Санкт-Петербургский государственный университет

Кафедра математической лингвистики

Направление: «Лингвистика»

Образовательная программа: «Прикладная и экспериментальная лингвистика»

Профиль: «Компьютерная лингвистика и интеллектуальные технологии»

**Кластеризация языковых выражений в корпусе текстов на основе стохастического ранжирования**

Выпускная квалификационная работа   
соискателя на степень магистра филологии

**Букия Григория Теймуразовича**

Научный руководитель

к.ф.н., доц. Митрофанова О.А.

Рецензент: Тарелкин А.В.,

руководитель группы инструментов

оценки качества, «Яндекс»

Санкт-Петербург

2016

# Введение

В последние годы электронные корпуса становятся всё объемнее и разнообразнее, а количество информации в интернете увеличивается в геометрической прогрессии. Такой объем данных зачастую не поддается ручной обработке. Возникает естественная задача автоматически упорядочить коллекцию, объединяя в одну группу тематически близкие документы. Стандартные методы кластеризации, решающие данную задачу, не позволяют, однако же, определить тему той или иной группы. Если определять тему вручную, можно столкнуться с рядом проблем. Во-первых, для этого потребуется прочитать несколько документов из каждой группы – это далеко не всегда удобно. Во-вторых, очень часто автоматическая классификация, учитывающая внутренние свойства документов, не вполне согласуется с классификацией, выполненной человеком. Поэтому в качестве меток для каждого класса удобно использовать ключевые слова или выражения, характерные именно для соответствующего класса.

Цель данной работы – решение задачи автоматической кластеризации новостных документов и расстановка списка тематических меток для каждого класса. В качестве тематических меток выделяются не только ключевые слова, но и биграммные конструкции, при этом учитывается степень связи элементов каждой биграммы.

Традиционные методы выделения тематических меток словам, отвечающим сразу нескольким классам ставят меньший вес, чем словам, характерным исключительно для данного класса. Для новостных документов это не вполне отвечает интуитивному представлению иерархии ключевых слов: такие темы, как *война в Сирии*, должны, как нам кажется, иметь больший вес, поскольку задают контекст всего документа или кластера. Поэтому в работе предлагается новый, двухэтапный подход к составлению списка тематических меток.

Работа состоит из трех глав в соответствии с решаемыми задачами.

Первая глава посвящена кластеризации. В ней определяются понятия классификации и кластеризации, описываются наиболее используемые методы кластерного анализа и предлагаются различные метрики для оценки качества кластеризации. В последнем разделе дается общий обзор применения методов кластеризации в лингвистических задачах.

Вторая глава описывает лингвистический контекст нашего исследования – грамматику конструкций. Это молодая отрасль науки, изучающая сложным образом организованные объекты языка, функционирующие как единое целое, и возникающие в ходе взаимодействия и взаимопроникновения явлений морфологического, синтаксического, семантического и других уровней языка. Теория, развитая в работах А. Стефановича и Ст. Гриса, используется в данной работе при извлечении осмысленных биграммых меток. В следующих разделах главы обозреваются научные работы в области проставления тематических меток и выделения ключевых слов.

Третья глава содержит эксперимент, соответствующий поставленной цели. Эксперимент проводится в несколько этапов. На первом этапе производится разделение документов на два кластера. На втором этапе выделяются ключевые слова, характерные для каждого кластера. На третьем этапе полученные списки ключевых слов расширяются биграммными конструкциями, и полученные конструкции оцениваются по степени тематической направленности. На четвертом этапе после повторной кластеризации небольшим группам документов проставляются тематические метки.

Для кластеризации документов используются методы, реализованные в библиотеке Scikit-learn языка Python. В ходе работы была написана программа, реализующая описанные эксперименты. Мы использовали наиболее популярные статистические критерии, необходимые для выделения ключевых слов и конструкций, описанные в монографии А.И. Кобзаря «Прикладная и математическая статистика».

Результаты, полученные в ходе выполнения данной работы, могут найти свое применение при разработке новостных порталов. Главная идея и отличительная особенность данной работы – двухэтапная кластеризация для выделения ключевых слов – основывается на идее условного разделения документов на два класса: «серьезные» и «несерьезные», лексика которых существенно отличается. Такое условие характерно именно для новостных корпусов.

Приступая к решению поставленных задач, отметим неоценимую помощь, оказанную доцентом кафедры математической лингвистики О.А. Митрофановой, при подготовке данной работы.

# Основные идеи и методы кластерного анализа

## История кластерного анализа

Классификация была издревле известна человечеству. Прообраз этого понятия можно найти в первых строках книги Бытия (Быт. 1:21). Известны классические примеры классификации у Платона и Аристотеля [Новая философская энциклопедия, 2001], однако систематизация процесса классификации долгое время не проводилась. В начале XIX века французский ботаник Огюстен Декандоль [Брокгауз, Ефрон 1907] предложил свою теорию классификации и систематизации, названную впоследствии таксономией. Декантоль стремился классифицировать все существующие растения, объединяя их в однородные группы разных уровней, образующих иерархическую структуру (вид, род, семейство, класс, отдел). Данный метод вскоре получил широкое распространение и за пределами биологии. Теперь он положен в основу иерархических методов кластеризации.

Немецким биологом Ф. Гейнке был предложен метод группировки объектов по нескольким признакам. Всякий новый объект принадлежал той группе, к центру которой он ближе всего – идея, легшая в основу метода k средних.

Пионером применения базовых принципов кластеризации считается польский антрополог К. Чекановский. В 1913 году он предложил идею «структурной классификации» [Плюта 1980]: выделять компактные группы объектов. Для этого он разработал и оригинальный метод, применяемый при диагонализации признаковой матрицы.

В 1925 году советским гидробиологом П.В. Терентьевым был разработан метод корреляционных плеяд [Терентьев 1959] – это по-видимому первый алгоритм, направленный на выявление групп тесно коррелирующих признаков. Идеи этого алгоритма легли в основу многих пороговых алгоритмов на графах, например метода связных компонент.

Термин кластерный анализ впервые применил английский ученый Р. Трион [Trion 1939].

В 50-х годах ситуация стала развиваться значительно быстрее. Появились ЭВМ, способные обрабатывать данные гораздо быстрее человека. Алгоритмы усложнялись и совершенствовались, объемы данных росли и вскоре кластерный анализ завоевал прочное место в ряду прикладных дисциплин. Появилась возможность обрабатывать такое количество информации, которое прежде было человеку не под силу.

Следующие два десятилетия считаются золотым веком кластерного анализа. Тогда были получены основные результаты, изучен метод k-средних, иерархические процедуры, диагонализация и пр. Важную роль в этом сыграли и советские ученые [Мандель 1988].

Сегодня существует не меньше сотни методов кластеризации, которые применяются в тех или иных задачах, однако нет ни одного универсального алгоритма. На практике приходится не только выбирать наиболее подходящий для данной задачи алгоритм, но и настраивать его параметры. Чтобы оценить качество кластеризации, необходимы некоторые метрики, описание которых мы дадим в предстоящем разделе.

## Кластеризация как нечеткая классификация

Кластерный анализ – бурно развивающаяся область прикладной математики, тем не менее, специалисты обычно не оперируют какими-либо строгими определениями, опираясь лишь на интуитивное представление о кластеризации. Кластеризация – это нечеткая классификация, некоторое её приближение, основанное на взаимном расположении классифицируемых объектов. В основном своем значении термин *cluster* может быть переведен как *скопление*. Близкие объекты относятся к одному классу, а далекие объекты, наоборот, – к разным классам. Для этого между объектами должно быть введено некоторое расстояние. Результат кластеризации существенно зависит от выбора метрики.

Чтобы сформулировать и другие параметры процедуры кластерного анализа, нам следует в первую очередь определить самое базовое понятие – классификацию.

### Ограничения, определяющие классификацию

Представим себе несколько объектов, про каждый из которых можно сказать, обладает ли он данным признаком или нет. В результате объекты распределяются на две группы – обладающие и не обладающие признаком. Такую операцию называют *дихотомией*. Иногда может существовать более двух состояний по отношению к признаку, например, цвет глаз у людей, – в таком случае принято говорить о *классификации на несколько групп*.

Обычно при классификации вводят некоторые ограничения. Во-первых, должно однозначно определяться состояние объекта по отношению к признаку, т.е. группы не могут пересекаться. Во-вторых, состояние может зависеть только от внутренних свойств объекта и не должно прямо зависеть от других выбранных объектов. Условно говоря, разделение чисел на четные и нечетные – это классификация, т.к. четность числа не зависит от того, какие числа выбраны для проверки, а распределение чисел на большие и меньшие не является классификацией, потому что не выполняется ни одно из условий. Классификацию как логическое действие называют *делением*, а как математическую операцию – *факторизацией*.

Простейший пример факторизации – остатки от деления по модулю. Здесь множество объектов классификации – это натуральные числа, а признак, относительно которого происходит классификация – это остаток от деления по некоторому модулю. Если число имеет остаток 1, то оно относится к первой группе, если остаток 2, то ко второй и т.д. Каждое число может иметь ровно один остаток, поэтому оба ограничения выполнены, т.е. перед нами действительно классификация.

### Формальное описание классификации

Теперь формализуем полученное интуитивное определение, чтобы в дальнейшем как можно нагляднее показать разницу между классификацией и кластерным анализом.

Представим некоторое множество , элементы которого мы будем называть *объектами*. На множестве объектов определим функцию , называемую *признаком*. Эта функция каждому объекту ставит в соответствие номер группы, к которой он принадлежит. Примеры признаковой функции – цвет глаз человека или остаток от деления по модулю для некоторого числа.

Все объекты множества , которые относительно признака принадлежат к одной и той же группе называют *классом* . Разбиение множества на непересекающиеся классы , и есть *классификация*.

### Нечеткая и зависимость параметра от выборки

Не всегда признаковую функцию классификация можно определить однозначно. В некоторых случаях граница между классами достаточно условна и определяется пороговым значением параметра. Пороговые значения требует, например, классификация боксеров по весовой категории на тяжелую, среднюю и легкую.

Зачастую нельзя выделить единственный признак, классифицирующий объекты, поэтому используют методы многомерной классификации и факторного анализа: выделяются наиболее значимые признаки.

При нечеткой классификации предполагается, что параметр каждого объекта определяется однозначно и не зависит от других объектов. Иногда задание такой функции невозможно в принципе, например, в задаче разбиения группы людей на дружественные компании. Здесь само отношение дружбы подразумевает участие нескольких объектов.

На практике, во многих задачах использовать модель нечеткой классификации хотя и возможно, но само формирование признаковых описаний представляет большие трудности. Так, например, гораздо проще сравнить два романа и определить, что они написаны одним и тем же человеком, чем описать, на основе каких признаков они схожи. Точно так же можно определить, что данное письмо является спамом, сравнив его с другими письмами, уже помеченными как спам, чем описать все возможные признаки рекламной рассылки. В всех этих случаях удобнее разбить множество объектов на «дружные компании», в каждой из которых будут похожие объекты, а объекты разных классов, наоборот, будут отличаться.

### Метрическое пространство объектов

Итак, в некоторых задачах нам необходимо ввести понятие *расстояния* между объектами. Двум объектам , поставим в соответствие некоторое число . Естественно задать расстояние так, что для него выполнены стандартные свойства метрики, в частности, расстояние объекта до самого себя должно быть нулевым. Чем ближе объекты, тем скорее их следует отнести в одну группу и наоборот, чем они дальше друг от друга, тем скорее их следует отнести в разные группы.

Чтобы оценить близость двух объектов, нам, вообще говоря, понадобится какое-то математическое представление пространства . Один из традиционных способов – представить каждый объект в виде числового вектора, характеризующего его свойства. В таком случае, естественно расстоянием между объектами считать евклидовым расстоянием между векторами.

В задаче разделения группы людей на дружественные компании удобнее использовать графовое представление пространства, объекты считая вершинами, а ребра устанавливая в случае выполнения условия (дружбы), возможно, с некоторым «весом». Тогда естественное расстояние между двумя объектами – длина кратчайшего пути между соответствующими вершинами графа.

### Кластерный анализ как поиск оптимальной классификации

Прежде чем формально определить кластерный анализ, введем еще одно понятие – *критерий качества классификации*  *относительно метрики* . Критерий качества должен отражать насколько объекты одного класса близки друг к другу, а объекты разных классов друг от друга далеки. Классификация, оптимизирующая функционал качества, таким образом, группирует каждое скопление объектов в соответствующий класс.

Наконец, можно сформулировать следующее определение: кластеризация – поиск классификации, оптимальной относительно критерия качества. То есть кластеризация , где

### Особенности кластеризации

Окончив формальное описание, мы можем сказать о некоторых отличительных особенностях кластеризации. Итак, было замечено, что каждый объект классифицируется однозначно и значение соответствующей признаковой функции определяется его внутренними свойствами и не зависит от всего множества объектов. Не так при кластеризации. Процесс минимизации критерия качества требует знания всего пространства объектов. Замена нескольких объектов в может привести и к изменению оптимальной классификации.

Кластеризация существенно зависит от заданной метрики и критерия качества . Но структура этой зависимости непредсказуема и универсальных параметров нет.

Наконец, отыскание глобального минимума критерия качества не всегда возможно, особенно, когда пространство объектов велико. В таком случае ведется поиск локального минимума одним из оптимизационных алгоритмов.

## Алгоритмы кластерного анализа

### Неявное задание критериев качества в алгоритме кластеризации

Как мы уже сказали, главной задачей кластерного анализа является поиск оптимальной классификации. Результат зависит от выбора критерия качества и способа отыскания его экстремума. Существует немало *алгоритмов кластеризации*, решающих эту задачу. Получая на входе метрическое пространство объектов, эти алгоритмы производят классификацию.

Выделяется как минимум три цели кластеризации [Воронцов 2015]:

1. описать структуру пространства объектов, разбивая его на классы похожих объектов и исследуя области сгущения (выделить темы),
2. сократить объем данных, выделяя наиболее типичных представителей каждого кластера,
3. выделить специфические объекты, не относящиеся ни к одному из кластеров.

В первом случае число кластеров (тем) должно быть невелико, тогда как во втором случае в каждый кластер должны попасть только наиболее близкие объекты – поэтому ясно, что разные задачи требуют специального подхода, и универсального подхода быть не может.

Алгоритмы кластеризации, как правило, имеют параметры регулировки, с помощью которых ведется поиск наиболее оптимального результата. Вид этих параметров различный – это может быть число кластеров, скорость сходимости или иные константы в зависимости от алгоритма.

Существует немало работ, посвященных описанию алгоритмов кластеризации [Jain A. и др. 1999]. Мы не ставим перед собой задачу повторить эти труды и описать все возможные алгоритмы, но лишь описываем наиболее стандартные и используемые, для того чтобы оправдать выбор методов, предпринятый нами в практической части работы.

### Алгоритм выделения связных компонент

Выделение связанных компонент [Лагутин 2003] – один из простейших методов кластеризации. Пространство представляется в виде графа: вершинами являются объекты, а каждому ребру соответствует число – длина пути между соответствующими объектами. Фиксируется некоторый параметр и перекрываются все те ребра, длина которых больше параметра . Граф распадается на компоненты связности (между вершинами различных компонент не может быть пути, проходящего по ребрам графа). Каждая компонента и будет искомым кластером.

Такой алгоритм имеет ряд недостатков. Во-первых, количество кластеров плохо регулируется параметром , и для решения задачи требуется запускать алгоритм несколько раз. Во-вторых, алгоритм чувствителен к исходным данным: даже наличие узкой перемычки между сгущениями или разреженного фона приводит к неадекватному результату.

### Алгоритм кратчайшего пути

Модификацией связных компонент является алгоритм кратчайшего пути. На первом шаге строится остовное дерево с минимальной суммарной длиной ребер: вершины-объекты соединяются ребрами так, что получившийся граф не содержит циклов, при этом минимизируется сумма длин ребер полученного дерева.

В этом случае число кластеров задается входным параметром. Из остовного графа удаляется ребро наибольшей длины, оставшийся граф распадается на компонент, которые и будут искомыми кластерами.

В отличие от предыдущего, алгоритм кратчайшего пути позволяет регулировать число кластеров. Но он по-прежнему чувствителен к исходным данным и дает неадекватный результат при разреженных данных. Еще один недостаток данного метода – низкая скорость поиска остовного дерева с минимальной суммарной длиной ребер.

### Алгоритм ФОРЭЛ (Формальный элемент)

Задается некоторая точка (формальный элемент) метрического пространства и параметр . Рассматриваются только те точки-объекты, расстояние от которых до меньше . Затем точка перемещается в центр тяжести выбранных объектов и операция повторяется до стабилизации в центре области сгущения.

Этот алгоритм имеет множество вариаций и модификаций [Загоруйко 1999], которые отличаются способом объединения окрестностей формальных элементов в кластеры, способами изменения параметра , способами выбора начального положения формальных элементов, однако построены на одном и том же принципе. Значительно усложняется процедура кластеризации, если пространство нелинейно. Доказано, что он сходится за конечное число шагов.

### Минимизация функционала качества

Мы уже описали, что задача алгоритма кластеризации – минимизировать некоторый критерий (функционал) качества, то есть подобрать оптимальную классификацию. Есть множество функционалов качества, среди которых нельзя назвать один «правильный». Отметим несколько естественных критериев.

Во-первых, среднее внутрикластерное расстояние должно быть как можно меньше:

С другой стороны, среднее межкластерное расстояние должно быть как можно больше:

Чтобы учесть и межклассовые, и внутриклассовые расстояния, в качестве функционала качества используют отношение этих величин: .

Для поиска минимума функционала можно использовать один из оптимизационных алгоритмов, например, EM-алгоритм и др.

### Разделение смеси распределений

Один из статистических методов кластеризации – разделение смеси распределений. Предполагается, что каждый кластер – реализация некоторой случайной величины с плотностью , а – неизвестная априорная вероятность появления объектов из кластера . Тогда плотность распределения объектов выборки равна смеси плотностей .

Как правило, предполагается, что каждый кластер описывается гауссовской плотностью, т.е. представляет собой эллипсоид, оси которого параллельны осям координат.

Разделение смеси – стандартная задача, для которой используют разные методы, в частности, EM-алгоритм, корректирующий параметры распределений. На первом этапе по формуле Байеса вычисляются скрытые переменные, значение которых равно вероятности того, что объект принадлежит кластеру . На втором этапе уточняются параметры каждого кластера с учетом скрытых переменных.

### Метод k-средних

Один из самых используемых алгоритмов кластеризации [Мандель 1988], метод k-средних, делит пространство объектов на кластеры близкого объема, минимизируя разброс – среднеквадратическое отклонение объектов кластера от соответствующего центра масс. Для его работы требуется задание числа кластеров.

Более формально, если – выборка пространства объектов, подбирается такой классификатор , что если – центры масс кластеров, то величина

достигает минимума.

Среднеквадратический разброс характеризует, насколько элементы одного кластера согласованы друг с другом. Однако такая метрика имеет ряд недостатков.

* Во-первых, при такой постановке предполагается, что кластеры выпуклы и изотропны, т.е. распределение значений внутри кластера не зависит от координат. Некоторые модификации позволяют ослабить это требование.
* Во-вторых, среднеквадратический разброс – ненормированная метрика. Известно лишь, что эталонным значением является нуль, а меньшее значение разброса предпочтительнее.
* Отдельной проблемой представляется работа с пространствами очень больших размерностей, тогда зачастую векторы сильно разрежены. Применение алгоритмов понижения размерностей (например, PCA) существенно сокращает скорость вычисления, зачастую при этом понижая шум.

Принцип поиска оптимального значения функционала качества похож на действие алгоритма Форэл. Рассматриваются приближенные значения центров кластеров. Каждый объект выборки относится к тому кластеру, к центру которого он ближе, затем значения центров пересчитываются.

Остановимся на этом более подробно. Сначала задаются некоторые априорные значения центров масс кластеров (в западной литературе центры масс кластеров называют центроиды): . Затем циклично повторяют следующие действия. На первом шаге каждый элемент выборки относят к некоторому кластеру в соответствии к ближайшим центроидом. На следующем шаге в каждом кластере заново считается центр масс и получается новая система центров . Затем принадлежность каждого элемента определяется заново и т.д. до сходимости – пока перемещение центроидов не станет меньше некоторого фиксированного порога. Доказано, что данный алгоритм сходится за конечное число шагов, хотя возможно ­– в точке локального, а не глобального минимума функционала качества.

Метод k-средних по сути представляет собой упрощенную версию EM-алгоритма при разделении смеси: в первом случае каждому объекту однозначно соответствует единственный кластер, тогда как во втором случае каждому объекту соответствует некоторое вероятностное распределение на всем множестве кластеров.

Этот метод – один из самых используемых в прикладных задачах, существует немало модификаций, позволяющих учесть тонкие настройки оптимизации или формы кластеров. Он крайне чувствителен к начальным приближениям центров кластеров вплоть до того, что разные стартовые позиции могут привести к разным локальным минимумам функционала качества. С этой проблемой есть разные способы борьбы, один из которых – параллельный запуск нескольких случайных приближений. Иной метод, предназначенный для решения проблемы поиска глобального минимума, называют схемой инициализации k-means++. В качестве априорного приближения выбираются наиболее удаленные друг от друга точки. Доказано, что таким путем получается заведомо лучшая кластеризация, чем при генерации случайных центров.

### Метод сигнала родства (Affinity Propagation)

Иную структуру имеет метод сигнала родства: объекты выборки обмениваются сигналами, характеризующими степень близости между ними, затем информация корректируется сообщениями, полученными от других источников. Данная операция повторяется до сходимости. В результате данные можно описать с помощью небольшого набора элементов, каждый из которых является наиболее характерным представителем своего кластера.

Метод сигнала родства не позволяет заранее задать количество кластеров. Используется он при выделении кластеров, описывающих как можно больше типов данных – внутри каждого кластера выбран типичный представитель, описывающий все элементы остальные элементы кластера.

Существенным недостатком сигнала родства является высокая сложность алгоритма – порядка , где – размер выборки, а – число итераций до сходимости. Это серьезное препятствие, если объем данных слишком велик (как часто бывает в лингвистике). В некоторых случаях можно достигнуть ускорения если матрицу исходных данный задавать в виде разреженной.

Итак, вернемся к описанию самого алгоритма. Сигнал, передаваемый между элементами и , бывает двух типов: выразительность (responsibility) и характерность (availability). Выразительность определяет насколько по сравнению с другими элементами элемент , будучи представителем, выражает признаки элемента . Характерность , наоборот, отражает, насколько элемент характерен для элемента как для представителя класса. В результате алгоритма представители класса выбираются в том случае, если элементы с одной стороны имеют достаточно много близких элементов, с другой – не описывают друг друга.

Формально выразительность элемента как представителя элемента задается следующим образом:

где – степень близости, а характерность, в свою очередь:

Начальные значения , считаются нулями, параметры пересчитываются до сходимости.

### Сдвиг среднего (Mean Shift)

Данный алгоритм кластеризации применяют для выделения контрастных пятен на однородном фоне в метрическом пространстве объектов. В частности, он используется для сегментации изображений. Подобно методу k средних, здесь итерационно строится оценка положения центроидов. После построения первичной оценки происходит пересчет и фильтрация.

Алгоритм основан на естественном предположении – центроиды (точки сгущения кластеров) находятся в экстремумах плотности. Форма кластера определяется ядром , как правило, рассматривают гауссово ядро: , предполагая кластеры имеющими нормальное распределение. Тогда, если обозначить кластер, соответствующий точке , а функцию сдвига рассмотреть как взвешенное среднее по объектам кластера:

Новое приближение центроида есть сдвиг прежней величины на функцию :

### Спектральная кластеризация (Spectral clustering)

Еще один алгоритм, который работает на графовой структуре данных, которая, между прочим, часто используется в лингвистике. Матрица данных (попарных расстояний между объектами) при этом должна быть разрежена. Спектральный метод требует задания количества кластеров. Стоит заметить, что с увеличением их числа качество кластеризации заметно понижается.

В основе спектрального метода лежит алгоритм нормализованного разделения (normalized cuts), основанного на вычислении собственного вектора симметричного нормализованного лапласиана, которому соответствует второй с конца по величине собственное число. При разделении объектов максимизируется вес кластеров.

Спектральный метод особенно распространен при обработке изображений.

### Иерархическая кластеризация

Отдельную группу алгоритмов представляют собой иерархические методы (методы таксономии). Они строят не один классификатор всего пространства, а целая система классификаторов. На первом уровне несколько типов объектов, на втором уровне каждый тип подразделяется еще на несколько и т.д. В результате получается иерархическое дерево – дендрограмма, в корне которого единичные кластеры, а вершина – всё пространство объектов.

Иерархические алгоритмы бывают двух видов – нисходящие и восходящие. Первые делят пространство на всё более мелкие части.

Вторые – восходящие – используются чаще и представляют некоторый интерес, потому что проделывают операцию, обратную традиционному разделению при кластеризации. Они объединяют кластеры.

На первом шаге каждый объект считается одним кластером, между которыми естественным образом задана функция расстояния: . Объединение происходит следующим образом. На каждом шаге два самых близких кластера и объединяются в единый кластер . Расстояние между и прочими кластерами пересчитывается по формуле, в самом общем виде предложенной Лансом и Уильямосом []:

где , , и – числовые параметры. Через это представление выражаются почти все разумные метрики пространства кластеров.

Чаще всего используются следующие способы вычисления расстояний :

1. Расстояние ближнего соседа:
2. Расстояние дальнего соседа:
3. Среднее расстояние:
4. Расстояние между центрами:
5. Расстояние Уорда:

### DBSCAN

Данный метод использует естественное представление о кластере как об области высокой концентрации элементов, за пределами которого пространство разрежено, т.е. концентрация элементов мала. Такой общий подход позволяет обнаруживать кластеры самой причудливой формы, тем более он не требует предположения о гауссовском распределении векторов.

В процессе работы алгоритма отбираются элементы, для которых считается плотность. Если плотность меньше критического уровня, элемент помечается как шум. В результате будет выделено несколько кластеров, элементы которого обладают высокой плотностью. Как правило, предельный объем кластера ограничивается.

Отметим еще, что в отличие от прочих алгоритмов, DBSCAN помечает некоторые элементы как выбросы. В зависимости от конкретной задачи это может способствовать сохранению стабильности при кластеризации, а может, наоборот, нарушить структуру кластеров.

## Оценка качества кластеризации

На первый взгляд, оценка качества кластеризации – задача почти тривиальная – достаточно сосчитать количество ошибочно классифицированных объектов. Но это не вполне верно.

В первую очередь, оценка не должна зависеть от объема выборки и конкретного кластера. Ведь десяток ошибочных элементов в кластере размером пятьдесят и тысяча – не одно и то же. Есть и другие разумные требование на оптимальную кластеризацию.

Если при тестировании алгоритма известна реальная классификация, к которой следует стремиться, то соответствующая оценочная метрика должна отражать, насколько прогнозное разбиение близко к эталонному. На практике чаще всего такая классификация известна лишь для небольшого числа объектов, поэтому разумно использовать другие метрики, которые оценивают качество классификации исходя из ее внутренней структуры. В сущности, они представляют собой разновидность тех самых критериев качества, которые алгоритм кластеризации должен оптимизировать.

Рассмотрим ряд таких метрик первого типа.

### Индекс Ранда

Итак, пусть имеется эталонная классификация объектов , с которой требуется сравнить прогнозную кластеризацию . Дополнительно определим:

* – число пар объектов, попавших в один класс и в один кластер,
* – число пар объектов, попавших в различные классы и различные кластеры,
* – общее число элементов.

Стандартный (смещенный) индекс Ранда (Rand index) задается по следующей формуле:

При совпадении прогнозной кластеризации и эталонной классификации он будет иметь единичный вес. С увеличением числа ошибок значение индекса уменьшается.

Эта метрика не зависит от «переименования» кластеров: если два объекта одного класса попали в один кластер, это хорошо вне зависимости от того, какой именно это кластер. Иными словами, не требуется никакого отождествления классов и кластеров для оценки. Использование пар позволяет учитывать размер соответствующих кластеров при подсчете ошибок.

Однако кластеризация, которая случайным образом определяет принадлежность к кластеру и никак не учитывает расстояние между объектами, согласно данной метрике будет иметь положительный вес. Естественным было бы требование для такой кластеризации иметь нулевую метрику.

Этому условию отвечает несмещенный индекс Ранда (Adjusted Rand index), задаваемый следующим образом:

Он требует неоднократного запуска процесса кластеризации.

Существенным достоинством индекса Ранда является то, что он не учитывает форму кластеров, поэтому разные подходы и методы кластеризации тем самым сравнимы между собой.

Несмещенный индекс Ранда принимает значения на отрезке : отрицательная величина говорит о неудачной кластеризации, значения около нуля говорят о том, что данный метод кластеризации не зависит от классификационных признаков. Значения, близкие к единице говорят об удовлетворительной кластеризации.

### Взаимная информация

Оценка взаимной информации – класс широко используемых метрик для измерения корреляции векторов. В рассмотренном случае она используется при сравнении эталонной и прогнозной классификации и характеризует их согласованность. Данный метод, как и предыдущий, не зависит от взаимного соответствия между классами и кластерами.

Существует две модификации взаимной информации – нормированная и несмещенная. Первая стала уже традиционной и хорошо описана в литературе, тогда как вторая всё чаще применяется на практике, когда используются приближенные методы кластеризации. Как и индекс Ранда, взаимная информация не зависит от формы и структуры кластеров.

Как стандартная взаимная информация, так и ее нормированная модификация являются смещенными оценками.

Обозначим два разбиения пространства объектов U и V. Тогда энтропия неравномерности для каждого разбиения задается по формуле:

где – вероятность того, что случайный элемент попадет в класс . Аналогичная величина определяется и для :

при этом вероятность попадания случайного элемента в класс обозначается величиной .

Взаимная информация определяется формулой:

где – вероятность того, что случайный объект попадет в оба класса и .

Сама по себе взаимная информация существенно зависит от абсолютных значений выборки. Поэтому используется нормированный вариант, определяемый следующим образом:

Обе метрики не учитывают стохастической составляющей и, кроме того, их значения возрастают при увеличении числа кластеров, хотя сама по себе кластеризация может быть неудачной.

Статистический метод усреднения был предложен в статье [Vinh, Epps & Bailey, 2009]. Нам понадобиться обозначить , . Тогда верна следующая оценка:

Используя эту величину, можно вычислить усредненную взаимную информацию:

### Точность, полнота и V-мера

Если известна эталонная классификация объектов, то можно определить ряд естественных метрик, характеризующих степень близости прогнозной классификации.

В статье [Rosenberg and Hirschberg, 2007] описываются два интуитивно понятных требования для оценки качества кластеризации.

Во-первых, каждый кластер должен содержать элементы одного класса. Во-вторых, все члены одного класса должны попасть в один кластер. Исходя из этого строятся следующие метрики.

Для эталонной классификации и прогнозной кластеризация обозначим величину, называемую условной энтропией разбиения:

а – энтропию класса, определенную по формуле:

где – общее число элементов, и – число элементов соответственно в классе и кластере . Аналогично зададим и .

Точность (homogeneity), характеризующая величину ошибки первого рода, т.е. количество элементов разных классов, попавших в один кластер, определяется по формуле:

Аналогично полнота (completeness), характеризующая, в свою очередь, величину ошибки второго рода – количество элементов разных классов, оказавшихся в одном кластере, – задается по формуле:

Их комбинация (-мера), учитывающая и величину полноты, и точности, является средним гармоническим:

Все величины имеют значения в пределах от нуля до единицы. Использование данных метрик позволяет более детально оценить проблемы той или иной кластеризации: ошибки первого и второго рода. Это полезно при регулировке параметров алгоритма.

С другой стороны, V-мера чувствительна по отношению к шуму (т.е. смещена). Случайное разбиение, таким образом, не всегда имеет нулевую метрику. Кроме того, все три метрики зависит от размера выборки, числа кластеров и структуры классификации.

Этой проблемой можно пренебречь, когда объем выборки более тысячи, а число кластеров менее 10.

### Коэффициент силуэта

Как мы уже замечали, если эталонная классификация неизвестна, следует использовать один из коэффициентов качества, описанных ранее. Такие метрики, естественно, весьма условны. Во-первых, даже эталонная классификация не всегда им удовлетворяет. Во-вторых, всякий алгоритм кластеризации минимизирует тот или иной функционал качества и, являясь эталонным для одного, может и не являться таковым для другого.

Как бы то ни было, критерии качества весьма показательны – они позволяют понять, насколько полученное разбиение однозначно. Среди прочих используется и коэффициент силуэта.

Задается две величины:

a – среднее расстояние между данным элементом и всеми точками соответствующего кластера.

b – среднее расстояние между данным элементом и всеми точками другого ближайшего к нему кластера.

Сам коэффициент определяется по формуле:

Чаще всего его значение оказывается в пределах отрезка , однако это выполнено не всегда. Он характеризует, насколько кластеры отделены друг от друга, иными словами, насколько кластеризация однозначна.

Он достойно работает, если кластеры выпуклы, но, если они более сложной формы, например, при использовании алгоритма DBSCAN, величина метрики падает, и он перестает отражать реальную ситуацию.

Мы достаточно исследовали алгоритмы кластеризации, осталось сказать, как этот чисто математический метод нашел свое применение в лингвистике и позволил решить ранее, казалось бы, неподъемные задачи.

## Кластеризация в лингвистике

Электронные корпусы текстов позволили широко использовать методы кластеризации для решения самых разных лингвистических задач. Классический пример применения кластеризации – статья [Schütze, 1998], в которой комбинация аггломеративного и EM-алгоритма использовалась для снятия лексической неоднозначности. Основное положение подобных методов заключается в том, что схожесть контекста дает основание считать одинаковым значения обеих лексем. Подобные подходы использовались в работах [Lin, 1998, 2002] и по-прежнему актуальны [McCarthy и др., 2016]. Обзор работ по применению методов кластеризации в задачах снятия лексической неоднозначности можно найти в статье [Navigli, 2009].

Оригинальная идея была предложена в статье [Biemann, 2006]. В ней решалась задача частеречной разметки корпуса на основе контекста и был предложен оригинальный метод графовой кластеризации, оптимизированный для данной задачи.

Нашел свое применение кластерный анализ и в задаче построения тонального словаря [Четверкин, 2013] по корпусу текстов.

Как и в нашей работе, нередко методы кластеризации используются в задачах тематического моделирования [Basu, Murphy, 2013]. В частности, их используют для классификации научных текстов, насыщенных специальной терминологией [Savova и др., 2005].

Особую роль методы кластерного анализа играют в компьютерной текстологии. В рамках исследований по определению силы связей между списками рукописей на основе данных об узлах разночтений, проводимых А.А. Алексеевым, Е.Л. Алексеевой (Кузнецовой) и Д.М. Мироновой, используется вариант аггломеративного кластерного анализа, в процессе которого близкие тексты объединяются в стемму [Миронова 2016].

В завершении главы скажем, что широкое применение кластерного анализа в лингвистике привело к появлению специальных библиотек, адаптированных для работы с текстовыми документами, в частности, уже упомянутой библиотеки Scikit-learn для языка Python, с которой мы работали во время проведения экспериментов.

# Лингвистические основания автоматической кластеризации текстов по ключевым словам и конструкциям

## Лингвистика конструкций и оценка связей в конструкциях

### Основные идеи лингвистики конструкций

Лингвистика конструкций – молодая отрасль науки, возникшая как альтернатива модульному, или атомарному подходу к явлениям языка и речи, опирающемуся на понятия основных единиц и отношений в рамках того или иного уровня языка. Ранее именно элементарные единицы и отношения привлекали внимание ученых, многим казалось, что именно в них кроются загадки естественного языка. Однако в языке и в текстах существуют конструкции – сложным образом организованные объекты, функционирующие как единое целое, возникающие в ходе взаимодействия и взаимопроникновения явлений морфологического, синтаксического, семантического и других уровней языка. В фокусе внимания лингвистики конструкций находятся воспроизводимые единицы текста, большие, чем слово: словосочетания разной степени устойчивости, фразеологизмы, идиомы, обороты, неоднословные целостности, явления «малого синтаксиса» и т.д. С одной стороны, этот класс явлений, наблюдаемых в корпусе, привлекает интерес компьютерных лингвистов, однако до сих пор представляет большую проблему в автоматической обработке корпусов текстов. Одна из причин заключается в том, что их природа и особенности функционирования не до конца изучены. Лингвистика конструкций предназначена как раз для того, чтобы отчасти восполнить возникший пробел.

Развитие лингвистики конструкций идет непростыми путями. Существует ряд отечественных и зарубежных теорий, среди которых обращают на себя внимание грамматика конструкций Ч. Филлмора [Fillmore 1988, Goldberg 1995, 2006, Tomasello 2003], теория лексических функций в рамках модели «Смысл <=> Текст» [Мельчук 1974/1999; ТКС 1984], конструктивный синтаксис Н.Ю.Шведовой [АК 1980], описание классов ядерных глагольных конструкций русского языка (моделей управления) и их трансформационных преобразований Ю.Д.Апресяна [Апресян 1967] и т.д. Есть немало последователей этого направления в современной российской лингвистике [Кузнецова 2007; Рахилина 2010; Овсянникова, Сай 2010; Оскольская, Овсянникова, Сай 2011].

Рассмотрим основополагающие идеи лингвистики конструкций, важные для решения наших исследовательских задач.

Признанным основоположником зарубежной школы исследователей конструкций является Ч. Филлмор, разработчик теории падежной грамматики [Филлмор 1981] и глава разработчиков первой в мире базы данных лексических конструкций FrameNet [Fillmore 1982]. В своих работах Ч. Филлмор и его коллеги опирались на очень общее определение конструкции: это языковое выражение, в котором наличествуют формальные или содержательные аспекты, которые невозможно вывести из значения или формы составных частей. Простейшим примером такой конструкции может служить выражение *железная дорога*, общее значение которой, подобно паззлу, хоть и состоит из отдельных частей, в сочетании воспринимается как неделимое. Лексические конструкции, согласно теории Филлмора, представляют собой единство формы и содержания. Форма конструкции фиксируется, во-первых, уже заданными элементами, а во-вторых, грамматическими, семантическими и пропозициональными ограничениями, которые естественным образом накладываются на переменные ячейки конструкции. Как правило, значение лексической конструкции не выводится из составляющих, хотя возможна, как более широкая, так и более узкая трактовка понятия: от свободных сочетаний лексических единиц до фразеологизированных структур.

Конструкциями по Ч. Филлмору могут являться языковые единицы любого уровня, если они обладают формой и содержанием: их элементами могут быть и морфемы, и слова, и целые предложения [Рахилина 2010]. В классической работе Ч. Филлмора и коллег [Fillmore et al. 1988] обсуждается многоуровневый анализ лексической конструкции *let alone* («не говоря уже о…») и продемонстрировано, как следует анализировать конструкции в качестве единицы грамматического описания на синтаксическом, семантическом и прагматическом уровне. С точки зрения синтаксиса в конструкции есть два обязательных компонента (A) *let* *alone* (B), требующие восполнения за счет обязательного третьего компонента (С), без которого конструкция не является законченной: например, *I doubt (C) he made colonel (A) let alone general (B).* На семантическом уровне данная лексическая конструкция может быть описана шкалой, или иерархией пропозиций (от более сильной (*to make colonel*) к более слабой (*to make general*). Прагматическая интерпретация конструкции связана с выразительностью отрицания, носителем которого является компонент (С).

Грамматика конструкций позволяет по-новому объяснить целый класс языковых выражений. Так, в своей пионерской работе [Goldberg, 1995] А.Голдберг предложила отказаться от традиционного взгляда на рамки глагольных валентностей: вместо того, чтобы приписывать именным группам, заполняющим позиции в рамке глагольных валентностей, статус глагольных актантов, исследовательница призывает их рассматривать наравне с самим глаголом как компоненты глагольных конструкций. Тем самым, в семантическом анализе глагольных контекстов появляется принципиально новое решение проблемы неоднозначности и разграничения глагольных значений. Например, следуя этой точке зрения, мы признаем, что в глагольных конструкциях бросать камни и бросаться камнями реализуются два разных значения, но они ассоциируются не с глаголом (что заставило бы нас в словаре представить два разных глагола с разными рамками валентностей), а с глагольной конструкцией (точнее, с двумя разными конструкциями для одного и того же глагола, отличающимися лишь оформлением актантов) [Рахилина 2010]. Не менее успешным оказывается данный подход при оценке степень приемлемости различных сочетаний слов в рамках конструкций. Всё это позволяет говорить о том, что лексические конструкции более гибки в описании нюансов языка, чем традиционная глаголоцентрическая модель. [Рахилина 2010].

### Статистический анализ структурной организации конструкций

В русле лингвистики конструкций существует течение, появившееся как ответ на запросы специалистов, работающих с корпусами текстов. Речь идет о так называемом коллострукционном анализе, исследовательской процедуре, разработанной А.Стефановичем и Ст. Грисом [Stefanowitsch, Gries, 2003] в качестве алгоритмической надстройки к теории. Коллострукционный анализ — это многоступенчатая процедура статистического анализа структурной организации конструкций. Основными этапами коллострукционного анализа являются колексемный анализ, различительный колексемный анализ и ковариация колексем.

Колексемный анализ имеет целью количественно оценить степень тяготения лексемы к тому или иному слоту конструкции. Колексемный анализ позволяет оценить силу взаимодействия пары факторов в конструкции. В качестве факторов можно рассматривать, например, лексему и некий грамматический признак конструкции, а силу взаимодействия данных факторов оценить с помощью точного критерия Фишера. Например, таким образом можно определить тенденции в употреблении глагольных конструкций с формами прошедшего времени того или иного глагола: в случае с русским глаголом *сказать* можно вычислить общую частоту его встречаемости в корпусе, частоту употребления в форме прошедшего времени, а также частоту всех глагольных лексем в корпусе и частоту всех их форм прошедшего времени. Частота употребления глагола *сказать* в форме прошедшего времени оказывается выше, чем «среднестатистическая», что свидетельствует о тяготении данного глагола к конструкциям прошедшего времени (прежде всего, к конструкциям, вводящим прямую и косвенную речь) [Рахилина 2010].

Различительный колексемный анализ предназначен для того, чтобы сравнивать схожие конструкции и выяснять степень тяготения лексемы к той или иной конструкции из набора. В работе [Gries, Stefanowitch 2004] проводится сопоставительный анализ дитранзитивной конструкции и конструкции с *to*: (*John sent Mary the book. John sent the book to Mary*.). Семантический характер различий между двумя конструкциями подтверждается тем фактом, что они притягивают к себе разные наборы лексем. В рассматриваемом случае глагол *give* проявляет тенденцию к употреблению в дитранзитивной конструкции.

И наконец, ковариация колексем определяется для того, чтобы оценить, силу связи между лексемами – кандидатами на заполнение разных слотов одной и той же конструкции. А.Стефанович и Ст. Грис в качестве иллюстрации приводят анализ каузативной конструкции с *into* в английском языке [Stefanowitch, Gries 2005], где есть один фискированный компонент и два свободных: например, *Newley had been tricked into revealing his hiding place*. Оказывается, что семантика конструкции накладывает ограничения на комбинацию пар глаголов, которые могут заполнить слоты конструкции. В частности, в первом слоте конструкции могут быть глаголы со значением оказания давления или уловки. Во втором слоте следует ожидать глаголы, обозначающие неприятное или нежелательное действие.

Наблюдения, сделанные разработчиками коллострукционного анализа, дают нам основания полагать, что статистические критерии являются мощным средством исследования внутренней организации конструкций, а также предоставляют богатый материал для разработки методов автоматической обработки данных о конструкциях.

Коллострукционный анализ применим для оценки сочетаемости слов. Сгенерированный текст (например, при машинном переводе) может быть грамматически корректен, но лишен, при этом, всякого смысла. Встречаются фразы, причины семантической несогласованности которых далеко не очевидны, а подчас и необъяснимы. Трудно объяснить, к примеру, разницу в оттенках значений слов *жаркий* и *горячий* (см. [Апресян 2010: 510]). Хотя на первый взгляд, эти слова можно назвать синонимами, существительные, с которыми они сочетаются, у каждого свои. Возникает вопрос: почему *горячая вода* и *горячий чайник* сочетаются, тогда как *горячая погода*, *горячее лето* уже нет? Универсальный ответ для любого словосочетания найти невозможно.

Поскольку лексическая конструкция определяется как словосочетание, которое встречается в речи, естественно предположить, что, если некоторая фраза встретилась в достаточно большом корпусе, ее можно считать лексической конструкцией. Здесь, впрочем, возникают определенные проблемы. На практике далеко не каждая биграмма является лексической конструкцией — это может быть описка, нарочное искажение, неверно снятая омонимия, наконец, оба слова могут относиться к разным лексическим конструкциям. Поэтому важно определить именно степень сочетаемости слов. Ясно, что совместная частота встречаемости не вполне отражает то, насколько данные слова характерны друг для друга, ведь одно из них может быть частотным само по себе. Например, фраза *хороший цвет* имеет большую частоту, чем фраза *пурпурный цвет*, при этом сочетаемость последней должна быть выше.

А. Стефанович и С. Грис [Stefanowitsch, Gries 2003] рассмотрели в качестве степени сочетаемости вероятность того, что события «встретилось слово *x*» и «встретилось слово *y*» зависимы. Для этого строилась статистическая модель, называемая таблицей сопряженности, а зависимость признаков оценивалась точным критерием Фишера. При таком подходе наибольшее значение получали устойчивые лексические конструкции.

Однако, если словосочетание отсутствует в корпусе, это вовсе не значит, что оно не является лексической конструкцией. Каким бы большим ни был корпус, нельзя ожидать, что он вместит в себя все мыслимые коллокации. Так, например, слово *краска* может встречаться с набором слов *красный*, *белый*, *зеленый*, но не встречаться со словами *бежевый* или *фиолетовый*. Возникает задача — как извлечь информацию о сочетаемости слов, если они не встречались в корпусе. Данная задача получила свое разрешение в нашей статье [Букия и др., 2015].

### Оценка взаимозаменяемости элементов конструкций

Для решения основных задач данного исследования мы разработали авторскую методику оценки взаимозаменяемости элементов конструкций, в которой развиваются идеи коллострукционного анализа А.Стефановича и Ст. Гриса [Bukia и др. 2015].

Итак, мы будем рассматривать текст как марковскую цепь нулевого порядка, считая, что каждое следующее слово зависит лишь от соседнего, а зависимостью от всего предшествующего текста мы пренебрегаем. Данное предположение можно обосновать тем, что корреляционная связь элементов внутри биграммной конструкции значительно сильнее, чем каждого элемента в отдельности со всем остальным текстом. Рассмотрим следующий пример:

*Вдоль реки протянулась забытая всеми железная дорога.*

Ясно, что конструкции *забытая всеми* и *железная дорога* воспринимаются как единое целое, их нельзя поэлементно разбить и заменить элементы синонимами, тогда как влияние остальных слов на эти конструкции значительно меньше, слово *протянулась* можно заменить на слово *пролегла*, слово *реки* можно заменить на *Волхова* и это едва ли заметно изменит значение каждой конструкции.

### Таблица сопряжённости

Обозначим за событие «встретилось слово » и за – «встретилось слово ». Если вероятность события при условии события равна самой вероятности события , мы полагаем, что фраза не является конструкцией. Это значит, что слово не характерное или не специфическое для и связь между этими словами отсутствует.

Наоборот, если после слова вероятность встретить слово заметно повышается, то можно с уверенностью говорить, что между этими словами есть связь. Это означает, что события и коррелированы, а величина корреляции отражает степень связи.

Из корпуса мы можем извлечь следующие статистические данные:

* Частота встречаемости первого слова ,
* Частота встречаемости второго слова ,
* Совместная частота встречаемости
* Общее количество слов в корпусе .

Чтобы оценить корреляцию событий и , нужно представить данные в виде таблицы сопряжённости и воспользоваться одной из стандартных статистик. Обозначим, как это принято за и отрицания соответствующих событий. Тогда таблица сопряжённости выглядит следующем образом:

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |

Здесь:

– частота события ,

– частота события «встретилось слово , а слово не встретилось,

– частота события «встретилось слово , а слово не встретилось,

– частота события «не встретилось ни , ни ».

Тогда для оценки корреляции можно воспользоваться следующими критериями.

### Оценка корреляции в таблицах сопряжённости

Мы оставим за скобками теоретическое описание данных критериев, сославшись лишь на монографию, посвященную статистическим критериям корреляции [Кендалл и др., 1973].

**Коэффициент ассоциации.**  Статистика критерия:

Принимает значения от 0, если слова не связаны, до 1 в случае полной тождественности. Эта оценка слишком приближенная, зато легко поддается вычислению.

**Коэффициент коллигации Юла.** Статистика критерия:

Между и существует связь:

**Коэффициент контингенции .** Статистика критерия:

Значения статистики от 0, если слова не связаны, и возрастает при наличии связности. Применим при больших значениях параметров, что, в случае с частотами, бывает далеко не всегда, особенно в корпусах небольшого размера.

**Точный критерий Фишера.** Статистика критерия:

С достоверностью проверяемые события коррелируют, значит в качестве степени сочетаемости здесь нужно положить . Он работает лучше, чем , когда параметры таблицы сопряженности малы.

**Быстрый критерий *z*.** Статистика критерия:

Статистика критерия применима, если или , зато он не требует больших вычислительных мощностей.

**-критерий Вулфа [Woolf, 1957].** Для четырехклеточных таблиц критерий Вулва наиболее теоретически обоснован.

Каждый критерий имеет свои внутренние особенности, поэтому окончательный выбор можно осуществить только при непосредственной проверке.

## Автоматическое выделение ключевых слов в документах

Существует немало методов автоматического извлечения ключевых слов как общего профиля, так и терминов, относящихся к определенной предметной области. Сравнение общих подходов можно найти в работе [Мирзагитова, 2014]. Вслед за большинством исследователей, под ключевыми словами мы будем понимать минимальный набор слов, наилучшим образом представляющий документ или коллекцию документов.

Спектр использования данной задачи в лингвистике чрезвычайно широк. Ключевые слова могут представлять документ для обеспечения быстрого поиска, для сортировки или группировки документов по смыслу или для установления тематической направленности. В узком смысле ключевые слова могут пониматься как термины, например, при автоматическом анализе научных работ. Подобным образом ключевые слова могут использоваться в новостных статьях для нахождения новостей близкой тематики.

Ввиду широкой трактовки самого понятия ключевых слов ему нельзя дать однозначного определения, которое не опиралось бы на интуитивное представление носителей языка. Это приводит к определённой проблеме при анализе качества методов. Как правило, про некоторые слова однозначно можно сказать, что они являются ключевыми, про другие – наоборот, что они однозначно ими не являются. Но существенная часть слов находится посередине и выбор их в качестве ключевых является в известном смысле субъективным и зависит от поставленной цели. В некотором случае, при работе над конкретной задачей исследователь может сузить границы неопределённости и подобрать удачный метод.

### Понятие ключевого слова

Ещё в 40-е годы прошлого века были первые попытки описать ключевые слова с точки зрения психолингвистики [Соколов 1941]. В современной науке есть несколько вариантов трактовки этого понятия.

В статьях [Камшилова 2013] и [Сахарный 1988] ключевые слова рассматриваются как разновидность «текста-примитива». С этой точки зрения набор ключевых слов:

* достаточны для восстановления исходного текста,
* используют наименьший возможный набор слов.

Для таких списков основополагающими являются не формальные (грамматические), а функциональные признаки (восприятие человеком).

Другая точка зрения постулирует, что максимизацию тематической составляющей и минимизацию связности между элементами.

В статье [Виноградова, Иванов 2015] отражены следующие представления о ключевых словах:

* совокупность ключевых слов должна представлять текст [Камшилова 2013];
* слова и конструкции, отражающие содержание документа [Шереметьева 2015];
* индексируемые слова, по которым может вестись поиск документа [Абрамов 2011];
* важные термины в документе, которые могут дать высокоуровневое описание содержания документа для читателя [Гринева 2009];
* неслучайно встречающиеся в документах слова и словосочетания, важные для рассматриваемой выборки (выдачи) в рамках общего массива документов [Большакова и др. 2011];
* слова, наиболее важные для решения поставленных в инструкции задач [Большакова и др. 2011].

Первые подходы, основанные на частотности слов, были предложены еще в середине прошлого столетия [Luhn 1957], с тех пор появилось немало новых алгоритмов, которые можно разделить на три категории: статистические, лингвистические и гибридные. Общий принцип всех этих методов можно сформулировать так:

1. Предварительная обработка текста. Исключение элементов маркировки, приведение слова к словарной форме, удаление стоп-слов, не несущих смысловой нагрузки (предлоги, союзы, частицы, местоимения, междометия и т. д.).
2. Отбор кандидатов в ключевые слова.
3. Фильтрация кандидатов в ключевые слова (анализ значимых признаков для каждого кандидата).
4. Отбор ключевых слов из числа кандидатов.

Первый этап, как правило, не отличается для различных методов и имеет немало стандартных реализаций. Наиболее существенные различия возникают на третьем этапе.

### Лингвистический подход

В прошлом столетии лингвистические методы были широко распространены [Солтон, 1979]. Как правило, они используют методы синтаксического и семантического анализа.

Это может быть описание правил построения текстов, разметка, словари терминов, антологии и прочее. Чаще всего такие методы недоступны для быстрой реализации, потому что требуют длительной подготовки и большого числа специалистов, ведь создание антологий – очень трудоёмкий процесс. Отдельные проблемы могут возникнуть из-за авторских прав при использовании готовых словарей. Как правило, лингвистические методы не универсальны и разрабатываются под конкретную задачу. Подобная технология использовалась при разработке автоматической информационной системы аудита нормативных документов [Баканова 2014].

Оригинальный способ обойти указанные проблемы был предложен в статье [Гринева 2009], в которой антология автоматически строится на основе корпуса статей Википедии. Тематический охват такой антологии был достаточно широк для охвата существенной части корпуса, при этом для работы системы не требовалось дополнительного обучения – термины выделяются на любом, сколь угодно маленьком корпусе, а внушительная система перекрёстных ссылок позволила достаточно неплохо оценить семантическую близость потенциальных ключевых слов.

В [Ефремова 2010] описывается способ извлечения с помощью лексико-синтаксических шаблонов, описывающих типичные терминологические конструкции. Такие системы особенно хорошо работают в текстах, имеющих жёсткую синтаксическую структуру, например, в научных статьях.

### Статистические методы

Статистические методы обладают широким спектром применения и, как правило, гораздо проще и дешевле в реализации, чем методы чисто лингвистические. С другой стороны, отсутствие лингвистической обработки приводит к тому, что полученные результаты нередко плохо согласованны, содержат шум и требуют дальнейшей обработки для практического использования.

Как правило, в статистических методах учитывается относительная частота встречаемости слов в документе. Слово, с одной стороны характерное для данной группы документов, с другой не характерное для всего корпуса, с большой вероятностью является тематическим. Традиционная статистика, оценивающая эту величину TF-IDF. TF (term frequency – частота слова) – отношение числа вхождения слова к общему количеству слов в документе:

Она характеризует, насколько данное слово является значимым для документа.

IDF (inverse document frequency – обратная частота документов) – инверсия частоты документов, содержащих данное слово:

Эта метрика уменьшает вес частотных слов языка, которые характерны для любых текстов коллекции.

Результирующая статистика является произведением TF и IDF, тем самым учитывает оба фактора [Маннинг 2011].

Сейчас появилось немало модификаций данного метода [Рубцова 2014].

В нашей работе составление тематического словаря производится оригинальным методом, который учитывает специфическую природу новостных документов.

### Гибридные методы

Часто для выделения ключевых слов используют комбинированный подход, пытаясь свести тем самым недостатки отдельных методов к нулю. В сущности, гибридный метод бывает двух типов, условно говоря, стати-лингвический и лингво-статистический.

В первом случае в качестве постобработки ключевых слов используют статистическую информацию об их встречаемости в корпусе, например, для сокращения перечня ключевых слов или для сортировки списка. Во втором случае используется лингвистическая обработка – морфо-синтаксические шаблоны, приведение к нормальному виду и прочее. В строгом смысле гибридных методов, в которых на равных участвует оба подхода, почти нет. Это обусловлено тем, что серьёзный лингвистический аппарат, как правило, сопряжён с высокими затратами на его создание, при этом существенно эффективность может и не повыситься. Поэтому, как правило, используются стандартные парсеры на этапе пред- и постобработки текста. Примеры таких подходов – алгоритмы C-Value, KEA, RAKE и др.

Мы тоже будем использовать простейший морфоанализатор, находящийся в открытом доступе, для повышения эффективности алгоритма.

Известны и другие подходы к выделению ключевых слов, например, интересные идеи предлагаются в работе [Лукашевич 2011], в которой используется спектральный анализ и всё более популярное вейвлет-преобразование для представления текстов. Множество идей основано не только на статистических, но и на алгебраических и графовых представлениях, среди которых заметен алгоритм TextRank, ранжирующий слова методом, аналогичным ранжированию web-страниц – путем запуска случайного блуждания и пересчета весов.

# АВтоматическая кластеризация текстов в новостном корпусе с назначением ключевых слов – меток кластеров

## Общие положения

Приступая к третьей главе нашей работы, скажем несколько слов о самом эксперименте.

Имеется корпус новостных документов, требуется кластеризовать его, построить тематический словарь и выставить метки кластерам документов.

При кластеризации корпуса текстов на небольшие группы, крупные темы, охватывающие сразу несколько групп, согласно стандартным метрикам будут иметь вес меньший, чем темы, соответствующие только одной группе. Это не отвечает интуитивному представлению ключевого слова: хотелось бы, чтобы глобальная тема *война в Сирии* имела больший вес, чем какая-нибудь локальная тема, задавая, тем самым, контекст для остальных ключевых слов. С другой стороны, попытка изменить это может привести к тому, что увеличится вес у слишком общих, частотных слова, охватывающих всю коллекцию.

Чтобы отделить зерна от плевел, мы разработали специальный двухступенчатый алгоритм, применимый в той или иной мере именно к новостным информационным сайтам.

### Основное предположение

Было замечено, что многие новостные порталы публикуют новости двух классов ­– серьезные и развлекательные. К первому типу относятся политические, финансовые новости. Ко второму – новости о звездах, «британские ученые доказали» и тому подобные материалы. Как оказалось, в тои или иной степени этой классификации отвечает большинство порталов.

Этой особенностью мы решили воспользоваться следующим образом: разбить методом кластеризации корпус на две части: серьезную и несерьезную и составить словарь ключевых слов, основываясь на следующем принципе: если слово характерно для одной группы и не характерно для другой, то оно скорее является тематическим и имеет больший вес. Такой метод позволил избавиться от нейтральных слов, при этом поднять вес крупным темам. Можно назвать эту идею центральной идеей всей работы.

### Применение

Может быть несколько вариантов применения данного метода. Прямое применение – организовать серию перекрестных ссылок, проставив метки каждому документу. Другой способ – составлять новостные карточки, объединяя новости с разных порталов и составлять краткое описание с помощью извлеченных ключевых слов подобно тому, как это делается порталом «Яндекс.Новости». Наконец, третий способ, как нам кажется, наиболее интересный и актуальный, – составление рекомендаций и ранжирование новых документов, исходя из предпочтений пользователя.

### Данные

В работе использовался корпус новостей портала Ruposters за май 2016 года, объемом 94 тысячи словоупотреблений и 428 документов. Каждый документ имеет пометку «news» или «life», означающую категорию новости.

Мы нарочно ограничили корпус документов – в задачах машинного обучения без учителя увеличение корпуса приводит лишь к улучшению работы алгоритма – это было проверено на корпусе новостей за 2015 год. Далеко не всегда имеется в распоряжении такой большой корпус. Поэтому мы рассмотрели пограничный случай: когда корпус мал, чтобы алгоритму было «слишком легко», но достаточно велик, чтобы оценка была статистически значимой.

Тексты были предобработаны с помощью модуля Pymorphy2, каждое слово было переведено в нормальную форму и получило метку части речи.

Моделирование произведено на языке Python с использованием дополнительных математических библиотек numpy, scipy, scikit-learn и некоторых технических модулей.

### План эксперимента

Таким образом, наш эксперимент состоит из четырех шагов:

1. Кластеризация новостного корпуса на два кластера, анализ кластеризации, подбор оптимальных параметров. Сравнение с эталонной классификацией.
2. Выделение ключевых слов, характерных для серьезного и несерьезного кластеров. Сравнение метрик, оптимизация алгоритма. Анализ списков.
3. Выделение конструкций, дополняющих список ключевых слов. Сравнение и анализ коллокационных метрик.
4. Кластеризация корпуса текстов на множество близких небольших групп. Анализ проставления меток кластеру.

## Кластеризация

### Основная цель данного эксперимента

Мы воспользовались порталом Ruposters, на котором разработчиками сайта уже произведена разметка: каждый документ, относящийся к серьезным и несерьезным новостям, размечен соответственно либо как news, либо как life. Эту классификацию будем считать эталонной и проверим, какие методы кластеризации при каких параметрах показывают лучшие результаты. Тем самым, при работе с другими порталами можно использовать оптимизированный алгоритм.

Итак, задача этого раздела отделить серьезные новости от несерьезных, приблизив параметры алгоритма к эталонным.

### Корпус

Представим ниже типичные заголовки обоих классов.

***Раздел life***

* Пользователи Сети нашли серьезное нарушение Джамалой правил "Евровидения";
* Звезда реалити-шоу "Дом-2" попала в психиатрическую больницу;
* Звезда "Дома-2" Ольга Бузова полностью обнажилась, шокировав своих поклонников;
* В Москве установят павильоны "Куку-вата";
* 10 самых откровенных нарядов Каннского кинофестиваля.

***Раздел news***

* Красноярские чиновники "погуляли" в командировках почти на 6 млн бюджетных рублей;
* Кудрин заявил о "последнем рубеже", которого достигла российская экономика;
* Глава Азербайджана объявил о ядерной угрозе со стороны Армении;
* Эрдоган захотел реформировать Совбез ООН;
* "Газпром" продал свои 37% в эстонской AS Eesti Gaas.

### Входные данные

В первую очередь, с помощью модуля Pymorphy2 производится предобработка корпуса, которая заключается в следующем:

1. Графематическая обработка, разделение слов и знаков препинания;
2. Приведение каждого слова к нормальному виду;
3. Выставление частеречных меток, отделенных от слова знаком ‘\_’;

Тем самым, каждый документ представлен в следующем виде:

*речь\_NOUN идти\_INFN о\_PREP хмель\_NOUN ,\_PNCT который\_ADJF придавать\_INFN алкогольный\_ADJF напиток\_NOUN характерный\_ADJF горьковатый\_ADJF вкус\_NOUN .\_PNCT он\_NPRO оказывать\_INFN благотворный\_ADJF воздействие\_NOUN на\_PREP организм\_NOUN ,\_PNCT благодаря\_PREP свой\_ADJF способность\_NOUN останавливать\_INFN рост\_NOUN бактерия\_NOUN и\_CONJ развитие\_NOUN болезнь\_NOUN .\_PNCT*

### Векторизация документов

Для кластеризации каждый документ представляется в векторном виде. При этом используется модель «мешка слов», то есть не учитывается порядок появления слов в документе. На данном этапе такое допущение вполне корректно – порядок слов будет учитываться при выделении конструкций.

Вектор документа имеет размерность, равную количеству уникальных слов в корпусе. Таким образом, каждому слову для данного документа соответствует некоторое число. Это стандартная в таких задачах модель представления в виде матрицы «слова на документы».

Функцию определим двумя способами:

1. – частота слова в документе ;
2. ­­– метрика TF-IDF.

### Метод кластеризации

Мы выбрали наиболее эффективные методы кластеризации, позволяющие задать фиксированное число кластеров.

* Метод k средних;
* Спектральный метод
* Иерархические методы

### Метрики оценки качества

Качество кластеризации оценивается стандартными критериями, описанными в первой главе:

* Несмещенный индекс Ранда
* Несмещенная взаимная информация
* Точность
* Полнота
* V-мера
* Коэффициент силуэта

### Выходные данные

Модуль строит 4 текстовых документа:

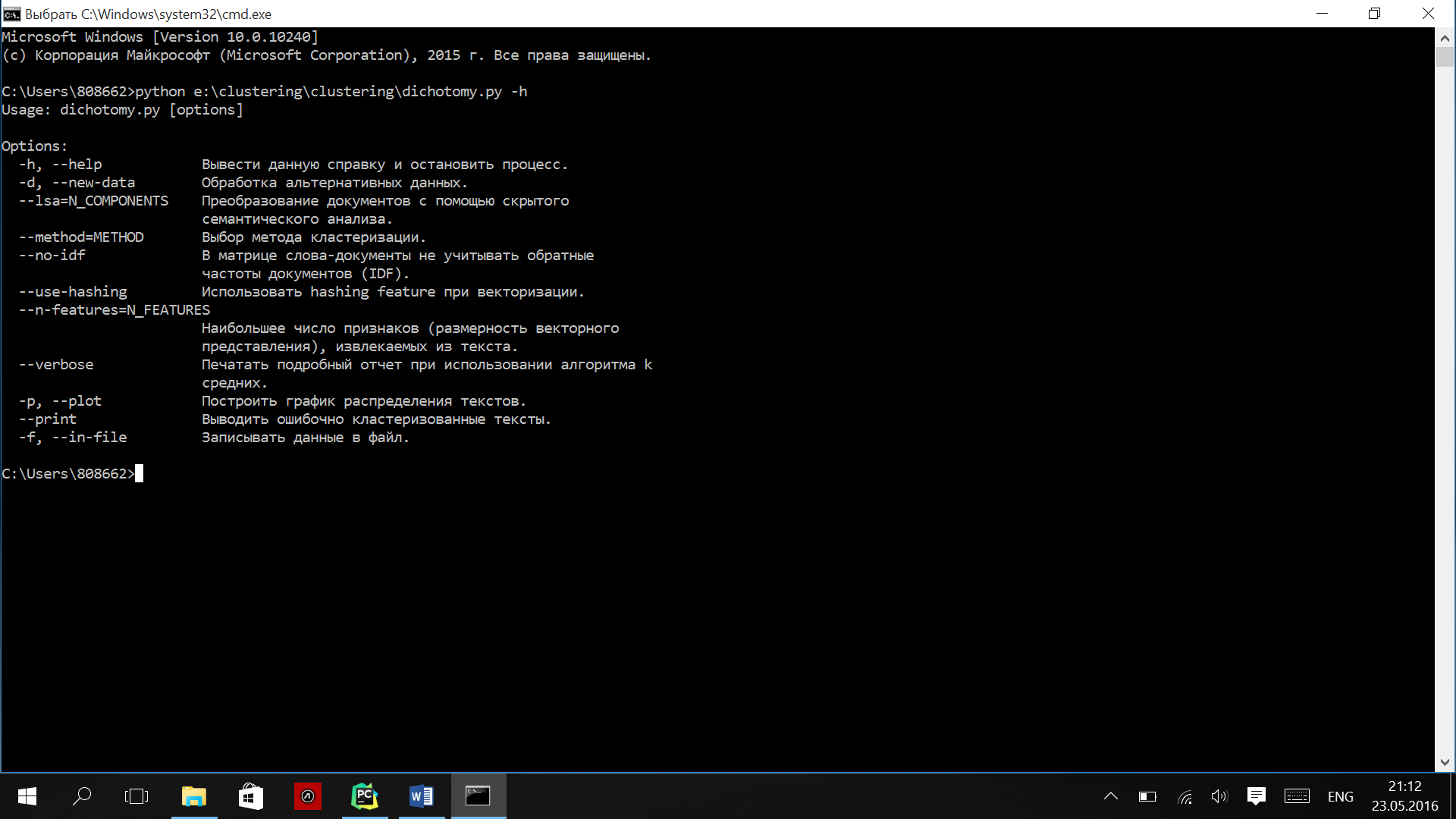
* Объединенные в один файл документы, попавшие в первый кластер;
* Объединенные в один файл документы, попавшие во второй кластер;
* Файл с ошибочными документами первого класса, попавшие во второй кластер;

Рисунок 1. Меню первого файла программы: опции позволяют настроить кластеризацию

* Файл с ошибочными документами второго класса, попавшие в первый кластер.

Первые два файла необходимы для дальнейшей работы программы, тогда как другие два вспомогательные и позволяют проанализировать ошибки алгоритма.

### Результат кластеризации

Результат помещен в двух таблицах – в первой даны значения метрик для частотной матрицы «слова на документы», а во второй – TF-IDF матрицы.

Таблица 1. Результаты кластеризации с использованием частотной матрицы «слова на документы»

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Rand | MI | h | c | v | s |
| k-mean | 0.838 | 0.757 | 0.757 | 0.779 | 0.768 | 0.013 |
| Spectral | 0.787 | 0.69 | 0.69 | 0.713 | 0.702 | 0.013 |
| Agglomerative | 0.698 | 0.58 | 0.58 | 0.596 | 0.588 | 0.01 |

Таблица 2. Результаты кластеризации с использованием метрики TF-IDF в матрице «слова на документы»

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Rand | MI | h | c | v | S |
| k-mean | **0.972** | **0.938** | **0.938** | **0.939** | **0.939** | 0.009 |
| Spectral | 0.754 | 0.681 | 0.681 | 0.72 | 0.7 | 0.008 |
| Agglomerative | 0.746 | 0.643 | 0.635 | 0.651 | 0.643 | 0.008 |

Эксперимент показал, что использование метрики TF-IDF существенно повышает качество кластеризации, особенно в случае метода k средних, который чувствителен к шуму. TF-IDF существенно снижает шум – слова, равномерно распределенные по документам и имеющие высокую частоту (например, предлоги), в этом случае существенно на кластеризацию не влияют.

Высокий результат метода k средних косвенно подтверждает главный тезис: лексика серьезного и несерьезного классов существенно отличается.

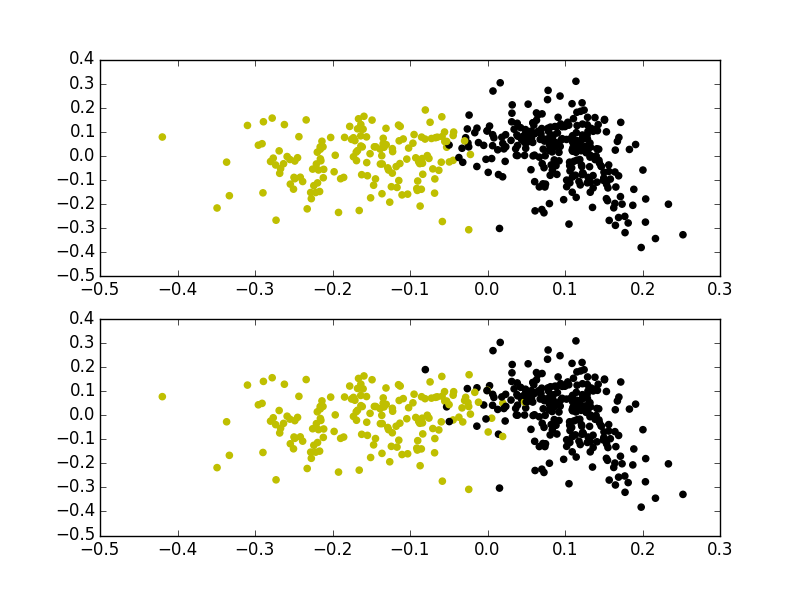


Рисунок 2. Распределение документов при частотной векторизации. Вверху кластеризация, внизу эталонная классификация. Черным цветом обозначены документы, помеченные как серьезные, серым цветом – как несерьезные.

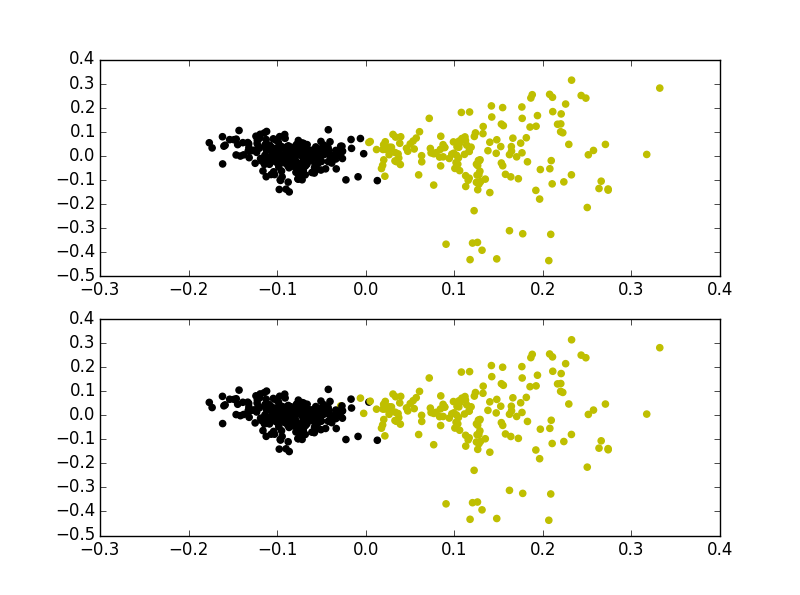


Рисунок 3. Распределение документов при TF-IDF векторизации. Вверху кластеризация, внизу эталонная классификация. Черным цветом обозначены документы, помеченные как серьезные, серым цветом – как несерьезные.

Заметно, что кластеры выпуклые и, при использовании TF-IDF, хорошо разделены. Это и объясняет полученный результат.

### Анализ ошибок

Ряд документов из пограничной области был неверно размечен методом k средних. Анализ их лексики показал, что в действительности эти документы можно отнести как к первой, так и ко второй группе. Продемонстрируем один из таких документов:

## *Иракские ополченцы, сражающиеся против террористов "Исламского государства", предложили своим подписчикам в Instagram самим решить судьбу пойманного ими боевика при помощи онлайн-голосования.*

## *"Юг Мосула. Вы можете проголосовать: убить его или отпустить. На голосование отводится один час. Мы сообщим о его судьбе через час", — написали авторы аккаунта @iraqiswat в Instagram, опубликовав фотографию пленного террориста.*

## Выделение ключевых слов

После кластеризации документов, произведенной в предыдущем разделе, мы хотим извлечь слова, которые характерны для первого класса, и, в то же время, не характерны для второго. Согласно гипотезе, это и будут ключевые слова. В качестве входных данных используются файлы, полученные на предыдущем этапе.

### Метод выделения ключевых слов

Здесь мы снова возвращаемся к модели таблицы сопряженности. Нам нужно оценить, в какой степени данное слово тяготеет к одному или другому документу. Обозначим:

– частота встречаемости слова в первом документе;

– частота встречаемости слова во втором документе;

– общая частота остальных слов в первом документе;

– общая частота остальных слов во втором документе;

и воспользуемся одним из критериев корреляции в таблице сопряженности, которые мы описали выше.

Сортируя слова, получим два списка ключевых слов, отвечающих каждой группе.

### Статистики корреляции

Мы будем использовать следующие статистики:

* Быстрый критерий z;
* Коллигация Юла
* G-критерий Вулфа
* Взаимная информация
* Ассоциация
* Точный критерий Фишера
* Хи-квадрат

Статистики, принимающие неограниченные значения нормируются стандартным преобразованием .

### Выходные данные

На выходе строятся два файла, отвечающие каждому кластеру. В каждом файле ранжированный по данной метрике список ключевых слов. Через табуляцию для каждого слова выводится его частота в первом кластере, во втором кластере и значение метрики.

### Результаты

Рисунок 4. Меню второго модуля программы. Опции позволяют регулировать длину списка.

***Быстрый критерий Z***

Одна из простейших метрик показала достаточно неплохие результаты. Она не очень чувствительна к общим словам (в топ-10 попало слово *как*), но в целом ранжирование можно назвать удовлетворительным.

Таблица 3. Топ-10 ключевых слов для быстрого критерия Z

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Life words | L | N | Вес | News words | L | N | Вес |
| мужчина\_NOUN | 73 | 3 | 0.889 | россия\_NOUN | 27 | 233 | 0.927 |
| life\_LATN | 55 | 0 | 0.881 | украина\_NOUN | 3 | 166 | 0.926 |
| женщина\_NOUN | 72 | 7 | 0.879 | президент\_NOUN | 1 | 124 | 0.916 |
| фильм\_NOUN | 53 | 3 | 0.869 | российский\_ADJF | 17 | 158 | 0.914 |
| девушка\_NOUN | 53 | 7 | 0.855 | сирия\_NOUN | 0 | 106 | 0.911 |
| учёный\_NOUN | 45 | 5 | 0.849 | сша\_NOUN | 15 | 132 | 0.906 |
| ранее\_ADVB | 78 | 22 | 0.848 | страна\_NOUN | 31 | 161 | 0.903 |
| как\_CONJ | 241 | 133 | 0.848 | крым\_NOUN | 2 | 86 | 0.899 |
| ruposters\_LATN | 57 | 11 | 0.847 | глава\_NOUN | 4 | 86 | 0.896 |
| картина\_NOUN | 30 | 0 | 0.845 | савченко\_Surn | 0 | 70 | 0.893 |

**Коллигация Юла, взаимная информация, ассоциация**

Эти три метрики словам, встретившимся лишь в одном классе, дают максимальный вес, поэтому слово *ячменный* будет всегда выше слова *Россия*.

Таблица 4. Топ-10 ключевых слов для неудачных критериев

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Life words | L | N | Вес | News words | L | N | Вес |
| a\_LATN | 3 | 0 | 1.0 | ячменный\_ADJF | 0 | 2 | 1.0 |
| персональный\_ADJF | 2 | 0 | 1.0 | вооружение\_NOUN | 0 | 10 | 1.0 |
| петербургский\_ADJF | 2 | 0 | 1.0 | катастрофа\_NOUN | 0 | 3 | 1.0 |
| печень\_NOUN | 2 | 0 | 1.0 | понижать\_INFN | 0 | 2 | 1.0 |
| жаркое\_NOUN | 2 | 0 | 1.0 | касьянов\_Surn | 0 | 3 | 1.0 |
| жанр\_NOUN | 2 | 0 | 1.0 | кастро\_Surn | 0 | 4 | 1.0 |
| пик\_NOUN | 2 | 0 | 1.0 | пономарев\_Surn | 0 | 10 | 1.0 |
| пикантный\_ADJF | 3 | 0 | 1.0 | понятный\_ADJF | 0 | 5 | 1.0 |
| пиксель\_NOUN | 2 | 0 | 1.0 | вооружённый\_ADJF | 0 | 15 | 1.0 |
| пингвин\_NOUN | 5 | 0 | 1.0 | пообещать\_INFN | 0 | 21 | 1.0 |

**G-критерий Вулфа**

Данный критерий показал наиболее хороший результат. G-критерий с одной стороны не очень чувствителен к частотным общим словам, с другой ­– к редким словам, встретившимся в одном классе текстов. На наш взгляд ранжирование в целом отвечает интуитивному представлению о ключевых словах.

Таблица 5. Топ-10 ключевых слов для G-критерия Вулва

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Life words | L | N | Вес | News words | L | N | Вес |
| мужчина\_NOUN | 73 | 3 | 0.988 | украина\_NOUN | 3 | 166 | 0.995 |
| life\_LATN | 55 | 0 | 0.986 | россия\_NOUN | 27 | 233 | 0.994 |
| женщина\_NOUN | 72 | 7 | 0.984 | президент\_NOUN | 1 | 124 | 0.993 |
| фильм\_NOUN | 53 | 3 | 0.982 | сирия\_NOUN | 0 | 106 | 0.992 |
| девушка\_NOUN | 53 | 7 | 0.976 | российский\_ADJF | 17 | 158 | 0.992 |
| учёный\_NOUN | 45 | 5 | 0.975 | сша\_NOUN | 15 | 132 | 0.990 |
| картина\_NOUN | 30 | 0 | 0.974 | крым\_NOUN | 2 | 86 | 0.990 |
| специалист\_NOUN | 33 | 2 | 0.973 | страна\_NOUN | 31 | 161 | 0.989 |
| ruposters\_LATN | 57 | 11 | 0.972 | савченко\_Surn | 0 | 70 | 0.989 |
| животный\_ADJF | 28 | 1 | 0.972 | глава\_NOUN | 4 | 86 | 0.988 |

**Точный критерий Фишера**

Точный критерий Фишера в данной задаче показывает вполне удовлетворительный результат, однако его ранжирование отличается от g-критерия. Он так же чувствителен к перекосам.

Таблица 6. Топ-10 ключевых слов для точного критерия Фишера

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Life words | L | N | Вес | News words | L | N | Вес |
| life\_LATN | 55 | 0 | 1.0 | украинский\_ADJF | 1 | 72 | 1.0 |
| мужчина\_NOUN | 73 | 3 | 1.0 | крым\_NOUN | 2 | 86 | 1.0 |
| женщина\_NOUN | 72 | 7 | 1.0 | ес\_NOUN | 0 | 57 | 1.0 |
| фильм\_NOUN | 53 | 3 | 0.999 | президент\_NOUN | 1 | 124 | 1.0 |
| девушка\_NOUN | 53 | 7 | 0.999 | трамп\_NOUN | 1 | 65 | 1.0 |
| картина\_NOUN | 30 | 0 | 0.999 | военный\_NOUN | 4 | 77 | 1.0 |
| учёный\_NOUN | 45 | 5 | 0.999 | украина\_NOUN | 3 | 166 | 1.0 |
| ruposters\_LATN | 57 | 11 | 0.999 | сирия\_NOUN | 0 | 106 | 1.0 |
| ранее\_ADVB | 78 | 22 | 0.999 | страна\_NOUN | 31 | 161 | 1.0 |
| как\_CONJ | 241 | 133 | 0.999 | российский\_ADJF | 17 | 158 | 1.0 |

**Хи-квадрат**

Стандартный критерий Хи-квадрат показал результат, близкий к тривиальной метрике z.

Таблица 7. Топ-10 ключевых слов для критерия хи-квадрат

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Life words | L | N | Вес | News words | L | N | Вес |
| мужчина\_NOUN | 73 | 3 | 0.984 | россия\_NOUN | 27 | 233 | 0.993 |
| life\_LATN | 55 | 0 | 0.981 | украина\_NOUN | 3 | 166 | 0.993 |
| женщина\_NOUN | 72 | 7 | 0.981 | президент\_NOUN | 1 | 124 | 0.991 |
| фильм\_NOUN | 53 | 3 | 0.977 | российский\_ADJF | 17 | 158 | 0.991 |
| девушка\_NOUN | 53 | 7 | 0.971 | сирия\_NOUN | 0 | 106 | 0.990 |
| как\_CONJ | 241 | 133 | 0.968 | сша\_NOUN | 15 | 132 | 0.989 |
| учёный\_NOUN | 45 | 5 | 0.968 | страна\_NOUN | 31 | 161 | 0.988 |
| ранее\_ADVB | 78 | 22 | 0.968 | крым\_NOUN | 2 | 86 | 0.987 |
| ruposters\_LATN | 57 | 11 | 0.967 | глава\_NOUN | 4 | 86 | 0.986 |
| картина\_NOUN | 30 | 0 | 0.965 | савченко\_Surn | 0 | 70 | 0.985 |

При работе с большими корпусами имеет смысл хранить только верхушку корпуса. Наш модуль позволяет обрезать списки ключевых слов автоматически, исходя из распределения или задавая параметры вручную, построив гистограмму.

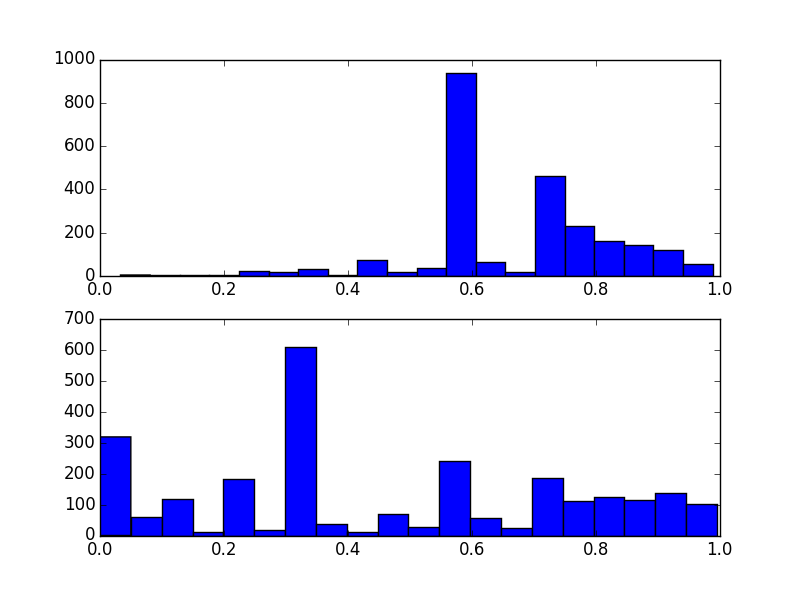


Рисунок 5. Гистограмма распределения весов ключевых слов для критерия Вулва. Вверху несерьезные, внизу серьезные документы. Наиболее удачные слова имеют вес, больший 0.7.

## Выделение конструкций

### Задача

В корпусе текстов выделить биграммы, лексемы которых с одной стороны обладают высокой степенью ковариации, с другой – являются ключевыми словами.

Униграммы не всегда в полной мере могут описать текст. Например, ключевое слово *Владимир* ни о чём не говорит, тогда как биграмма *Владимир Путин* является важной меткой. В данной работе рассматриваются только конструкции-биграммы, однако предложенный метод без труда обобщается и на более сложные конструкции.

В качестве входных данных используются файлы, полученные на предыдущих этапах.

### Выделение тематических конструкций

Выше мы подробно описывали, как, развивая идеи Стефановича и Гриса, можно выделять биграммы, лексемы которой обладают высокой ковариацией, то есть характерны друг для друга.

Интуитивно ясно, что если биграмма является конструкцией по Стефановичу, то она оказывается лучшей меткой для статьи, чем каждая ее лексема в отдельности. При этом, даже если одна из них не попала в список ключевых, вполне возможно, что вся конструкция, тем не менее, может являться меткой.

Простейшая функция, отвечающая этим параметрам, задается следующим образом: вес биграммы равен сумме весов ее лексем, умноженный на их корреляцию. Иными словами, если вес слова обозначить за , а меру связи за , вес биграммы равен:

Как и в предыдущем разделе, мы рассмотрим те же метрики для таблиц сопряженности, однако теперь победители будут совсем другие. Попрежнему метрики, принимающие неограниченные значения нормируются стандартным преобразованием .

### Шаблоны

Поскольку мы имеем данные о частях речи, можно применить стандартные шаблоны, чтобы получившиеся конструкции были более осмысленными. В большей степени это необходимо для того, чтобы сократить пространство, однако даже такая простейшая обработка помогает избавиться от незавершенных конструкций (вроде, *пригласил в*). Используются следующие шаблоны:

* Прил. + сущ.
* Сущ. + гл.
* Гл. + сущ.
* Имя + фам.
* \* + латиница

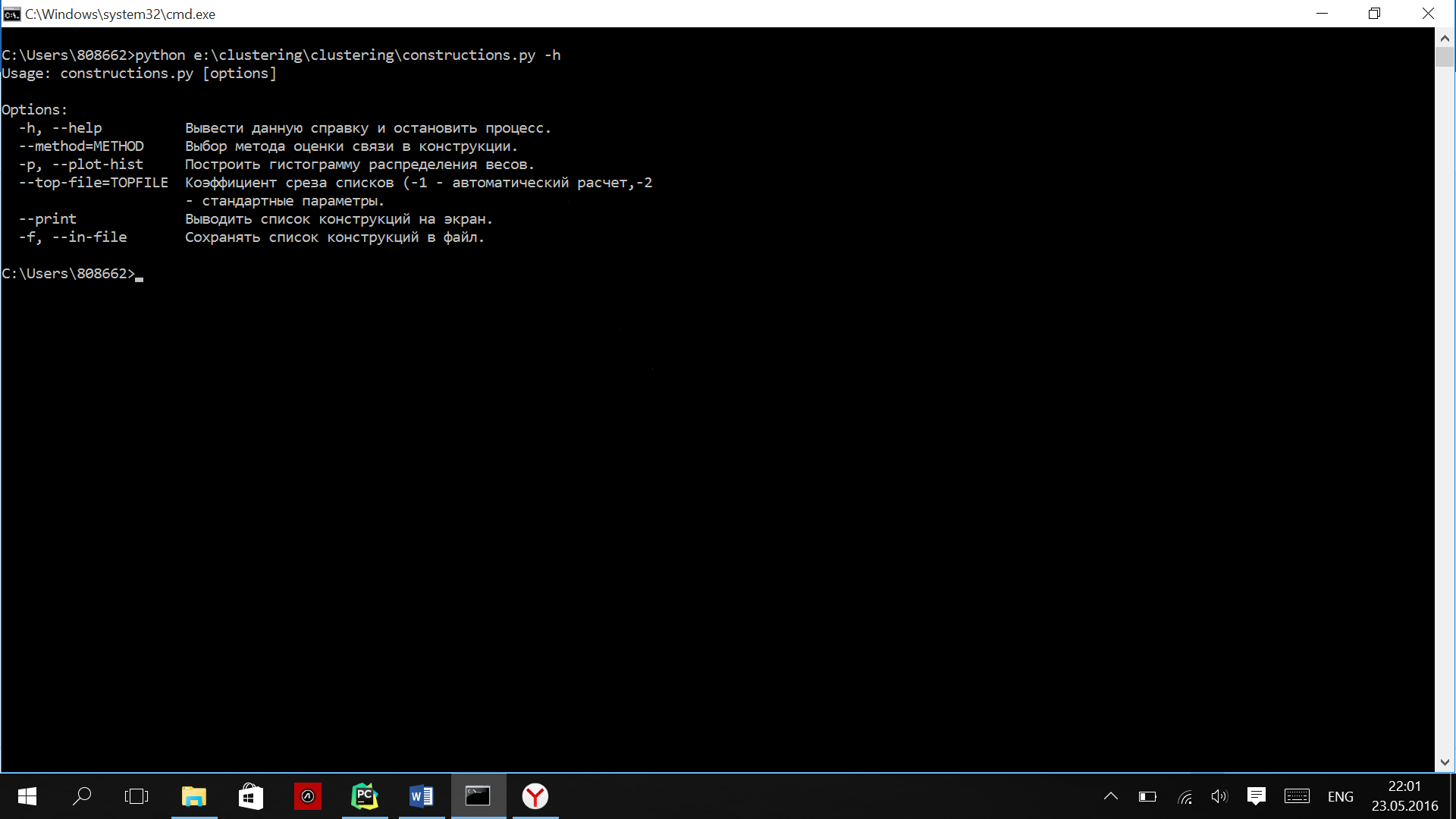
и т.п. комбинации.

Рисунок 6. Меню модуля выделения конструкций

### Выходные данные

На выходе получается список биграмм, который отдельной процедурой объединяется с униграммами и упорядочивается.

### Результаты

Итак, теперь поочередно применим различные варианты меры связи и сравним списки. Как и в случае с ключевыми словами, ниже будут представлены десять биграмм из каждого кластера.

***Быстрый критерий Z***

Данный метод показал себя не лучшим образом – его достоинство в этой задаче обратилось в недостаток. Он слишком чувствителен к частотным словам, в результате чего наибольший вес получили биграммы, состоящие из самых частотных слов.

Таблица 8. Конструкции, полученные методом z

|  |  |  |  |
| --- | --- | --- | --- |
| Life | Вес | News | Вес |
| свой мужчина | 1.885 | российский украина | 1.910 |
| свой женщина | 1.882 | российский сирия | 1.908 |
| свой фильм | 1.880 | российский президент | 1.906 |
| свой снимка | 1.857 | российский глава | 1.903 |
| свой тело | 1.849 | российский военный | 1.897 |
| свой лента | 1.847 | российский политика | 1.894 |
| свой собака | 1.845 | российский мид | 1.894 |
| свой поклонник | 1.843 | российский власть | 1.892 |
| свой песня | 1.841 | российский правительство | 1.891 |
| свой композиция | 1.841 | российский лидер | 1.891 |

***Коллигация Юла***

Данный метод, как нам кажется, выделяет наилучшие биграммы, которые действительно можно назвать тематическими.

Таблица 9. Конструкции, полученные методом коллигации Юла

|  |  |  |  |
| --- | --- | --- | --- |
| Life | Вес | News | Вес |
| ruposters life | 1.953 | исламский государство | 1.896 |
| rolling stones | 1.791 | минский соглашение | 1.858 |
| нижний бельё | 1.790 | барак обама | 1.839 |
| домашний питомец | 1.782 | петр порошенко | 1.831 |
| семейный пара | 1.775 | владимир путин | 1.828 |
| откровенный фотосессия | 1.774 | олег пешков | 1.797 |
| разбитый сердце | 1.772 | иностранный министр | 1.792 |
| метаболический синдром | 1.755 | дипломатический представительство | 1.784 |
| сексуальный фантазия | 1.744 | иностранный дело | 1.782 |
| психотропный вещество | 1.736 | вооружённый сила | 1.781 |

**G-критерий Вулва**

Здесь частотность существенно повлияла на веса: очевидно фразы *данный момент*, *данное заболевание* попали в верхушку списка из-за этого.

Таблица 10. Конструкции, полученные критерием Вулва

|  |  |  |  |
| --- | --- | --- | --- |
| Life | Вес | News | Вес |
| ruposters life | 1.956 | российский военный | 1.950 |
| хороший возраст | 1.874 | украинский власть | 1.938 |
| данный момент | 1.867 | владимир путин | 1.932 |
| откровенный фотосессия | 1.852 | исламский государство | 1.929 |
| знаменитый актёр | 1.835 | российский мид | 1.917 |
| пользователь youtube | 1.832 | американский президент | 1.913 |
| хороший фильм | 1.823 | европейский страна | 1.912 |
| данный заболевание | 1.814 | сирийский оппозиция | 1.912 |
| оригинальный фильм | 1.811 | иностранный министр | 1.911 |
| обычный мужчина | 1.808 | российский авиация | 1.903 |

**Взаимная информация**

Взаимная информация страдает тем же недугом, однако сортирует слова иначе.

Таблица 11. Конструкции, полученные с помощью взаимной информации

|  |  |  |  |
| --- | --- | --- | --- |
| Life | Вес | News | Вес |
| ruposters life | 1.955 | украинский власть | 1.938 |
| хороший возраст | 1.888 | владимир путин | 1.937 |
| сексуальный певица | 1.887 | сирийский оппозиция | 1.934 |
| данный заболевание | 1.887 | исламский государство | 1.933 |
| откровенный фотосессия | 1.884 | иностранный министр | 1.930 |
| оригинальный фильм | 1.882 | турецкий власть | 1.921 |
| обычный мужчина | 1.881 | российский военный | 1.920 |
| сексуальный тело | 1.880 | сирийский войско | 1.916 |
| женский тело | 1.880 | сирийский народ | 1.916 |
| пользователь youtube | 1.877 | российский авиация | 1.914 |

**Ассоциация**

Распределение коэффициента ассоциации оказалось очень близко к сортировке методом взаимной информации.

Таблица 12. Конструкции, полученные методом ассоциации

|  |  |  |  |
| --- | --- | --- | --- |
| Life | Вес | News | Вес |
| ruposters life | 1.959 | владимир путин | 1.939 |
| хороший возраст | 1.888 | исламский государство | 1.936 |
| откровенный фотосессия | 1.884 | сирийский оппозиция | 1.931 |
| данный заболевание | 1.876 | иностранный министр | 1.930 |
| пользователь youtube | 1.876 | украинский власть | 1.925 |
| данный момент | 1.874 | российский авиация | 1.916 |
| оригинальный фильм | 1.873 | мирный гражданин | 1.909 |
| обычный мужчина | 1.867 | сирийский народ | 1.908 |
| знаменитый актёр | 1.867 | турецкий власть | 1.905 |
| начать эксперимент | 1.8656 | верховный депутат | 1.902 |

**Точный критерий Фишера**

Метод, предложенный Стефановичем и Гриссом показал неудовлетворительный результат. Частотные слова слишком сильно увеличивают вес конструкции.

Таблица 13. Конструкции, полученные точным критерием Фишера

|  |  |  |  |
| --- | --- | --- | --- |
| Life | Вес | News | Вес |
| ruposters life | 1.959 | российский президент | 1.983 |
| хороший фильм | 1.918 | российский военный | 1.979 |
| свой тело | 1.912 | украинский президент | 1.976 |
| свой поклонник | 1.905 | сирийский президент | 1.975 |
| обычный мужчина | 1.905 | российский мид | 1.972 |
| свой композиция | 1.903 | российский лидер | 1.968 |
| данный заболевание | 1.903 | украинский политика | 1.967 |
| оригинальный фильм | 1.900 | украинский власть | 1.965 |
| хороший возраст | 1.896 | российский глава | 1.964 |
| начать эксперимент | 1.888 | украинский мид | 1.962 |

**Хи-квадрат**

Еще один традиционный метод ничем не выделяется от остальных.

Таблица 14. Конструкции, полученные методом Хи-квадрат

|  |  |  |  |
| --- | --- | --- | --- |
| Life | Вес | News | Вес |
| ruposters life | 1.959 | российский военный | 1.973 |
| хороший возраст | 1.895 | украинский власть | 1.962 |
| хороший фильм | 1.888 | российский мид | 1.959 |
| данный заболевание | 1.888 | российский войско | 1.947 |
| откровенный фотосессия | 1.886 | американский президент | 1.946 |
| оригинальный фильм | 1.884 | европейский страна | 1.946 |
| обычный мужчина | 1.883 | сирийский президент | 1.945 |
| пользователь youtube | 1.881 | владимир путин | 1.942 |
| данный момент | 1.880 | сирийский оппозиция | 1.942 |
| начать эксперимент | 1.875 | турецкий власть | 1.939 |

Как оказалось, коэффициент коллигации Юла – наиболее подходящий критерий для оценки сочетаемости в адаче выделения ключевых слов.

На гистограмме видно, что наш способ взвешивания биграмм может быть доработан. Некоторые биграммы имеют вес, больший единицы, при этом явно уступают лучшим униграммам, вес которых всегда меньше единицы. Поэтому возможна настройка алгоритма с помощью дополнительных коэффициентов, полученных эмпирически.

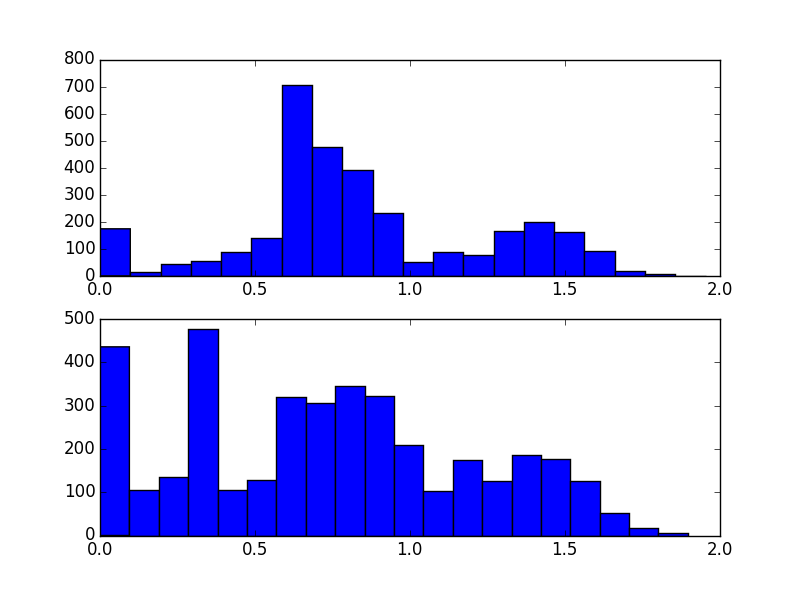


Рисунок 7. Гистограмма распределения тематических меток для метода Юла

## Выделение тематических меток в кластере новостных документов

### Общее положение

Чтобы иллюстрировать качество выделения тематических меток для группы документов, произведем кластеризацию любым методом и для произвольной группы извлечем ключевые слова. При этом вес каждой тематической метки умножается на частоту в кластере. Ниже представлены тексты одного кластера, состоящего из четырех документов, с выделенными ключевыми словами.

### Результат

***Ключевые слова****:*

турция 6.89853420862

сша 5.94330227089

президент 4.96131582517

страна 4.94784744148

обама 4.89944481432

война 3.91119922742

право 3.82381747822

беженец 3.77026358358

болгария 3.43676387354

год 3.39483088399

международный право 3.07828953878

европа 2.85682103533

османский империя 2.35259785334

россия 1.98925601691

сирия 1.98588165513

глава 1.97779589388

ес 1.97308425211

государство 1.9715097254

правительство 1.95040262693

анкара 1.94853523608

исламский государство 1.89643005104

штаты 1.88154406128

встреча 1.85888711847

эрдоген 1.81639801475

освобождение 1.79199105972

заявление 1.73539433659

интерес 1.72973210732

безвизовый режим 1.7187762516

оппозиционный газета 1.68356065688

международный сообщество 1.64126028237

российский флаг 1.60639565483

дипломатический миссия 1.58962410986

турецкий власть 1.57835846238

турецкий сторона 1.47742446702

русский армия 1.47685128012

международный принцип 1.4736234574

турецкий войско 1.46084793496

османский иго 1.43954482025

официальный представитель 1.43750799581

ополченец 1.43359953676

американский президент 1.40762378158

русский солдат 1.3939957058

официальный встреча 1.38029651251

конкретный страна 1.32705925561

этот война 1.26007091627

газета zaman 1.24067373402

европейский конвенция 1.2320622458

империя 1.18551520529

саудовский аравия 1.18551520529

этот страна 1.10822728857

американский правительство 1.00306018498

Количество ключевых слов регулируется в зависимости от требований в конкретной задаче. Представим тексты целиком с тем, чтобы ярче проиллюстрировать, насколько тематические метки удачны.

***Текст 1***

Официальный представитель МИД Болгарии Бетина Жотева опровергла информацию о приглашении президента Турции Реджепа Тайипа Эрдогана на празднование 138-ой годовщины освобождения от османского ига, передает ТАСС.

"В официальных мероприятиях принимают участие главы дипломатических миссий, аккредитованных в Софии", - сказала Жотева.

Он также назвала слухи о приглашении на празднование президента Турции "грубой манипуляцией, не соответствующей действительности".

День освобождения от османского ига Болгария ежегодно отмечает 3 марта. В торжествах участвуют десятки тысяч человек. Многие приходят на мероприятия с болгарскими и российскими флагами, отдавая дань уважения местным ополченцам и русским солдатам, погибшим за освобождение страны.

В 1396-1878 Болгария входила в состав Османской империи. После подавления Турцией апрельского восстания в Болгарии в 1876, а также обращения боснийских повстанцев к сербам о помощи, Сербия и Черногория объявили Османской империи войну. Россия вступила в войну 24 апреля 1877 года, которая вошла в историю как русско-турецкая война 1877-1878 годов. Русской армии и ополченцам удалось разгромить турецкие войска. В Болгарии эта война названа Освободительной.

***Текст 2***

Президент США Барак Обама отказал главе Турции Реджепу Тайипу Эрдогану в личной встрече в США, пишет The Wall Street Journal.

Источники в американском правительстве рассказали, что президент Турции попросил Обаму о встрече в штате Мэриленд, где Эрдоган будет открывать мечеть, построенную на деньги Анкары. Однако глава Белого Дома отклонил предложение Эрдогана. Кроме того, Обама не планирует никаких официальных встреч с ним в формате "один на один".

В Белом Доме подчеркнули, что отказ президента "не следует расценивать, как проявление неуважения" и напомнили, что он уже встречался с Эрдоганом в ноябре на саммите "Большой двадцатки", а в феврале провел с ним телефонную беседу.

8 марта официальный представитель Госдепартамента США Джон Кирби прокомментировал давление турецких властей на оппозиционную газету Zaman. В своем заявлении он раскритиковал руководство Турции и отметил, что оно игнорирует сигналы международного сообщества, нарушает международные принципы и собственную конституцию. За несколько недель до этого Эрдоган выступил с обвинениями в адрес США и России.

**Текст 3**

Президент США Барак Обама в интервью The Atlantic назвал "нахлебниками" некоторых союзников Вашингтона в Европе и на Ближнем Востоке.

По мнению Обамы, ряд государств пытаются за счет США решить собственные проблемы. Он не стал называть конкретные страны, однако упомянул Саудовскую Аравию — этой стране, по его мнению, следует найти способ уживаться с Ираном.

Затрагивая тему отношений с Тегераном, американский президент подчеркнул, что США не может предпочитать интересам Ирана интересы своих союзников, поскольку страны долгое время враждовали.

Барак Обама также упомянул ситуацию в Сирии. Он подчеркнул, что гордится тем, что в 2013 году США не начали прямой военный конфликт с правительством Сирии, несмотря на то, что на карту был поставлен его личный авторитет и доверие к Америке.

Террористическая группировка "Исламское государство", по мнению Обамы, не является серьезной угрозой для Соединенных Штатов. Более существенной проблемой президент США назвал изменение климата.

***Текст 4***

Управление Верховного комиссара ООН по делам (УВКБ) беженцев назвало незаконным соглашение между Евросоюзом и Анкарой, согласно которому мигрантов будут возвращать обратно в Турцию, передает Reuters.

В заявлении отмечается, что "быстрая сделка" ЕС по массовой отправке беженцев в Турцию нарушает их права на защиту, гарантированные европейским и международным правом.

Региональный директор УВКБ по Европе Винсент Коштель заявил, что планы Европы на добровольной основе переселить 20 тысяч беженцев за два года являются сомнительными. Он рассказал, что Турция в прошлом приняла очень небольшое количество беженцев из Афганистана и Ирака — в районе 3% от общего потока.

"Массовая высылка иностранцев запрещена в соответствии с Европейской конвенцией по правам человека. Соглашение о высылке беженцев в третью страну не согласуется ни с европейским, ни с международным правом", — сказал он на пресс-конференции.

Ранее лидеры Евросоюза приветствовали предложение Анкары забрать всех мигрантов, попадающих в Европу с территории Турции и дали принципиальное согласие на требования турецкой стороны дать больше денег, ускорить переговоры о вступлении в ЕС и безвизовом режиме.

### Оценка

Анализ этого и других кластеров показал, что метод хорошо ранжирует тематические слова и может быть доработан с учетом специфики применения – в самом алгоритме предусмотрена гибкая регулировка параметров. Мы планируем на основе данного ранжированного списка строить рекомендации новостей и для этого оптимизировать параметры алгоритма.

# ЗАКЛЮЧЕНИЕ

В данной работе была представлена автоматическая кластеризация документов с присвоением тематических меток. Для проведения экспериментов была написана компьютерная программа, позволяющая кластеризовать документы и выделять тематические метки из корпуса новостных текстов, используя множество тонких настроек алгоритма.

Мы считаем, что поставленная цель достигнута – документы кластеризованы, конструкции выделены, тематические метки расставлены. Развивая идеи Стефановича, мы смогли извлечь достаточно неплохие тематические конструкции, которые почти не требуют постобработки. Но это не значит, что на этом следует остановиться.

Полученные результаты можно назвать промежуточными и, в зависимости от дальнейшего направления работы, их можно адаптировать в ту или иную сторону, поэтому в программу уже заложена возможность тонкой настройки на всех уровнях работы.

В дальнейшем мы планируем приступить к разработке рекомендательных систем для сайта Рупостерс. После проведения дополнительных экспериментов и подготовки черновой версии алгоритма мы планируем связаться с руководителем портала и предложить внедрить алгоритм на сайт, но это потребует дополнительные ресурсы.

А данную работу можно считать оконченной.

# Список литературы

Новая философская энциклопедия в 4-х томах / под ред. Степина В.С. — М.: Ин-т философии РАН, Нац. обществ.-науч. фонд, Мысль, 2000—2001.

Энциклопедический словарь Брокгауза и Ефрона в 86 т. — СПб. АО «Ф. А. Брокгауз — И. А. Ефрон», 1890—1907.

Плюта В. Сравнительный многомерный анализ в экономических исследованиях — М.: Статистика, 1980. — 152 с.

Терентьев П. В. Метод корреляционных плеяд / Вестник ЛГУ. — 1959. — № 9. — С. 137—141.

Trion R. G. Cluster analysis. — L.: Ann Arbor Edwards Bros. — 1939. — 139 p.

Мандель И.Д. Кластерный анализ — М.: Финансы и статистика, 1988 — 176 с.

Воронцов К.В. Машинное обучение (курс лекций) — публикация на сайте <http://www.machinelearning.ru>

Jain A., Murty M., Flynn P. Data clustering: A review // ACM Computing Surveys. — 1999. — Vol. 31, no. 3. — Pp. 264–323.

Лагутин М. Б. Наглядная математическая статистика. — М.: П-центр, 2003.

Загоруйко Н. Г. Прикладные методы анализа данных и знаний. — Новосибирск: ИМ СО РАН, 1999.

Rosenberg A. and Hirschberg J. Detecting pitch accent using pitch-corrected energy-based predictors. — In Interspeech, 2007

Четвёркин И. И. Кластеризация оценочных слов по тональности на основе марковских цепей // Новые информационные технологии в автоматизированных системах. Вып. 16 / 2013.

Шмулевич, М. М., Пивоваров, В. С., Киселев, М. В. Метод кластеризации текстов, учитывающий совместную встречаемость ключевых терминов, и его применение к анализу тематической структуры новостного потока, а также ее динамики. – Уральский федеральный университет, 2005

Aggarwal, Charu C., ChengXiang Zhai A survey of text clustering algorithms // Mining Text Data, 77–128. Springer, 2012.

Savova, Guergana, T. Pedersen, A. Purandare, A.Kulkarni Resolving ambiguities in biomedical text with unsupervised clustering approaches – University of Minnesota Supercomputing Institute Research Report, 2005

Schütze, H. Automatic word sense discrimination // Computational linguistics 24 / 1. – MIT Press, 1998. С. 97–123.

Navigli, R. Word Sense Disambiguation: A Survey // ACM Computing Surveys 41 (2). – ACM, 2009. C. 1–69.

Lin, D.; Pantel, P. Discovering word senses from text // 8th International Conference on Knowledge Discovery and Data Mining (KDD). - Edmonton, Canada, 2002. C. 613–619.

Lin, D. Automatic retrieval and clustering of similar words // 17th International Conference on Computational linguistics (COLING). - Montreal, Canada, 1998. C. 768–774.

Филлмор Ч. Дело о падеже; Дело о падеже открывается вновь. - В кн.: Новое в зарубежной лингвистике, вып. X. - М., 1981

Fillmore Ch.J. 1982c – Frame semantics // Linguistics in the morning calm: Selected papers from the SICOL-1981. – Seoul: Hanship, 1982. – P. 111- 137.

Fillmore, Ch. J.; Kay, P.; O’Connor, M. C. 1988. Regularity and idiomaticity in grammatical constructions: The case of LET ALONE // Language, 64.3, 501—538

Goldberg, A. 1995. Constructions: A Construction Grammar Approach to Argument Structure. Chicago: University of Chicago Press.

Stefanowitsch, A.; Gries, S. Th. 2003. Collostructions: investigating the interaction between words and constructions // International Journal of Corpus Linguistics 8.2, 209—43.

Gries, Stefan Th., Anatol Stefanowitsch 2004. Extending collostructional analysis: a corpus-based perspective on ‘alternations’. International Journal of Corpus Linguistics 9 (1). Pp. 97–129

Stefanowitsch, A., Gries, S. Th. 2005. Covarying collexemes // Corpus Linguistics and Linguistic Theory 1.1, 1—43.

Woolf B. The log likelihood ratio test [the G-test]. Methods and tables for tests of heterogeneity in contingency tables // Ann. Human Genetics. 1957. V. 21. P. 397-409.

Кендалл М. Дж., Стьюарт А, Статистические выводы и связи. — М.: Наука, 1973.

Luhn H.P. A Statistical Approach to Mechanized Encoding and Searching of Literary Information. IBM Journal of Research and Development. 1957, vol. 1, no. 4, pp. 309–317.

Солтон Д. Автоматическое индексирование и реферирование // Динамические библиотечно-информационные системы. - Москва: Изд. Мир, 1979. - С. 90–132.

Соколов, А.Н. Внутренняя речь и понимание // Ученые записки государственного научно-исследовательского ин-та психологии. – М., 1941. - Т.2. - С. 99-146

Камшилова, О.Н. Малые формы научного текста: ключевые слова и аннотация (информационный аспект) //Известия Российского государственного педагогического университета им. А.И. Герцена. - 2013. - № 156. - С. 106-117.

Сахарный, Л.В. Набор ключевых слов как тип текста /Л.В. Сахарный, А.С. Штерн //Лексические аспекты в системе профессионально-ориентированного обучения иноязычной речевой деятельности. – Пермь: Перм. политехн. ун-т, 1988. - С. 34−51

Виноградова Н.В., Иванов В.К. Современные методы автоматизированного извлечения ключевых слов из текста // Информационные ресурсы России. – Москва : ФГБУ "Российское энергетическое агентство" Минэнерго РФ, 2015.

Шереметьева, С.О. Методы и модели автоматического извлечения ключевых слов / С.О. Шереметьева, П.Г. Осминин //Вестник Южно-Уральского государственного ун-та. - 2015. - № 1, т.12. - С. 76-81.

Абрамов, Е.Г. Подбор ключевых слов для научной статьи //Научная периодика: проблемы и решения. - 2011. - № 2. - С. 35−40.

Гринева, М. Анализ текстовых документов для извлечения тематически сгруппированных ключевых терминов /М. Гринева, М. Гринев //Труды Института системного программирования РАН. Т.16. - 2009. - С. 155-165.

Большакова Е.И., Клышинский Э.С., Ландэ Д.В. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика – Москва, 2011

Баканова, Н.Б. Исследование ключевых слов как инструмент оптимизации управления электронными документами [Электронный ресурс] /Н.Б. Баканова, И.В. Усманова //Современные проблемы науки и образования: электрон. науч. журн. – 2014. - № 2.

Н.Э. Ефремова, Е.И. Большакова, А.А. Носков, В.Ю. Антонов Терминологический анализ текста на основе лексико-синтаксических шаблонов – Компьютерная лингвистика и интеллектуальные технологии: по материалам ежегод. Междунар. конф. «Диалог‘2010». Вып. 9(16). - Москва: Изд-во РГГУ, 2010. - С. 124-129.

Маннинг, К.Д. Введение в информационный поиск: пер. с англ. – Вильямс, 2011

Рубцова, Ю.В. Методы автоматического извлечения терминов в динамически обновляемых коллекциях для построения словаря эмоциональной лексики на основе микроблоговой платформы Twitter // Доклады ТУСУРа. – Томск, 2014. - № 3(33). – С. 140-144

Лукашевич, Н.В. Комбинирование признаков для автоматического извлечения терминов /Н.В. Лукашевич, Ю.М. Логачев // Вычислительные методы и программирование. Т.11. - 2010. - С. 108-116.

Миронова Д.М. Автоматизированная классификация древних рукописей по степени текстовой близости (На материале 525 списков славянского Евангелия от Матфея XI-XVI вв.). Автореф. дис. канд. филол. наук. СПб., 2016.