

Further hypotheses development and data preparation

After not very impressing results on clusterization of students based on their interests, we have decided to have a look from a different angle: to divide students into certain categories and check, whether characteristics of students of one category are different from another's. To get any specific features the further analysis was necessary, and the next part of our work was devoted to testing a number of hypotheses in tree possible dimensions:

1) Specific features for students from different years of study.
2) Specific features for students from different concentrations.
3) Specific features for students with different academic progress and success.

First of all, we had to mark the existing students in accordance with these dimensions.

To mark students basing on their concentrations and academic progress, we performed certain manipulations with the data received from GSOM administration. First, we needed to obtain a list of all the courses contained in the initial dataset. To do this we dropped duplicates and found out that there were 182 unique course titles. Next, we researched all the existing study plans

and manually marked the courses depending on whether the course was elective (1) or not (0). Also we assigned each of them to one of the following fields: ‘fin’ (finance), ‘mark’ (marketing), ‘hr’ (human resources), ‘it’ (information management), ‘log’ (logistics), ‘gmu’ (public administration), ‘lang’ (foreign languages) and ‘gen’ for all the courses that could not fall into any of the aforementioned categories.

Next, the table with categories was concatenated with the initial dataset. Later it allowed us to count average marks of each student in every category. To do so another change was to be made. In the initial dataset the marks appeared as follows: Excellent, Good, Satisfactory, etc., and were accompanied with additional ECTS grading scale marks (A, B, C, D, E, F). According to the GSOM rules[4], the following translation scale is applied:

- Excellent, A: 5.0
- Excellent, B: 4.7
- Good, B: 4.3
- Good, C: 4.0
- Good, D: 3.7
- Satisfactory D: 3.3
- Satisfactory, E: 3.0
- Passed, A: 5.0
- Passed, B: 4.5
- Passed, C: 4.0
- Passed, D: 3.5
- Passed, E: 3.0

As the initial table consisted of not only positive, but also negative marks and no-show cases, the latter were excluded, as they would have influenced the analysis (a student might have 2 negative marks on 1 course, followed by a positive one). After that we used the study plans codes to differentiate between different programs. Also, we used group codes to distinguish between different concentrations of Management program. As a result, we have found 842 students in total, 71 of which are from Financial Management concentration, 70 - from Marketing concentration, 39 - from Human Resources Management concentration, 56 - from Information Management concentration, 60 - from Logistics concentration, 93 - from Public Administration program and 140 - from International Management program. The rest are 1st and 2nd year students of Management program, as they are not distributed into concentrations yet. For better clusterization results we filled in the missing values with column means and normalized the table contents. After

[4] https://gsom.spbu.ru/files/folder/prikaz_sistema_ocenivaniya_i_srednij_ball.pdf

that we applied the so-called elbow method (Figure 5), which is used to determine the number of clusters in the dataset.

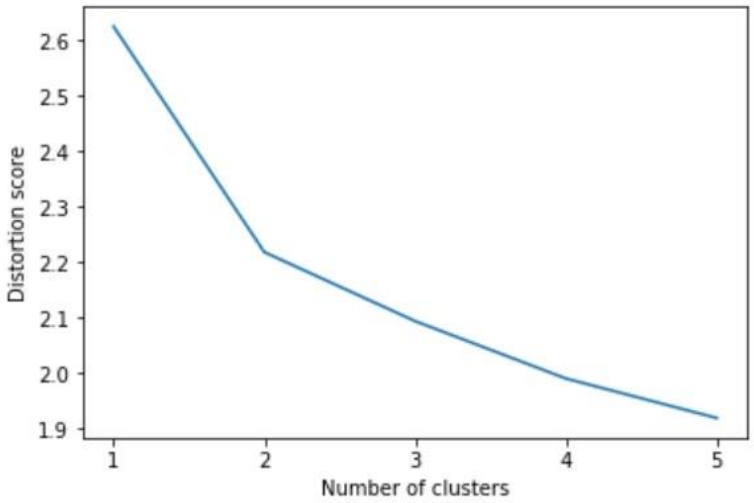


Figure 5. Visual representation of the applied elbow method

The place on the plot where the ”elbow" breaks down tells the number of clusters that the data is divided into. After performing k-means clustering with k=2, the following results were received (Table ?), and the students were assigned with cluster codes (0 or 1).

Table ?. Mean grades of two detected clusters by groups of subjects

Cluster / Group of subjects	Finance	Marketing	HR	IT	Logistics	PA	Languages	General courses	Number of students in cluster
High	4.27	4.36	4.24	4.22	4.22	4.10	4.39	4.40	346
Low	3.65	3.90	3.77	3.55	3.88	4.01	3.58	3.77	496

To no surprise, there is one cluster of a better performing students and another one consisting of those who perform worse in every group of subjects. Small difference in clusters’ mean grades for Public Administration courses could be explained by the fact that this group of subjects is very specific and is met only among students of Public Administration concentration, and, as the missing values of the majority (i.e. students from programs other than Public Administration) were assigned with column means, which influenced the overall means of each of the clusters for this particular group of subjects.

After all mentioned above manipulations we got students labeled by:

1) Year of study:
    a) First-year students
    b) Second-year students

c) Third-year students
d) Fourth-year students

2) Concentration:

a) General management (1st and 2nd-year students): “1”
b) Financial Management: “2”
c) Marketing: “3”
d) Human Resources Management: “4”
e) Information Management: “5”
f) Logistics: “6”
g) Public Administration: “7”
h) International Management: “8”

3) Academic progress:

1) Better performance: “0”
2) Lower performance: “1”

All further hypotheses testing was fulfilled in the frameworks of these three dimensions. The main source of hypothesis was expert opinion on how students tend to behave depending on certain conditions and GSOM administration personnel and professors were attracted as experts. After a set of tests and discussion the following hypotheses were formulated:

1. Students tend to change their interests during studying in GSOM. The initial assumption is that after graduation of secondary school people are less professionally oriented and by the completion of their bachelors they should be more interested in profession.

2. Students of different concentrations show different communication patterns. That was based on the conjecture that despite the fact than managerial studies imply relatively high level of extraversion and communication level there might be some difference within different concentrations as this concentrations may require slightly different personality traits.

3. Well-performing students possess specific features that could be found by analyzing their interests and communication patterns. In general, that is a reverse approach to the one implemented in k-means clustering based on interests and attempt to compare their average scores.

4. There is a significant difference in interests between students studying on different concentrations. It is quite similar to the assumption concerning differences in their communication patterns with additional initial guess that their professional interests may somehow be traced via social media analysis.

5. Students of other concentrations who choose electives of a certain field may be more interested in this field than those students who study on a concentration, corresponding to this field.

2.2. Hypotheses testing and results

Students interests and year-of-study

Hypothesis 1: GSOM influences students' interests.

The full version of hypothesis sounded like "As people study in GSOM they tend to change their opinion and/or interests possibly to professional sphere, so there should be difference in what is interesting for 1st and 4th-year students".

So, to test it we should have looked carefully through remarkable words describing the students of different years of study. The examples of words with their relatives wights are presented in the table below (Table ????):

Table ??? Specific words describing first-year and last-year students with their values.

Year of education	Specific words for first-year students			Specific words for last-year students		
Words	'battlegrounds'	'лупно'	'егэ'	'excel'	'экспертный'	'карьерный'
Year 1	6.13	53.15	46.13	9.55	0.98	1.13
Year 4	0.00	0.00	16.85	29.99	4.14	5.59

To no surprise, the most popular words for freshmen were those related to studies and student life itself like 'management', 'GSOM', 'school', etc. and a set of specific words describing their field of interests due to their age and current life experience namely 'ege', 'oge' 'rule', 'will' and so on that are in general illustrate their immediate past in secondary school. At the same time, there was a peculiar set of words relevant only for first-year students such as 'lupno' and 'pupno' describing modern 'memes' that are popular mostly among people under 20 (people of the last year of study do not know such terms at all) and some words and collocations related to popular computer game 'PlayerUnknown's Battlegrounds' that are quite specific namely for GSOM freshmen. As the absence of the feature is the feature itself, we should also mention the words that are less specific for freshers. To some surprise, they tend not to be interested in future career or professional development as they show a dramatic decline in frequency of words related to famous brands and big companies as their future possible employees. At the same time, despite being fresh graduates from secondary school, they demonstrate less interest to other higher education institutions in St.Petersburg and Moscow that could be explained by their initial intention to enter namely to GSOM to get higher education or, on the other hand, to by secondary, professional or 'corporate' interest to other universities due to interorganizational contacts among elder students.

In general, we can make a conclusion that younger students tend to be more interested to entertaining content, their recent day-to-day reality as school students and quite less interested in career related topics.

As for the last-year-students, they show themselves as much more ‘neutral’ with only general words describing their interests and career and professionally oriented, the most specific words characterizing them are ‘work’, ‘sf’ (name of business oriented online educational institution), ‘Saint Petersburg’, ‘GSOM’, etc. They tend to demonstrate more impersonal interests like ‘cinema’ instead of any peculiar genres, names or links. There are quite less specific words that may characterize namely the eldest students as all ‘general’ terms are on average level and the most ‘weighty’ are quite neutral.

The interesting finding is a set of words those values gradually change with the year of study, the examples of this words are presented below (Table ????)

Table ???. Example of words those values gradually change depending on the year of study.

Words	‘петербург’	‘время’	‘вшм’	‘excel’	‘вакансии’	‘бизнес’
Year 1	24.87	27.27	26.33	7.45	15.08	28.74
Year 2	28.34	29.85	28.06	16.08	17.30	33.75
Year 3	35.17	35.56	32.45	17.00	22.49	37.96
Year 4	41.17	38.23	36.20	19.44	23.16	39.19

All these words relate to job, academic and professional topics that serves as an additional proof to the initial hypothesis that studying at GSOM influences student interests directing them into more professional and business related field.

Concerning second-year and third-year students that initially were beyond the framework hypothesis and the detailed information on whom would not be provided here, we can say that they are characterized with the highest interest to social life (they tend to be more interested in social projects, volunteering, etc - related words have the highest relative values in comparison with 1st and 4th year students). Interest to big companies and only in the context of internships is also their distinctive feature (they tend to have words of company names with the same frequencies as the internship-related words). And a common secondary interest among middle-year students to educational institutions in Saint Petersburg that reflected in the increased number of university names mentions may be explained as creation a network of professional contacts/interests.

So, to cut a long story short we can conclude that:

1) Freshmen tend to like entertaining content (‘pupa-lupa’ memes, ‘BattleGround’ computer game, etc) in comparison to their elder colleagues, no career issues are spread among younger students, they still tend to stick ‘school’ topics (‘ege’, teenage humor, etc.)
2) Last-year students are very ‘neutral’: only general staff or some career related topics
3) Interest to professional and career issues definitely correlates with the year of study and GSOM changes students minds from ‘general business’ to more specific ones, there is also strong dependency between interest to St.Petersburg and the year-of-study: higher the year, higher the interest
4) Middle year students have the highest interest to social projects, interest to internships in big companies (only them and only internships, not positions) and secondary linkages with other educational institutions.

Students interests and profile

As a next step we decided to distinguish between students of different profiles, so our next hypothesis was: there is a significant difference in interests between students studying on different profiles or programmes.

To put this hypothesis on test some additional data preparation was needed to be done - we had to assign students a certain label based on their direction of study. We used group and study plans codes (for example, “Б01-вшм” in a group code means that this group is from Information Management profile, and students of Public Administration profile have “5072” in their study plan codes). It is worth noting that Management program students’ distribution to profiles happens only after the end of the 2nd year - before it all students of this program study on a general track, whereas students of International Management and Public Administration programs choose their study direction while applying to study at the School. Considering this, we divided students into 8 categories:

1) 1st and 2nd year students of Management program.
2) Students of Financial Management concentration.
3) Students of Marketing concentration.
4) Students of Human Resource Management concentration.
5) Students of Information Management concentration.
6) Students of Logistics concentration.
7) Students of Public Administration program.
8) Students of International Management program.

Concerning the results, on the very top of keywords of all concentrations are news, which could indicate that students use this social network as a source of information about current situation, which seems quite logical for representatives of generation Z. This generation is also known as a main producer and consumer of memes, and this is also the case in our findings. Surprisingly, students of Human Resource Management concentration are much more indifferent to memes than others. Another quite obvious result is that GSOM students of all profiles are interested in topics related to their direction of education - such keywords as business, management, GSOM itself (and everything related to it) are also among the top interests of all profiles. All students also seem to be actively seeking work-related information.

Continuing with leisure-related topics, all concentrations are interested in videos, especially students of HRM and Logistics. Photos are also among top interests of all groups, as well as languages.

Communities devoted to movies are popular among three categories of students which include 1st and 2nd year students (i.e. 1, 7 and 8). Since we are considering undergraduate students, these 3 categories also account Unified State Exam (ЕГЭ) as one of their top-interests.

The table below contains some of the keywords that distinguish one or a few concentrations from the others (keywords are either **unique** or much *more common* than for other concentrations).

Table 3. Keywords distinguishing different concentrations and programs

Concentration (category)	Number of students	Keywords
1. 1st and 2nd year students of Management program	315	*мемы*/*memes*, *фильмы*
2. Students of Financial Management concentration	71	**кейс**/**case**, *искусство*
3. Students of Marketing concentration	70	*мемы*/*memes*, **sf**
4. Students of Human Resource Management concentration	39	*фото*/*фотографии*, *видео*, *искусство*/*art*, **дизайн**, *музыка*/*music*, **vogue**, **стиль**
5. Students of Information Management concentration	56	*мемы*/*memes*, *искусство*, *music*
6. Students of Logistics concentration	60	*мемы*/*memes*, **карьера**, *видео*, *music*, **market**, **zenit**

7. Students of Public Administration program	93	*фильмы*, **ielts**
8. Students of International Management program	140	*мемы*/*memes*, *фильмы*, **english**, **usa**

1st and 2nd year students of Management program basically do not show any specific features, which could be explained by the fact that this group is the most numerous and diverse. Students of Financial Management concentration possess only one, but very remarkable feature: this is the only concentration with 'case' among its top-interests, which implies that they are the most active in regards to case competitions. This is not surprising, as, according to the administration of GSOM, for many years the best-performing students have been choosing this concentration.

Marketing students demonstrate diverse interests: on the one hand, they are more than average subscribed to groups with memes, but on the other hand, the abbreviation 'sf' appears in their subscriptions, which quite certainly means SF Education - "an online university for white-collar workers", that prepares "people for work in finance, consulting, management and business analytics"[5]. It is almost impossible to imagine modern marketing without some data-handling IT skills, so this is a sign for administration - probably, Marketing concentration needs more courses related to analytics.

Students of Human Resource Management, besides their low interest in memes, are much more strongly interested in design, vogue and style, and also show above the average interest in photo, video, music and art. These peculiarities could be probably explained by the fact that there are more females studying on this profile.

Students from Information Management like memes, and, surprisingly, art and music, whereas students from Logistics like not only memes and music, but also video, and some special features: career, themarket (which is a brand-name clothes marketplace), and Zenit, that most likely stands for the name of the football club. Students of International Management program are predictably interested in English and the USA, while those studying Public Administration are subscribed to the communities related to IELTS[6].

[5] An online education platform that is specialised in training specialists from scratch on short-term and annual programs and provides to start a career in Finance and related fields. URL: https://sf.education/.
[6] The International English Language Testing System (IELTS) measures the language proficiency of people who want to study or work where English is used as a language of communication. URL: https://www.ielts.org/.

These findings support our initial hypothesis: there are some differences in interests of students of different concentrations, some of which are quite predictable, and some of which are unexpected.

Students interests and profile vs. electives

While comparing students of different concentrations, we have come up with another idea. According to GSOM rules the distribution on profiles is conducted on a ranking basis, which means that students with higher GPA after 2 years of study are given a priority to choose the concentration they will be studying in 3rd and 4th years. This leads to the situation, when some "popular" concentrations (among those usually Financial Management and Marketing) receive students with higher GPAs, while students with worse academic performance have to join other concentrations even in case they are not interested in them. Another important point is that every semester during 3rd and 4th years GSOM offers students some electives - courses, that they can freely choose regardless of their concentration. Therefore, we have decided to test the hypothesis: students of other concentration who choose electives of a certain field may be more interested in this field than those students who study on the concentration, corresponding to this field.

At first we have conducted some research on the given data. Unfortunately, most of the electives belong to the general group of courses, which is hard to assign to certain sphere or concentration, and among specialised ones there are more electives related to IT. What is more, such courses are popular not only among Information Management students, but among others as well, so we have decided to compare 2 groups: 56 students of Information Management and students of other concentrations, that chose IT electives during their studies (152 in total).

Then we counted the average marks for IT courses only. For Information Management students it was exactly 4.0, and for the rest - 4.27. Research on interests of these two groups has shown, that although the second group has no IT-related words among the top keywords, 'business' and 'management' appear in subscriptions of this group much more frequently than among Information Management students. The obtained results confirm the words of the GSOM representative, that the Information Management concentration is not one of the most popular among students choosing their future study track, meaning that those who study on it are less determined towards content related to their studies. Additionally this outcome may underpin the idea that while choosing the concentration students probably are not quite satisfied with Information Management concentration's curriculum content and prefer to opt for other concentrations instead, continuing to be interested in IT-related courses.

Student profiles and communication patterns

Going further to define specific features of student profiles we decided to go beyond the 'interest' profiling and turn to other psychological characteristics, namely, communication patterns. In social networks analysis they are traditionally associated with how 'open' and active the person is, so, the following features describing the level of extraversion were chosen:

1) Average number of friends within group to define the visible range of social contacts
2) Average number of 'followers' to define the scope of potential audience of profile
3) The ratio of friends and followers quantity as secondary measure for 'openness' of person (the initial desire to accept the other person's initiative to interact)
4) The frequency of publication on personal webpage as general as a general measure to estimate the activity of users
5) The share of fully or partly closed profiles within group to gauge the 'openness' of profile.

The initial hypothesis was that students from more 'human oriented' concentrations such as marketing and HR tend to be more sociable and open, whereas students with more 'technical' fields of study such as logistics, finance and IT management should be less communicative.

To test the hypothesis, the numerical values for mentioned above features were calculated and then tested with ANOVA method for statistical significance of the observed results. In the table below (Table 4) you may see the results of average values for these features for each concentration as well as and the respective p-values.

Table 4. Numerical values for communication features and their statistical significance.

Concentration	Number of friends (average)	Number of followers (average)	Friends/ Followers ratio (average)	Number of publications within last 12 months (average)	Share of private profiles
General management	255.9	207.5	3.19	5.22	0,14
Finance	290.1	173.4	2.79	3.86	0,08
Marketing	253.3	141.7	4.89	2.94	0,18
HR	352.8	236.6	2.23	6.37	0,18
Information management	293.6	180.0	2.53	4.34	0,21
Logistics	264.2	186.4	7.51	3.92	0,18
Public administration	232.3	168.9	3.19	2.71	0,16
International	293.3	198.7	2.46	4.68	0,20

management					
p-value	0.097	0.488	0.285	0.212	0.273

As far as we can see here, the only feature with visible statistical significance is number of friends that users on average have (we regard p-value < 0.1 as acceptable). At the same time, we can still use the other numbers as a 'reference point' for further conclusion and assume the existence of certain tendencies. As the additional proof to this presumption we can regard the fact that the group of general management students that actually consists of first- and second-year students share the average values of the proposed features, so, then further chose their concentration later, they will present the same outcomes as we can see now.

Despite our expectations, the initial guess on higher values for more 'extraversive' concentrations was only partly true: students who have chosen the human resource management as their field of study tend to demonstrate more sociable behavioral pattern in social network as they have the highest rates for three of five features. At the same time, the other 'social oriented' concentration is marketing, the students have the least or almost the least rates related to extrovert communicative patterns: the have much less friends and tend to post less (more than two times) frequently than their colleagues from HR management.

The same unexpected contingency was the fact that such 'average' concentration towards which there were no initial proposal regarding being different from others, turned to be less sociable in comparison even with more technical fields. Namely students from this field of study share the same communicative patterns with their counterparts from marketing despite the initial assumption that these concentrations are usually chosen by different types of personalities.

As for the conjecture on less sociable or more close personalities of students who have chosen more numerical and less creative concentrations such as logistics, information management or finance, it also turned to be fallacious: these students tend to be show quite ordinary results. None of their rates (except friends/followers ratio for logistics and the share of closed profiles for finance) lies on the edge of the range of values. Furthermore, in some cases, 'openness' features tend to be higher than for the other concentrations, for example, finance students represent two times less than average rate of private profiles. At the same time, students from logistics and information management tend to be more private as they illustrate higher values on friends/followers ratio and private profiles share.

In general, we can conclude that our initial guess on the presence of dependencies between the concentration of a student and some communication patterns is true, as all defined features show similar tendencies, however, the detailed hypothesis on the nature of these dependencies was

mostly incorrect and students from quantitative concentrations do not different from others and marketing students are, vice-versa, the least sociable at least in terms of social media interactions.

Specific features of well-performing students

As more detailed approach to discovering specific features of certain groups of students turned out be more fruitful than general clustering, we decided to look at the same dimensions (communicative patterns and interests sets) to the students in the context of their academic performance as we on the first step failed to trace any statistically significant evidence of differences in average score of students from different clusters. For now, let us state hypothesis as follows:

1) Well performing students should differ from their lower performing colleagues in terms of communicative patterns, however the nature of this difference is not obvious.

In other words, we recognize the fact that the values should be different, but we can not predict in advance if well performing students tend to be more sociable or not.

2) Students with better academic progress should be more interested in professional development or academic related issues than students with lower average score.

Both hypotheses were tested in the same way as we previous ones: with the help of tf_idf of set of words regarding the subscriptions and the set of numeric values on communicative features, the analysis is performed with a breakdown on two clusters defined with k-mean clustering as it is mentioned above.

Let us look at the most common and specific words for both clusters (Table 5):

Table 5. Examples of the most common and the most specific words and their relative weights in context of academic performance.

Topic	Academic or professional				Entertaining and Internet				Social life		
Term	SF[7]	бизнес	excel	вакансии	калья н	memes	‘battlegrounds’[8]	анекдоты	волонтеры спбгу	мероприятия	молодежь
High achievers	18.33	37.45	22.31	20.46	0.84	15.12	0.097	0.91	4.86	33.12	2.94
Low achievers	10.14	31.17	16.33	14.75	3.62	20.88	4.63	4.84	1.61	24.24	1.21

[7] ‘SF education’ is an online education platform for young people that proposes courses for those who is interested in career in consulting, finance etc. (URL: ‘https://vk.com/sfeducation’)

[8] ‘Battlegrounds’ is an online multiplayer battle royale game (‘URL: https://www.pubg.com/’)

In general, it is obvious that people with different academic progress really have differences in what they prefer to read in social media. High achievers are more interested in business related issues, business events, studies and so on whereas people who perform less successfully tend to like entertaining content and activities such as memes, Internet-related topics, video games, etc. The presented above words are just an examples but they outline the general distinctions of two student profiles, for examples, pretty the same results illustrate the words like ‘internships’, ‘big companies’, ‘case’, ‘leadercup’ etc. for professional sphere and ‘humor’, ‘lupa’ etc. for entertaining content. This does not actually mean that well-performing students tend to be interested only in business events, career development and so on, they still have high frequencies of words related to leisure activities and amusement, but the most popular terms for them here are ‘music’, ‘cinema’ and other neutral notions. At the same time, we should mention that such categories as leadership, entrepreneurship, projects, etc. have relatively similar values for both groups that can evidence the fact that lower performing students are more interested in developing their own business rather than in building their careers in large corporations.

The very meaningful finding is that people who have higher academic achievements tend to more associate themselves with GSOM and University in comparison to poor-performing students. The relative weights and frequencies of words related to the School is presented below (Table 6):

Table 6. Words related to GSOM and their relative weights in context of academic performance.

Term	‘GSOM’	‘School’	‘ВШМ’	‘школа менеджмента’
Cluster 0: High achievers	24.47	32.62	33.75	31.13
Cluster 1: Low achievers	18.88	23.23	24.01	24.68

As we can see from the table, for students with higher academic progress, the values are 1.3 - 1.5 times more than for worse performing students and this tendency is relevant for any way of naming the School regardless of language and formality. The possible explanation for the observed phenomenon is that people from less performing group are simply less interested in all academic related topics and their motivation to obtain higher education lies in other dimension not in curiosity or desire to get professional skills. Another option is that students feel depressed due to their poor academic results and subconsciously tend to avoid negative feelings associated with School digging into more entertaining issues. However, this question lies beyond the framework of the research and cannot be tested with available data and expertise.

Our second assumption was that people with higher academic performance might differ in terms on communicative patterns but the sense of these distinctions is unclear. So, let us check this supposition and find the main features that may characterize each group in the frameworks of proposed above dimensions (Table 7):

Table 7. Numerical values for communication features and their statistical significance in the context of academic progress

Academic cluster	Number of friends (average)	Number of followers (average)	Friends/ Followers ratio (average)	Number of publications within last 12 months (average)	Share of private profiles
Cluster 0: High achievers	259.6	207.5	2.54	4.92	0.15
Cluster 1: Low achievers	278.1	173.4	4.20	5.11	0.16
p-value	0.246	0.093	0.045	0.873	0.896

As we can see, the differences in communication patterns are less significant rather than differences in students' interests: all the features show almost the same rates with very low measure in statistical significance. The only observed distinction is related to the number of followers and theration of friends and followers: less performing students tend to have less followers and similar number of friends, that can be explained in two ways. On the one hand, less performing students are more open to new acquaintances and less choosy about their virtual connections and friends, as their level of ambitions is lower or they try to 'compensate' their lower academic position with the number of people they are connected with that is associated with the degree of 'popularity' and can be regarded as a measure of personal success. On the other hand, high achievers might be more conscious on their virtual connections and prefer to 'hold a distance' with people with whom they are not acquinced well enough, due to some 'status' issues or simple due to being less open to new contacts. Anyway, this hypothesis cannot be tested without deeper psychological investigation but might be a fruitful topic for further researches regarding communication pattern of students with different academic progress.

To conclude, we can define the existing differences between these two groups of students as follows:

1. In terms of communication patterns, both clusters are very similar.
2. The only great difference is that underperforming students tend to have more friends and less followers.
3. Interests explain academic progress: poor-performers are interested in entertaining content.

4. Poor-performers prefer to consume content related internet, jokes, memes, etc. with no social or cultural activities.

5. Poor-performers are interested in career and success in terms of leadership status or entrepreneurship but not in professional managerial development.

6. Well-performers are interested in research, career events, professional growth.

7. Well-performers much stronger associate themselves with GSOM and University.

2.3. Summary of Chapter 2

The chapter 2 was the gradual implementation of concepts and methods mentioned in the theoretical part of the research, its main purpose was to analyse data from VKontakte social network, reveal hidden patterns and interpret some insights.

Firstly we described the ways of collecting data and outlined the main characteristics of data sources and obtained dataset such such privacy issues, poor structure and the problem of data incompleteness. The data was collected from multiple sources namely internal GSOM tables with students studying plans and academic performance and external sources such as information from students' pages in VK. Next, the descriptive data analysis was implemented in order to reveal some socio-demographic characteristic of students and estimate perspectives for further direction of the research. That stage also included an attempt to build a k-means clustering model based mainly on the features describing students interests (obtained with tf-idf NLP method applied to set of documents got from students subscriptions) that ended with the conclusion that students are quite homogeneous in terms of their interests and another approach to revealing peculiarities is required.

For further work, the set dimensions within which students would be considered and list of hypotheses to test was elaborated. The dimensions included year of study, concentration and academic progress cluster mined with application of k-means clustering on students' scores. The five proposed hypotheses included several assumptions based on the idea of existence of dependencies between labeling student according to certain dimension and possible differences in interests and student communication patterns. For example, the guess that students with higher academic performance may have different interests in social media in comparison to their lower performing colleagues or an presumption that people who chose HR or marketing as their major are characterized with higher level of extraversion.

As a result, we got of set of findings concerning the way student behave in social networks that can shed a light on the hidden part of student life for the administration and make some adjustments in internal processes within GSOM. The possible managerial implications of the obtained results are proposed in the next part of the paper.

Conclusion

Main findings

Trends in modern education are moving towards further ‘onlinization’ of education process and related changing in interaction between students and educational institutions. To build a firmer connection with undergraduates, the explicit image of the student is required.

The main purpose of the research was to define specific features by analyzing data from students’ profile pages in Vkontakte and provide GSOM with insights based on it. The study started with the theoretical overview of all relating topics such as the modern trends in education, approaches to social networks analysis and student profiling and data mining tools used for these purposes. Applying such data mining and statistical methods as tf-idf, k-means clustering, descriptive statistics we obtained the following results.

1) Descriptive statistics revealed that undergraduate students possess certain patterns in terms of privacy and completeness of information. They are ready to indicate their real name, gender, place of study and location but tend to hide their real age and more private issues like religious and political views or habits. That could be generalized to the theory that students might be ready to state only materialized issues that can be a basis for building connections or may serve as a tool to certain group recognition.
2) The statistics of used devices and stated links to other social networks reveal that students tend to use mobile phones more than laptop or desktop devices for communication in social networks that may serve an assumption that they tend to consume graphical content rather than text.
3) Attempts to implement k-means clustering based on students interest or wider range of features together with the results of exploratory data analysis nudged us to the conclusion that undergraduate students of GSOM are quite homogeneous and in order to obtain more meaningful outcomes they must be researched on a lower level, in a framework of certain dimensions. Upon the whole, ‘students are just students’ who are interested in business, education, entertainment, and Saint Petersburg, but the delightful finding is that they associate themselves with GSOM and show their affiliation to GSOM.
4) Studying in the Graduate School of Management influences people's’ interests and these changes are happening gradually from year to year. Thus, freshmen are interested more in entertaining content and have not fully left their school identity that includes specific humor, relevance of Unified State Exam (ЕГЭ) topic, teenagers’ memes and so on, whereas the elder students prefer more general and professional topics in social media and are much closer to the image of a young white-collar. Moreover, middle course students

demonstrate interest to volunteering, social project and extracurricular activities while their younger and elder colleagues are relatively indifferent to these issues.

5) There are some differences in communication features that describe people from different concentrations and our hypothesis on the existence of some dependencies between this two variable war true, but the nature of the relation was stated erroneously as the students from quantitative concentrations do not different from others and marketing students that were regarded as extraverts are, vice-versa, the least sociable at least in terms of social media interactions, though the hypothesis on the higher level of extraversion of their colleagues from HR proved to be true.
6) Some findings concerning dependencies of students concentrations and set of words characterizing each of them also were got, that dependencies are not so direct and evident as it we supposed and finally we can conclude that GSOM students of all concentrations are interested in topics related to their direction of education, work-related information and all of them are keen on videos. At the same time, among interesting peculiarities we can state that Financial Management students are interested in cases and case championships, HRM students are the most creative ones and like beauty, creativity and design related topics, Logistics students demonstrate their interest to marketing and 'Zenit' (Russian football team), Public Administration students are quite interested in IELTS exam. Findings based on these specifics can be considered in further work of GSOM administration.
7) Another result of the analysis that is important to mention was obtained by comparing interests of students from Information Management concentration and students of other concentrations that chose elective courses from IT in Management department during their studies. Latter, although not showing concerns about IT-related issues as we initially expected, have demonstrated much stronger interest in business and management, which supports established in GSOM practice when top performing (and thus more motivated towards studies) students opt for Financial Management and Marketing concentrations.
8) The last but not least implication of the study is that we managed to reveal certain peculiarities of students with higher academic progress. Among the specific communication patterns is the fact that higher performing students tend to have more followers and less friends in comparison with their lower-performing colleagues. At the same time, we can conclude that are some large differences in interests of lower and higher performing students: students with higher GPA are quite more interested in professional topics, entrepreneurship, research, career events, etc. and more close associate themselves

with GSOM whereas lower performing students prefer entertaining content such as memes, jokes and video games.

Managerial implications of the results and proposals

As for the practical application of the results of conducted analysis, the following propositions could be elaborated.

As all of the current GSOM undergraduate students belong to the Generation Z, they are used to consuming information in video format (Baron, 2020). This is supported by the results of our research: 'video' is among the top interests of all categories of students, regardless of their concentration or academic performance. What is more, 75% of students log in Vkontakte from mobile devices, and on their Vkontakte pages they indicate Instagram profile more often than other social networks. After a brief study of the GSOM communities in Vkontakte[9] we can judge that format of video communication on this platform has room for growth. However, recently on GSOM Instagram page[10] videos began to appear, so we can advise to continue this practice and, possibly, spread it on Vkontakte. This recommendation is also relevant with regard to attracting enrollees in the upcoming years, as they also a part of this generation and thus share the same attitude.

Our second suggestion is closely related to the first one - it implies encouraging User-Generated Content by engaging current students in some kinds of contests aimed at promotion of GSOM. This can potentially help increase recognition of the School and attract more attention to its undergraduate programs (https://blog.neongoldfish.com/five-keys-to-successful-marketing-for-the-education-sector). The effect will be even more significant if it is possible to recognize and engage influencers (opinion leaders) from among students. More on this in the next part of the conclusion.

The next proposition is connected with the interest of Marketing students in online analytics courses. After analysis of the study plans, we found out that Marketing concentration is lacking courses directly linked with marketing analytics. Therefore, probably after a deeper study such applied courses could be added either to the Marketing concentration curriculum or to the electives, so that all committed students could opt for them.

Global and Russian education trends as well as the prospecting development strategy of GSOM, coupled with the interest shown by students of all concentrations in IT courses, allow us to recommend a fairly obvious thing – increasing the share of applied IT-related courses. Perhaps

[9] https://vk.com/gsom.spbu; https://vk.com/gsom_abiturient; https://vk.com/careercenter_gsom
[10] https://www.instagram.com/gsom.spbu/

at first in the form of electives, and subsequently - compulsory courses for all concentrations. This solution may even have a double effect, and here is why. The ability to study in an intercultural environment remained one of the advantages of GSOM for many years, which has always been attracting Russian enrollees as well. Thus, given the growing interest in digitalization worldwide, GSOM, by expanding the offer of IT-related courses, at first, will attract more not only Russian, but also foreign students (both permanent and exchange), which, in turn, will make GSOM even more popular among applicants from Russia.

Another suggestion is connected with our difficulties with classification students according to their interests and building a clear picture of student interests of different concentrations. The fact that the our analysis revealed 'only general business' interests of undergraduates regardless of their concentration and year of study it might be useful to implement a set of trainings on building personal professional brand from middle courses as a part of career definition and career path building measures. According to the research provided by the largest Russian recruitment platform HeadHunter (hh.ru), 84% (hh.ru, 2019) of employers admitted the fact that they check the candidate profile in social networks and a low quality or controversial content in personal page can be a reason for employment rejection (hh.ru, 2019). In this case, building a proper 'professional profile' in social media can increase chances of young specialist to get desired position or, with further additional work, expand the network of business contacts and become a part of professional community. For GSOM administration that means that increased number of young professionals from GSOM in social networks may positively influence GSOM own brand in social media.

Prospects for future research

As the research is at the edge of information technologies, management and social studies, the prospects for the further research lies in the cope of these disciplines.

The main managerial and analytical way for further research is observing the revealed patterns in dynamics as the limitations of social networks data structure do not allow to obtain historical data on moment of subscription or time changes in the number of friends and followers. Applying the same algorithms to the data of future students will help to prove (or reject) the revealed specifics. The most prominent example is changes in users' subscriptions between their applicant and freshman status. Here the question could be formulated like 'to what extend GSOM impacts the student interests in terms of attracting their attention to professional sphere and in what year of study the influence if the most significant?'

Another prospect at the edge of managerial and information technology studies is building a social graph to reveal the key nodes (the most influential students) that is defined by the number of intersections of friend and followers and. in the most complicated models, the data concerning

‘likes’, comments and page visits. The managerial implication in this case would be building a certain type of relations with these students in order to ensure more efficient bilateral communication between student community and administration beyond the frameworks of existing institutions like Student Council or head girls.

Some further research regarding marketing and brand management of the university beyond the proposed communication recommendation also would be useful to conduct. We found out that higher performers tend to show their affiliation with GSOM while the lower performing students tend to do that much less, so, the public opinion on the School in social media is mostly influenced by the picture of better performing students. In this case, the research regarding the other factors influencing GSOM image in social networks seems to be valuable as if the picture of well performing students is really a crucial factor, the work on the enhancing or reshaping their image would help to ensure higher GSOM brand recognition and appreciation in social media.

For social and psychological sciences, there are several issues related to internal psychological drivers explaining the observed phenomena. For example, underlying mechanisms causes distinction in the number of followers and the ratio of friends and followers: lower performing students tend to have less followers and similar number of friends. According to our suggestion, it is explained either by the compensatory mechanisms to reimburse the less success results in professional sphere with personal popularity in social sphere or by specific internal characteristics of higher achievers like consciousness in choice of contacts or desire to highlight their more successful status with social distance. Similar socio-psychological issues is hidden beyond the less extroversive than expected profile of marketing students.

List of references

Agranovich, M. (9 December 2019 г.). *What specialties were the most popular in universities in 2019*. Получено 8 May 2020 г., из Rossiyskaya Gazeta: https://rg.ru/2019/12/09/na-kakie-specialnosti-v-vuzah-byl-samyj-bolshoj-spros-v-2019-godu.html

Ahmed, F., & Abulaish, M. (2013). A generic statistical approach for spam detection in Online Social Networks. *Computer Communications, Volume 36, Issues 10–11*, 1120-1129. Получено 18 May 2020 г., из http://www.sciencedirect.com/science/article/pii/S0140366413001047

Aikins, E. D., Adu-Oppong, A. A., & Darko, G. M. (2018). University Education as a Service: Issues of Quality. *Journal of Social Sciences & Humanities Research*, 1-5.

Alekseev, A., & S., N. (2017). Word Embeddings for User Profiling in Online Social Networks. *Computación y Sistemas, 21(2)*, 203-226. Получено 18 May 2020 г., из https://dx.doi.org/10.13053/cys-21-2-2734

Anjaria, M., & Guddeti, R. (2014). A novel sentiment analysis of social networks using supervised learning. *Soc. Netw. Anal. Min. 4, 181*. Получено 18 May 2020 г., из https://doi.org/10.1007/s13278-014-0181-9

Barbon, S., Igawa, R. A., & Bogaz Zarpela, o. B. (2017). Authorship verification applied to detection of compromised accounts on online social networks. *Multimedia Tools and Applications, 76(3)*, 3213–3233. Получено 18 May 2020 г., из https://proxy.library.spbu.ru/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=inh&AN=16696584&lang=ru&site=eds-live&scope=site

Baron, J. (1 June 2020 г.). *The Key To Gen Z Is Video Content*. Получено из Forbes: https://www.forbes.com/sites/jessicabaron/2019/07/03/the-key-to-gen-z-is-video-content/

Biner, P., & Dean, R. (1997). Profiling the Successful Tele-Education Student. *Distance Education Report, 1(2), 1-3*. Получено 18 May 2020 г., из https://eric.ed.gov/?id=EJ546274

Bishop, C. (2006). *Pattern Recognition and Machine Learning.* Springer.

Burns, V., Blumenthal, A., & Sitter, K. C. (2018). How Twitter is changing the meaning of scholarly impact and engagement: Implications for qualitative social work research. *Qualitative Social Work, [s. l.], v. 19, n. 2*, 178. Получено 18 May 2020 г., из http://search.ebscohost.com/login.aspx?direct=true&db=edb&AN=142316441&lang=ru&site=eds- live&scope=site

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J., & Zanasi, A. (1998). *Discovering Data Mining: From Concept to Implementation.* New Jersey: Prentice Hall.

Camama, R. A. (2019). Q-DAR: quick disaster aid and response model using Naive Bayes and Bag-of-words algorithm., (стр. 012051). doi:10.1088/1757-899X/482/1/012051

Cambridge University Press. (б.д.). *Cambridge Business English Dictionary*. Получено 18 May 2020 г., из https://dictionary.cambridge.org/dictionary/english/profiling

Chaudhary, P. (2019). Flood-Water Level Estimation from Social Media Images.

Cui, R., Agrawal, G., & Ramnath, R. (2020). Tweets can tell: activity recognition using hybrid gated recurrent neural networks. *Social Network Analysis and Mining, 10(1)*. doi:10.1007/s13278-020-0628-0

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, (стр. 886-893). San Diego, CA, USA. Получено 19 May 2020 г., из

https://www.webcitation.org/6DvoEMDXQ?url=http://lear.inrialpes.fr/people/triggs/pubs/Dalal-cvpr05.pdf

Darcan, O., & Badur, B. (2012). Student profiling on academic performance using cluster analysis. *Journal of e-Learning & Higher Education, 2012, e1-8*. Получено из https://ibimapublishing.com/wp-content/uploads/articles/JELHE/2012/622480/622480.pdf

Desai, A., Shah, N., & Dhodi, M. (2016). Student profiling to improve teaching and learning: A data mining approach. *In 2016 International Conference on Data Science and Engineering (ICDSE)*, (стр. 1-6). Получено 18 May 2020 г., из https://ieeexplore.ieee.org/abstract/document/7823947

Dougnon, R., Fournier-Viger, P., & Nkambou, R. (2015). Inferring User Profiles in Online Social Networks Using a Partial Social Graph. *Advances in Artificial Intelligence. Canadian AI 2015. Lecture Notes in Computer Science, vol 9091*. Получено 18 May 2020 г., из https://link.springer.com/chapter/10.1007/978-3-319-18356-5_8

Dumbre, P., & Ingawale, V. (2016). Student Profiling: Importance in Enhancing Student Employability.

Eddey, P., & Baumann, C. (2009). Graduate Business Education: Profiling Successful Students and Its Relevance for Marketing and Recruitment Policy. *Journal of Education for Business, 84(3)*, 160-168. doi:10.3200/JOEB.84.3.160-168

Fire, M., Kagan, D., Elyashar, A., & Elovici, Y. (2014). Friend or foe? Fake profile identification in online social networks. *Social Network Analysis and Mining, 4(1)*, 1-23.

Fishman, T., Ludgate, A., & Tutak, J. (16 March 2017 г.). *Success by design. Improving outcomes in American higher education*. Получено 2 May 2020 г., из Deloitte: https://www2.deloitte.com/us/en/insights/industry/public-sector/improving-student-success-in-higher-education.html#

Fok, e. a. (2014). Data Mining Application of Decision Trees for Student Profiling at the Open University of China. *2014 IEEE 13th International Conference on Trust, Security and Privacy in Computing and Communications*, (стр. 732-738). Beijing. doi:10.1109/TrustCom.2014.96

Goldberg, Y., & Levy, O. (2014). word2vec Explained: deriving negative-sampling word-embedding method. Получено 18 May 2020 г., из arXiv:1402.3722

Graduate School of Management. (14 February 2020 г.). *The first meeting of the GSOM Strategy Commission under the Advisory Board.* Получено 7 May 2020 г., из Graduate School of Management: https://2025.gsom.spbu.ru/tpost/s16900sgpk-the-first-meeting-of-the-gsom-spbu-strat

Graduate School of Management. (13 April 2020 г.). *Updating GSOM SPbU Strategy: Interview with the Director.* Получено 15 May 2020 г., из Graduate School of Management: https://gsom.spbu.ru/en/all_news/event2020-04-13/

Hawtin, M., & Percy-Smith, J. (2007). Community Profiling : A Practical Guide. Получено 18 May 2020 г., из http://proxy.library.spbu.ru:2124/login.aspx?direct=true&db=nlebk&AN=234615&lang=ru&site=eds-live&scope=site

hh.ru. (25 March 2019 г.). *Not invited to work? Perhaps the problem is in your profile on social networks*. Получено 1 June 2020 г., из HeadHunter: https://spb.hh.ru/article/24311

hh.ru. (7 May 2019 г.). *Social networks for a careerist: how to create a professional image for yourself*. Получено 2 June 2020 г., из HeadHunter: https://spb.hh.ru/article/24624

Howard, C., Boettcher, J. V., Justice, L., Schenk, K. D., & Rogers, P. L. (2005). *Encyclopedia of Distance Learning.* Idea Group Inc (IGI).

Jamil, J., & Shaharanee, I. (2017). Student profiling on university co-curriculum activities using data visualization tools. *AIP Conference (Vol. 1905, No. 1, p. 040013).* AIP Publishing LLC. Получено 18 May 2020 г., из https://aip.scitation.org/doi/abs/10.1063/1.5012254

Jha, S. K., & Shorko, S. (2018). A novel CALM algorithm in student profiling. *Computer Applications in Engineering Education, 26(4)*, 841-851. Получено 18 May 2020 г., из https://onlinelibrary.wiley.com/doi/abs/10.1002/cae.21926

Jiamthapthaksin, R., & Aung, T. H. (2017). User preferences profiling based on user behaviors on Facebook page categories. *2017 9th International Conference on Knowledge and Smart Technology (KST)* (стр. 248–253). Knowledge and Smart Technology (KST). doi:10.1109/KST.2017.7886077

Joo, S.-I. (2014). ASM-Based Objectionable Image Detection in Social Network Services. *International Journal of Distributed Sensor Networks*, 1-10. Получено 17 May 2020 г., из https://proxy.library.spbu.ru/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=iih&AN=100534677&lang=ru&site=eds-live&scope=site

Khrapov, P., & Stolbova, V. (2019). Mathematical Modelling of the News Spreading Process in Social Networks. *Sovremennye informacionnye tehnologii i IT-obrazovanie = Modern Information Technologies and IT-Education*, 225-231. Получено 18 May 2020 г., из https://cyberleninka.ru/article/n/mathematical-modelling-of-the-news-spreading-process-in-social-networks

Kirschner , P., & Karpinski, A. (2010). Facebook and academic performance. *Computers in Human Behavior 26*, 1237–1245.

Kotler, P., & Keller, K. (2006). *Marketing management.* New Jersey and London: Prentice-Hall.

Lowe, D. (1999). Object recognition from local scale-invariant features. *Proceedings of the Seventh IEEE International Conference on Computer Vision*, (стр. 1150-1157). Kerkyra, Greece. Получено 18 May 2020 г., из https://ieeexplore.ieee.org/document/790410

Mehmet, S., & Bulent, S. (2004). Survey over image thresholding techniques and quantitative performance evaluation. *Journal of Electronic Imaging 13(1)*, 146–165. Получено 18 May 2020 г., из https://www.spiedigitallibrary.org/journals/Journal-of-Electronic-Imaging/volume-13/issue-1/0000/Survey-over-image-thresholding-techniques-and-quantitative-performance-evaluation/10.1117/1.1631315.short?SSO=1

Mironova, K. (23 October 2019 г.). *Foreigners are doing better.* Получено 8 May 2020 г., из Kommersant: https://www.kommersant.ru/doc/4134138

Mitchell, T. (1997). *Machine Learning.* New York, NY, United States: McGraw-Hill, Inc.

Moules, J. (2 June 2019 г.). *How business schools compete in a disrupted market.* Получено 10 May 2020 г., из Financial Times: https://www.ft.com/content/6a77610e-76f2-11e9-b0ec-7dff87b9a4a2

Ng, I. C., & Forbes, J. (2009). Education as Service: The Understanding of University Experience Through the Service Logic. *Journal of Marketing of Higher Education*, 1-26.

Nwangwa, K., Yonlonfoun, E., & Omotere, T. (2014). Undergraduates and Their Use of Social Media: Assessing Influence on Research Skills. *Universal Journal of Educational Research 2(6)*, 446-453.

Paul, J., Baker, H., & Cochran, J. (2012). Effect of online social networking on student academic performance. *Computers in Human Behavior*, 2117-2127.

Peters, K., Smith, R. R., & Thomas, H. (2018). *Rethinking the Business Models of Business Schools: A Critical Review and Change Agenda for the Future.* Emerald Group Publishing.

Richthammer, C., Netter, M., Riesner, M., Saenger, J., & Pernul, G. (2014). Taxonomy of Social Network Data Types. *EURASIP Journal on Information Security*.

Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation, Vol. 60 No. 5*, 503-520. Получено 18 May 2020 г., из https://doi.org/10.1108/00220410410560582

Ruwandika, N. D., & Weerasinghe, A. R. (2018). Identification of Hate Speech in Social Media. *2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer).* Advances in ICT for Emerging Regions (ICTer). doi:10.1109/ICTER.2018.8615517

Stahl, F., Gaber, M., & Adedoyin-Olowe, M. (2014). Survey of data mining techniques for social media analysis. *Journal of Data Mining & Digital Humanities 2014*.

Strang, K. D. (2008). Quantitative online student profiling to forecast academic outcome from learning styles using dendrogram decision models. *Multicultural Education & Technology Journal*. Получено 18 May 2020 г., из https://www.emerald.com/insight/content/doi/10.1108/17504970810911043/full/html

Tacchini, E., Ballarin, G., Della Vedova, M. L., Moret, S., & de Alfaro, L. (2017). Some like it hoax: Automated fake news detection in social networks. Получено 18 May 2020 г., из https://arxiv.org/abs/1704.07506

Tcheremisova, I. (2016). Content analysis of pages of active users of the Vkontakte social network. *VestnikVolGU. Series 11. Natural sciences. №2 (16)*. Получено 18 May 2020 г., из https://cyberleninka.ru/article/n/kontent-analiz-stranits-aktivnyh-polzovateley-sotsialnoy-seti-vkontakte

Tempelaar, D., Rienties, B., Mittelmeier, J., & Nguyen, Q. (2018). Student profiling in a dispositional learning analytics application using formative assessment. *Computers in Human Behavior, 78*, 408-420. Получено 18 May 2020 г., из https://www.sciencedirect.com/science/article/pii/S0747563217304776

Thomas, H. (2007). Business school strategy and the metrics for success. *Journal of Management Development 26(1)*, 33-42.

Tutubalina, E., & Nikolenko, S. (2018). Exploring convolutional neural networks and topic models for user profiling from drug reviews. *Multimedia Tools and Applications: An International Journal, 77(4)*, 4791. doi:10.1007/s11042-017-5336-z

Vasanthakumar, G. U. (2015). UP3: user profiling from profile picture in multi-social networking. Получено 19 May 2020 г., из https://proxy.library.spbu.ru/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=inh&AN=15888229&lang=ru&site=eds-live&scope=site

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (стр. I-I). Kauai, HI, USA: CVPR 2001. Получено 18 May 2020 г., из https://ieeexplore.ieee.org/abstract/document/990517

Wiedmann, K., Buxel, H., & Walsh, G. (2002). Customer profiling in e-commerce: methodological aspects and challenges. *Journal of Database Marketing 9 (2)*, 170–184.

Woodgates, P. (15 March 2018 г.). Universities must innovate to adapt and succeed. London, United Kingdom. Получено 04 05 2020 г., из

https://www.timeshighereducation.com/hub/pa-consulting/p/universities-must-innovate-adapt-and-succeed

Xiao, C., Freeman, D., & Hwa, T. (2015). Detecting Clusters of Fake Accounts in Online Social Networks. *In Proceedings of the 8th ACM Workshop on Artificial Intelligence and Security (AISec '15)* (стр. 91–101). Association for Computing Machinery. doi:https://doi.org/10.1145/2808769.2808779

Xu, D., Wang, H., & Su, K. (2002). Intelligent student profiling with fuzzy models. *35th Annual Hawaii International Conference on System Sciences (pp. 8-pp).* Получено 18 May 2020 г., из https://ieeexplore.ieee.org/abstract/document/994005

Yousefi Nooraie, R., & al., e. (2020). Social Network Analysis: An Example of Fusion between Quantitative and Qualitative Methods. *Journal of Mixed Methods Research, 14(1),* 110-124. Получено 18 May 2020 г., из http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ1237127&lang=ru&site=eds-live&scope=site

Appendix. Code.

```
#getting main profile data
import numpy as np
import vk
import pandas as pd

session = vk.Session(access_token='token')
vkapi = vk.API(session)

df = pd.read_csv('ids.csv', header = 0)
user_ids = df.squeeze()

users = vkapi.users.get(v=5.103,
                        user_ids=user_ids,
                        fields='sex,bdate,city,country,home_town,education,universities,
                        'schools,last_seen,occupation,relation,personal,connections,activities,
                        'interests,music,movies,tv,books,games,about,quotes,career')

print(users)
users = users[1:]
[print(user) for user in users]
```

```
# transliteration and dead profiles removal
data['full_name'] = data.last_name + ' ' + data.first_name
data.drop(data[data.first_name == 'DELETED']. index, axis=0, inplace=True)

from transliterate import translit, get_available_language_codes
for i in range (0,len(data)):
    try:
        data['full_name'][i] = translit(data['full_name'][i], 'ru')
    except: KeyError
```

```
#getting main features as dataframe columns
for i in range (0,len(data)):
    try:
        data.loc[i, 'languages'] = data.loc[i,'personal']['langs']
        data.loc[i, 'people_main'] = data.loc[i,'personal']['people_main']
        data.loc[i, 'life_main'] = data.loc[i,'personal']['life_main']
        data.loc[i, 'smoking'] = data.loc[i,'personal']['smoking']
        data.loc[i, 'alcohol'] = data.loc[i,'personal']['alcohol']
        data.loc[i, 'occupation'] = data.loc[i,'occupation']['id']
        data.loc[i, 'occupation'] = data.loc[i,'occupation']['type']
        data.loc[i, 'school_spec'] = data.loc[i,'schools']['speciality']
        data.loc[i, 'school_type'] = data.loc[i,'schools']['type']
        data.loc[i, 'last_platform'] = data.loc[i,'last_seen']['platfrom']
        data.loc[i, 'city_name'] = data.loc[i,'city']['id']
        data.loc[i, 'country_name'] = data.loc[i,'country']['id']
        data.loc[i, 'platform'] = data.loc[i,'last_seen']['platform']
    except: TypeError
```

```
 #getting users list of subscriptions
import time
session = vk.Session(access_token='token')
vkapi = vk.API(session)

groups = {}
private_profile = 0
inactive_profile = 0

for us_id in user_ids:
    try:
        time.sleep(0.5)
        groups[us_id] = vkapi.groups.get(v = 5.103,
                        user_id =  us_id,
                       extended = 1,
                      fields = 'description, status')
    except vk.exceptions.VkAPIError as e:
        if e.code == 30 or e.code == 7:
            private_profile = private_profile + 1
        if e.code == 18:
            inactive_profile = inactive_profile + 1
        else:
            private_profile = private_profile + 1
print("loaded_profiles: ", len(groups))
print("private profiles: ", private_profile)
print("inactive_profiles: ", inactive_profile)
```

```
id_text_dict = {} #dict for saving pairs of ids as keys and text(group_name + group_decription) as values
for i in range(len(groups)):
    id_text_dict[list(groups.keys())[i]] = ""
    for j in range(len(groups[list(groups.keys())[i]]['items'])):
        if 'name' in groups[list(groups.keys())[i]]['items'][j].keys()
        and 'description' in groups[list(groups.keys())[i]]['items'][j].keys():
            id_text_dict[list(groups.keys())[i]] = id_text_dict[list(groups.keys())[i]]
            + groups[list(groups.keys())[i]]['items'][j]['name'] +" "
            + groups[list(groups.keys())[i]]['items'][j]['description']
        else:
            id_text_dict[list(groups.keys())[i]] = id_text_dict[list(groups.keys())[i]]
            + groups[list(groups.keys())[i]]['items'][j]['name'] + " " + " "
```

```
#text cleaning function
import re
import string
import demoji
def clean_text (text):
    text = text.lower()
    text = re.sub(r'https?:\/\/.*[\r\n]*', ' ', text) #remove http-urls
    text = re.sub(r'\S*@\S*\s?',' ', text) #remove emails
    text = re.sub(r'\S*.ru\S*\s?', ' ', text) #remove other urls.ru
    text = re.sub(r'\S*.com\S*\s?', ' ', text) #remove other urls.com
    text = text.replace('—',' ') #remove dash
    text = re.sub('-', ' ', text) #replace hypthen with space
    text = re.sub(r'\d+', ' ', text)
    text = text.replace('\n',' ') #remove \n
    text = "".join(c for c in text if c not in string.punctuation and c not in extra_signs) #remove punctuation
    text = ''.join([i for i in text if not i.isdigit()]) #remove digits
    text = demoji.replace(text,' ') #remove emoji
    text = text.strip() #remove extra spaces
    text = re.sub(r'   +',' ', text) #remove extra spaces
    return (text)
```

```
#Lemmatization
from nltk.corpus import stopwords
from pymystem3 import Mystem
#all signs and words to remove
my_stop_words = ['instagram', 'insta', 'fb', 'facebook', 'whatsapp', 'telegram', 'youtube', 'twitter', '
                 'vk', 'весь', 'это', 'свой', 'группа', 'сообщество', 'паблик', 'самый', 'club', 'который',
                 'наш', 'студент',
                'спбгу', 'спбга',
                'санкт', 'спбгу', 'студентов','также','которые','россии','свои','новости','группы','день',
                'наши','каждый','правила', 'жизни','мемы','нам','сообщества', 'можете','бизнес'
                'петербург', 'петербурге', 'людей',
                'петербурга', ]
extra_signs = ['«', '»']
mystem = Mystem()
all_stopwords = stopwords.words("russian") + stopwords.words('english') + my_stop_words

def preprocess_text(text):
    tokens = mystem.lemmatize(text)
    tokens = [token for token in tokens if token not in all_stopwords
              and token != " "]
    text = " ".join(tokens)

    return text
```

```
#tokenization
from nltk.tokenize import word_tokenize
df_gr['groups_text_tokens']=df_gr['groups_text']
for i in range(0,len(df_gr)):
    df_gr['groups_text_tokens'][i] = word_tokenize(df_gr['groups_text_tokens'][i])
```

```
#vectorizer
tfidf = TfidfVectorizer(
    max_features = 4000,
    stop_words = all_stopwords
)
tfidf.fit(df_gr.groups_text2)
text = tfidf.transform(df_gr.groups_text2)
df_gr['groups_text2'] = df_gr['groups_text']
```

```
#define optimal number of clusters
def find_optimal_clusters(data, max_k):
    iters = range(2, max_k+1, 1)

    sse = []
    for k in iters:
        sse.append(kmeans(n_clusters=k,  random_state=552).fit(data).inertia_)
        print('Fit {} clusters'.format(k))

    f, ax = plt.subplots(1, 1)
    ax.plot(iters, sse, marker='o')
    ax.set_xlabel('Cluster Centers')
    ax.set_xticks(iters)
    ax.set_xticklabels(iters)
    ax.set_ylabel('SSE')
    ax.set_title('SSE by Cluster Center Plot')

find_optimal_clusters(text, 20)
```

```
#clusteting algorithms
from sklearn.cluster import KMeans
clusters = KMeans(n_clusters=3, init='k-means++', n_init=10, max_iter=300, tol=0.0001,  verbose=0,
                  random_state=552, copy_x=True, n_jobs=None, algorithm='auto').fit_predict(text)
```

```
# getting key words
def get_top_keywords(data, clusters, labels, n_terms):
    df = pd.DataFrame(data.todense()).groupby(clusters).mean()

    for i,r in df.iterrows():
        print('\nCluster {}'.format(i))
        print(', '.join([labels[t] for t in np.argsort(r)[-n_terms:]]))

get_top_keywords(text, clusters, tfidf.get_feature_names(), 30)
```

```
#setting posts from the wall
class posts_features():
    def __init__(self, posts):
        self.post_text = []
        self.repost = []
        self.date = []
        self.attachments = []
        self.postType = []
        self.id = []
        self.length = len(posts)

        for post in posts:
            if 'owner_id' in post:
                self.id.append(post['owner_id'])
            else:
                self.id.append("no data")

            if 'text' in post:
                self.post_text.append(post['text'])
            else:
                self.post_text.append("no data")

            else:
                self.repost.append("no data")

            if 'date' in post:
                self.date.append(datetime.datetime.fromtimestamp(post['date']))
            else:
                self.date.append("no data")

            if 'attachments' in post:
                self.attachments.append(post['attachments'])
            else:
                self.attachments.append("no data")

            if 'post_type' in post:
                self.postType.append(post['post_type'])
            else:
                self.postType.append("no data")

    def export2DataFrame(self):

        return pd.DataFrame(np.array([self.id, self.post_text,self.repost,
                                      self.date,self.attachments,self.postType]).transpose(),
                            index = range(self.length),
                            columns = ['id','text','copy_history','date','attachments','post_type'])

pf = posts_features (vkapi.wall.get(owner_id = str(user_ids[8]),
                                    count=100,
                                    v = 5.103,)['items'])
pf.export2DataFrame()
```

```
pfs = []
private_profile_w = 0
inactive_profile_w = 0
for us_id in user_ids:
    try:
        tmpResponse = vkapi.wall.get(owner_id = str(us_id),
                                     count=100,
                                     v = 5.103)
        if tmpResponse['count']:
            pfs.append(posts_features(tmpResponse['items']).export2DataFrame())
        else:
            pfs.append(pd.DataFrame(np.array([[us_id], [""],[""],[""],[""]]).transpose(),
                                    columns = ['id','text','copy_history','date','attachments','post_type']))
        time.sleep(0.4)
    except vk.exceptions.VkAPIError as e:
        pfs1.append("private profile")
        if e.code == 30 or e.code == 7:
            private_profile_w = private_profile_w + 1
        if e.code == 18:
            inactive_profile_w = inactive_profile_w + 1
print("loaded_profiles: ", len(pfs))
print("private profiles: ", private_profile_w)
print("inactive_profiles: ", inactive_profile_w)
```

```
#getting list of friends
session = vk.Session(access_token='token')
vkapi = vk.API(session)

friends = {}
private_profile_f = 0
inactive_profile_f = 0

for us_id in user_ids:
    try:
        time.sleep(0.4)
        friends[us_id] = vkapi.friends.get(v = 5.103,
                    user_id =  us_id)
    except vk.exceptions.VkAPIError as e:
        if e.code == 30 or e.code == 7:
            private_profile = private_profile_f + 1
        if e.code == 18:
            inactive_profile = inactive_profile_f + 1
        else:
            private_profile = private_profile_f + 1
print("loaded_profiles: ", len(friends))
print("private profiles: ", private_profile_f)
print("inactive_profiles: ", inactive_profile_f)
```

```
#getting lists of followers
session = vk.Session(access_token='token')
vkapi = vk.API(session)

followers = {}
private_profile_fl = 0
inactive_profile_fl = 0

for us_id in user_ids:
    try:
        time.sleep(0.35)
        followers[us_id] = vkapi.users.getFollowers(v = 5.103,
                    user_id =  us_id)
    except vk.exceptions.VkAPIError as e:
        if e.code == 30 or e.code == 7:
            private_profile = private_profile_fl + 1
        if e.code == 18:
            inactive_profile = inactive_profile_fl + 1
        else:
            private_profile = private_profile_fl + 1
print("loaded_profiles: ", len(followers))
print("private profiles: ", private_profile_fl)
print("inactive_profiles: ", inactive_profile_fl)
```

```
#examples of stats operations
df_friends.groupby (by = 'cluster_lbl').friends_count.mean()
df_followers.groupby (by = 'cluster_lbl').fllwrs_count.mean()
df_friends.groupby (by = 'cluster_lbl').friends_count.mean()
flwr = []
for i in range(1,8):
    flwr.append(df_followers.loc[df_followers['profile'] == float(i)]['fllwrs_count'])
stats.f_oneway(*flwr)
df_followers.groupby (by = 'profile').ff_ratio.mean()
ff_rat = []
for i in range(1,8):
    ff_rat.append(df_followers.loc[df_followers['profile'] == float(i)]['ff_ratio'])
stats.f_oneway(*ff_rat)
```

```
#student grades processing examples
file1 = '....\\MiBA Final project\\выгрузка.xlsx'
file2 = '...\\MiBA Final project\\courses.xlsx'

# Load spreadsheet: xls
xls1 = pd.ExcelFile(file1)
xls2 = pd.ExcelFile(file2)
df1 = xls1.parse(0)
df2 = xls2.parse(0)
df_merged = pd.merge(df1, df2, on=(df1.columns[4]), how='left')
df_merged.columns = ['Surname', 'Name', 'Group', 'Curriculum', 'Course',
                     'Grade1', 'Grade2', 'Semester', 'Date', 'Elective', 'Field']
df_merged['FullName'] = df_merged.Surname + ' ' + df_merged.Name
df_merged['Grade'] = df_merged.Grade1 + ', ' + df_merged.Grade2

# Changing grades to numeric values
df_full['Grade'].replace({'Отлично, A': 5.0, 'Отлично, B': 4.7, 'Хорошо, B': 4.3,
                          'Хорошо, C': 4.0, 'Хорошо, D': 3.7, 'Удовлетворительно, D': 3.3,
                          'Удовлетворительно, E': 3.0, 'Зачтено, A': 5.0,
                          'Зачтено, B': 4.5, 'Зачтено, C': 4.0, 'Зачтено, D': 3.5, 'Зачтено, E': 3.0}, inplace=True)
df_full['Grade'] = df_full['Grade'].astype(float)
```

```
# Measuring academic performance of students of a Financial Management specialization
finance1 = finance.groupby(['FullName', 'Field'])['Field'].count().to_frame(name = 'NumberOfCourses').reset_index()
finance2 = finance.groupby(['FullName', 'Field'])['Grade'].mean().to_frame(name = 'Average').reset_index()
finance_merged = pd.merge(finance1, finance2, on=('FullName', 'Field'), how='left')
# Measuring academic performance of all GSOM students at once
df_full1 = df_full.groupby(['FullName', 'Field'])['Field'].count().to_frame(name = 'NumberOfCourses').reset_index()
df_full2 = df_full.groupby(['FullName', 'Field'])['Grade'].mean().to_frame(name = 'Average').reset_index()
df_full_merged = pd.merge(df_full1, df_full2, on = ('FullName', 'Field'), how = 'left')
# Pivoting df_full2 for the further analysis
df_full2_pivoted = pd.pivot_table(df_full2, values = 'Average', index = 'FullName', columns = 'Field').reset_index()
print(df_full2_pivoted.head())
```

```
# Applying the elbow method to check how many clusters there are
from matplotlib import pyplot as plt
import seaborn as sns
from scipy.cluster.vq import kmeans, vq

distortions = []
num_clusters = range(1, 6)
scaled_features = ['scaled_fin', 'scaled_gen', 'scaled_gmu', 'scaled_hr',
                   'scaled_it', 'scaled_lang', 'scaled_log', 'scaled_mark']

# Create a list of distortions from the kmeans function
for i in num_clusters:
    cluster_centers, distortion = kmeans(df_full2_pivoted[scaled_features], i)
    distortions.append(distortion)

# Create a data frame with two lists - num_clusters, distortions
elbow_plot = pd.DataFrame({'num_clusters': num_clusters, 'distortions': distortions})

# Creat a line plot of num_clusters and distortions
sns.lineplot(x='num_clusters', y='distortions', data = elbow_plot)
plt.xticks(num_clusters)
plt.show()
```

```
# Creating centroids with kmeans for 2 clusters
cluster_centers,_ = kmeans(df_full2_pivoted[scaled_features], 2)

# Assigning cluster labels and print cluster centers
df_full2_pivoted['cluster_labels'], _ = vq(df_full2_pivoted[scaled_features], cluster_centers)
print(df_full2_pivoted.groupby('cluster_labels')[scaled_features].mean())

# Plotting cluster centers to visualize clusters
df_full2_pivoted.groupby('cluster_labels')[scaled_features].mean().plot(legend=True, kind='bar')
plt.show()
```