SIBUR CHALLENGE 2020

1 место в задаче 2

Продажи

Александр Желубенков
zhelubenkovalexandr@gmail.com

Lamoda, Data Science Team Lead

Компоненты решения

1. Предобработка и токенизация названий
2. Кластеризация графа названий
3. Факторы и их способ подсчета
4. Обучение модели и финальный прогноз

Предобработка и токенизация

- **Нормализация названий**
  - транслит: интертекс → intertex
  - диакритика(пакет unidecode): özşahin → ozsahin
  - удаление слов в скобках
  - оставляем только [0-9a-z]

- **Фильтрация гео**
  - spacy модель en_core_web_lg
  - пакеты geonamescache, pycountry

- **Фильтрация legal entities**
  - https://en.wikipedia.org/wiki/List_of_legal_entity_types_by_country

Предобработка и токенизация

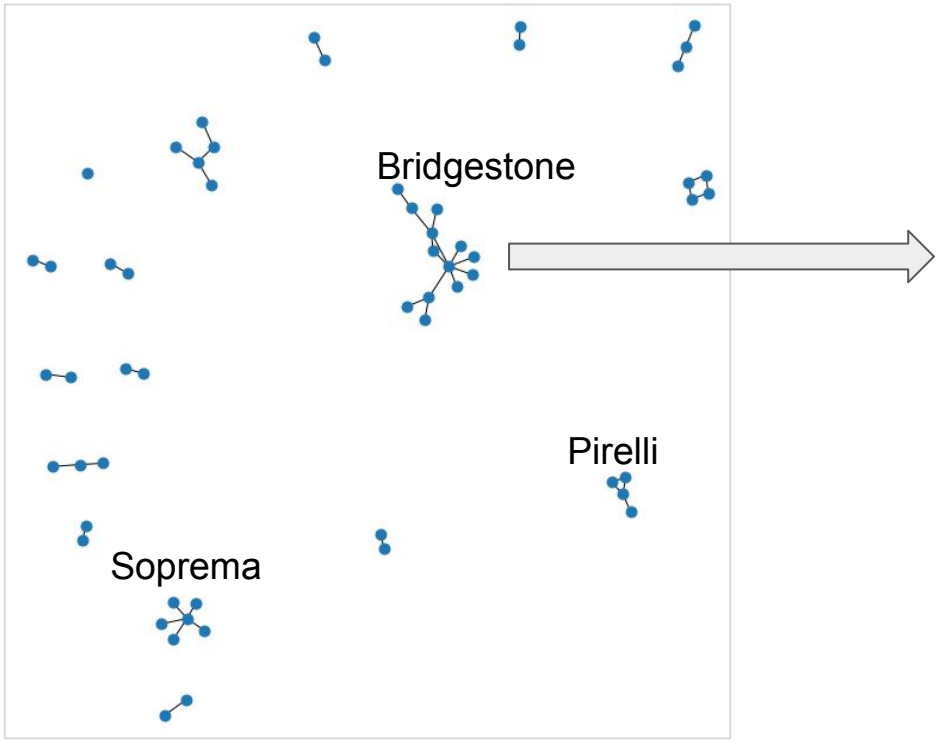
Примеры работы:

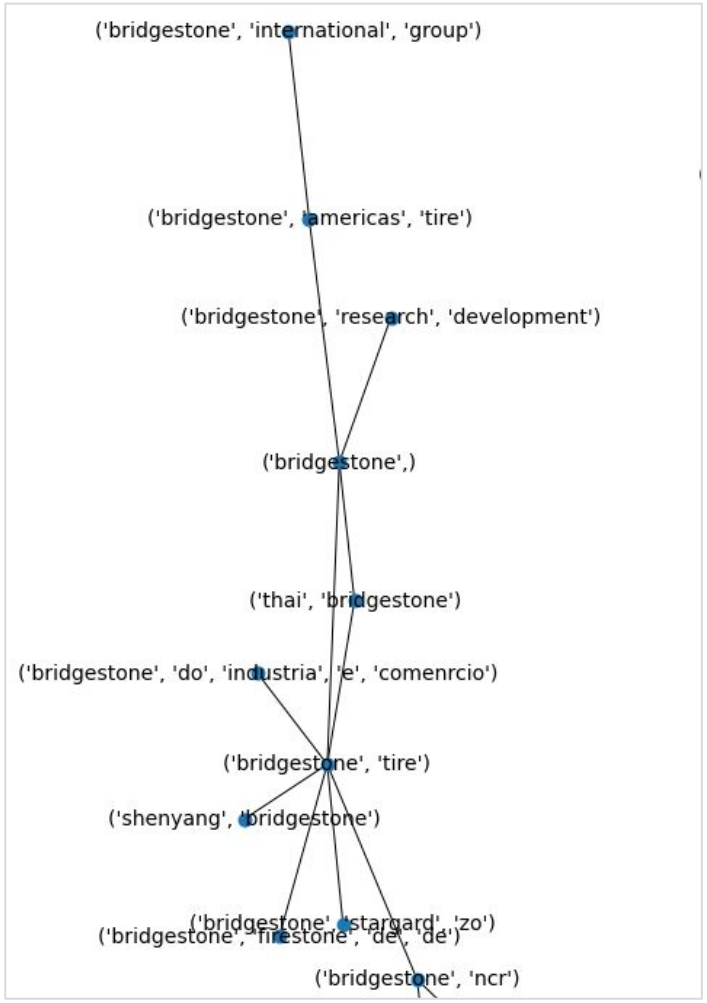
- "Bridgestone India Pvt., Ltd."
  → ("bridgestone")
- "Beijing Zhongyi Rongda Tech Trading Co., Ltd."
  → ("zhongyi", "rongda", "tech")
- "ООО"ЭЛ КЕРАМИКА""
  → ("el", "keramika")

x - legal entity

x - geo entity

Кластеризация графа названий



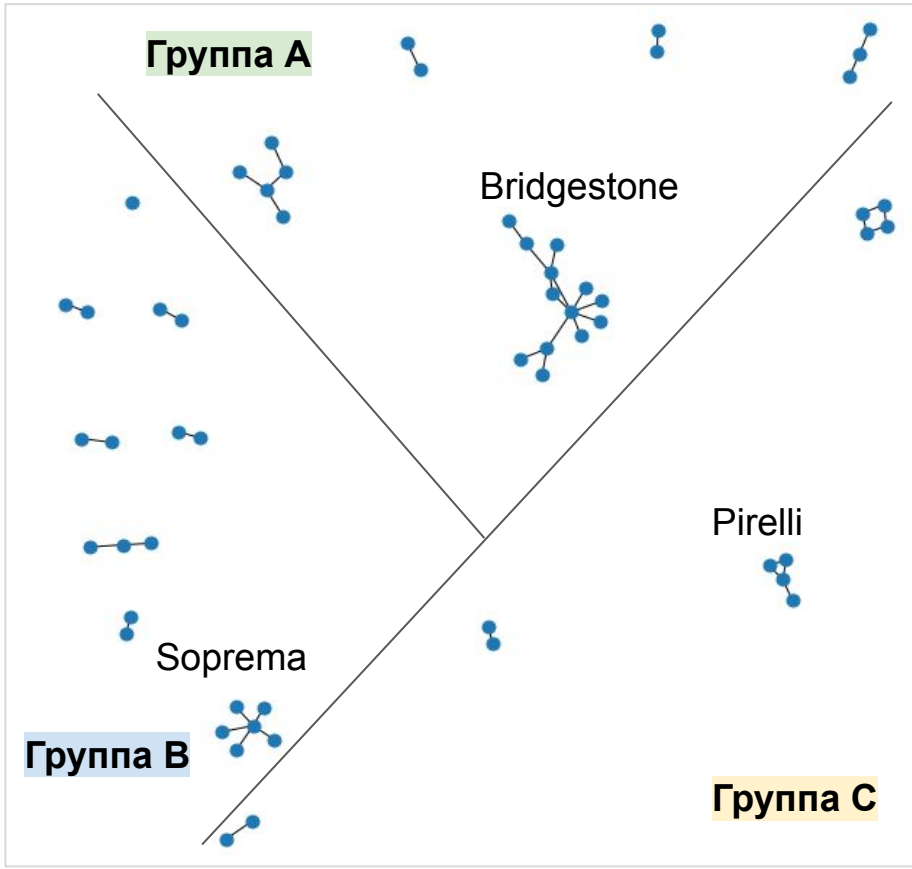


Граф названий:

- **вершины**: токенизированные названия
- **ребра**: связки (name_1, name_2, is_duplicate=1)

Кластера - это компоненты связности.

Разбиение тренировочной выборки на фолды



Каждый кластер целиком попадает в одну из трех групп (GroupKFold). Это важно как для тренировки модели, так и для локальной кросс-валидации.

Факторы и их подсчет для train-выборки

Две группы факторов

- Группа 1: считается для пары названий (name_1, name_2):
  - объединение/пересечение по токенам (с расширением и без);
  - LCS(Longest Common Subsequence) для названий, для первых N слов названий;
- Группа 2: считается для пары названий на основании частотных словарей
  - вероятность того, что токен НЕ значимый - в какой доле позитивных/негативных пар его было можно или нельзя выкидывать.

Важно: когда считаем факторы для группы A частотные словари насчитываем на группах B и C - **чтобы не было лика**

Факторы и их подсчет для train-выборки

Пример

Кластер: ("bridgestone"), ("bridgestone", "automotive"), ("bridgestone", "synthetic", "rubber")

- **bridgestone** - значимое слово (его выкидывать нельзя)
- rubber, automotive, synthetic - НЕ значимые слова

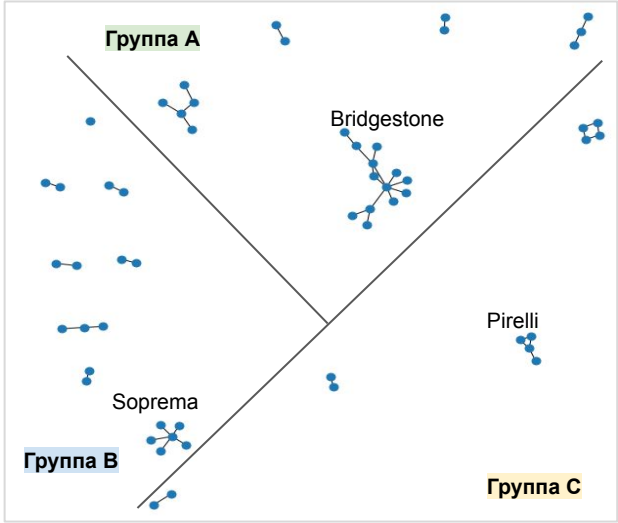
ТОП НЕзначимых слов из частотных словарей, построенных на train:

- logistics
- rubber
- warehouse
- polymers
- chemicals
- industries

Итого:

- "**Sumitomo** Rubber Industries", "**Saiko** Rubber"
  → большой штраф
- "**Sumitomo** Rubber Industries", "**Sumitomo** Warehouse"
  → маленький штраф

Факторы и их подсчет для test-выборки



Когда строим факторы для тренировочной выборки частотные словари считаются три раза:

- для группы A на основании групп B,C
- для группы B на основании групп A,C
- для группы C на основании групп A,B

Факторы "Группы 2" на тестовой выборке считаем, используя все три частотных словаря.

Важно: в качестве значения фактора берем медиану - так распределение значений факторов в тесте у нас НЕ поедет (будет такое же, как на тренировочной выборке).

Обучение модели и финальный прогноз

В качестве модели используем CatBoostClassifier.

Основные параметры:

```
params = {
    "iterations": 100,
    "learning_rate": 0.03,
    "depth": 6,
    "auto_class_weights": 'SqrtBalanced'
}
```

Финальный прогноз: топ-1600 пар с наибольшим score:

- public: 0.885 (1 место)
- private: 0.898 (1 место)

pair_id	name_1	name_2	is_duplicate_predict
115	NOKIAN TYRES PLC	ООО "НОКИАН ТАЙЕРС"	1
284	Alliance Polychem	Aks Polychem Pvt., Ltd.	1
307	Canadian Saddlery & Supply Inc.	Canadian Saddlery & Supply In	1
363	Trelleborg Ysh Sa De Cv	Trelleborg Engineeed Products A	1
635	Repsol Ypf Lubricantes Y	Repsol Lubricantes Y Especialidades	1
774	Henkel Industrie Ag No 3 5 Th Floor	Henkel Corporation	1
800	Bridgestone (Wuxi) Tire.Co. Ltd.	Bridgestone Firestone Venezolana C	1
830	Mol Logistics (Usa.) Inc.Dallas	Mol Logistics (Deutschland) Gmb H	1
982	Sanyo Energy(Beijing) Co. Ltd.	Sanyo E & E S.A. De C.V.	1
1065	Sanyo Corporation Of America	Sanyo Corp	1

...