Министерство образования и науки Российской Федерации
Московский физико-технический институт (государственный университет)

Выпускная квалификационная работа магистра по направлению
03.04.01 Прикладные математика и физика

Тематическое моделирование финансовых потоков корпоративных клиентов банка по транзакционным данным

Шишкина Вера Сергеевна
Группа 374

Научный руководитель: Воронцов К.В., д.ф-м.н.

Москва, 20.06.2019

Аннотация

В данной работе предлагается способ применения тематических моделей для выявления схожих по видам деятельности корпоративных клиентов банка на основе транзакций. Для нахождения интерпретируемых векторных представлений компаний, в которых каждая компонента соответствует виду экономической деятельности, предлагается применить технику аддитивной регуляризации тематических моделей, которая ранее применялась в основном для анализа текстовых коллекций. Степень схожести компаний определяется через расстояния между полученными векторными представлениями. В данной работе проводится сравнение двух типов регуляризованных мультимодальных тематических моделей в контексте данной задачи - классическая тематическая модель (TM) и Word Network модель (WNTM). Также в рамках данной задачи показана возможность использования метрик устойчивости моделей в качестве максимизируемого функционала при обучении, что избавляет от необходимости проведения разметки при переходе к новым данным.

Оглавление

1 Введение

В данной работе предлагается способ применения тематических моделей для выявления скрытых видов деятельности корпоративных клиентов банка. Полученные таким образом темы позволяют определить схожие по видам деятельности компании и компании-конкуренты, что является конечной целью данной работы. Информация о схожих компаниях позволяет банку рекомендовать своим клиентам успешные бизнес стратегии компаний, занимающихся той же деятельностью, а также в целом оценивать состояние отраслей и компаний, входящих в нее. Для крупного бизнеса, с которым работают клиентские менеджеры банков, и информацию о котором можно найти в сети, эта задача не представляет интереса. Для малого бизнеса информация о видах деятельности зачастую отсутствует, или не соответствует действительности. Поэтому в данной задаче необходимо выявить деятельность компаний из микро-малого сегмента. Отсутствие разметки и необходимое условие интерпретируемости выделенных видов деятельности в данной задаче логично подводят к использованию тематических моделей.

Тематическое моделирование - статистическая обработка естественного языка с целью выделения тематик текстовых документов. Это направление активно развивается с 90х годов и сейчас тематические модели применяются для широкого круга задач: выявление трендов в новостных потоках, архивах научных статей((Wang and Blei, 2011), Янина and Воронцов (2016)), многоязычный информационный поиск(Zhang et al. (2010)), поиск тематических сообществ в соцсетях (Ramage et al. (2010)), категоризация документов (Xie and Xing (2013)), тематическая сегментация текстов, тегирование текстов, выявление спама и др.. Вероятностная тематическая модель (probabilistic topic model) выявляет тематику коллекции документов, представляя каждую тему дискретным распределением вероятностей слов, а каждый документ - дискретным распределением вероятностей тем. Тематическое моделирование похоже на мягкую кластеризацию документов: каждый кластер - одна тема, но преимуществом тематических моделей по сравнению различными методами кластеризации является интерпретируемость кластеров, а именно наличие описаний тем с помощью слов. Хорошая тематическая модель выделяет в коллекции однородные но различные по смыслу темы, каждая из которых интерпретируема человеком.

Предлагается применить тематические модели для построения интерпретируемых векторных представлений компаний (эмбеддингов), в которых каждая компонента соответствует виду экономической деятельности. Для применения тематических моделей в данной задаче проводятся следующие аналогии: компания является аналогом документа, а экономический вид деятельности - темы. В отличии от классических тематических моделей для текстовых документов, здесь для компании нет соответствующего ей представления в виде слов, с помощью которых при описании тем словами и появляется свойство интерпретируемости. В данном случае текст, из слов которого предполагается получить описание ви-

дов деятельности, на самом деле соответствует транзакциям между клиентами банка (текст платежных поручений), а не самим компаниям. Далее полученные с помощью тематического моделирования векторные представления анализируются с точки зрения близости друг к другу, и таким образом выделяются похожие компании. Полагая, что конкурентами для компании являются фирмы, которые занимаются тем же видом деятельности и сбывают свой товар или услуги на том же рынке, что и сама компания, можно дополнительно проанализировать множество её покупателей и выделить для нее компании-конкуренты.

2 Тематические модели

Рассмотрим подробнее устройство тематических моделей. Введем обозначения:

- $D = \{d_1, d_2, \ldots, d_N\}$ - коллекция текстовых документов.
- $W = \{w_1, w_2, \ldots, w_M\}$ - множество всех слов из всех документов.
- T - конечное множество тем, $t \in T$ - скрытая переменная.

Таким образом данные представляются в виде упорядоченных троек (d_i, w_i, t_i), для $i = \overline{1, K}$. Предполагается, что темы t - скрытая переменная, которая порождает пары наблюдаемых величин (w_i, d_i). Тогда модель, при условии независимости распределения слов от распределения документов, имеет вид:

$$p(w, d) = \sum_t p(w, d|t)p(t) = \sum_t p(w|t)p(d|t)p(t) = \sum_t p(w|t)p(t|d)p(d),$$
$$p(w|d) = \frac{p(w, d)}{p(d)} = \frac{\sum_t p(w|t)p(t|d)p(d)}{p(d)} = \sum_t p(w|t)p(t|d).$$

Введем обозначения, принятые в тематическом моделировании:

- $\phi_{wt} = p(w|t)$ – матрица размера $M \times |T|$, с нормированными на единицу столбцами, которая соответствует распределению слов в темах;
- $\theta_{td} = p(t|d)$ – матрица размера $|T| \times N$, с нормированными на единицу столбцами, которая соответствует распределению тем в документах;
- $\Omega = \Phi, \Theta$ – совокупность матриц ϕ_{wt} и θ_{td} – параметры модели;
- $X = (n_{dw})_{D \times W} = (d_i, w_i)_i^n$ – матрица размера $N \times M$, значения соответствуют числу вхождений слова w в документ d – наблюдаемые данные;
- $Z = (t_i)_i^n$ – вектор размера n, значения соответствуют теме, которая породила слово w_i в документе d_i – скрытые данные.

Тогда окончательно модель порождения данных выглядит следующим образом:

$$p(w|d) = \sum_t \phi_{wt}\theta_{td}. \tag{2.1}$$

Распределение $p(w|d)$ – распределение слов в документах – известно. Задача состоит в том, чтобы найти такие $p(w|t)$ и $p(t|d)$, что $\sum_t p(w|t)p(t|d)$ будет равна известному $p(w|d)$, то есть научить модель генерировать коллекцию. Таким образом речь идет об обучении без учителя.

Полученные при обучении модели распределения $p(w|t)$ и $p(t|d)$, обеспечивают особенности тематических моделей: $p(w|t)$ позволяет интерпретировать темы, так как каждая тема имеет описание словами, а $p(t|d)$ доли участия различных тем при формировании документа. Другими словами, тематическая модель позволяет получить для документов интерпретируемые векторные представления.

Регуляризация

Задачу тематического моделирования коллекции документов можно трактовать как стохастическую факторизацию матрицы совстречаемости слов и документов $X \in R^{|W|\times|D|}$:

$$\begin{cases} X = \Phi\Theta \\ \Phi, \Theta \text{ - стохастические} \end{cases}$$

Такая задача имеет бесконечно много решений, так как $X = \Phi\Theta = \Phi(S^{-1}S)\Theta$, где ΦS^{-1} и $S\Theta$ - стохастические матрицы. Соответственно, задача является некорректной по Адамару, и необходимо добавить дополнительные ограничения на решение задачи, то есть добавить регуляризацию - оптимизацию дополнительных критериев, называемых регуляризаторами. Обозначим линейную комбинацию регуляризаторов –

$$R(\Omega) = R(\Phi, \Theta) = \sum_{i=1}^{r} \tau_i R_i(\Phi, \Theta).$$

2.1 EM алгоритм

Будем использовать принцип максимума логарифма правдоподобия: $\ln(p(X|\Omega)) = \ln \sum_z (p(X, Z|\Omega))$.

$$\ln(p(X|\Omega)) = \sum_z q(Z) \ln p(X|\Omega) = \sum_z q(Z) \ln \frac{p(X, Z|\Omega)}{p(Z|X, \Omega)} = \sum_z q(Z) \ln \frac{q(Z)p(X, Z|\Omega)}{q(Z)p(Z|X, \Omega)} =$$

$$= \sum_z q(Z) \ln p(X, Z|\Omega) - \sum_z q(Z) \ln q(Z) + \sum_z q(Z) \ln \frac{q(Z)}{p(Z|X, \Omega)} \qquad (2.2)$$

Последний член в сумме - KL-дивергенция $KL(q(Z)||p(Z|X, \Omega)) \geq 0$, тогда разность первых двух слагаемых $L(q, \Omega) = \sum_z q(Z) \ln p(X, Z|\Omega) - \sum_z q(Z) \ln q(Z)$ – нижняя оценка логарифма правдоподобия $\ln(p(X|\Omega))$. Тогда будем максимизировать эту нижнюю оценку попеременно по каждому параметру.

E-шаг: Максимизация $L(q, \Omega)$ по q:

$$\sum_z q(Z) \ln p(X, Z|\Omega) - \sum_z q(Z) \ln q(Z) = \sum_z q(Z) \ln(p(Z|X, \Omega)p(X|\Omega)) - \sum_z q(Z) \ln q(Z) =$$

$$= \sum_z q(Z) \ln p(Z|X,\Omega) + \sum_z q(Z) \ln p(X|\Omega) - \sum_z q(Z) \ln q(Z) =$$
$$= \sum_z q(Z) \ln p(Z|X,\Omega) + Const - \sum_z q(Z) \ln q(Z) \to \max_q$$

$$\sum_z q(Z) \ln p(Z|X,\Omega) - \sum_z q(Z) \ln q(Z) = -\sum_z q(Z) \ln \frac{q(Z)}{p(Z|X,\Omega)} \to \max_q$$

$$\sum_z q(Z) \ln \frac{q(Z)}{p(Z|X,\Omega)} \to \min_q . \tag{2.3}$$

Так как $KL(q(Z)||p(Z|X,\Omega)) \geq 0$, то точным решением E-шага будет:

$$q(Z) = p(Z|X,\Omega). \tag{2.4}$$

M-шаг: Максимизация $L(q,\Omega)$ по Ω:

$$L(q,\Omega) = \sum_z q(Z) \ln p(X,Z|\Omega) - \sum_z q(Z) \ln q(Z) \to \max_\Omega$$

$$\sum_z q(Z) \ln p(X,Z|\Omega) \to \max_\Omega . \tag{2.5}$$

Итого с добавлением регуляризаторов $R(\Omega)$ получаем:

$$\left[\begin{array}{ll} \text{E-шаг:} & q(Z) = p(Z|X,\Omega) \\ \text{M-шаг:} & \sum_z q(Z) \ln p(X,Z|\Omega) + R(\Omega) \to \max_\Omega \end{array} \right. \tag{2.6}$$

Получим формулы для вычисления обоих шагов. E-шаг в силу независимости элементов выборки:

$$q(Z) = p(Z|x,\Omega) = \prod_{i=1}^n p(t_i|d_i,w_i) = \prod_{i=1}^n p(\phi_{w_i t_i}, \theta_{t_i d_i}). \tag{2.7}$$

M-шаг:

$$\sum_{Z \in T^n} q(Z) ln p(X,Z|\Omega) + R(\Omega) \to \max_\Omega$$

$$\sum_{(t_1,t_2,\ldots,t_n) \in T^n} \prod_{k=1}^n p(t_k|d_k,w_k) \sum_{i=1}^n \ln p(d_i,w_i,t_i|\Omega) + R(\Omega) \to \max_\Omega$$

$$\sum_{i=1}^n \sum_{t_1 \in T^n} \cdots \sum_{t_n \in T^n} \prod_{k=1}^n p(t_k|d_k,w_k) \ln p(d_i,w_i,t_i|\Omega) + R(\Omega) \to \max_\Omega$$

$$\sum_{i=1}^n \sum_{t \in T^n} p(t|d_i,w_i) \ln p(d_i,w_i,t|\Omega) + R(\Omega) \to \max_\Omega$$

$$\sum_{d \in D} \sum_{w \in W} \sum_{t \in T} n_{dw} p(t|d,w) \ln(\phi_{wt}\theta_{td}) + R(\Phi,\Theta) \to \max_{\Phi,\Theta}$$

$$\sum_{w,t} n_{wt} \ln \phi_{wt} + \sum_{d,t} n_{td} \ln \theta_{td}) + R(\Phi,\Theta) \to \max_{\Phi,\Theta} . \tag{2.8}$$

Применим условия Каруша-Куна-Таккера к лагранжиану

$$\mathrm{L}(\Phi,\Theta) = \sum_{w,t} \ln \phi_{w,t} - \sum_t \lambda_t \left(\sum_w \phi_{w,t} - 1 \right) + \sum_{d,t} n_{t,d} \ln \theta_{t,d} - \sum_d \mu_d \left(\sum_t \theta_{td} - 1 \right) + R(\Phi,\Theta).$$

Условия ККТ для стационарной точки лагранжа:

$$\begin{array}{ll} \frac{\partial \mathrm{L}}{\partial \phi_{wt}} = \frac{n_{wt}}{\phi_{wt}} + \frac{\partial R}{\partial \phi_{wt}} - \lambda_t = 0 & \frac{\partial \mathrm{L}}{\partial \theta_{td}} = \frac{n_{td}}{\theta_{td}} + \frac{\partial R}{\partial \theta_{td}} - \mu_d = 0 \\ \left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right)_+ = \lambda_t \phi_{wt} & \left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right)_+ = \mu_d \theta_{td} \\ \phi_{wt} = norm_{w \in W}\left(n_{wt} + \phi_{wt}\frac{\partial R}{\partial \phi_{wt}}\right) & \theta_{td} = norm_{w \in W}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right) \end{array} \tag{2.9}$$

2.2 Мультимодальная тематические модели

Помимо информации содержащейся в текстах документов может быть доступна дополнительная информация о самих документах, называемая метаданными. Эту дополнительную информацию позволяют учитывать мультимодальные тематические модели. Примером метаданных является информация об авторах документа, о журнале, в котором был напечатан документ, о происхождении авторов, и тд.. Метаданные улучшают качество выделения тематик документов, а также тематические модели позволяют предсказывать пропущенные метаданные.

Каждый тип метаданных образует отдельную модальность со своим словарем. Пусть M - множество модальностей. Каждая модальность имеет свой набор токенов W_m, $m \in M$. Эти множества могут пересекаться. Мультимножество, равное арифметической сумме этих множеств, обозначим W. Модальность токена $w \in W$ будем обозначать $m(w) \in M$. Тематическая модель для каждой модальности имеет вид определенной выше тематической модели 2.1:

$$p(w|d) = \sum_t \phi_{wt}\theta_{td}, w \in W_m, d \in D. \tag{2.10}$$

Каждой модальности m соответствует стохастическая матрица $\Phi_m = (\phi_{wt})_{|W_m| \times |T|}$. Совокупность матриц Φ_m, если их записать в столбец, образует матрицу Φ размера $|W| \times |T|$. Распределение тем в каждом документе является общим для всех модальностей.

Мультимодальная модель строится путем максимизации взвешенной суммы логарифмов правдоподобия модальностей и регуляризаторов. Веса τ_m позволяют сбалансировать модальности по их важности с учетом их частотности в документах.

$$\begin{array}{l} \sum\limits_{m \in M} \tau_m \sum\limits_{d \in D} \sum\limits_{w \in W_m} n_{dw} \ln \sum\limits_{t \in T} \phi_{wt}\theta_{td} + R(\Phi, \Theta) \to \max\limits_{\Phi, \Theta} \\ \sum_{w \in W_m} \phi_{wt} = 1; \quad \phi_{wt} \geq 0; \quad \sum_{t \in T} \theta_{td} = 1; \quad \theta_{td} \geq 0; \end{array} \tag{2.11}$$

2.3 WNTM модель

Word Network Topic Model (WNTM, (Zuo et al., 2016)) - другой тип тематической модели. Принципиальное отличие этой модели от описанной выше TM состоит в использовании контекстных представлений слов. В отличии от TM модели, которая предсказывает документы, WNTM модель предсказывает совстречемости

слов. Не смотря на это WNTM модель может быть применена для выделения тематик документов, если темами считать, выделяемые WNTM моделью, скрытые группы слов. Рассмотрим устройство модели WNTM на примере моделирования коллекции документов только с одной модальностью - словами в текстах документов. Очевидным образом изложенные ниже рассуждения применяются независимо для всех модальностей в модели.

Документы представляют собой упорядоченные наборы слов. Будем называть контекстом слова w при заданной ширине окна h неупорядоченное мультимножество слов, таких что они в коллекции отстояли от слова w не дальше, чем на h слов -

$$C_w = \bigcup_{d \in D} \bigcup_{i \in ind(w,d)} \{w_{i-h}, w_{i-h+1}, \ldots, w_{i-1}, w_{i+1}, \ldots, w_{i+h-1}, w_{i+h}\}, \tag{2.12}$$

где $ind(w, d)$ - множество индексов слова w в документе d. Для каждого слова w из множества W всех слов в коллекции найдем контекст этого слова C_w.

Так как построение контекстного описания слов происходит по всей коллекции, то можно считать, что вклад контекста C_w слова w равен вкладу самого слова w. Таким образом предполагается, что

$$p(t|d) = \sum_w p(t|w)p(w|d), \tag{2.13}$$

где $p(t|w)$ соответствует вероятности появления темы t в контексте C_w.

Такой подход позволяет использовать матрицы типа $words \times words$, а не $words \times documents$, как в модели выше. Это означает, что матрица будет меннее разреженной для коллекции коротких текстов. Также такая модель в отличии от TM модели, выявляет малопредставленные в коллекции темы.

3 Постановка задачи

Целью исследования было построить векторные представления для корпоративных клиентов банка при помощи тематических моделей по транзакционным данным и внешним данным о компаниях так, что каждая компонента векторного представления интерпретируется как вид экономической деятельности. Затем на основе полученных профилей компаний определить, какие компании похожи, и какие являются конкурентами.

3.1 Данные

3.1.1 Информация о компании.

Для решения задачи имеются данные двух типов. Первый тип информации - информация о самой компании.

ИНН	Название	КПП	ОКАТО	ОКВЭДы	Сегмент
7707083893	ПАО "СБЕР-БАНК РОССИИ"	775001001	45293554000	64.19	фин. институты

Таблица 3.1. Фирма.

В таблице 3.1 представлена та информация о фирме, которая использовалась при решении данной задачи. Здесь

- ИНН - индекс налогоплательщика (10 цифр для компаний, 12 - для индивидуальных предпринимателей), использовался только в качестве идетификатора компании.
- название - два типа названия (полное и сокращенное), использовались только для интерпретации результатов.
- КПП - код причины постановки, различные части этого кода отображают различную информацию о постановке на учет в налоговую. Здесь была использована только информация о регионе (первые две цифры), в котором компания вставала на учет. Эта информация необходима для выявления филиалов одной компании, а также для выделения компаний, находящихся в заданном регионе.
- ОКАТО - код окато использовался также для выявления компаний в регионе. В нем закодирована информация о юридическом адресе компании. В совокупности с кодом КПП эта информация позволяет понять находится ли фактически компания в регионе, в котором зарегистрирована, и сколько у компании филиалов.
- ОКВЭДы - коды из всероссийского классификатора видов экономической деятельности (имеет иерархическую структуру). Компания при регистрации обозначает один основной вид деятельности и несколько дополнительных. Информация о виде деятельности компании из ОКВЭДа, который она указала, не является достоверной по нескольким причинам. Может меняться деятельность компании, или компания может изначально указать неверный ОКВЭД. Информация об указанных компаниями ОКВЭДах использовалась в моделировании.
- сегмент - обозначение класса бизнеса: микро, малый, крупный, финансовые институты и тд. Эти данные использовались для выделения компаний микро-малого бизнеса, а также при обучении модели.

Для того чтобы получить множество компаний для формирования векторных представлений производилась фильтрация компаний. Множество компаний, полученных после этой фильтрации, назовем ядром. В ходе этой фильтрации:

1. отбираются компании только из заданного региона (в данном случае из Нижегородской области);
2. из отобранных оставляются только компании из сегментов микро и малого бизнеса;

3.1.2 Транзакции

Вторым типом информации являются транзакции. Транзакции - тип информации, который отображает направленное взаимодействие объектов. В данном случае объектами являются компании, а транзакциями - платежи и переводы между компаниями. Таким образом здесь транзакции - совокупность данных таких как, покупатель, продавец, дата проведения платежа, сумма платежа, назначение платежа. Будем называть компании участвующие в одной транзакции (покупатель и продавец) контрагентами. Соответственно, если компания A продала товар компании B, то компания A является контрагентом компании B, а компания B является контрагентом A. В таблице 3.2 можно видеть пример транзакции.

Покупатель	Продавец	Назначение	Сумма	Дата
1111111111	222222222222	Оплата по сч. №520 от 28.04.2018 г. за покраску. В том числе НДС 18% - 9907.63 рублей.	64950.00	2018-04-28 15:12:56

Таблица 3.2

Для транзакций также применяются фильтры:

1. отбираются транзакции из заданного временного промежутка;
2. удаляются транзакции с крупнейшими банками России. Транзакции с банками составляют более 50% от числа всех транзакций. При этом наиболее часто встречающиеся слова в текстах платежей этих транзакций являются техническими, "банковскими" словами, которые никак не описывают экономическую деятельность фирмы. Эти слова скорее описывают множество продуктов банка, которыми пользуется компания, или идентифицируют всякого рода технические платежи (перевод денег, банковское обслуживание и т. д.). Такие транзакции можно удалить без вреда для модели и значительно сократить размер коллекции. Отфильтровываются транзакции с семью крупнейшими банками в России (ИННы и рейтинг с сайта banki.ru). Транзакция удаляется, если хотя бы один из контрагентов является банком.

Итого сперва отбирается некоторое множество компаний (ядро) посредством фильтраций на основе информации о самих компаниях, затем производится фильтрация множества транзакций. Последним шагом для получения объектов для моделирования является совмещение этих двух множеств. Каждой компании из ядра сопоставляются транзакции из множества фильтрованных транзакций, такие что компания в них участвует либо как покупатель, либо как продавец. Таким образом компания, помимо свойств обозначенных в таблице 3.1, описывается совокупностью транзакций, в которых она участвовала.

3.2 Модель

Для построения векторных представлений компаний предлагается использовать два типа тематических моделей – мультимодальные регуляризованные TM и WNTM. Проведем аналогии между задачей выделения тематик в текстовых документах и задачей определения видов деятельности компаний. Будем считать документами компании, темами - виды деятельности (например, производство мебели, услуги такси и т.д.), словами в документах - слова из текстов назначения платежей. В данной задаче необходимость использовать несколько типов информации очевидна. Таким образом мы приходим к необходимости использовать мультимодальные тематические модели. Пример двух модальностей в нашей задаче: слова из платежей, в которых компания выступала как продавец, и в которых компания выступала как покупатель - две разные модальности. Ожидается, что слова в одной модальности будут сильно отличаться от слов для другой для каждой конкретной модальности. Например, компания покупает лес, а продает мебель. Причем вид деятельности в основном определяется тем, что компания продает. Разделение слов на две модальности позволяет нам придавать больший вес одной модальности, то есть части слов. Так мы можем настроить модель, чтобы она больше внимания уделяла словам из транзакций-продаж, но в то же время учитывала слова из транзакций-покупок. Также в качестве модальностей можно брать списки покупателей и продавцов, ОКВЭДы компании, ОКВЭДы продавцов и др. Мультимодальная модель помимо использования более полной информации о документах (компаниях), еще и более интерпретируемая. Теперь для каждой модальности будет свое распределение токенов в теме. Таким образом мы надеемся получить интерпретацию темы по нескольким типам информации. Например добавив модальность с ОКВЭДами, мы сможем не только по словам понять, что продают компании принадлежащие к этому виду деятельности, но и посмотреть какой ОКВЭД наиболее сильно проявил себя в этой теме, а в идеальной ситуации, какой ОКВЭД соответствует теме.

3.2.1 Мультимодальная TM

В таблице 3.3 перечислены модальности, которые были выделены для решения задачи. Эти модальности охватывают наиболее доступную и достоверную информацию о компаниях. Также эта информация позволяет провести некоторый анализ. Информация о покупателях, например, поможет выявить не только похожие компании, но и компании-конкуренты, так как конкурентами считаются компании, производящие один и тот же продукт и сбывающие его на одном рынке. Информация об ОКВЭДах существенно облегчит задачу интерпретации тем, так как на выходе модели мы получим распределение ОКВЭДов в теме, а для каждого ОКВЭДа в общедоступном справочнике есть сформулированное текстовое описание.

3.2.2 WNTM модель

В тематической мультимодальной модели TM каждая модальность представляет собой мешок слов - неупорядоченное множество. То есть не сохраняется никакой информации о последовательности транзакций во времени. Разумным следующим шагом для улучшения качества модели будет добавление некоторой информации о временном контексте фирмы. Таким образом мы приходим к идее использования

Таблица 3.3. Модальности тематической модели.

Название	Пример значений	Описание модальности
buyers	111222333444, 1122334455	Покупатели данной фирмы (ИННы)
buyokv_0	A, Q	ОКВЭД покупателей данной фирмы, 0 уровень (буквы)
buyokv_1	A.2, Q.86	ОКВЭД покупателей данной фирмы, 1 уровень (буква и цифра через точку)
buyokv_2	A.2.40, Q.86.10	ОКВЭД покупателей данной фирмы, 2 уровень (буква и две цифры через точку)
buywords	покраска, аудит	Слова текстов платежей из транзакций, где фирма была продавцом (слова)
sellers	111222333444, 1122334455	Продавцы данной фирмы (ИННы)
sellokv_0	J, D	ОКВЭД продавцов данной фирмы, 0 уровень (буквы)
sellokv_1	J.62, D.35	ОКВЭД продавцов данной фирмы, 1 уровень (буква и цифра через точку)
sellokv_2	J.62.01, D.35.12	ОКВЭД продавцов данной фирмы, 2 уровень (буква и две цифры через точку)
sellwords	фреза, алюминий	Слова текстов платежей из транзакций, где фирма была покупателем (слова)
okv_0	H	Основной ОКВЭД фирмы, 0 уровень (буквы, одно значение для каждой фирмы)
okv_1	H.52	Основной ОКВЭД фирмы, 1 уровень (буква и цифра через точку, одно значение для каждой фирмы)
okv_2	H.52.29	Основной ОКВЭД фирмы, 2 уровень (буква и две цифры через точку, одно значение для каждой фирмы)
all_okv_0	M	Дополнительные ОКВЭДы фирмы, 0 уровень (буквы, много значений для каждой фирмы)
all_okv_1	M.71	Дополнительные ОКВЭДы фирмы, 1 уровень (буква и цифра через точку, много значений для каждой фирмы)
all_okv_2	M.71.12	Дополнительные ОКВЭДы фирмы, 2 уровень (буква и две цифры через точку, много значений для каждой фирмы)

WNTM модели - тематической модели, аналогичной мультимодальной TM, но в которой каждая модальность фирмы несет информацию о временном контексте компании. WNTM строится не по частотам токенов в документах, а по частотам локальной совместной встречаемости пар токенов. Этим данная модель похожа на Word2Vec.

Разберем подробно способ получения таких модальностей. Для каждой фирмы f можно составить упорядоченный по времени список компаний L_f, с которыми она взаимодействовала (контрагентов). Составим подобные упорядоченные списки для каждой компании в выборке. Обозначим L множество полученных списков. Рассмотрим некоторую фирму X и соответствующий ей документ упорядоченных контрагентов L_X из L (см. рис.3.1). В WNTM модели формируется псевдодокумент не для компании X, а для компании E, из сделок, которые другие компании совершали с теми же контрагентами, что и компания E.

- Рассмотрим случайную фирму E из списка L_X и обозначим ее как центральную. Будем называть контекстом центральной фирмы все фирмы, которые

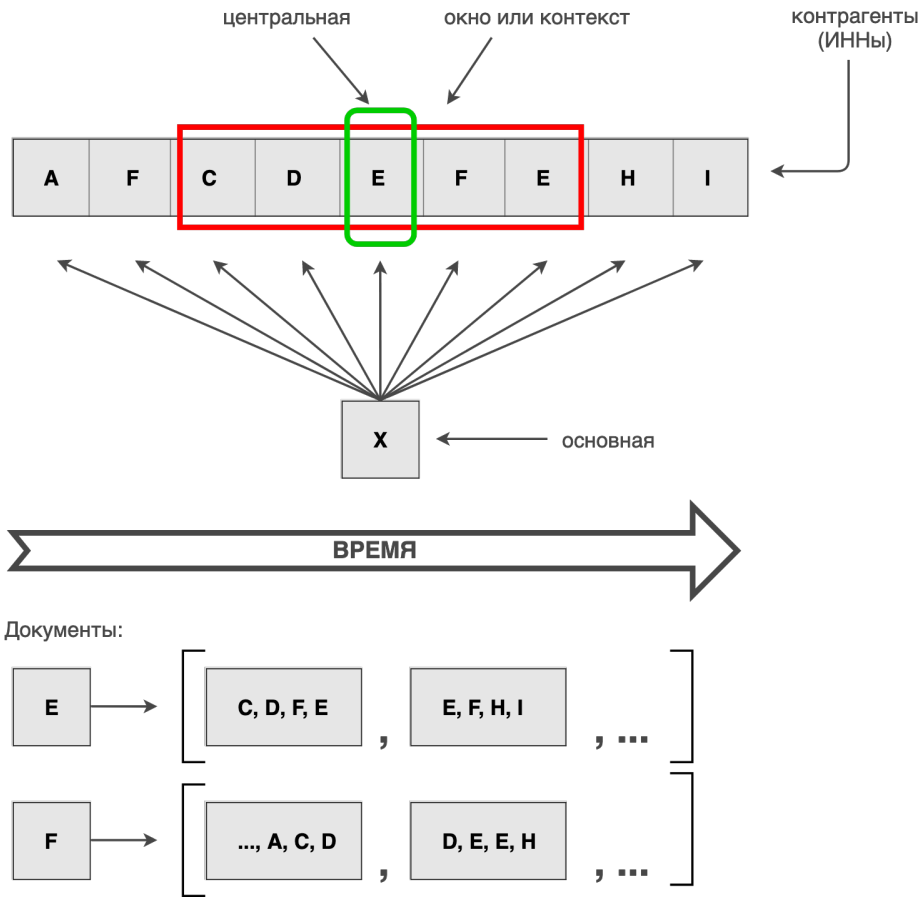


Рис. 3.1. Процесс получения wntm-модальностей.

встречались в окне фиксированного размера слева или справа от рассматриваемой центральной фирмы (фирмы C, D, F, G при окне ширины 2 на рис.3.1).

- Будем называть псевдо-документом K_E для данной центральной фирмы E совокупность фирм из всех контекстов данной фирмы в коллекции L. Такой псевдо-документ формирует модальность контрагентов для модели WNTM.
- Другие модальности формируются из этих псевдо-документов, но в качестве значений модальностей используются не сами компании из контекстов, а их некоторые описания: ОКВЭДы, слова текстов платежей, сегмент, регион.

Все опробованные wntm модальности описаны в таблице 3.4. В WNTM моделях есть возможность комбинировать обычные модальности из таблицы 3.3 с WNTM модальностями из таблицы 3.4.

3.3 Разметка и метрики

Тематическое моделирование - обучение без учителя, то есть разметка не требуется, однако тематические модели обучаются генерировать коллекцию документов и сами по себе никак не гарантируют выделение похожих документов, целью же данной задачи является выявление похожих компаний. Таким образом, необходимо проверить, насколько расстояния между векторными представлениями ком-

Таблица 3.4. Модальности wntm модели.

Название	Описание модальности	Пример значений модальностей
text_buyer, text_seller	Слова текстов платежек из транзакций, где фирма была продавцом / покупателем (слова)	покраска, аудит
agent_buyer, agent_seller	Продавцы / покупатели данной фирмы (ИННы)	111222333444, 1122334455
wntm_agent	Контрагенты в контексте центральной фирмы, когда основная фирма является или продавцом, или покупателем (ИННы)	111222333444, 1122334455
wntm_agent_seller	Контрагенты в контексте центральной фирмы-продавца, когда основная фирма является покупателем (ИННы)	111222333444, 1122334455
wntm_agent_buyer	Контрагенты в контексте центральной фирмы-покупателя, когда основная фирма является продавцом (ИННы)	111222333444, 1122334455
wntm_text	Тексты платежей в контексте центральной фирмы, когда основная фирма является или продавцом, или покупателем (слова)	покраска, аудит
wntm_text_seller	Тексты платежей в контексте центральной фирмы-продавца, когда основная фирма является покупателем (слова)	покраска, аудит
wntm_text_buyer	Тексты платежей в контексте центральной фирмы-покупателя, когда основная фирма является продавцом (слова)	покраска, аудит
wntm_okved0_seller, wntm_okved1_seller, wntm_okved_seller	ОКВЭДы 0, 1 и 2 уровня контрагентов в контексте центральной фирмы-продавца, когда основная фирма является покупателем (буквы и цифры через точку)	D, D.35, D.35.12
wntm_okved0_buyer, wntm_okved1_buyer, wntm_okved_buyer	ОКВЭДы 0, 1 и 2 уровня контрагентов в контексте центральной покупателя, когда основная фирма является продавцом (буквы и цифры через точку)	J, J.62, J.62.01
wntm_okved_all_seller	Дополнительные ОКВЭДы 2 уровня контрагентов в контексте центральной продавца, когда основная фирма является покупателем (буквы и цифры через точку)	J.62.01
wntm_okved_all_buyer	Дополнительные ОКВЭДы 2 уровня контрагентов в контексте центральной покупателя, когда основная фирма является продавцом (буквы и цифры через точку)	J.62.01
wntm_region_seller, wntm_region_buyer	Аналогично предыдущим модальностям - регион контрагента (первые две цифры из ОКАТО)	22, 45, 01
wntm_segment_seller, wntm_segemnt_buyer	Аналогично предыдущим модальностям - сегмент контрагента (слова микро, малый и малые)	микро, малый

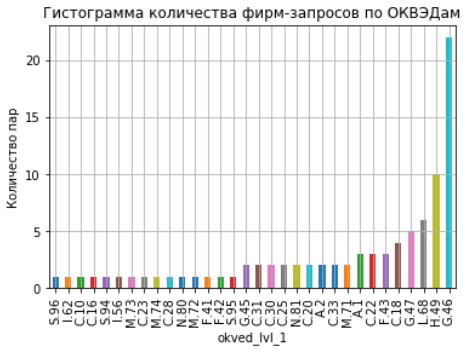
паний, полученными при помощи тематического моделирования, соответствуют похожести компаний. Назовем списком похожих для заданной компании – список компаний, упорядоченный по увеличению расстояния между эмбеддингами заданной компании и остальных компаний в коллекции.

3.3.1 Разметка

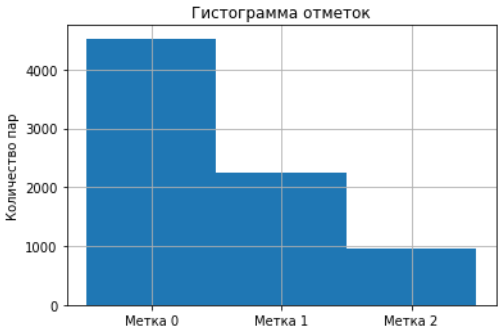
Для оценки качества модели по выявлению похожих была произведена разметка пар компаний. Разметка производилась следующим образом: случайно выби-

ралась компания (компания-запрос) из коллекции, для этой компании строился список похожих на нее компаний по косинусному расстоянию между их эмбеддингами, полученными с помощью некоторой модели. Для того чтобы разметка была получена более-менее равномерная, для построения списков использовалось несколько простых моделей:

- Tf-Idf. Эмбеддинги - векторы длины размера словаря. Компонента вектора для фирмы - tf-idf значение для слова этой компоненты.
- Счетчики контрагентов. Эмбеддинги - векторы длины количества фирм в коллекции. Компонента вектора для фирмы - количество раз когда фирма провзаимодействовала с фирмой этой компоненты.
- Суммы контрагентов. Эмбеддинги - векторы длины количества фирм в коллекции. Компонента вектора для фирмы - общая сумма денег (или среднее по транзакциям) в транзакциях, когда фирма провзаимодействовала с фирмой этой компоненты.
- Word2Vec. Эмбеддинги - не интерпретируемые векторы заданной длины. Модель на вход для обучения принимает документы - списки контрагентов компаний, упорядоченные по времени, и ширину окна.



(a) Распределение компаний в коллекции по видам деятельности. Основано на ОКВЭДах, которые указала компания.



(b) Распределение пар по типам меток во всей разметке.

Рис. 3.2

Далее полученный с помощью одной из этих моделей список из 60 наиболее похожих на фирму-запрос размечается. Похожесть пары компаний оценивается по трем параметрам: названия компаний, товарные слова в текстах платежей, ОКВЭДы. Если компания из списка похожа на компанию-запрос, то такой паре (компания, компания-запрос) ставится метка 1, если не похожа - метка 0. Если разметчик не уверен, то метка 2. Таким образом получается множество пар с метками трех типов.

Так как в коллекции преобладали компании связанные с транспортными услугами и арендой помещений (см рис. 3.2a), то в разметке, собранной с помощью случайного выбора компаний-запросов, преобладали компании именно с этими

видами деятельности. Это означает, что такая разметка показывает, насколько хорошо модель определяет компании, занимающиеся именно этими двумя видами деятельности. В связи с этим, аналогичным образом что и первая, была получена вторая разметка, но для этой разметки компании-запросы (25 компаний) выбирались из различных видов деятельности вручную. Эта разметка позволяет адекватно оценить качество модели для большего количества видов деятельности, в том числе и для видов деятельности, которые мало приставлены в коллекции.

3.3.2 Метрики

Необходимо проверить гипотезу, что наиболее похожие компании будут наиболее близкими по косинусному расстоянию их эмбеддингов, то есть в списке похожих компаний в топе будут находиться действительно похожие компании на компанию-запрос.

ROC-AUC на основе разметки

В качестве основной метрики при оценке качества моделей использовалась ROC-AUC-метрика на размеченных списках. С помощью некоторой модели строились списки похожих для фирм-запросов из разметки. Далее оставлялись в списках только фирмы, для которых есть метка для пары (фирма, фирма-запрос) из разметки. Таким образом получается список фирм, упорядоченный по увеличению косинусного расстояния между эмбеддингами фирм и эмбеддингом фирмы-запроса. Причем у каждой фирмы в этом списке есть метка, похожа она в действительности на фирму-запрос, или нет. Далее можно посчитать AUC для этого списка полагая, что мера близости - 1-мера расстояния $(1 - \cos(e_f, e_0))$. Усредним полученные значения ROC-AUC по всем спискам из разметки и получим Macro ROC-AUC. Если же мы все списки для фирм-запросов объединим в один большой список и отсортируем по косинусному расстоянию, посчитаем ROC-AUC, то получим Micro ROC-AUC. Сам алгоритм подсчета ROC-AUC был использован из пакета scikit-learn.

Метрики ранжирования списков на основе разметки

Для подсчета таких метрик также строятся списки похожих фирм с косинусной мерой расстояния для фирм-запросов из разметки. В каждом полученном списке напротив фирм ставится метка из разметки (0 - непохожи, 1 - похожи, 2 - не уверен и еще одна метка, которая говорит о том, что данной пары фирм в разметке нет). Таким образом, для каждой фирмы-запроса получается ранжированный список из меток разметки.

Метрика Rprecision@k показывает долю количества фирм, обозначенных в разметке как похожие (метка 1), среди первых k фирм списка к количеству похожих фирм в разметке для этой фирмы. Затем полученные значения усредняются по спискам для фирм-запросов из разметки.

$$Rprecision@k = \sum_{f_{query}} \frac{\sum_{i=1}^{k} [mark(f_{query}, f_i) = 1]}{\sum_{f_{markup}} [mark\,(f_{query}, f_{markup}) = 1]},$$

где f_{query} - фирма-запрос, f_i, $i = 1, \ldots, k$ - фирма на i-том месте в ранжированном списке похожих для фирмы f_{query}, полученном с помощью модели, $mark(f_{query}, f_i)$ - значение разметки для пары фирмы-запроса f_{query} и фирмы f_i.

Для подсчета метрика MeanAvgPrecision@k(list) в каждом списке похожих для фирм-запросов остаются только те компании, для которых есть оценка в разметке (метка 0 или 1). Полученный ранжированный список нулей и единиц оценивается следующим образом:

$$MeanAvgPrecision@k(list) = \sum_{f_{query}} \left(\sum_{i=1}^{len(list)} mark(f_{query}, f_i) * \frac{\sum_{j=1}^{i} mark(f_j, f_i)}{i} \right),$$

где $len(list)$ - длина получившегося списка (после выбрасывания некоторых пар не из разметки) для f_{query}.

Чем выше указанные метрики для модели, тем более похожими на разметку получаются списки.

Иерархия

Классификатор ОКВЭДов имеет иерархическую структуру. Эта метрика для оценки качества модели не показывает напрямую, насколько хорошо модель группирует похожие компании. Она позволяет оценить, насколько те виды деятельности, которые выделяет модель соответствуют классификатору ОКВЭД, а именно его иерархии. В идеальном случае наша модель должна, во-первых, каждую тему (вид деятельности) привязать к одному ОКВЭДу, причем не обязательно, чтобы для всех тем ОКВЭДы были одинакового уровня, а во-вторых, выполнить иерархию классификатора ОКВЭДов. Будем считать, что иерархия ОКВЭДов сохраняется, если при доминировании темы t в распределении тем для дочернего ОКВЭДа, тема t присутствует и в распределении тем для родительского ОКВЭДа. Определим плохость сохранения моделью иерархии ОКВЭДов:

$$H = \sum_{d \in d_{level}} \frac{V_d}{V_r} \frac{KL(p(t|d)||p(t|r_t))}{KL(p(t|d)||p(t|r))},$$

где KL - дивергенция Кульбака-Лейблера, V_d и V_r - веса дочернего и родительского ОКВЭДов в коллекции соответственно (зависят от нормировки в модели, например, сумма денег во всех транзакциях, где участвовала компания с таким ОКВЭДом), d_{level} - обозначение уровня иерархии к которому принадлежит дочерний ОКВЭД, $p(t|d)$ - распределение тем для дочернего ОКВЭДа, $p(t|r)$ - распределение тем для родительского ОКВЭДа, который является ближайшим по косинусному расстоянию к дочернему ОКВЭДу d из ОКВЭДов принадлежащих более высокому (родительскому) уровню иерархии, $p(t|r_t)$ - распределение тем для ОКВЭДа, являющегося родительским по отношению к d.

Таким образом эта величина описывает насколько в среднем по всем ОКВЭДам из дочернего уровня дочерний ОКВЭД ближе к истинному родительскому ОКВЭДу, чем к любому другому из родительского уровня иерархии. Если эта величина близка к 0, модель хорошо выполняет иерархию ОКВЭДов, если к 1, то плохо.

К сожалению эта метрика может вычисляться только как побочная, дополнительная, так как в следствии особенностей формулы, метрика может вырождаться даже при условии адекватной модели. Также как выяснилось в ходе проведения экспериментов, сам классификатор видов деятельности может иметь нелогичности, и если дочерний ОКВЭД привязывается не к своему родительскому, это не всегда плохо, и на взгляд человека решение модели может показаться даже лучшим, чем соответствие классификатору видов деятельности.

Устойчивость

Еще одна метрика, которая измеряет не качество поиска похожих компаний, а параметры самой модели. В модель могут подаваться различные данные. Это могут быть различные регионы, временные промежутки, разные компании по количеству транзакций, или среднему чеку. При этом модель должна работать для любых допустимых входных данных и не должна сильно меняться при незначительном изменении входных данных. Это свойство моделей называется устойчивостью. Подробнее об экспериментах и методах оценки оценки устойчивости моделей будет рассказано в разделе 4.3.

4 Эксперименты

4.1 Выделение товарных слов из текстов платежных поручений

Для улучшения качества интерпретируемости выделяемых тем, или экономических видов деятельности, было произведено выделение товарных слов из текстов платежных поручений. Текст назначения платежа обычно выглядит следующим образом:

Оплата по сч. №520 от 28.04.2018 г. за покраску. В том числе НДС 18% - 9907.63 рублей.

Очевидно, что некоторое понимание о виде деятельность продающей компании нам дает только слово "покраску". Остальные слова лишь будут мешать построению качественной модели. Будем называть товарными словами слова, соответствующие товарам или услугам, характеризующим деятельность фирмы-продавца. Так как тексты назначения платежей обеспечивают интерпретируемость тем, необходимо выделить товарные слова, и затем лематизировать их, чтобы модель распознавала слова "покраску"и "покраски как одно и то же слово.

4.1.1 Ручная разметка товарных слов

Чтобы говорить о качестве моделей для выделения товарных слов, необходима разметка. Разметка формировалась следующим образом:

- Для формирования разметки использовались тексты платежных поручений, семплированые из коллекции всех платежных поручений для Нижнего Новгорода. Общее количество размеченных данных составляет 4610 тысяч платежных поручений.

- При разметке текст рассматривается как последовательность слов, часть из которых является товарными словами. При этом слова, которые не являются товарными в контексте данного платежного поручения (например, фигурирующее в тексте название фирмы “ООО Семейная Аптека”), не считаются товарными при разметке.

- При разметке из платежного поручения выделяются товарные слова и сохраняются в исходной форме, без применения лематизации и иных средств, изменяющих форму слова.

- Таким образом, для каждого оригинального текста платежного поручения получается набор слов из него, которые считаются товарными.

4.1.2 Модели выделения словаря товарных слов

Для выделения товарных слов из текстов платежных поручений было опребованно два метода. Один (Word2Vec) основан на гипотезе схожестивекторных представлений тованых слов по сравнению с фоновыми словами, а второй (Контекстный алгоритм) на том, что все платежные поручения имеют один и тот же шаблон.

Алгоритм, основанный на word2vec

1. Первоначально обучается word2vec типа CBOW (Continuous Bag-of-Words Model) на коллекции текстов платежных поручений. Результатом обучения word2vec модели являются две матрицы векторных представлений U (основная матрица) и V (т. н. “негативная” матрица).

2. При помощи соответствующих скалярных произведений и норм (возведения в экспоненту с последующей нормировкой) получается величина, характеризующую вероятность нахождения заранее выбранного слова в рассматриваемом контексте.

3. Например, имеется следующий контекст: “по, счету, сумма, рублей”, и необходимо предсказывать вероятность слова "товар". Тогда искомой величиной будет косинусное расстояние между вектором, полученным усреднением векторов контекстных слов из V, и вектором слова “товар” из U. Также можно брать все векторные представления из одной матрицы, а не из разных.

4. Вводится небольшое множество положительных слов, таких как "товар”, “авто". Далее для слова платежного поручения вычисляется две величины: косинусное расстояние слова до множества положительных слов и максимальная вероятность нахождения слова из этого множества при условии его контекста. Если обе эти величины больше порогов (они подбирались экспериментально) рассматриваемое слово считается товарным. Этот шаг повторяется для всех текстов поручений в коллекции.

Контекстный алгоритм

1. С помощью регулярных выражений выявляются “гарантированные” товарные слова из всех платежных поручений.

2. Для каждой фирмы агрегируются все платежные поручения, в которых она выступает как продавец.

3. Выявляется некоторое локальное множество товарных слов: либо все слова, встретившиеся в тех платежках из агрегированных, в которых с помощью регулярных выражений удалось выделить товары, либо из тех же самых слов только те, у которых документная частота (document frequency) в данном агрегированном множестве (т.е. доля платежек фирмы, в которых слово фигурирует) выше по средней документной частоты по всем фирмам.

4. Далее происходит второй “пробег” по платежным поручениям фирмы, в которых первоначально с помощью регулярных выражений не удалось найти товарные слова, и ищутся слова из полученного ранее локального множества товарных слов.

Метрика качества выделения товарных слов

Для оценки качества выделения товарных слов моделями на разметке были рассчитаны классические метрики точности, полноты и F1 мера. Рассматриваются размеченные платежные поручения (оригинальный текст и выделенные разметчиком товарные слова) и товарные слова, полученные из этих же текстов каким-либо алгоритмом.
Сравниваются множество истинных товарных слов с множеством товарных слов, полученных в результате алгоритма. Для каждой платежки считаются следующие величины:

FN - алгоритм не считает истинное товарное слово товарным;

TP - алгоритм считает истинное товарное слово товарным;

TN - алгоритм не считает слово товарным и оно не является таковым (его нет в истинных товарных словах);

FP - алгоритм считает слово товарным, хотя оно не является таковым (его нет в истинных товарных словах).

Далее вычисляется точность и полнота по следующим формулам:

$$Precision = \frac{TP}{TP + FP}, Recall = \frac{TP}{TP + FN}.$$

Для каждого текста вычисляются описанные метрики и усредняются по всем текстам платежек.

Сравнение качества разных моделей выделения товарных слов

В таблице 4.1 представлены результаты метрик для каждой из описанных моделей, причем модель с выделением товарных слов при помощи регулярных выражений представлена в двух вариантах. В одном совершается только один проход с поиском слов из регулярных выражений (пункты 1-3 в 4.1.2), а во втором используется дополнительный проход с выделением уже найденных слов в транзакциях без наличия регулярных шаблонов (пункты 1-4 в 4.1.2).

Алгоритм	Precision	Recall
W2V	0.71	0.71
Контекстный	0.70	0.70
Контекстный с агрегацией	0.75	0.90

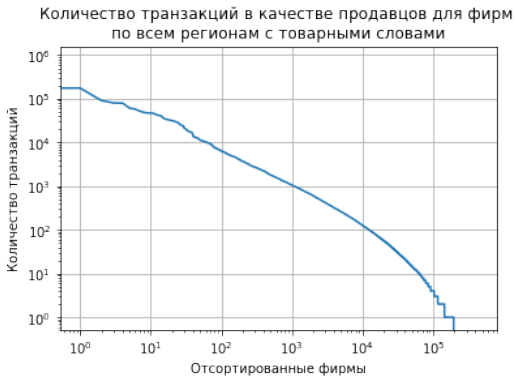
Таблица 4.1. Качество моделей выделения товарных слов.

4.2 Эксперименты с моделями

В общем случае запуск модели происходит следующим образом:

1. Выгрузка всех транзакций для всех компаний из Нижнего Новгорода за определенный промежуток времени (полгода, год).
2. Удаление транзакций с крупнейшими банками.
3. Фильтрация компаний по минимальному количеству продаж за этот промежуток.
4. Выделение товарных слов в текстах оставшихся транзакций.
5. Выбор модальностей (либо для TM, либо для WNTM). Сбор данных в vowpal wabbit файл.
6. Выбор параметров модели: количество тем, количество фоновых тем, количество итераций EM алгоритма.
7. Подбор оптимальных с точки зрения AUCa на разметке весов модальностей (без включения регуляризаторов).
8. Подбор весов регуляризаторов с использованием полученных весов модальностей.
9. Финальный запуск модели с выбранными параметрами модели и весами модальностей и регуляризаторов. Сохранение модели.

На шаге 3 порог отсечения компаний - около 20 транзакций в качестве продавца. Этот порог соответствует резкому увеличению количества транзакций, то есть граница хвоста распределения количества компаний от



количества транзакций. На рисунке 4.1 представлено распределение количества транзакций в качестве продавца в коллекции. График распределения количества транзакций в качестве покупателя в коллекции имеет аналогичный вид.

4.2.1 Подбор числа тем

В таблице 4.2 представлены результаты исследования на оптимальное количество тем для TM модели. Наилучший результат был получен для 400 тем. Этот результат представляется разумным, потому что ОКВЭДов второго уровня порядка 700. Для WNTM модели результат почти не менялся в широком диапазоне количества тем, поэтому для единообразия было принято решение все следующие модели и TM, и WNTM строить для одинакового количества тем - 400.

Число тем	Случайная разметка		Разнообразная разметка		Вся разметка	
	AUC	MAP	AUC	MAP	AUC	MAP
200	0.832	0.582	0.755	0.599	0.781	0.545
300	0.818	0.564	0.744	0.591	0.771	0.54
400	0.849	0.585	0.757	0.583	0.808	0.557
500	0.833	0.563	0.743	0.596	0.787	0.557
600	0.823	0.568	0.765	0.623	0.786	0.569

Таблица 4.2

4.2.2 Подбор весов модальностей

Обе модели TM и WNTM являются мультимодальными, и не все модальности в одинаковой степени важны для выделения видов деятельности. Тематическое моделирование позволяет учитывать степени важности отдельных модальностей. Так, например, по предположению деятельность компании в качестве продавца намного больше характеризует вид деятельности компании, чем деятельность в качестве покупателя. Но для многих модальностей угадать соотношения важности информаций проблематично. Часто важной оказывается неожиданный тип информации. Поэтому был создан алгоритм для автоматического подбора весов модальностей.

Предполагается, что итоговые оптимальные веса являются произведением важности информации в модальности и доли токенов модальности в коллекции. Домножение на долю токенов модальности в коллекции обеспечивает одинаковый вклад модальностей в работу EM алгоритма. Начальные веса берутся как величина, обратная средней мощности документа в коллекции по рассматриваемой модальности (если счетчиками являются не встречаемости токенов, а какие-то другие числовые характеристики, то берется сумма этих характеристик по документу). Например, это может быть среднее количество слов в платежках компании. Нижеприведенный алгоритм подбирает те веса модальностей, которые отвечают за важность информации и являются лишь одной из двух частей итоговых весов.

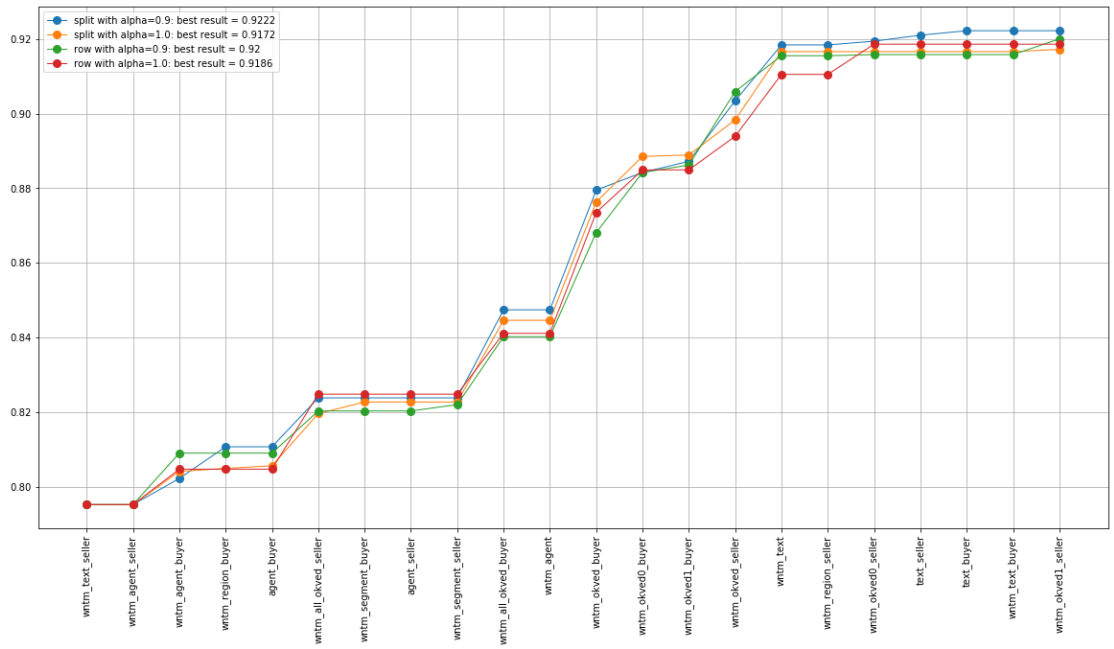


Рис. 4.2. Изменение метрики AUC в ходе итерации алгоритма по подбору весов модальностей.

Алгоритм использует жадную пошаговую аддитивную стратегию максимизации заданного функционала качества (например, качество ROC-AUC списков похожих компаний).

Шаг 1. Выбираются две модальности, и подбирается выпуклая комбинация весов (важности) этих модальностей, дающая максимальное значение функционала. Таким образом подбирается λ_1 такое, что весами модальностей являются значения $(1 - \lambda_1)$ и λ_1 соответственно. При этом здесь и далее модель, дающая максимум функционала, получила на вход веса $(1 - \lambda_1) * w_2$ и $\lambda_1 * w_1$ соответственно (здесь w_1 и w_2 доля токенов соответствующей модальности в коллекции).

Шаг 2. Соотношение между этими двумя модальностями фиксируется, и подбирается λ_2 такое, что выпуклая комбинация первых двух модальностей и некоторой новой модальности дают максимальное значение функционала. Таким образом новые веса равняются $(1 - \lambda_1) \cdot (1 - \lambda_2)$, $\lambda_1 \cdot (1 - \lambda_2)$ и λ_2 соответственно.

Шаг k. Зафиксировано соотношение между первыми k модальностями. Подбирается k такое, что выпуклая комбинация первых k модальностей с новой модальностью дает наилучшее значение функционала.

Для поиска наилучшего значения важности *lambda* для новой модальности значение либо перебирается по сетке, либо используется метод золотого сечения. Также для уравновешивания весов есть возможность сжимать каждый следующий отрезок, на котором идет поиск , на некоторую постоянную величину, так как с добавлением в алгоритм каждой следующей модальности, значения весов важности становятся все более сгруппированы вокруг нуля.

Были проведены эксперименты с различными методами поиска оптимального λ для каждой модальности:

- Деление отрезка методом золотого сечения. Всего 8 различных λ для одной модальности. Область поиска каждой следующей модальности сжималась в $\alpha = 0.9$ раз.
- Деление отрезка методом золотого сечения. Всего 8 различных λ для одной модальности. Область поиска каждой следующей модальности не изменялась.
- Проход по равномерной сетке. Всего 12 различных λ для одной модальности. Область поиска каждой следующей модальности сжималась в $\alpha = 0.9$ раз.
- Проход по равномерной сетке. Всего 12 различных λ для одной модальности. Область поиска каждой следующей модальности не изменялась.

На графике 4.2, показано, как менялось значение метрики AUC в зависимости от шага алгоритма (добавления модальности). Здесь split означает деление отрезка методом золотого сечения, row - проход по сетке, α - коэффициент сжатия отрезка. Так как качество полученных наилучших моделей примерно одинаковое, то было решено пользоваться методом золотого сечения, так как этот способ требует меньших затрат по времени.

Тематическая модель

Модальность	λ
@buyers	0,009130199
@sellers	0,182680704
@buywords	0,164847525
@sellwords	0,003202765
@all_okv_0	0,016169312
@all_okv_2	0,583230986
@okv_0	0,03283851
@sellokv_0	0,0079

Таблица 4.3. Веса ТМ модели.

Подбор регуляризаторов проводился для всех 16 модальностей описанных подробно в пункте 3.2.1. Результаты работы алгоритма представлены в таблице 4.3. Как видно из таблицы, алгоритм подбора весов модальностей отобрал ровно половину значимых модальностей. Остальные модальности либо не влияют на модель, либо вносят шум. Как и предполагалось информация о контрагентах полностью осталась. Причем информация о продавцах оказалась на два порядка важнее, чем информация о покупателях, что не интуитивно. Зато информация о словах из транзакций, где компания продавала гораздо важнее информации о словах из транзакций, где компания покупала. То есть модель считает, что товары, которые компании производят, и соответственно потом продают, определяют схожесть компаний, и схожесть компаний слабо зависит от того, какие товары компании покупают. Это имеет смысл. Если две компании покупают пирожки на обед своим сотрудникам, то это не значит, что они похожи. Одна из них может производить одежду, а другая быть типографией. Но если две компании продают одежду, логично считать их виды деятельности похожими.

WNTM модель

Подбор модальностей проводился для всех 22-х модальностей описанных в пункте 3.2.2. В таблице 4.4 представлены результаты работы алгоритма для WNTM модели. Здесь численно отображено оптимальное соотношение влияния информации из той или иной модальности на модель. Таким образом самой ценной информацией оказалась информация об ОКВЭДах. wntm_okved0_seller для компании отображает ОКВЭДы тех компаний, которые покупали у того же поставщика товаров или услуг, что и сама компания. То есть если сама компания покупает лес у компании А в июне 2019 года, то в модальность попадут ОКВЭДы компаний, которые покупали у компании А в июне (не обязательно лес).

Модальность	λ
wntm_okved0_seller	0,364
wntm_okved0_buyer	0,149
wntm_okved_seller	0,275
wntm_okved_buyer	0,112
wntm_agent	0,029
text_seller	0,071

Таблица 4.4. Веса WNTM модели.

4.2.3 Подбор весов регуляризаторов

Схожим образом с подбором весов модальностей подбираются веса регуляризаторов. Регуляризаторы, которые использовались в моделировании:

- Сглаживания/Разреживания ϕ:

$$p_{wt} \propto (n_{wt} \mp \tau)$$

- Сглаживания/Разреживания θ:

$$p_{td} \propto (n_{td} \mp \tau * dict[t])$$

- Сглаживания $p_{tdw} = norm_t(\phi_{wt}\theta_{td})$:

$$p_{tdw} \propto (p_{tdw} - \tau)$$

- Декорреляция ϕ:

$$p_{wt} \propto (n_{wt} - \tau * p_{wt} * \sum_{s \in T, s \neq t} (p_{ws}))$$

Здесь $\tau \in \mathbb{R}_+$ - коэффициент регуляризации, $dict[t]$ - множество тематик, для которых необходимо провести регуляризацию. Этот параметр позволяет ввести фоновые темы в модель путем сглаживания небольшого числа тем и одновременного разреживания оставшихся. В контексте решаемой задачи фоновой темой могут быть технические платежи с банком, налоги, пошлины и т.д., оплата комунальных услуг и транспортных перевозок. Эти платежи свойственны всем компаниям и никак не характеризуют их деятельность.

На вход алгоритму для подбора регуляризаторов подается список регуляризаторов (список может быть случайно перемешан по желанию пользователя). Для каждого регуляризатора указано, сколько нужно сделать итераций и в каких пределах искать коэффициенты регуляризации для каждого из них. Алгоритм аналогичен алгоритму подбора весов модальностей, поэтому кратко опишем его особенности.

Шаг 1. Модель инициализируется, или предварительно обученная модель подается на вход в качестве параметра. Выбирается первый регуляризатор $R1$ из списка, и методом золотого сечения ведется поиск τ_1 для этого регуляризатора такого, чтобы качество модели с этим регуляризатором было больше качества модели без него с тем же количеством итераций. Если такого коэффициента нет, то лучшей считается модель без регуляризатора. Иначе - модель с регуляризатором.

Шаг k. Аналогичным образом происходит подбор коэффициентов τ_k для нового регуляризатора R_k с лучшей моделью на предыдущем шаге. При этом качество получаемых моделей сравнивается уже не с моделью без регуляризаторов, а с лучшей моделью на предыдущем шаге, то есть сравнение происходит с моделью, с которой начался подбор. При этом если в качестве одного из возможных значений τ_k будет задан 0, то одной из моделью при подборе будет модель без регуляризаторов, и тогда сравнение будет также и с ней.

TM модель

Выяснилось, что полезными регуляризаторами для мультимодальной TM модели будут сглаживание p_{tdw} на первых нескольких итерациях, а затем декорреляция матрицы ϕ. Веса этих регуляризаторов в каждом конкретном случае приходится подбирать заново, поэтому их значения здесь не приводятся.

WNTM модель

Для WNTM модели, как выяснилось, регуляризаторы не только не дают улучшения с точки зрения метрики AUC на разметке, но и портят модель.

4.3 Анализ устойчивости моделей

Устойчивость модели — свойство модели, характеризующее ее способность обеспечить отклонение результатов не более чем на допустимо малую величину в условиях наличия определенных возмущений на входных данных.

Так как данная задача является обучением без учителя, нет возможности оценить отклонение результата на возмущенных данных относительно истинного. Поэтому в данной работе предлагается оценить различие результатов двух моделей, одна из которых строится на имеющихся данных, а для другой имеющиеся данные изменяются. Например, изменением данных может быть удаление части транзакций для компании, или удаление информации об ОКВЭДе. Так как ОКВЭД компания указывает всегда, то всевозможное разнообразие исходных данных можно описать всевозможными наборами транзакций для компаний. Будем рассматривать в качестве возмущения удаление части транзакций для компаний. Тогда ожидается, что эмбеддинг для такой компании с частью транзакций будет близок к эмбеддингу компании с полным набором транзакций, то есть компания будет похожа на саму себя после выкидывания транзакций. Тогда, как следствие, компания с частью транзакций будет похожа на компании, похожие на полную компанию, а значит модель сильно не поменяется, и ее можно будет назвать устойчивой.

Алгоритм проведения эксперимента для анализа устойчивости модели:

1. Из компаний отобранных для моделирования отбираются несколько компаний (1000 компаний), у которых транзакций в качестве продавца не меньше 200, и которые не участвовали в разметке на схожесть компаний как компании-запросы.

2. Транзакции для каждой из отобранных компаний делятся на две части случайным образом в заданном соотношении (были опробованы следующие соотношения: 25/75, 50/50).

3. Создаются две новые псевдокомпании для каждой компании. Эти псевдокомпании являются копиями исходной компании (те же ОКВЭДы, названия и другая информация о самих компаниях), за исключением того, что каждой из них приписывается только одна из частей всех транзакций компании. Например, одной компании могло достаться 75% транзакций, а другой - оставшиеся 25%, если деление на шаге 2 производилось в отношении 25/75.

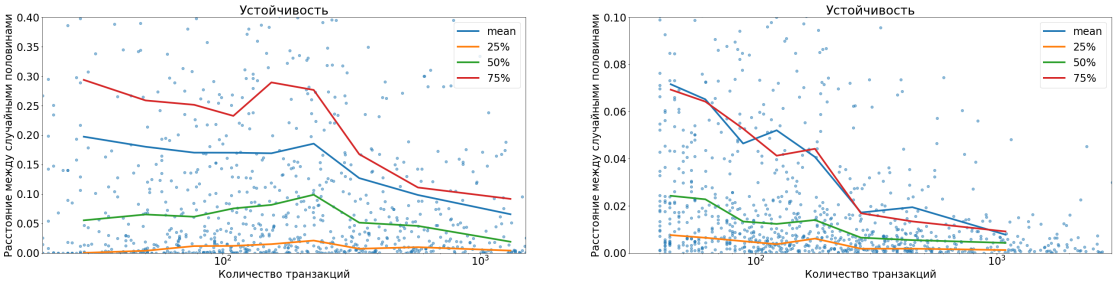
4. Далее в данных каждая из отобранных на шаге 1 компания, заменяется двумя псевдокомпаниями, которые получились из нее на шаге 3. В моделировании сама исходная компания не участвует.

5. Обучается модель. Для каждой компании, которая была разделена на две, выбирается одна из псевдокомпаний. Для этой псевдокомпании строится список похожих компаний.

6. В получившемся списке похожих компаний для псевдокомпании определяется порядковый номер псевдокомпании, которая является второй частью той же исходной компании, что и псевдокомпания, для которой строился список.

7. Измеряется средний порядковый номер второй псевдокомпании в списке первой для разделенной компании.

8. Измеряется среднее косинусное расстояние для пары псевдокомпаний среди разделенных компании.



(a) TM. (b) WNTM.

Рис. 4.3. Здесь компаниям соответствуют точки, с координатами: по одной оси - расстояния между двумя частями компаний, а по другой - количество транзакций с компанией в данных. Всего здесь отображено 1000 точек. Для обеих моделей точки соответствуют одному и тому же набору компаний с одинаковым их разбиением на пары псевдокомпаний.

На рисунках 4.3 видно, что в среднем расстояния между псевдокомпаниями небольшие - от 0.2 до 0.1 для TM модели, и от 0.08 до 0.01 для WNTM модели. Для WNTM модели в среднем расстояние уменьшается с ростом числа транзакций для компании, и достаточно наличия всего лишь порядка сотни транзакций у компаний для устойчивости. Для TM модели расстояния также уменьшаются с ростом числа транзакций, но для нее даже для компаний с порядка тысячи транзакциями еще нельзя уверенно говорить об устойчивости. Таким образом мы выяснили, что модели обладают некоторой устойчивостью по отношению к входным данным. Теперь необходимо выяснить, насколько устойчивость модели согласуется с её качеством. То есть, можно ли утверждать, что чем качественнее относительно нашей разметки модель, тем она устойчивее.

Алгоритм проведения эксперимента для анализа корреляции устойчивости и качества модели:

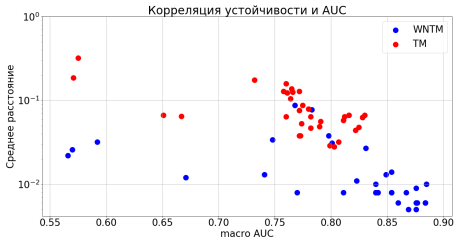
1. Отбирается 1000 компаний для исследования на устойчивость, таких, что у них не менее 200 транзакций в качестве продавца.

2. Строится несколько моделей с делением отобранных компаний пополам по алгоритму представленному выше.

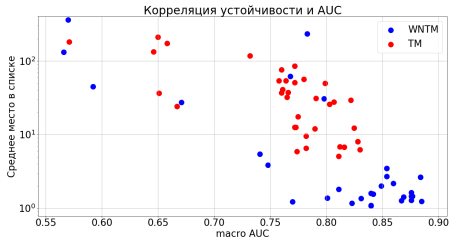
3. Для этих моделей измеряется метрика AUC на разметке похожих и метрики устойчивости - среднее место в списке второй половины компании, среднее косинусное расстояние до второй половины.

4. Измеряется корреляция AUC и метрики устойчивости.

Таким образом было получено по 30 точек (моделей) для каждого типа модели (WNTM и TM). От модели к модели менялось количество тем, модальностей, веса модальностей, наличие и веса регуляризаторов. Результаты представлены на графиках 4.4a и 4.4b.



(a) Зависимость среднего расстояния в парах псевдокомпаний от AUC.



(b) Зависимость среднего порядка одной псевдокомпании в списке второй от AUC.

Рис. 4.4. Исследование корреляции устойчивости и AUC.

Для оценки корреляции использовался коэффициен корреляции Кендалла. В таблице 4.5 отображены коэффициенты корреляции между метрикой AUC на разметке похожих и n - средним местом одной половины компании в списке другой, или d - средним расстоянием между двумя половинами компаний. Для вычисления коэффициента Кендалла точки сортируются по одной координате, затем

оценивается упорядоченность точек по второй координате. Подсчитывается сумма совпадений P и сумма инверсий Q.

$$t = \frac{2(P-Q)}{n(n-1)} = 1 - \frac{4Q}{n(n-1)} = \frac{4P}{n(n-1)} - 1$$

Нулевая гипотеза отвергается при $n \geq 10$, если

$$|\tilde{\tau}| = \left|\frac{t}{\sqrt{D_\tau}}\right| = \left|\frac{t\sqrt{9n(n-1)}}{\sqrt{2(2n+5)}}\right| \geq \Phi_{1-\alpha/2}.$$

Модель	Кендалл (t)	
	n	d
TM	-0.689	-0.286
WNTM	-0.330	-0.578

Таблица 4.5

В таблице 4.6 отображены расчёты соответствующих величин. Учитывая, что $\Phi_{99,99\%} = 3,715$, получается, что для $\tilde{\tau}_n^{TM}$ и $\tilde{\tau}_d^{WNTM}$ выполняется, что $|\tilde{\tau}| \geq \Phi_{99,99\%} = \Phi_{1-0,02\%/2}$. Следовательно, при уровне значимости $\alpha = 0,02\%$ нулевая гипотеза отвергается, то есть в данном случае можно с большой уверенностью говорить о корреляции. Аналогично $\tilde{\tau}_n^{WNTM} = 2.414 \geq 2,326 = \Phi_{99,00\%} = \Phi_{1-2\%/2}$ и $\tilde{\tau}_d^{TM} = 2.050 \geq 2,000 = \Phi_{97,50\%} = \Phi_{1-5\%/2}$. В данном случае получаем $\alpha = 2\%$ и $\alpha = 5\%$ соответственно.

Наличие корреляции метрики AUC с метриками устойчивости модели позволяет при переходе к новым данным (при изменении региона, или промежутка времени) не проводить разметку компаний по похожим для подсчета метрики AUC, а использовать в качестве максимизируемой величины устойчивость модели. Причем для WNTM моделей лучше обращать внимание на расстояние между псевдокомпаниями, а для тематической мультимодальной смотреть на порядок одной псевдокомпании в списке другой.

Модель	$\tilde{\tau}$	
	n	d
TM	4.938	2.050
WNTM	2.414	4.232

Таблица 4.6

5 Результаты

В данной главе представлены результаты работы моделей - примеры списков похожих компаний и темы, которые выделили модели. В этих моделях использовались подборы регуляризаторов и весов модальностей, строилось 400 тем (см.4.2.1), так что эти модели можно назвать оптимальными. В качестве алгоритма выделения товарных слов использовался контекстный алгоритм (см.4.1.2).

5.1 Тематическая модель

Модель выполнила 10 итераций EM алгоритма, использовала модальности, отобранные с помощью алгоритма подбора весов модальностей. На первых 4 итерациях использовалось разреживание p_{tdw}, на оставшихся декорреляция ϕ и разреживание θ.

Рис. 5.1. Список компаний, построенный тематичесокй моделью.

	inn	distance	okved	full name	short name	okved description
0	5260078907	-1.000000	18.12	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "РЕКО"	ООО "РЕКО"	Прочие виды полиграфической деятельности
1	5262019248	0.082538	18.13	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "ПОЛИПРИНТ"	ООО "ПОЛИПРИНТ"	Изготовление печатных форм и подготовительная деятельность
2	5263092402	0.216615	58	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "САРДАНА"	ООО "САРДАНА"	Деятельность издательская
3	5258049779	0.232764	18.12	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "АЛЬФА-ГРАФИКА"	ООО "АЛЬФА-ГРАФИКА"	Прочие виды полиграфической деятельности
4	5262151694	0.255244	58	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "РЕДАКЦИЯ "БИЗНЕС-КОНВЕЙЕР"	ООО "РЕДАКЦИЯ "БИЗНЕС-КОНВЕЙЕР"	Деятельность издательская
5	5257169047	0.276772	58.19	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "ПРИНТ-ЦЕНТР ВОСТОК"	ООО "ПРИНТ-ЦЕНТР ВОСТОК"	Виды издательской деятельности прочие
6	5261090826	0.277089	73.11	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "КОМПАНИЯ "АРТПРОМ-НН"	ООО "КОМПАНИЯ "АРТПРОМ-НН"	Деятельность рекламных агентств
7	5262324185	0.280034	73.11	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ РЕКЛАМНО-ПРОИЗВОДСТВЕННАЯ КОМПАНИЯ "РТНН"	ООО РПК "РТНН"	Деятельность рекламных агентств
8	5258125282	0.280365	46.49.33	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "ПРЕЗЕНТАЙМ"	ООО "ПРЕЗЕНТАЙМ"	Торговля оптовая писчебумажными и канцелярскими товарами
9	5260273785	0.281553	73.11	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "ПРОМ АРТ"	ООО "ПРОМ АРТ"	Деятельность рекламных агентств
10	5257084266	0.289283	46.6	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "ТОРГОВЫЙ ДОМ"НЕФТЕХИММАШ" КРАСНЫЙ ОКТЯБРЬ"	ООО "ТД "НЕФТЕХИММАШ" КО"	Торговля оптовая прочими машинами, оборудованием и принадлежностями
11	5260432876	0.295866	32.12.5	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "ЮВЕЛИРПРОМСЕРВИС"	ООО "ЮВЕЛИРПРОМСЕРВИС"	Производство ювелирных изделий, медалей из драгоценных металлов и драгоценных камней
12	5260416063	0.295896	73.11	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "РЕКЛАМНО-ПРОИЗВОДСТВЕННАЯ КОМПАНИЯ "АЛЬКОР"	ООО РПК "АЛЬКОР "	Деятельность рекламных агентств

Рис. 5.2. Тема, построенная TM моделью.

	@buywords	@sellwords	@buyokv_2	description
0	('купля', 0.057)	('строить', 0.038)	('G.45.32', 0.171)	Торговля розничная автомобильными деталями, узлами и принадлежностями
1	('имущество', 0.034)	('право', 0.032)	('H.49.4', 0.064)	Деятельность автомобильного грузового транспорта и услуги по перевозкам
2	('недвижимый', 0.034)	('система', 0.023)	('F.41.20', 0.062)	Строительство жилых и нежилых зданий
3	('строить', 0.026)	('страхование', 0.016)	('G.45.20', 0.052)	Техническое обслуживание и ремонт автотранспортных средств
4	('дать', 0.019)	('код', 0.015)	('F.43.11', 0.047)	Разборка и снос зданий
5	('право', 0.019)	('вывоз', 0.015)	('N.78.10', 0.046)	Деятельность агентств по подбору персонала
6	('вино', 0.018)	('монтаж', 0.015)		
7	('кровля', 0.018)	('текстильный', 0.014)		
8	('стройматериал', 0.017)	('комплекс', 0.012)		
9	('мещера', 0.017)	('транспортный', 0.011)		
10	('доплата', 0.016)	('штраф', 0.011)		
11	('санитарный', 0.016)	('земельный', 0.011)		
12	('гос', 0.015)	('юридический', 0.011)		
13	('страница', 0.015)	('подготовка', 0.011)		
14	('чистка', 0.014)	('обяз', 0.01)		
15	('гсм', 0.013)	('стр', 0.01)		

5.2 WNTM модель

Модель выполнила 10 итераций EM алгоритма, использовала модальности, отобранные с помощью алгоритма подбора весов модальностей. Без регуляризаторов.

5.3 Сравнение моделей

Здесь в таблице 5.1 приведены результаты расчета метрик для двух основных моделей TM и WNTM, и для бейзлайновых моделей. Бейзлайны в основном представляют собой one-hot энкодинг компаний по оквэдами. Contragent (buyers) - модель суммы контрагентов, Word2Vec - модель на контрагентах (корпус для обу-

Рис. 5.3. Список компаний, построенный WNTM моделью.

	inn	distance	okved	full name	short name	okved description
0	5260078907	-1.000000	18.12	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "РЕКО"	ООО "РЕКО"	Прочие виды полиграфической деятельности
1	5243033423	0.038590	18.12	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "СМАРТ ПРИНТ"	ООО "СМАРТ ПРИНТ"	Прочие виды полиграфической деятельности
2	5262019248	0.042066	18.13	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "ПОЛИПРИНТ"	ООО "ПОЛИПРИНТ"	Изготовление печатных форм и подготовительная деятельность
3	5261079325	0.063731	18.1	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "В ТОЧКУ"	ООО "В ТОЧКУ"	Деятельность полиграфическая и предоставление услуг в этой области
4	5248031162	0.084885	58.14	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "ЭТС"	ООО "ЭТС"	Издание журналов и периодических изданий
5	5262286660	0.101454	18.1	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "АРТ ЗАВОД"	ООО "АРТ ЗАВОД"	Деятельность полиграфическая и предоставление услуг в этой области
6	5260416063	0.106526	73.11	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "РЕКЛАМНО-ПРОИЗВОДСТВЕННАЯ КОМПАНИЯ "АЛЬКОР"	ООО РПК "АЛЬКОР "	Деятельность рекламных агентств
7	5260306504	0.111791	18.1	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ " РЕГИОН"	ООО "РЕГИОН"	Деятельность полиграфическая и предоставление услуг в этой области
8	5261069729	0.117150	18.1	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "ГЛАГОЛ"	ООО "ГЛАГОЛ"	Деятельность полиграфическая и предоставление услуг в этой области
9	5257161471	0.119153	18.12	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "КОННОВ"	ООО "КОННОВ"	Прочие виды полиграфической деятельности
10	5259116700	0.129604	18.12	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "БЕРДС МЕДИА"	ООО "БЕРДС МЕДИА"	Прочие виды полиграфической деятельности
11	5259101020	0.131756	18.1	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "БРАВО"	ООО "БРАВО"	Деятельность полиграфическая и предоставление услуг в этой области
12	5254026795	0.132532	18.12	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "ИНТЕРКОНТАКТ"	ООО "ИНТЕРКОНТАКТ"	Прочие виды полиграфической деятельности
13	5262151694	0.134319	58	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "РЕДАКЦИЯ "БИЗНЕС-КОНВЕЙЕР"	ООО "РЕДАКЦИЯ "БИЗНЕС-КОНВЕЙЕР"	Деятельность издательская
14	5262059868	0.136013	73.11	ОБЩЕСТВО С ОГРАНИЧЕННОЙ ОТВЕТСТВЕННОСТЬЮ "АВТОГРАФ"	ООО "АВТОГРАФ"	Деятельность рекламных агентств

Рис. 5.4. Тема, построенная WNTM моделью.

	wntm_text_seller	text_seller	wntm_text	wntm_text_buyer	wntm_okved1_seller	description
0	('аренда', 0.382)	('аренда', 0.634)	('аренда', 0.156)	('энергия', 0.074)	('L.68', 0.436)	Операции с недвижимым имуществом
1	('помещение', 0.186)		('помещение', 0.076)	('эл', 0.06)		
2			('связь', 0.029)	('водоснабжение', 0.041)		
3			('нежилой', 0.028)	('водоотведение', 0.039)		
4			('водоотведение', 0.017)	('вывоз', 0.033)		
5			('газопровод', 0.016)	('связь', 0.032)		
6			('эл', 0.016)	('газопровод', 0.03)		
7			('то', 0.015)	('газа', 0.026)		
8			('газ', 0.014)	('то', 0.026)		
9			('поставка', 0.013)	('технический', 0.019)		
10			('газа', 0.013)	('газ', 0.016)		
11			('с-но', 0.011)	('поставка', 0.013)		

чения - упорядоченные во времени контрагенты компании для всех компаний в коллекции), TF-IDF (buywords) - каждой компании сопоставлен вектор из tf-idf счетчиков слов из платежных поручений продаж компании. Последние три модели - те же самые модели, которые использовались при разметке похожих пар компаний в разделе 3.3.1.

Модель	micro AUC	macro AUC	MAP
TM	0,856	0,811	0,678
WNTM	0,893	0,876	0,699
Word2Vec	0,733	0,714	0,551
Contragent (buyers)	0,585	0,645	0,536
ContragentOKVED, level=0	0,699	0,653	0,517
ContragentOKVED, level=1	0,757	0,668	0,534
ContragentOKVED, level=2	0,796	0,702	0,572
OKVED, level=0	0,812	0,757	0,592
OKVED, level=1	0,838	0,774	0,645
OKVED, level=2	0,735	0,700	0,659
TF-IDF, (buywords)	0,857	0,786	0,647

Таблица 5.1. Сравнение моделей.

6 Заключение

В данной работе показана возможность использования мультимодальных тематических моделей для анализа транзакционных данных, а именно выявления видов деятельности компаний на основе их транзакций. Полученные результаты доказывают корректность использования таких данных, как транзакции, для тематического моделирования компаний, а предложенные алгоритмы подбора весов модальностей и регуляризаторов облегчают работу с большим количеством гиперпараметров модели, свойственным многомодальным моделям. Также построена модель WNTM, которая использует не только факт наличия транзакций у компаний, как в TM модели, но и упорядоченность транзакций во времени.

И по метрикам, описанным в пункте 3.3.2, то есть по качеству списков похожих, и по интерпретируемости тем, лучше всего себя показала смешанная модель с WNTM и TM модальностями. Немного хуже проявила себя чистая WNTM модель (результаты не указаны в таблице), и значимо хуже классическая мультимодальная TM. Однако надо учитывать, что для WNTM модели используется большее число модальностей, их формирование занимает много времени, и среднее количество токенов одной модальности в документе возрастает пропорционально ширине окна. Все это заметно замедляет построение WNTM модели в сравнении с TM моделью. Поэтому выбор той или иной модели зависит от приоритета - скорость или качество.

Также для обеих моделей была показана их устойчивость при наличии достаточно большого числа транзакций у компаний и рассчитана степень корреляции устойчивости с качеством моделей. Полученные результаты показывают возможность использования в качестве максимизируемого функционала в алгоритмах для подбора модальностей и регуляризаторов метрики устойчивости вместо метрик ранжированности списков похожих. Это, как минимум, означает, что при пе-

реходе к новым данным можно значительно уменьшить количество размечаемых данных и рассчитывать вместо этого дополнительно метрику устойчивости, а как максимум избавиться от разметки вообще.

Литература

Daniel Ramage, Susan Dumais, and Dan Liebling. Characterizing microblogs with topic models. In Fourth International AAAI Conference on Weblogs and Social Media, 2010.

Chong Wang and David M Blei. Collaborative topic modeling for recommending scientific articles. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 448–456. ACM, 2011.

Pengtao Xie and Eric P Xing. Integrating document clustering and topic modeling. arXiv preprint arXiv:1309.6874, 2013.

Duo Zhang, Qiaozhu Mei, and ChengXiang Zhai. Cross-lingual latent topic extraction. In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 1128–1137. Association for Computational Linguistics, 2010.

Yuan Zuo, Jichang Zhao, and Ke Xu. Word network topic model: a simple but general solution for short and imbalanced texts. Knowledge and Information Systems, 48 (2):379–398, 2016.

А.О. Янина and К.В. Воронцов. Мультимодальные тематические модели для разведочного поиска в коллективном блоге. Машинное обучение и анализ данных, 2(2):173–186, 2016.