Тематическое моделирование финансовых потоков корпоративных клиентов банка по транзакционным данным

Шишкина Вера Сергеевна

Научный руководитель: Воронцов К.В.

16 апреля 2019 г.

Overview

Постановка задачи

Данные

Транзакция:

- b - фирма покупатель
- s - фирма продавец
- sum - сумма денег
- nazn - назначение платежа
- data - время проведения платежа

Фирма (ммб - до 100 работников, 400 млн рублей годовой выручки):

- name - название компании
- okveds - одновной и дополнительный оквэды
- adresses - юридический и фактический адреса компании
- anket - Заполненая работниками компании анкета(вид деятельности, оборот и др.)

Постановка задачи

Найти

Задача:

- Определить настоящий основной вид деятельности фирмы. (Может не соответствовать оквэду)
- Определить схожие компании.
- Определить компании-конкуренты.

Постановка задачи

Обработка данных

Предварительная обработка транзакций:

- Выделение компаний из одного региона (Нижний Новгород)
- Выделение временного промежутка (6 месяцев)
- Выделение компаний, которые за данный временной промежуток выступали в качестве продавца не менее заданного количества раз
- Удаление транзакций с крупнейшими банками

Итого 16 385 фирм.

Предварительная обработка текстов:

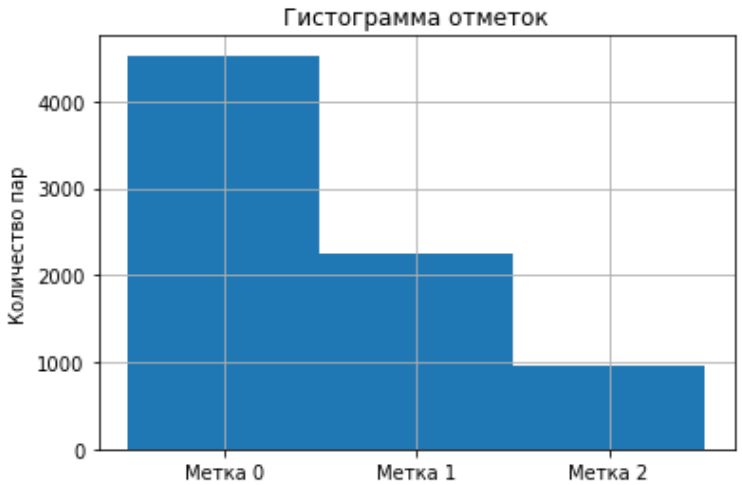
- Выделение товарных слов (регулярные выражения)
- Лематизация

Длина словаря товарных слов: 20,130
Количество транзакций с непустыми текстами платежей - 66%:

Разметка

Методика сбора разметки:

- Выбираются несколько бейзлайнов.
- В каждом бейзлайне случайным образом выбирается фирма.
- Для этой фирмы строится список на нее похожих.
- Каждой фирме в этом списке ставится оценка: 1 - похожа на фирму-запрос, 0 - не похожа, 2 - разметчик не уверен.



Метрики
AUC

$$MicroAUC = AUC(\cup_f list(f))$$

$$MacroAUC = \frac{\sum_{f \in F} AUC(list(f))}{|F|}$$

Метрики
MAP

Precision - доля схожих в ранжированном списке длины k.

$$P(f, k) = \frac{[mark = 1]}{k}$$

Average Precision - доля схожих в ранжированном списке длины k.

$$AP(f) = \sum_k P(f, k)$$

Mean Average Precision - доля схожих в ранжированном списке длины k.

$$MAP = \frac{\sum_{f \in F} AP(f)}{|F|}$$

Метрики

Иерархия окведов

$$H = \sum_{o_d} \frac{V(o_d)}{V(o_r)} \frac{KL(o_d|real(o_r))}{KL(o_d|nearest(o_r)}$$

где V - мощность токена в коллекции (или сумма мощностей нескольких токенов). Если метрика приближается к 0, это хороший случай, если близко к 1 - плохой.

Решение задачи

Мультимодальная модель

Тематическая модель(BigArtm): документ - фирма.
Модальности:

- @sellers - контрагенты, у которых фирма покупала.
- @buyers - контрагенты, которым фирма продавала.
- @sellwords - слова из текстов транзакций, где фирма выступала как покупатель.
- @buywords - слова из текстов транзакций, где фирма выступала как продавец.
- @okv0, @okv1, @okv2 - основные оквэды фирмы разных уровней.
- @buyokv_0, @buyokv_1, @buyokv_2 - основные оквэды контрагентов-покупателей разных уровней.
- @sellokv_0, @sellokv_1, @sellokv_2 - основные оквэды контрагентов-продавцов разных уровней.
- @all_okv_0, @all_okv_1, @all_okv_2 - дополнительные оквэды фирмы различных уровней.

Решение задачи

Подбор модальносей

Жадная пошаговая аддитивная стратегия для подбора весов модальностей тематической модели, максимизирующих заданный функционал качества.

- Выбираются две модальности, и подбирается выпуклая комбинация весов этих модальностей, дающая максимальное значение функционала. $(1 - \lambda_1), \lambda_1$ - значения весов.
- Соотношение между этими двумя модальностями фиксируется, и подбирается λ_2. $(1 - \lambda_1) * (1 - \lambda_2), \lambda_1 * (1 - \lambda_2)$ и λ_2 - новые веса.
- Зафиксировано соотношение между первыми k модальностями. Подбирается λ_k такое, что выпуклая комбинация первых k модальностей с новой модальностью дает наилучшее значение функционала.

Используется метод золотого сечения. Для уравновешивания весов сжимается каждый следующий отрезок, на котором идет поиск λ, на некоторую постоянную величину.

Решение задачи

Подбор регуляризаторов

- На вход алгоритму подается список регуляризаторов. Для каждого регуляризатора указано, сколько нужно сделать итераций, и в каких пределах искать параметр τ.
- Для каждого регуляризатора в списке (список может быть случайно перемешан по желанию пользователя) методом золотого сечения ведется поиск τ этого регуляризатора такого, чтобы заданная метрика была максимальной.
- Если с помощью этого регуляризатора при некотором τ удалось получить качество лучше, чем в модели без него (модели, которая дообучалась с данным регуляризатором), то модель с наилучшим качеством запоминается, и алгоритм переходит к следующему регуляризатору.

Результаты

Метрики

OLD-VALID		DIFF-VALID		ALL-VALID	
AUC	MAP	AUC	MAP	AUC	MAP
0.815(0.747)	0.566	0.883(0.885)	0.797	0.838(0.781)	0.651

Таблица: TM

OLD-VALID		DIFF-VALID		ALL-VALID	
AUC	MAP	AUC	MAP	AUC	MAP
0.826(0.751)	0.535	0.748(0.711)	0.556	0.733(0.714)	0.551

Таблица: W2V

Результаты

Списки

	@buywords
0	хлеб:8 хлебобулочный:32
2	хлеб:246 хлебобулочий:1 хлебобуть:67 разовый:4 социальный:1 изд:34 питание:98 хлебобулочный:33 питаня:2 питать:1 конт:6 мол:1 приказ:1 доплата:1 булочный:57 помощь:1 столовый:6 издеть:8 хлебо:2 правовой:6
3	хлеб:184 питать:8 лысковский:244 база:113 хлебобуть:252 булочный:22 булочн:23 изд:257 спорт:2 хлебоб:1 кредиторский:11 письмо:2 физ:2 фев:1 подарок:1 код:367 тмц:113 энергия:1 конди:22 металл:1 ребёнок:8 янв:1 доплата:2 нопо:8 авг:1 ком:1 хлебобулочный:407 возмездный:2 питание:141 переменный:3 хлебобулочий:2 чёрный:1 филиал:99 агент:1 сост:2 отдых:113 кондитерский:48 поставщик:244 питаня:1 торт:2 кафе:18 апр:1 хлебо:45
4	лаваш:10
5	хлеб:58 хлебобулочный:139 булочный:5 поставщик:57 питание:3 ассортимент:39 розница:23 хлебо:5 упд:54
6	хлеб:71 гос:66 основной:1 хлебобулочный:144 изд:42 гараж:6 хлебоб:42 доплата:6 склад:14 оказать:2 рамка:66 нежилой:16 закрытие:1 обора:66 просрочить:16 питание:85
7	питание:13 хлебобулочный:1
8	атс:12 доплата:2 новгород:6 автомобиль:19 оказать:1 нижегородский:12 филиал:12 электроэнергия:16 энергия:2 предост:6 стоянка:31 предоставть:1
9	хлеб:1 питание:2 кредиторский:1 хлебобулочный:4

Результаты

Списки

	okved_description
0	Производство хлеба и мучных кондитерских изделий, тортов и пирожных недлительного хранения
1	Аренда и управление собственным или арендованным недвижимым имуществом
2	Производство хлеба и мучных кондитерских изделий, тортов и пирожных недлительного хранения
3	Производство хлеба и мучных кондитерских изделий, тортов и пирожных недлительного хранения
4	Производство хлеба и мучных кондитерских изделий, тортов и пирожных недлительного хранения
5	Производство хлеба и мучных кондитерских изделий, тортов и пирожных недлительного хранения
6	Производство хлеба и мучных кондитерских изделий, тортов и пирожных недлительного хранения
7	Производство хлеба и хлебобулочных изделий недлительного хранения
8	Аренда и управление собственным или арендованным недвижимым имуществом
9	Производство хлеба и мучных кондитерских изделий, тортов и пирожных недлительного хранения
10	Производство хлеба и мучных кондитерских изделий, тортов и пирожных недлительного хранения
11	Производство хлеба и мучных кондитерских изделий, тортов и пирожных недлительного хранения

	@full_name_firm	topic_id	distance	topic_prob	jaccard
0	ООО "ПЕКАРЬ"	topic_266	-1	0,714003265	1
1	ООО "ОПТИМУМ"	topic_266	0,014630854	0,999999881	0
2	ПО "ШАТКОВСКИЙ ХЛЕБ"	topic_266	0,016693532	0,731024623	0
3	АО "ЛЫСКОВСКИЙ ХЛЕБОЗАВОД"	topic_266	0,017666042	0,817571759	0,01
4		topic_266	0,020416677	0,784514904	0,1
5	ОАО "Колос-3 "	topic_266	0,024374962	0,875129759	0,02
6	ООО "БОРСКИЙ ХЛЕБ"	topic_266	0,025421917	0,732770264	0,03
7		topic_266	0,028946996	0,688109398	0
8	ООО "АЛЬТЕРНАТИВА НН"	topic_266	0,030113816	0,767230868	0
9	ООО "НОВОСЕЛИЦКИЙ ХЛЕБ"	topic_266	0,038313448	0,809339881	0
10	ООО "НАШ ХЛЕБ"	topic_266	0,042741179	0,704532146	0
11	ООО "БОЛДИНСКИЙ ХЛЕБ"	topic_266	0,044238746	0,640407324	0

Результаты

Темы

	@buywords	@okv_2	okv_description
0	[0.112, 'код']	[0.856, 'C.10.71']	Производство хлеба и мучных кондитерских изделий, тортов и пирожных недлительного хранения
1	[0.09, 'хлеб']	[0.048, 'L.68.20']	Аренда и управление собственным или арендованным недвижимым имуществом
2	[0.04, 'хлебобулочный']	[0.034, 'L.68.2']	Аренда и управление собственным или арендованным недвижимым имуществом
3	[0.039, 'питание']	[0.019, 'C.20.59']	Производство прочих химических продуктов, не включенных в другие группировки
4	[0.032, 'хлебо']	[0.008, 'C.10.1']	Переработка и консервирование мяса и мясной пищевой продукции
5	[0.03, 'хлебобуть']	[0.006, 'C.25.62']	Обработка металлических изделий механическая
6	[0.027, 'кредиторский']	[0.005, 'C.10.72']	Производство сухарей, печенья и прочих сухарных хлебобулочных изделий, производство мучных кондитерских изделий, тортов, пирожных, пирогов и бисквитов, предназначенных для длительного хранения
7	[0.025, 'доплата']	[0.003, 'C.14.19']	Производство прочей одежды и аксессуаров одежды
8	[0.024, 'стоянка']	[0.002, 'N.77.29']	Прокат и аренда прочих предметов личного пользования и хозяйственно-бытового назначения
9	[0.023, 'филиал']	[0.002, 'C.29.31']	Производство электрического и электронного оборудования для автотранспортных средств
10	[0.022, 'нежилой']	[0.001, 'C.22.23']	Производство пластмассовых изделий, используемых в строительстве

Заключение

- Для анализа транзакционных данных применима мультимодальная тематическая модель.
- Полученнай модель устойчива: при отбрасывании некоторого количества транзакций у некоторой фирмы, полученная псевдофирма похожа по косинусному расстоянию эмбеддинга на изначальную.
- Около трети тем привязались к одному оквэду второго уровня (вероятность больше 0.5). Еще около трети имеют два доминирующих оквэда, в основном схожих по смыслу.
- На данный момент проводится исследование WNTM модели, которая уже дала результат лучше, чем обычная тематическая модель.

План исследования

- Провести исследование WNTM модели. Проверить устойчивость.
- Использовать улучшенные способы выделения товарных слов. Добавить исправление опечаток.
- Использовать эвристики для балансировки тем.