Аналитика товарных наименований из чеков

выполнил: Тимофей Смирнов

Научный руководитель: Зимин Степан Михайлович

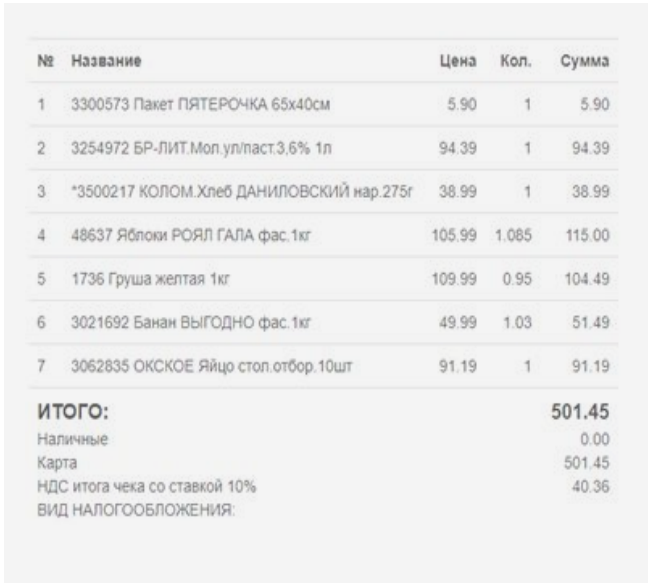
Откуда появляются чеки

Цель работы:

Получение продуктового каталога на основе наименований товаров из чеков

ИМЕЮЩИЕСЯ ДАННЫЕ

- 30 млрд чеков
- 1 млрд уникальных наименований



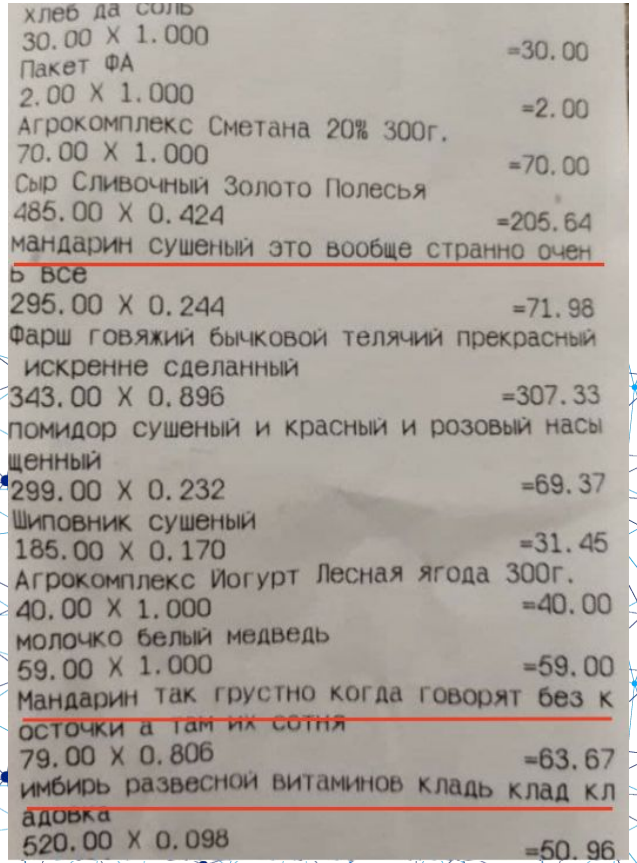
№	Название	Цена	Кол.	Сумма
1	3300573 Пакет ПЯТЕРОЧКА 65х40см	5.90	1	5.90
2	3254972 БР-ЛИТ.Мол.ул/паст.3,6% 1л	94.39	1	94.39
3	*3500217 КОЛОМ.Хлеб ДАНИЛОВСКИЙ нар.275г	38.99	1	38.99
4	48637 Яблоки РОЯЛ ГАЛА фас.1кг	105.99	1.085	115.00
5	1736 Груша желтая 1кг	109.99	0.95	104.49
6	3021692 Банан ВЫГОДНО фас.1кг	49.99	1.03	51.49
7	3062835 ОКСКОЕ Яйцо стол.отбор.10шт	91.19	1	91.19

ИТОГО: 501.45
Наличные 0.00
Карта 501.45
НДС итога чека со ставкой 10% 40.36
ВИД НАЛОГООБЛОЖЕНИЯ:

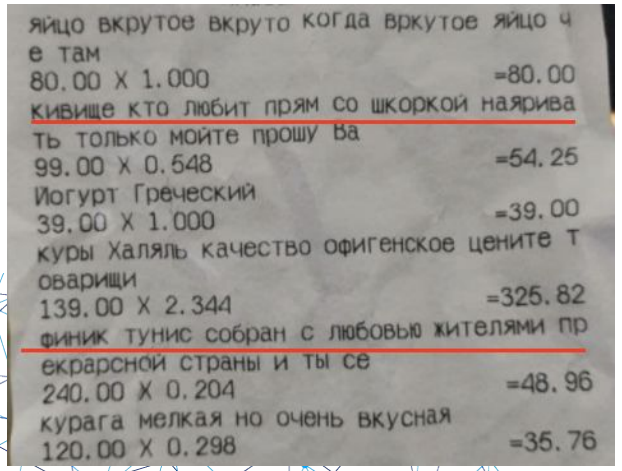
Категория 1-го уровням	Категория 2-го уровня	Бренд	Размеры
Упаковочные пакеты	Упаковочные пакеты	Пятерочка	----
Продукты питания	Молочные продукты и яйца	Брест-литовское	1 л
Продукты питания	Хлеб, выпечка, сладости	Коломенское	275 г
Продукты питания	Овощи, фрукты, ягоды, зелень, грибы	----	1кг
Продукты питания	Овощи, фрукты, ягоды, зелень, грибы	----	1кг
Продукты питания	Овощи, фрукты, ягоды, зелень, грибы	----	1кг
Продукты питания	Молочные продукты и яйца	Окское	10 шт

Почему анализ чеков — это сложно

яйцо вкрутое вкруто когда вркутое яйцо ч
е там
78.00 X 1.000 =78.00
Свекла или бурак почему бурак если свекл
а
26.90 X 0.714 =19.21
Сыр Пружаны Белорусы могут делать вкусно
474.00 X 0.344 =163.06
Фисташки турецкие вызывают зависимость у
никальный вкус
989.00 X 0.182 =180.00
хурма вяжет или не вяжет вот в чем вопро
с
95.00 X 0.856 =81.32
Апельсин кто почистил тот и сьел это спр
аведливо
99.00 X 1.426 =141.17
В любой сложной ситуации ешьте морковь и
днем и ночью
28.90 X 0.742 =21.44
йогурт клубника 300мл. агрокомпл.
40.00 X 1.000 =40.00



хлеб да соль
30.00 X 1.000 =30.00
Пакет ФА
2.00 X 1.000 =2.00
Агрокомплекс Сметана 20% 300г.
70.00 X 1.000 =70.00
Сыр Сливочный Золото Полесья
485.00 X 0.424 =205.64
мандарин сушеный это вообще странно очен
ь все
295.00 X 0.244 =71.98
Фарш говяжий бычковой телячий прекрасный
искренне сделанный
343.00 X 0.896 =307.33
помидор сушеный и красный и розовый насы
щенный
299.00 X 0.232 =69.37
Шиповник сушеный
185.00 X 0.170 =31.45
Агрокомплекс Йогурт Лесная ягода 300г.
40.00 X 1.000 =40.00
молочко белый медведь
59.00 X 1.000 =59.00
Мандарин так грустно когда говорят без к
осточки а там их сотня
79.00 X 0.806 =63.67
имбирь развесной витаминов кладь клад кл
адовка
520.00 X 0.098 =50.96



яйцо вкрутое вкруто когда вркутое яйцо ч
е там
80.00 X 1.000 =80.00
кивище кто любит прям со шкоркой наярива
ть только мойте прошу Ва
99.00 X 0.548 =54.25
Йогурт Греческий
39.00 X 1.000 =39.00
куры Халяль качество офигенское цените т
оварищи
139.00 X 2.344 =325.82
финик тунис собран с любовью жителями пр
екрасной страны и ты се
240.00 X 0.204 =48.96
курага мелкая но очень вкусная
120.00 X 0.298 =35.76

Почему анализ чеков — это сложно


КАССОВЫЙ ЧЕК

Цена	Скидка	Цена со скидкой	Кол-во	Итого	НДС
САРАФ.Мол.отб.паст.д3.4-4%930мл					
62.99		62.99 *	1.000	62.99	Б
ККК Хлеб ДОНСКОЙ форм.700г					
36.99		36.99 *	1.000	36.99	
LANDERS Сыр ЛЕГКИЙ 30% 230г					
159.99		159.99 *	1.000	159.99	
Томаты черри ПРЕМИУМ крас.250г					
139.99		139.99 *	1.000	139.99	
ИНДАНА Гвоздика целая 15г					
84.19		84.19 *	1.000	84.19	
*ALP.GOLD Шок.Д.Б.П.кл.н.б.150г					
99.99		99.99 *	1.000	99.99	
*MARVEL Патч НАЧИНКА 1шт					
55.00	55.00	0.00 *	1.000	0.00	

*ALP.GOLD Шок.Д.Б.П.кл.н.б.150г

ALP.GOLD Шок.Д.Б.П.кл.н.б.150г

Alpen Gold Шоколад Десерт «Безе Павлова» с клубничной начинкой безе, 150 гр

Бывают случаи гораздо хуже:

TEXTILES M* ECH.JHELAM DUO BL.MAR.BE
Магнит металлический стразы "Щит Оружие", шт.
ЛОМТ.КУР.КАВКАЗ.35Г
5 2456525618707 НБ д/ росписи/D_M/54193
ЦЕЛЬ Т 2, 2МЛ №5

Предобработка данных и токенизация

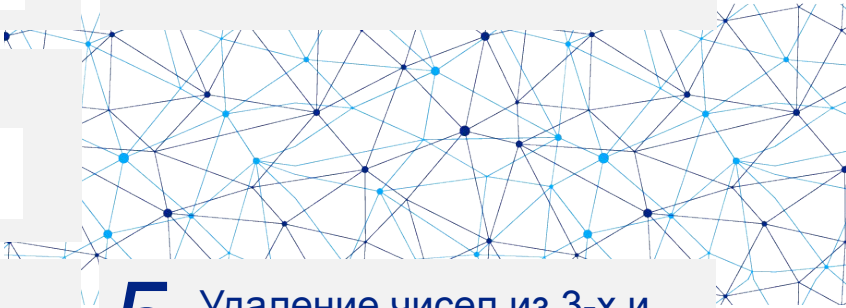
1 Разбиение CamelCase слов

```
"НапитокКока-КолаЧерри" -> "Напиток Кока-Кола Черри"
"КабернеФанКр150м" -> "Каберне Фан Кр150м"
```

2 Приведение к нижнему регистру

3 «Стандартные чистки»

```
"коптильня380х280х170 (сталь 0,8мм)" ->
"коптильня 380 280 170 (сталь 0.8мм)"
```



4 Выделение единиц объёма/кол-ва/массы

```
"нап вин тамянка белый п/сл 1.0 л 11%" → "1.0л"
"оджахури свин шеи 350 гр" → "350гр"
```

5 Удаление чисел из 3-х и более символов (номенклатура/штрихкоды)

6 Удаление пунктуации

7 Транслитерация

8 Лемматизация

Транслитерация

1. Английские слова с 1 кириллическим символом
2. Русские слова с 1-4 латинскими символами
3. Разделение слов состоящих как из кириллицы, так и из латиницы

```
"артикул" → "артикул"
"1 набор нав jardi нвмф" → "1 набор нав  jardi нвмф"
```

Сравнение с BPE токенайзером

	Исходное наименование	Наш токенайзер	BPE токенайзер
0	СидрЯблочныйИгрЖемч.п/сл.0.75л ск.10%	сидр яблочный игра жемча полусладкий ск 10	сидр яблочный игр жемч . п / сл . 0 . 75л ск ....
1	ПакетПодарБлест39,5х31,5х10,5смИМП	пакет подар блест 39 5 31 5 имп	пакет подар блест 39 , 5х 31 , 5 х10 , 5с мимп
2	Марм.УДАРНИЦАябл.325 101643	мармелад ударница аябл	марм . ударница ябл . 325 1016 43
3	Напиток энерг.Бёрн яблоко киви 0,33л	напиток энерг берн яблоко киви	напиток энерг . бёрн яблоко киви 0 , 33л
4	3,5х35 Саморез по ГК Част.шаг чёрн.фосф.	3 5 35 саморез гк частый шаг черн фосф	3 , 5х 35 само рез по гк част . ша г чёрн . фо...
5	мин. вода Зелёный городок н/г.1,5л	мина вода зеленый городок негазированный	мин . вода зелёный городок н / г . 1 , 5л
6	Пиво тёмное "Балтика Портер"№6 с/т. 0,47л.	пиво темный балтика портер 6 т	пиво тёмное " балтика порт ер " № 6 с / т . 0 ...
7	Пиво светлое Гёссер алк.4,7% 0,45л ж/б	пиво светлый госсер алк 4 7 б	пиво светлое г ё ссер алк . 4 , 7 % 0 , 45л ж / б
8	Добрая Бурёнка молоко ультрапаст.2,5% 0,950г.	добрый буренка молоко ультрапаст 2 5	добрая бурёнка молоко ультрапаст . 2 , 5 % 0 ,...
9	4,2х70 Саморез по ГК Част.шаг чёрн.фосф.	4 2 70 саморез гк частый шаг черн фосф	4 , 2х 70 само рез по гк част . ша г чёрн . фо...
10	Ликёр ИСТОРИЯ ЛЮБВИ СЛИВОЧНЫЙ ВКУС 0,5л	ликер история любовь сливочный вкус	ликёр исто рия люб ви сливочный вкус 0 , 5л
11	4381(А)Био-йогурт пит. "Активиа" 2,1% Печёная ...	биойогурт пит активиа 2 1 печеный груша 5 злак...	4 381 (а) био - йогурт пит . " активиа " 2 ,...
12	Молоко Кубанская бурёнка пастериз. 2,5% 800мл...	молоко кубанский буренка пастериз 2 5 м	молоко кубанская бурёнка пастериз . 2 , 5 % 80...

Восстановление сокращений

Первая идея - проверка правописания

DeepPavlov Spell Checker/ PyEnchant: “шамп для волос” - “штамп для волос”

Дополнение слов Autocomplete:

```
autocomplete.predict("волос", "шамп") → [('шампунь', 253)]
autocomplete.predict("мангал", "шамп") → [('шампур', 8)]
autocomplete.predict("гриб", "шамп") → [('шампиньон', 13)]
autocomplete.predict("вино", "шамп") → [('шампанское', 3), (шампань, 1)]
```

```
autocomplete.predict("шамп") → [('шампунь', 15374), ('шампанское', 680),
('шампур', 245), ('шампань', 174), ('шампиньон', 119), ('шампанский', 32),
('шампуневый', 12)]
```

Составление продуктового каталога

Было спаршено два крупных онлайн-магазина

- >2 млн товаров с разметкой этих магазинов
- 200 категорий в составленном каталоге
- 100к наименований на разметку
- 100к сложных наименований на разметку

Целевые метрики:

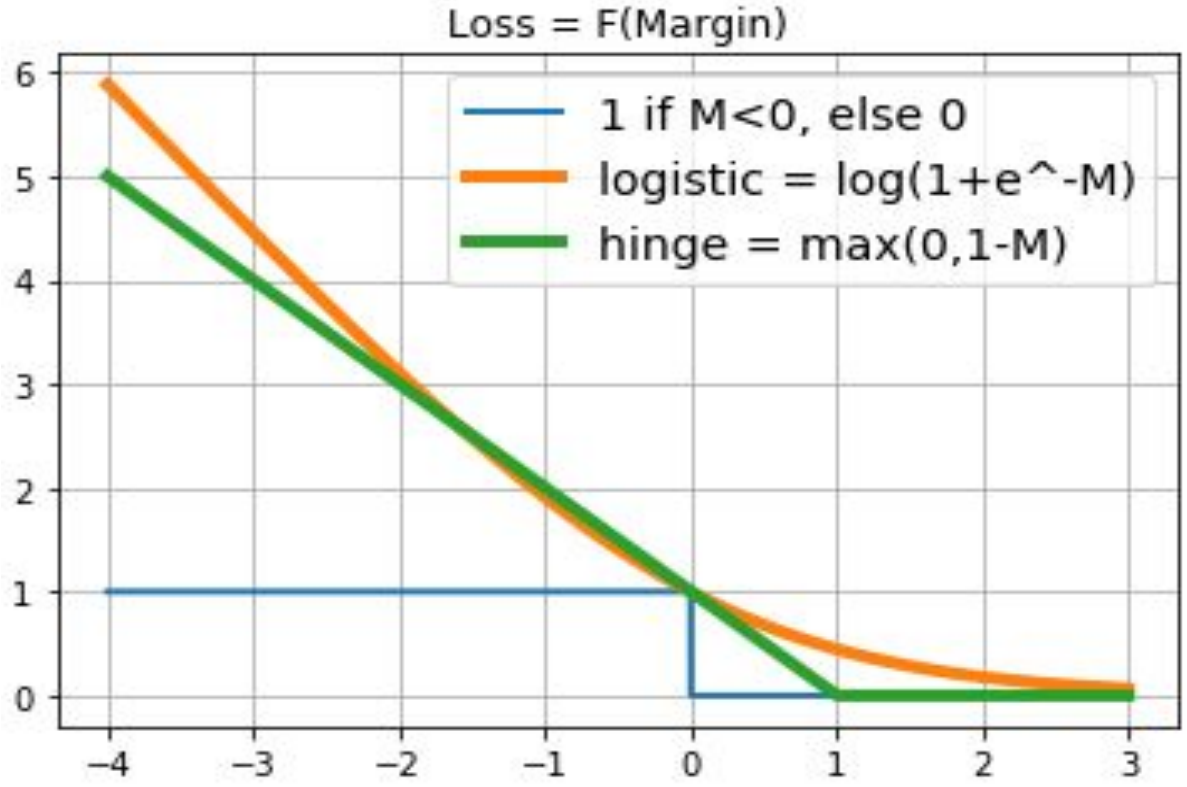
- Accuracy

$$Weighted\ F1\ Score = \sum_{i}^{N} W_i * F_i score$$

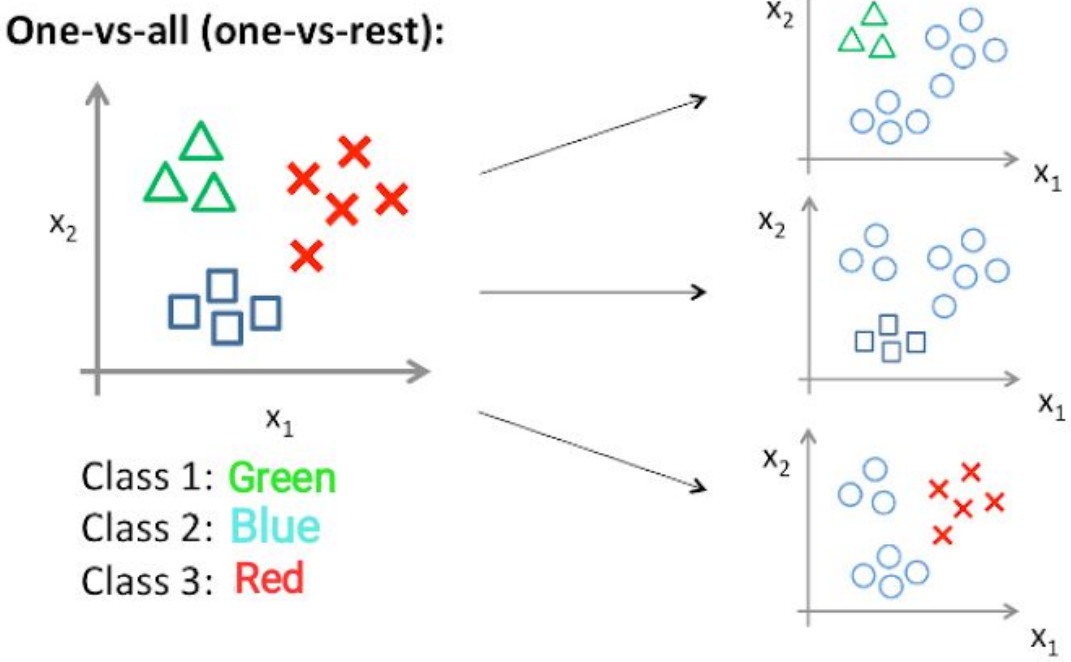
Простые модели

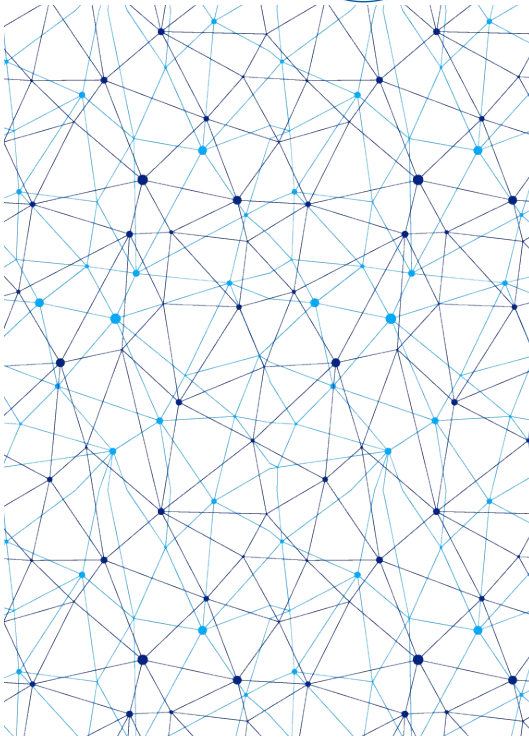
Метрика	Данные	SVM + наш токенайзер	SVM + BPE	Logistic Regression
Accuracy на категориях 1-го уровня	easy	95.4	95.5	92.7
	hard	85.4	86	76.2
Accuracy на категориях 2-го уровня	easy	88.1	88.5	83.7
	hard	71.4	72.8	61.4
Weighted F1	easy	87.9	88.3	83
	hard	71.2	72.5	60.7

Разница SVM и Logistic Regression



Подходы к многоклассовой классификации

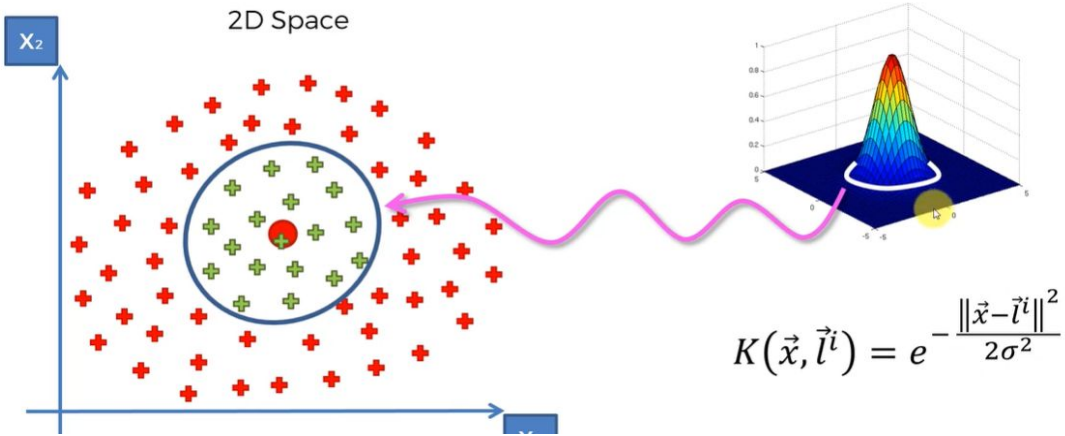




Еще один возможный фактор

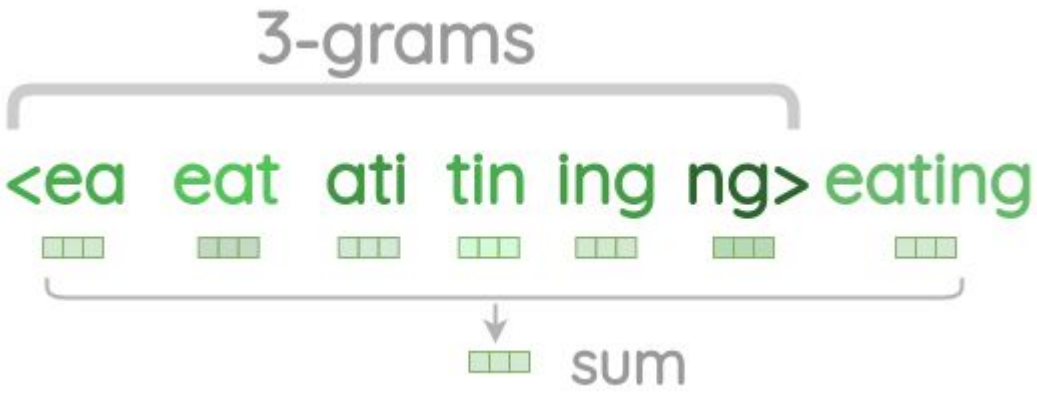
Последующие улучшения

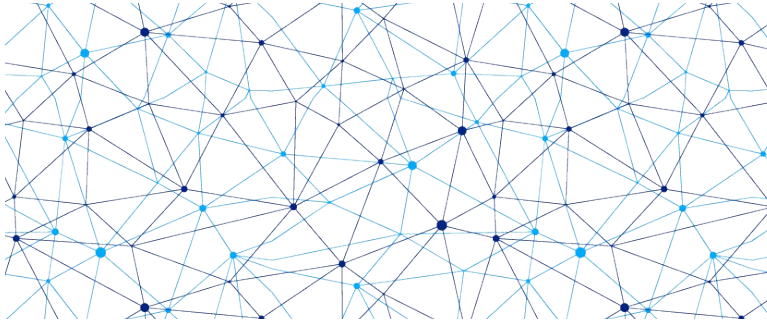
- Обучение нескольких классификаторов
  - снижение variance
- Генерация RFF признаков
  - апроксимация ядра
- Эксперименты с векторизатором
  - Добавление посимвольных tf-idf векторов
- Подбор гиперпараметров
- Использование других признаков, генерация фичей

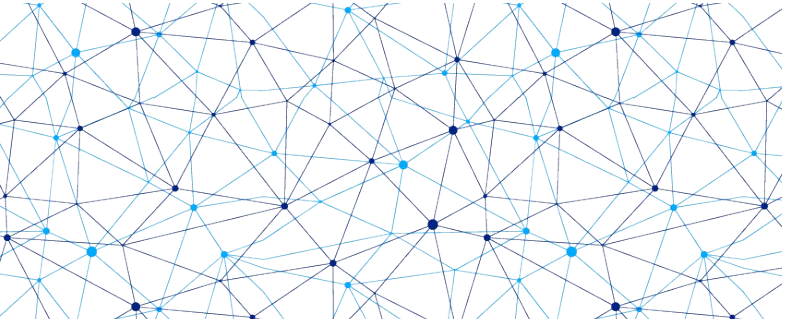


$$K(\vec{x}, \vec{l}^i) = e^{-\frac{\|\vec{x}-\vec{l}^i\|^2}{2\sigma^2}}$$

Fasttext

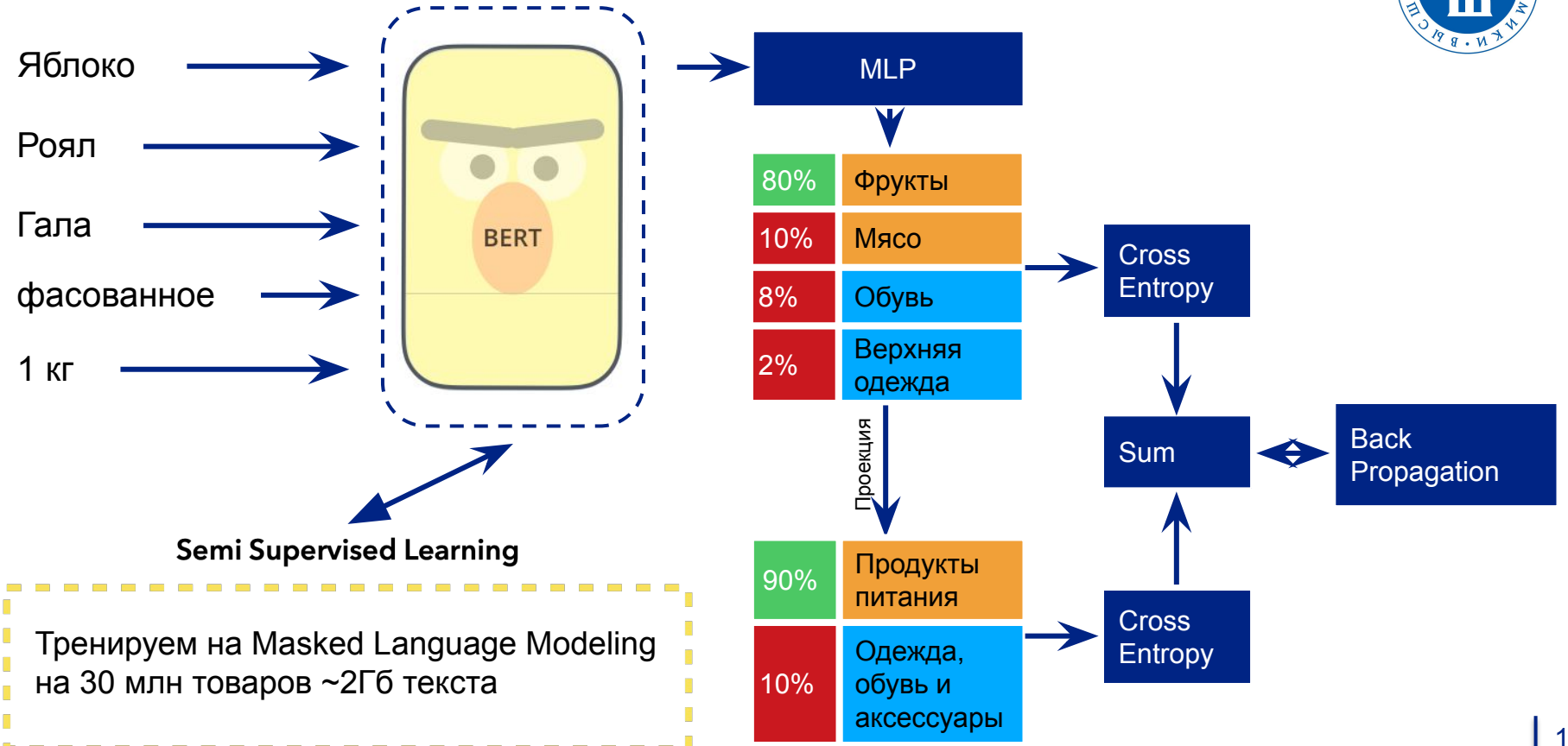

3-grams
<ea eat ati tin ing ng> eating
sum





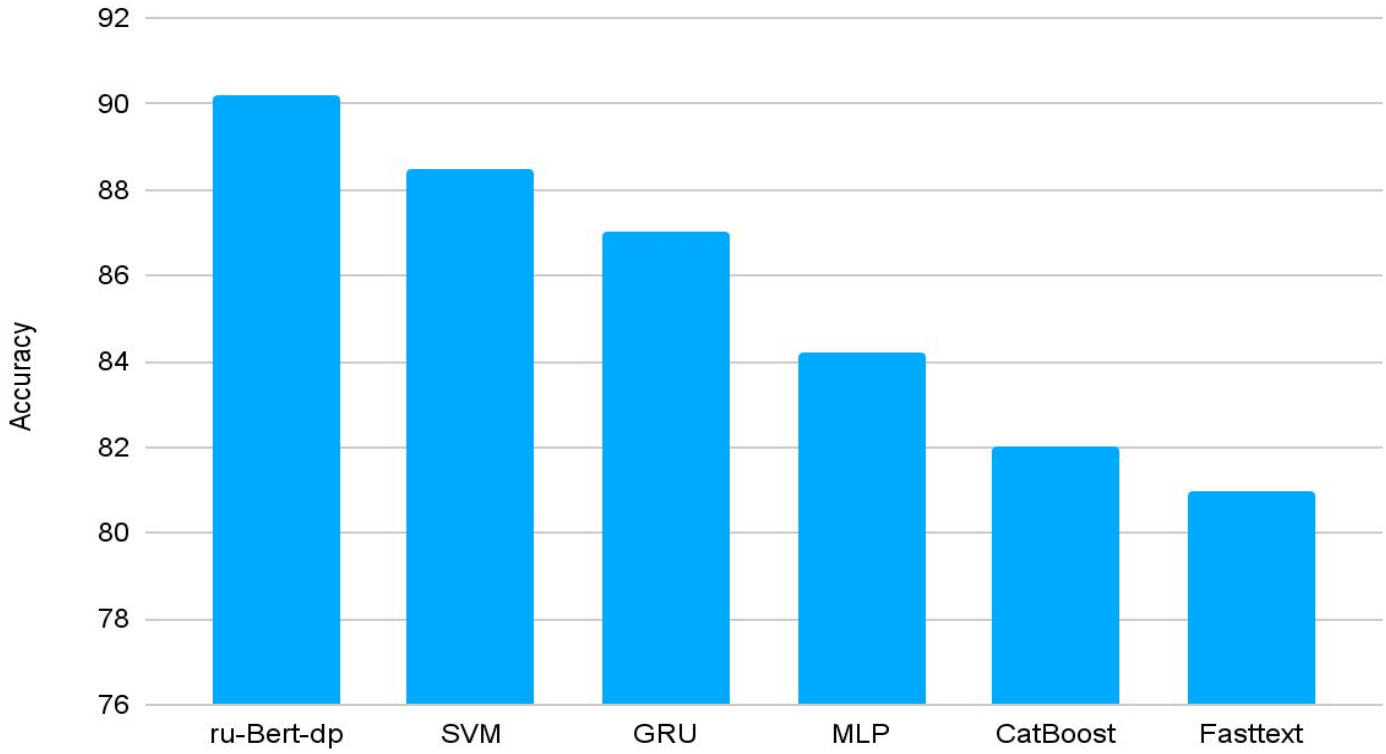

actual context
negative context
am
food
paris
earth
eating
dot product
0.7
0.4
0.5
0.3
sigmoid
bring to 1
bring to 0
goal

Нейронные сети для классификации

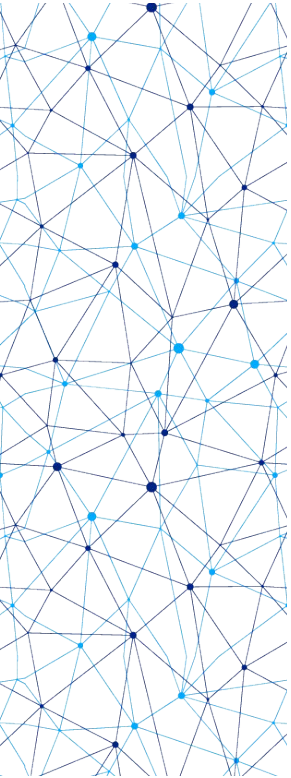


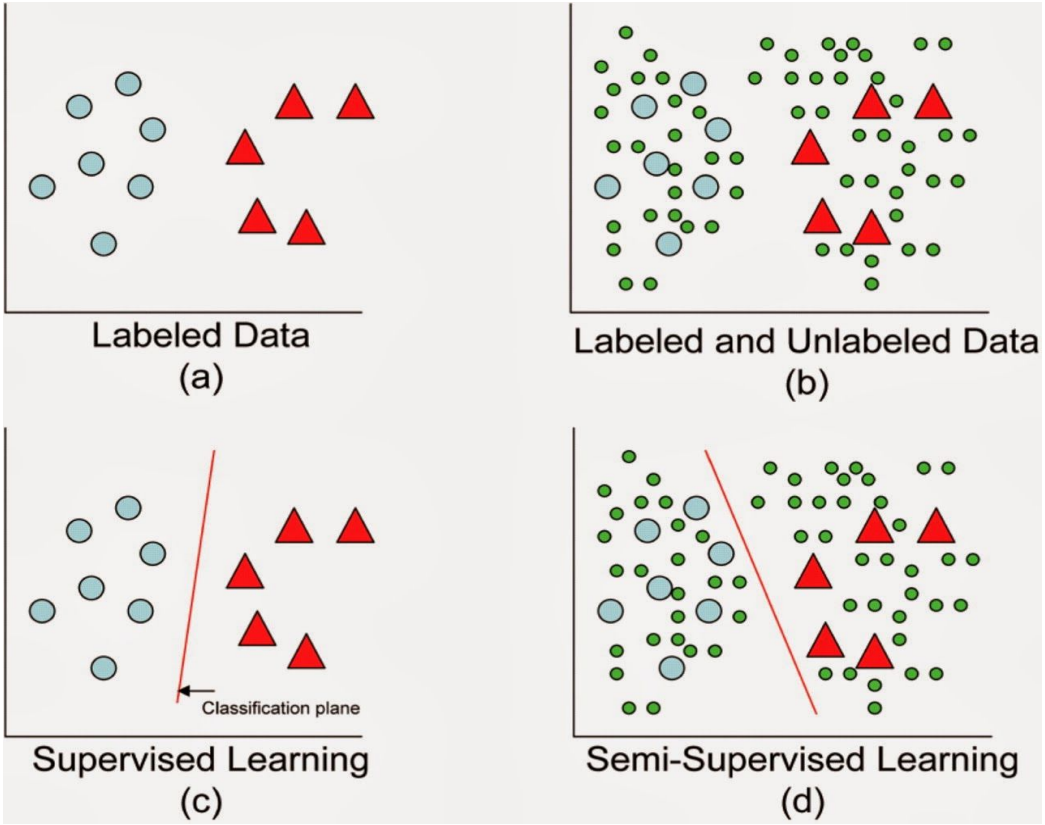
Тренируем на Masked Language Modeling на 30 млн товаров ~2Гб текста

Сравнение сложных моделей



Semi-Supervised Learning





Выделение брендов из товарных наименований

Первый этап - выделение по списку:

- Собран список наиболее популярных брендов (~45к)
- Фильтрация слов, которые могут не являться брендом (~33к)
- Выделение при помощи алгоритма Ахо-Корасика

Проблемы:

- не выделяет незнакомые бренды;
- не способен выделять бренды, которые могут быть включены в название не в качестве бренда (“Россия”, “Я”, “Морской”, ...)

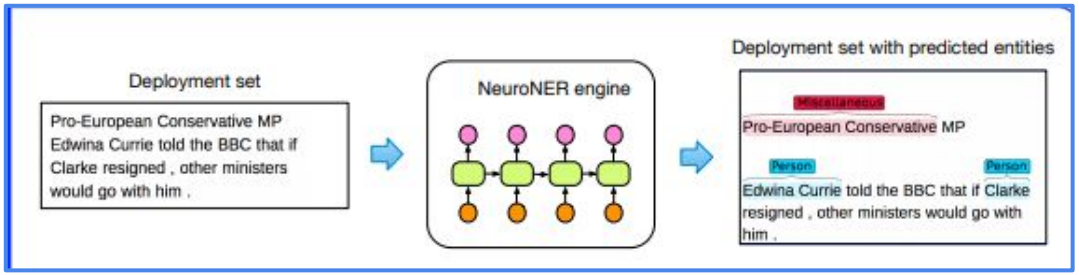
Покрытие около 30%

Модель выделения брендов

Выборка для обучения:

- результаты предыдущего этапа
- данные спаршенные с сайтов двух крупных онлайн магазинов

Neuro NER:



The NER engine's ANN contains three layers:

- Character-enhanced token-embedding layer,
- Label prediction layer,
- Label sequence optimization layer.

Bi-LSTM-CRF:

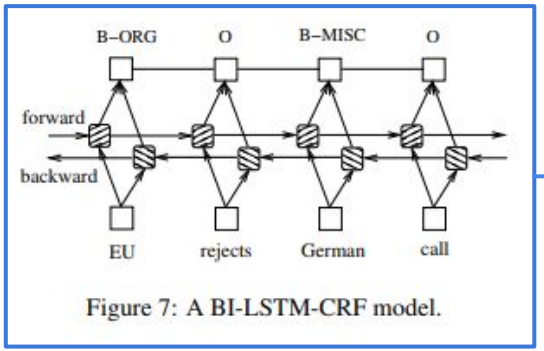


Figure 7: A BI-LSTM-CRF model.

Брест-литовское BRAND молоко ультрапастеризованное

Коломенское BRAND хлеб Даниловский нарезной

*http://neuroner.com

**https://arxiv.org/pdf/1508.01991.pdf

Использование брендов в бизнесе

- Статистические сводки для производителей по аудитории, потребляющей их продукцию
- Прогнозирование спроса на товары данного бренда
- Популярность бренда по регионам
- Маркетинговые рекомендации для производителей
- Аналитика продуктовой корзины потребителя

Использование чеков

- Рекомендации товаров и услуг для клиентов
- Оценка склонности к взятию банковского продукта
- Оценка кредитного риска
- Партнёрские программы с магазинами