САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ФАКУЛЬТЕТ ПРИКЛАДНОЙ МАТЕМАТИКИ–ПРОЦЕССОВ УПРАВЛЕНИЯ

КАФЕДРА ТЕХНОЛОГИИ ПРОГРАММИРОВАНИЯ

Марулин Дмитрий Алексеевич

Выпускная квалификационная работа бакалавра

Методы Фильтрации объявлений по аренде недвижимости

Направление 01.03.02

Прикладная математика и информатика

Научный руководитель,
кандидат физ.-мат. наук,
доцент
Добрынин В.Ю.

Санкт-Петербург

2018

Содержание

Введение

Российский рынок недвижимости находится в уникальном состоянии. Реалии таковы, что при аренде квартиры, офисного помещения или любой другой недвижимости, в большинстве случаев мы будем иметь дело с риелторами или агентами. Как правило, это означает, что нам дополнительно нужно будет оплачивать их услуги, стоимость которых может доходить до 50% от ежемесячной оплаты за снимаемую квартиру или помещение.

Причиной этому служит то, что отечественный рынок недвижимости довольно молодой - за всю историю России, включая советский период, массового владения частной собственностью в стране не существовало. Из-за этого российский рынок недвижимости остается не урегулированным законодательно. По сравнению с нашей страной в Европе и США рынок недвижимости существовал столетиями, за которые выработались правила рынка, требования к профессии, четко разделены функции специалистов. Например, в США брокеры занимаются поиском клиентов и продажей недвижимости, а юристы и нотариусы занимаются оформлением сделок. В России риелторы ищут объекты, рекламируют их, организуют показы недвижимости, ведут переговоры между продавцом и покупателем(арендодателем), занимаются юридическим оформлением сделки технической организацией. При этом, если в США брокеры несут ответственность, вплоть до уголовной, когда российские риэлторы в худшем случае получают негативную реакцию клиента или попадают в черный список, размещенный на одном из профессиональных сайтов.

По данным опроса Национального агентства финансовых исследований (НАФИ), в 2016 году 35% россиян не доверяют агентствам недвижимости, а еще 41% затруднились четко определить свое отношение к этой профессии. Часто используются различные уловки для привлечения потребителей.

Например, делается объявление для очень хорошего объекта со всеми условиями, с низкой стоимостью, но на деле этого объекта либо нет совсем и никогда не было, либо объект давно сдан и объявление используется только для привлечения внимания.

Обман и мошенничество на отечественном рынке недвижимости, к сожалению, очень распространены. По данным Росреестра, за 2016 год признаны недействительными 6% сделок с недвижимостью. Это действительно огромная цифра, особенно учитывая какие суммы стоят на кону при таких сделках.

Предпочтительнее для покупателя будет вести переговоры напрямую с собственником. Хотелось бы, просматривая объявления по недвижимости, быстро фильтровать их и находить объявления от собственников.Многие сайты объявлений недвижимости предлагают такую опцию, однако это не помогает исключать фальшивые объявления, сделанные специально для привлечения внимания. Конечно, арсенал уловок мошенников далеко этим не ограничивается, но даже имея возможность отсеивать такие фальшивые объявления может помочь многим людям и сохранить огромные суммы денег.

Постановка Задачи

Целью является построение с помощью методов машинного обучения классификатора, который мог бы различать объявления, написанные агентами от объявлений собственников. Объявления включают в себя такую информацию как: фотографии, описание, адрес, номер телефона. Список на этом не заканчивается – на разных сайтах, таких как, Авито, Циан, Яндекс.Недвижимость и других, параметры могут различаться. Однако из всего этого набора параметров самым информативным является описание.

В рамках данной работы делается предположение, что агенты и риэлторы при написании своих описаний для объектов, пользуются своим профессиональным языком, который не будет заметен для обычного человека. Рассматривая объявления агентов и собственников можно отметить заметные различия. В своих описаниях обычные арендодатели чаще допускают ошибки, обильно используют сокращения и слова в верхнем регистре. Также заметно отличие в использовании отдельных синтаксических структур.

Для того чтобы мы могли замечать такие отличия в объявлениях, нужно построить вероятностную тематическую модель. В такой модели каждый документ описывается распределением на множестве тем, а каждая тема описывается как распределение на множестве слов.Построить модель - значит построить две матрицы: матрицы описывающую распределение множества тем над множеством слов коллекции и матрицу описывающую распределение множества документов над множеством тем.

Для решения этой задачи потребуется большой набор данных объявлений риэлторов и собственников. Объявления были взяты с сайтов агентств недвижимости по Москве и Санкт-Петербургу. Объявления собственников были получены с помощью программы РиэлтСкан. Объявления длина описания, которых составляет менее 300 символов не

рассматривались, поскольку из них нельзя выделить достаточно необходимой информации.

В данной работе используется модель LDA(Latent Dirichlet Allocation). Методы основанные на этой модели очень популярны и эффективны для таких задач, как определение скрытых тем, различных аспектных терминов и многие другие. Этот метод позволяет использование неразмеченных коллекций документов и при добавление новых документов в коллекцию не требуется перестроение всей модели. Все это позволяет сэкономить огромное количество времени. Рассматривается стандартная реализация модели LDA, построенная с помощью сэмплирования по Гиббсу.

Построив модель на некоторой выборке агентских объявлений, к ней добавляется заранее отложенный набор объявлений агентов. Для этого не нужно перестраивать модель, достаточно лишь рассчитать необходимые коэффициенты для всех новых документов и добавить их в матрицы.

После того, как была построена модель, ее необходимо оценить. Самым лучшим критерием в данном случае является перплексия(Perplexity). Такой критерий позволяет оценить связность документов коллекции, сходство этих документов по темам, которые в них присутствуют.

Далее проводится оценка построенной модели с помощью перплексии. Меньшая перплексия означает, что добавленные документы тематически похожи на те что были там до этого, большая перплексия – документы не совпадают с, построенной до этого, моделью. Пороги для перплексии нужно вычислять эмпирически отдельно для каждой новой выборки агентских объявлений. Также строится ещё одна такая же вспомогательная модель с объявлениями собственников и считается перплексия. Процесс повторяется для нескольких выборок агентов и собственников, высчитывается перплексия. Сравниваются значения перплексий, вычисляются пороги значений. Если такие пороги вычислить не получается, например, когда

многие значения перплексии для агента и собственника чередуются, - модель отбрасывается и строится другая на новой выборке.

Такой процесс проводится для всех выборок объявлений агентов. Далее проверяется точность моделей. Берутся отдельные объявления агентов и собственников, добавляем их к нашей модели, высчитываем перплексию. По уже подсчитанным для каждой модели порогам можно понять, что модель “предполагает”, что это за объявление. Перебрав таким образом наборы объявлений двух категорий, можно будет оценить точность построенных моделей.

Наилучшие модели можно будет объединить в один большой классификатор. Также интересно было бы посмотреть как вела бы себя модель, построенная на большом наборе объявлений, включающий в себя все выборки.

Помимо этого хотелось бы также сравнить эффективность данного подхода с другими методами машинного обучения. В рамках данной работы построенный классификатор будет сравниваться с более простыми алгоритмами, такими как наивный байесовский классификатор и с мультиграммной моделью.

Обзор литературы

В данной работе была использована следующая литература:

1. Воронцов К.В. Вероятностное тематическое моделирование.

Данная статья является хорошим введением в вероятностное тематическое моделирование. Описаны задачи, которое оно решает, основные модели, критерии качества.

2. David Blei Latent Dirichlet Allocation

В данной работе предлагается алгоритм построения вероятностной тематической модели LDA, рассматриваются её реализации, теоретическое обоснование, преимущество и недостатки.

3. Heinrich G. Parameter estimation for text analysis

В данном исследовании рассматриваются различные тематические модели, способы их построения. Также рассматриваются такие понятия, как сэмплирование по Гиббсу и перплексия и как с помощью них можно построить тематическую модель и оценить ее.

4. Tutubalina E. Target-Based Topic Model for Problem Phrase Extraction.

Данная работа является примером применения методов, основанных на модели LDA. В ней решается задача определения дефекта, ориентируясь на отзывы потребителей.

5. L. Azzopardi, M. Girolami & K. van Risjbergen. Investigating the relationship between language model perplexity and IR precision-recall

В данном исследовании связь между значением перплексии и такими показателями как точность и полнота в информационном поиске.

Глава 1. Анализ предметной области

1.1. Обзор существующих решений

Машинное обучение на рынке недвижимости - это уже давно не новость. Это например, отдельные программы для оценки стоимости объектов по описанию и фотографиям, различные алгоритмы для предсказания цен на недвижимость и многие другие.

Задача, которую эту работа пытается разрешить весьма специфична. На российском рынке есть несколько программ, которые предоставляют риэлторским агентствам базы данных собственников и возможность получать получать объявления в реальном времени с самых крупных сайтов, например ЦИАН или Авито.

В частности, можно рассмотреть самую известную такую программу РиэлтСкан. Эта программа собирает все новые объявления с самых крупных сайтов недвижимости, это уже вышеупомянутые Авито, Циан, Яндекс.Недвижимость и многие другие. При помощи простых фильтров и обширной, заранее накопленной базы объявлений программа находит похожие объявления, смотрит какие из них актуальны и с помощью простых фильтров и условий принимает решение - объявления помечается как агентское или как от собственника. Проблема такого решения в том, что она построено полностью на огромной базе данных, которую нужно постоянно обновлять и следить за актуальностью.

В такой задаче, единственное что можно рассматривать и изучать - это описания таких объявлений. Таким образом, задача сводится к анализу текста и извлечения информации. В данной работе делается предположение, что при описании объектов агенты пользуются особым профессиональным языком, какие-то слова используются чаще, какие-то реже. Это могут быть какие-то уникальные фразы и синтаксические структуры, свойственные только

риелторам. Объявления собственников же могут значительно отличаться, это может быть, например отсутствие знаков пунктуации, грамматические ошибки, слова в верхнем регистре.

Для примера рассмотрим по одну описанию из каждой категории объявлений. Вот такое объявление взято с сайта агентства недвижимости Александр:

“Предлагается для проживания отличная квартира в северной части города. Сделан отличный ремонт, установлена кожаная итальянская мебель, есть вся необходимая техника. В каждой комнате установлено по телевизору. Для Вашей безопасности в квартире установлена сигнализация. Две комнаты смежные и одна изолированная. Метро в пешей доступности, рядом ТРК Норд и ТРК Гранд Каньон. Тихий, спокойный район и благоустроенный двор.”

А вот пример объявления собственника, взятого с сайта Авито:

“Сдам 1-к квартиру посуточно 39 метров на 11 этаже в 25-этажного кирпичного дома. Отличная квартира с ремонтом в новом доме. В квартире чисто уютно, чистое постельное белье, полотенца, средства гигиены. Также иметься вся не обходимая техника(телевизор, фен, чайник, холодильник, стиральная машинка, печь СВЧ, утюг.). Интернет WI-FI. До центра 10-15 мин. на машине. В 5 мин. ходьбы находиться ТЦ ЕВРОПОЛИС. Ценна может меняться в зависимости выходных и праздничных дней.”

Оба объявления во многом похожи, но есть заметные отличия. Собственник в своём объявлении зачастую обращается от своего лица, поэтому в таких объявлениях чаще встречаются такие слова как “сдам”, “сдаю”. Агенты же в своих объявлениях, чаще используют фразы “сдаётся”, “предлагается” и т.д. Помимо этого, в объявлении собственника было сделано несколько грамматических ошибок, в то время, как в агентском объявлении их нет. Также, агент в своем описании не упоминает, сколько

комнат в квартире, какая жилплощадь, какой этаж. Очень заметно также такое стилистическое отличие - “в пешей доступности” в первом объявлении и “в 5 мин. ходьбы” во втором. Это лишь те малые отличия, которые заметные с первого взгляда.

Существует несколько способов решения подобных задач. В последнее время очень популярным является модель латентного размещения Дирихле(LDA) и построенные на ней методы. Такие методы позволяют избежать больших затрат времени и помогают находить скрытые переменные, Хорошим примером применения таких методов является работа [11], где производится анализ отзывов покупателей, с целью определения дефектов в продукте, на который они жаловались.

Для выделения такого языка, необходимо построить тематическую вероятностную модель. Модель LDA подходит лучше всего, поскольку он позволяет строить модели на некотором наборе документов и быстро

1.2. Используемые данные

Для решения данной задачи требуются два набора данных – объявления собственников и объявления агентов. Как уже упоминалось ранее, в этой работе используются только описания. Объявления брались с разных сайтов в сети Интернет. При этом, нужно отметить что правила оформления описаний объектов на разных сайтов могут значительно различаться. Поэтому не рассматривались слишком короткие описания в одно или два предложения.

Объявления от собственников были получены с помощью программы РиэлтСкан. Этой программой пользуются многие агентства и риелторы. Она быстро собирает объявления с разных источников рассчитывает вероятность собственника и предоставляет удобный интерфейс для работы с объявлениями. РиэлтСкан использует простые фильтры и обширную базу

данных для расчета этой вероятности, никакие методы машинного обучения при работе этой программы не применяются.

Объявления от агентов взяты с сайтов различных агентств недвижимости по Москве и Санкт-Петербургу. Это такие агентства как: Apple Real Estate, Миэль, Инком-Недвижимость, ЭВО, БЕСТ-Недвижимость, Домострой, Арсенал-Холдинг и многие другие.

Итого имеем 10367 объявлений от собственников, 4984 объявления от агентов. Объявления агентов поделены на 31 отдельную выборку, каждая из которых состоит только из объявлений определённого агентства. самая малая сборка насчитывает 87 объявлений, самая большая – 432 объявления. В среднем 103 объявления на выборку.

Самое маленькое объявление насчитывает 336 символов, самое длинное – 3983 символа В среднем длина объявлений составляет 1399 символов.

При этом из текстов были удалены лишь предлоги и все слова были переведены в нижний регистр. Стемминг не был проведен, поскольку предполагается, что при таком малом размере документов он приведёт лишь к ухудшению результатов. Например, в рассмотренном ранее примере при проведении стемминга мы больше не сможем различить слова “сдаю”, “сдам” и “сдаётся”.

Глава 2. Теоретическое описание метода

2.1. Вероятностные тематические модели

Вероятностные тематические модели применяются для анализа коллекций текстовых данных. Они предназначены для определения тем t в документах d, и того, какие слова w образуют эти темы. В таких моделях темы представляют собой некоторые распределения над множеством всех слов коллекции, а документы - распределения над множеством тем.

Для тематических моделей делаются следующие предположения:

- не важен порядок документов в коллекции
- не важен порядок слов в документе
- разные формы слова воспринимаются как одно и то же слово
- исключены стоп-слова

Тематическая модель описывает вероятность появления слов, опираясь на гипотезу условной независимости $p(w|t) = p(w|d, t)$ и на формулу полной вероятности

$$p(w|d) = \sum_{t \varepsilon T} p(t|d)p(w|t).$$

Для того, чтобы построить тематическую вероятностную модель для коллекции документов D - необходимо построить две матрицы - матрица

- $\Phi = ||p(w|t)||$, описывающая связь между темами t и словами w и матрица
- $\Theta = ||p(t|d)||$, описывающая связь между документами d и темами t.

Для более сложных моделей многие такие предположения заменяются более реалистичным. В данной задаче, ввиду малого размера текстов, с

которыми нам приходится работать, список стоп-слов состоял только из предлогов.

2.2. LDA(Latent Dirichlet Allocation)

Существует множество методов и алгоритмов для построения вероятностных тематических моделей. В данной работе для построения тематической модели используется алгоритм LDA [2]. LDA (Latent Dirichlet Allocation) – генеративная модель. Она предполагает, что в наборе каждый документ представляет из себя смесь некоторых тем. Основная задачей этого является алгоритма – нахождение скрытых тем, с помощью которых можно раскрыть значение текста, который. С помощью модели LDA можно объяснить схожесть частей текста.

LDA очень тесно связан с вероятностным латентным семантическим анализом (Probabilistic Latent Semantic Analysis, PLSA) и устраняет основные его недостатки. При добавлении нового документа d в коллекцию распределение $p(t|d)$ в методе PLSA невозможно вычислить по тем же формулам, что и для остальных документов, не перестраивая всю модель заново. LDA же позволяет вычислять значения только для новых документов, сэкономив при этом время и ресурсы. То есть при добавлении новых документов нам не нужно пересчитывать полностью обе матрицы, мы подсчитываем по тем же правилам коэффициенты только для новых документов и вставляем их в новую матрицу. Это позволяет быстро и эффективно оценивать совпадают ли распределения в добавленном наборе документов с теми, что есть в изначальной модели.

В модели LDA задается количество тем K, количество документов M, гиперпараметры α и β для распределений Дирихле. Для их вычисления используется сэмплирование по Гиббсу.

Для каждой темы k строится распределение φ_k – слов по теме k, из которых затем строится матрица Φ:

$$\varphi_k = Dir(\beta)$$

Для каждого документа d строится распределение θ_d – тем в этом документе, из которых затем строится матрица Θ:

$$\theta_d = Dir(\alpha)$$

При этом алгоритм проходит огромное количество раз по всем документам. При построении моделей в данной задаче количество итераций составляет 3000. Это число было вычислено эмпирически и на при при таком количестве итераций модели показали наилучшие результаты. При других итерациях меньшее количество моделей достигали точности > 60% и в целом показали себя значительно хуже.

Количество тем также определялось эмпирически в результате многих экспериментов. Наилучшим количеством тем для данной задаче оказалось 120. При этом на маленьком количестве тем, значения таких тем могут быть достаточно понятны. Так одна тема может включать частое использование таких слов как “холодильник”, “телевизор”, “диван”, “кровать”, а другая “город”, “дом”, “район”,”метро” и можно понять содержание такой темы и её значение. При большем количестве тем, понять суть такой темы бывает сложно, но для данной задачи это не важно.

2.3. Сэмплирование по Гиббсу

Необходимо по набору документов построить модель LDA, то есть определить из текстов скрытые темы, их распределение по документам, а также распределение слов по темам. Для этого можно воспользоваться Сэмплированием по Гиббсу - это алгоритм генерации совместного

распределения множества скрытых тем. Основная идея этого метода заключается в выборе произвольной темы и изменению ее в зависимости от всех остальных. В исследовании [1] подробно описан один из алгоритмов сэмплирования по Гиббсу, который, в частности, использовался в данной работе.

Определим следующие параметры:

- $n_m^{(k)}$ - количество слов в документе m темы k;
- n_m - общее количество слов в документе m;
- $n_k^{(t)}$ - количество слов t в теме k
- n_k - общее количество слов в теме k

Алгоритм, описанный в работе [1] можно описать вкратце следующими шагами:

1. Обнуляются параметры $n_m^{(k)}$, n_m, $n_k^{(t)}$, n_k
2. Для каждого документа $m \in [1; M]$, для каждого слова n $\in [1; N_m]$ генерируются индексы $z_{m,n}$. В соответствии с ними увеличиваем параметры $n_m^{(k)}$, n_m, $n_k^{(t)}$, n_k

   Индексы $z_{m,n}$ подсчитываются из мультиномиального распределения

   $$z_{m,n} = Mult(m)$$
3. Для каждого документа $m \in [1; M]$, для каждого слова n $\in [1; N_m]$ уменьшаем на единицу $n_m^{(k)}$, n_m, $n_k^{(t)}$, n_k, подсчитывается новая тема k1 при помощи вероятностей принадлежности слов к темам, для слова на позиции n. При помощи $z_{m,n}$ увеличиваются значения $n_m^{(k)}$, n_m, $n_k^{(t)}$, n_k.

При вычислении темы *k1*, используется формула для вероятности того, что слово i в документе m принадлежит теме *k,* которую можно найти в

работе [1]:

$$p(z_i = k) = \frac{n^{(t)}_{k,\neg i} + \beta}{\sum_{t=1}^{V} (n^{(t)}_{k,\neg i} + \beta)} \cdot \frac{n^{(t)}_{m,\neg i} + \alpha}{(\sum_{t=1}^{V} n^{(t)}_{m,\neg i} + \alpha) - 1}$$

После этого шага рассчитываются распределения по следующим формулам:

- Распределение слов по темам

$$\varphi_{k,t} = \frac{n^{(t)}_{k} + \beta}{\sum_{t=1}^{V} (n^{(t)}_{k} + \beta)}$$

- Распределение тем по документам:

$$\theta_{k,t} = \frac{n^{(k)}_{m} + \alpha}{\sum_{k=1}^{K} (n^{(k)}_{m} + \alpha)}$$

2.4 Перплексия

После того, как модель была построена необходимо оценить её. Существует несколько способов оценки качества построенной модели. Наиболее распространенным критерием является перплексия(perplexity), равная экспоненте от минус усредненного логарифма правдоподобия. Перплексия - мера того, насколько хорошо модель предсказывает детали тестовой коллекции (чем меньше перплексия, тем лучше модель).

В работе [1] описано как можно вычислить значение перплексии по следующей формуле :

$$P = exp\ -\ \frac{\sum_{m=1}^{M}\sum_{t=1}^{V} n_m^{(t)} log(\sum_{k=1}^{K} \varphi_{k,t} \cdot \theta_{m,k})}{\sum_{m=1}^{M} N_m}$$

В данной работе перплексия вычисляется для вспомогательных моделей, в которые после первого построения модели, вставляется некоторый заранее отложенный набор документов. Поскольку используется алгоритм LDA, можно быстро, не перестраивая всю модель это сделать, подсчитав нужные коэффициенты матрицы по тем же формулам. Далее необходимо оценить эту модель с помощью перплексии.

Такая оценка говорит нам о том насколько тематически связаны или разрозненны документы. Маленькая перплексия говорит о связности этих документов, это значит что тематический состав документов похож. Большое значение перплексии означает противоположное.

В случае данной задачи маленькая перплексия означает, что модель считает, что добавленные объявления - принадлежат агентам, а большая - собственникам. Но вот оценки значений и границ, с помощью которых можно понять к какому классу отнести объявления приходится вычислять эмпирически, с помощью заранее отложенных объявлений. Для каждой такой проверки бралось по двадцать объявлений из обеих категорий.

В некоторых случаях получается, что значения для многих объявлений из тестовой выборки значения перплексии очень плотно сгруппированы, даже для объявлений из разных категорий. В таком случае получается, что модель считает, что все объявления принадлежат одной категории. Такие модели отсеиваются на этом этапе и далее не рассматриваются.

Можно рассмотреть это более подробно на примере. Скажем, модели уже построены, теперь необходимо вычислить пороги с помощью перплексии. Берутся поочередно объявления агентов и затем объявления

собственников, подставляются в модель, рассчитываются значения перплексии. Рассмотрим это на 30 объявлениях - 15 агентских, 15 от собственников.

Таблица 1. Значения перплексии для выбоки “Миэль”

Как можно видеть из таблицы 1 большинство объявлений собственников имеют более высокое значение перплексии по сравнению с объявлениями агентов. Это хороший случай, когда мы можем наглядно видеть такую тенденцию. Это говорит о том, что модель работает хорошо и способна замечать отличия между объявлениями. Пороги можно грубо оценить в 1500 для собственников и 1400 для агентов, выше верхнего порога - объявление помечается как от собственника , меньше нижнего порога - объявление от агента.

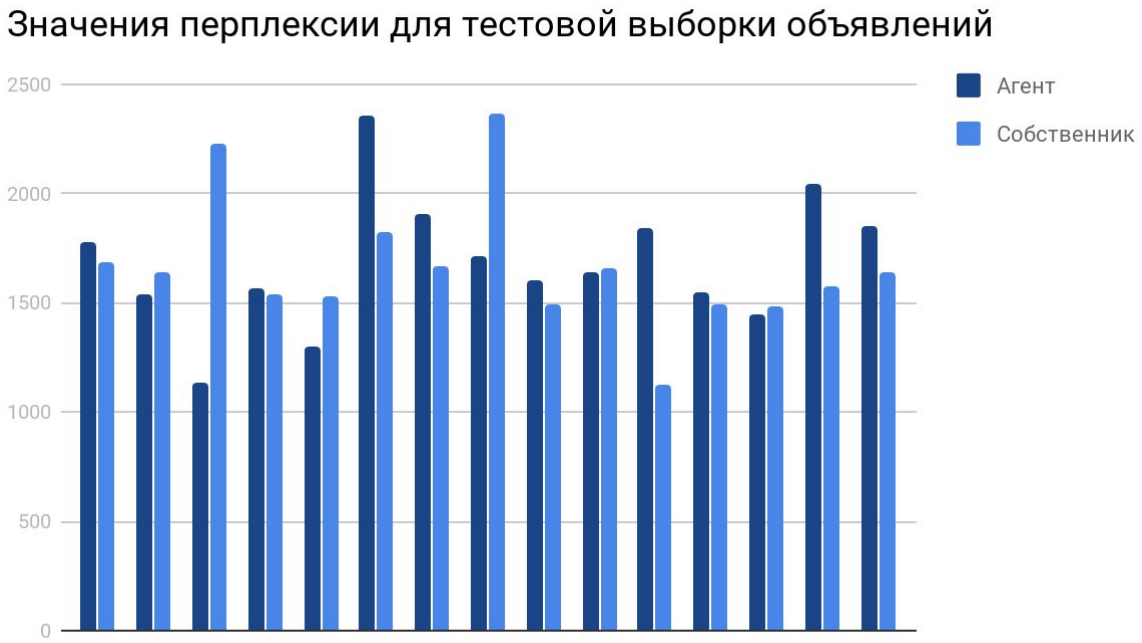

Таблица 2. Значения перплексии для выбоки “ЭВО”

Это плохой случай, когда модель плохо справляется с определением принадлежности объявления. Такая модель не может определить где собственник, а где агент, поэтому и пороги в таком случае вычислять не имеет смысла.

2.5. Описание процесса работы

Весь алгоритм можно описать с помощью рисунка 1. Через такой алгоритм проходят все выборки. На первом шаге важно отделить небольшой набор объявлений, который затем будет использоваться для вычисления порогов.

Рис. 1 Схема работы алгоритма

Глава 3 Имплементация и результаты

3.1. Инструменты разработки

Были написаны вспомогательные программы в интерактивной оболочке Jupyter Notebook на языке программирования Python. Также некоторые части программы были написаны в среде разработке PyCharm. Были использованы следующие библиотеки:

- gensim - библиотека Python для моделирования, тематическое моделирование документов и извлечения подобия с больших корпусов.
- nltk - пакет библиотек и программ для символьной и статистической обработки естественного языка.
- scikit-learn - библиотека для машинного обучения

В рамках данной работы были написаны вспомогательные программы для получения объявлений с сайтов агентств недвижимости. Была написана функция для вычисления перплексии в соответсвии с формулами, описанными в параграфе 2.3. и работе[1]. Был использован и модифицирован стандартный метод построения модели LDA из библиотеки gensim. Была написана функция, которая позволяла к существующей модели быстро добавлять наборы документов, не перестраивая при этом всю модель.

3.2. Эксперименты

В результате проведенной работы:

- Для всех 31 выборок агентских объявлений были построены вероятностные тематические модели при помощи алгоритма LDA.
- Затем были вычислены пороги значений перплексии для объявлений агентов и объявлений собственников. Те модели для которых, такие

пороги вычислить не удалось были отброшены.

- Оставшиеся модели были проверены на отдельных объявлениях агентов и собственников, были подсчитаны значения перплексии и с помощью вычисленных порогов модели “решали” были ли это объявления, написанные агентами или собственниками. Модели были проверены на 300 объявлениях агентов и 300 объявлениях собственников.
- Была подсчитана общая точность для каждой модели.

В результате экспериментов:

- Самая эффективная модель достигла точности в 72%. Это оказалась модель, построенная на выборке агентства недвижимости Александр
- Самая слабая модель достигла точности 39%. Это оказалась модель, построенная на выборке агентства недвижимости Сити-Недвижимость
- Для 11 моделей не удалось высчитать пороги перплексии для агентов и собственников. Данные модели оказались неработоспособными и далее не рассматривались.
- Количество моделей, которые достигли точности выше 60% - девять.

Более точные сведения о результатах приведены в таблице 3:

Название	Количество объявлений в выборке	Точность
Александр	302	79.3%
ЖилСтройСервис	336	75.1%
АэНБИ	287	69.5%
МГСН	350	64.2%
TWEED	194	62.9%
THE MOSCOW CITY	229	62.3%
Мира	295	61.5%
Простор	153	60.7%

Таблица 3. Лучшие модели и их результаты

Рассмотрим самую успешную модель более подробно. Из двух категорий лучше модель определяла, какие объявления были написаны агентами. Точность на таких объявлениях у данной модели достигла 89.1%, точность на объявлениях собственников достигла 69.3%.

3.3. Сравнение

Для того, чтобы оценить как хорошая построенная модель справляется с работой, можно сравнить с более простыми методами машинного обучения.

Самым простым таким методом является наивный байесовский классификатор. Обучать его можно на отдельных выборках агентов, на их комбинации и на общем наборе всех объявлений риэлторов. Тестовая выборка такая же, как и для моделей - 300 объявлений из одной категории 300 объявлений из другой.

При всех проведенных экспериментах с байесовским классификатором,

добиться точности выше 60%. Самый лучший результат достиг 58%. При этом он одинаково плохо показал себя на объявлениях агентов и на объявлениях собственников.

Так же как часть эксперимента, по выборке, модели которых дали лучший результат были отстеммированы и по ним были еще раз построены модели. Целью было выяснить, как сильно это может сказаться на результат. Для отстеммированных объявлений получили следующий результат:

Название	Точность
Александр	63.5%
ЖилСтройСервис	60.3%
АэНБИ	58.9%
МГСН	59.2%
TWEED	54.5%
THE MOSCOW CITY	57.1%
Мира	53.2%
Простор	50.4%

Таблица 4. Лучшие модели при стеммированных документах

Как можно видеть во всех случаях, стемминг приводил только к ухудшению результата. Это подтверждает, сделанные ранее предположения о том, что данный процесс лишь ухудшит модель.

Выводы

В результате проведенной работы были собраны данные для построения моделей. Были изучены методы решения подобных задач. Были построены тематические модели для каждой выборки. С помощью перплексии была оценена точность таких моделей. В результате наилучшая модель достигает 79.3% точности, что тем не менее намного лучше результатов, полученных при использовании более простых алгоритмов.

С помощью данного метода уже можно отсеивать большое число фальшивых объявлений, что поможет обычным людям чувствовать себя более комфортно на рынке недвижимости.

Заключение

Удалось построить метод оценивания объявлений по аренде недвижимости, который не требует обширной базы данных, актуальность которой необходимо постоянно поддерживать.

В дальнейшем этот метод можно следующими способами:

- найти или накопить более крупные выборки объявлений для улучшения показателей модели
- объединить лучшие модели в один классификатор для улучшения точности

Список литературы

1. Heinrich G. Parameter estimation for text analysis. Technical report, 2005.
2. D. Blei, A. Ng & M. Jordan. Latent Dirichlet allocation. In Advances in Neural Information Processing Systems 14. MIT Press, Cambridge, MA, 2002
3. T. Hofmann. Probabilistic latent semantic analysis. In Proc. of Uncertainty in Artificial Intelligence, UAI'99. Stockholm, 1999. URL http://citeseer.ist.psu. edu/hofmann99probabilistic.html.
4. Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, Richard Harshman. Indexing by Latent Semantic Analysis // JASIS (41) 1990 pp. 391-407.
5. L. Azzopardi, M. Girolami & K. van Risjbergen. Investigating the relationship between language model perplexity and IR precision-recall measures. In Proc. SIGIR. 2003.
6. M. Steyvers & T. Griffiths. Latent Semantic Analysis: A Road to Meaning, chap. Probabilistic topic models. Laurence Erlbaum, 2007.
7. X. Wei & W. B. Croft. LDA-based document models for ad hoc retrieval. In Proc. SIGIR. 2006.
8. Воронцов К.В. Вероятностное тематическое моделирование. http://www.machinelearning.ru/wiki/images/f/fb/VoronMLTopicModels.pdf
9. A. McCallum, X. Wang & A. Corrada-Emmanuel. Topic and role discovery in social networks with experiments on Enron and academic email. Journal of Artificial Intelligence Research, 30:249–272, 2007.
10. Moghaddam S., Ester M. On the design of LDA models for aspect-based opinion mining. Proceedings of the 21st ACM international conference

on Information and knowledge management. – ACM, 2012., pp. 803-812.
11. Tutubalina E. Target-Based Topic Model for Problem Phrase Extraction. Advances in Information Retrieval, 2015, pp. 271-277