Построение семантических векторных представлений текстовых документов

Студент 4 курса 3 группы
Синявский Тимур Владимирович

Задачи

- Изучение алгоритмов **эмбеддингов**.
- Разработка алгоритма **тематического моделирования**.
- Реализация системы для **применения** разработанного алгоритма.

Data

date	header	document	tags
2017-07-06T21:35:00+03:00	Тренер "Шахтера": Оправдываться не хочу. Все в...	главный тренер солигорский шахтер олег кубарев...	['футбол']
2017-07-07T09:25:00+03:00	"Зацветет" ли каменная роза на ул. Комсомольск...	план восстановление рисунок пока художник илья...	['архитектура', 'живопись', 'ЖКХ']
2017-07-07T09:27:00+03:00	Фотофакт. Скамейка в виде пожарной машины появ...	областной управление мчс день пожарный служба ...	['министерства']
2017-07-06T22:11:00+03:00	Станислав Драгун дебютировал за БАТЭ в матче с...	чемпион беларусь бате воспользоваться пауза че...	['футбол', 'БАТЭ']
2017-07-06T22:28:00+03:00	Генпрокурор Украины пообещал открыть уголовное...	генпрокуратура украина открывать уголовный про...	['Ситуация в Украине', 'государственные перево...
2017-07-06T22:48:00+03:00	"Славия" выбыла в 1/16 Кубка Беларуси, проигра...	аутсайдер высокий лига мозырский славия выбыва...	['футбол', 'Чемпионат Беларуси по футболу']
2017-07-06T23:05:00+03:00	Александр Бурый не вышел во 2-й раунд парного ...	белорусский теннисист александр бурый выходить...	['теннис']

126000 records

Text preprocessing

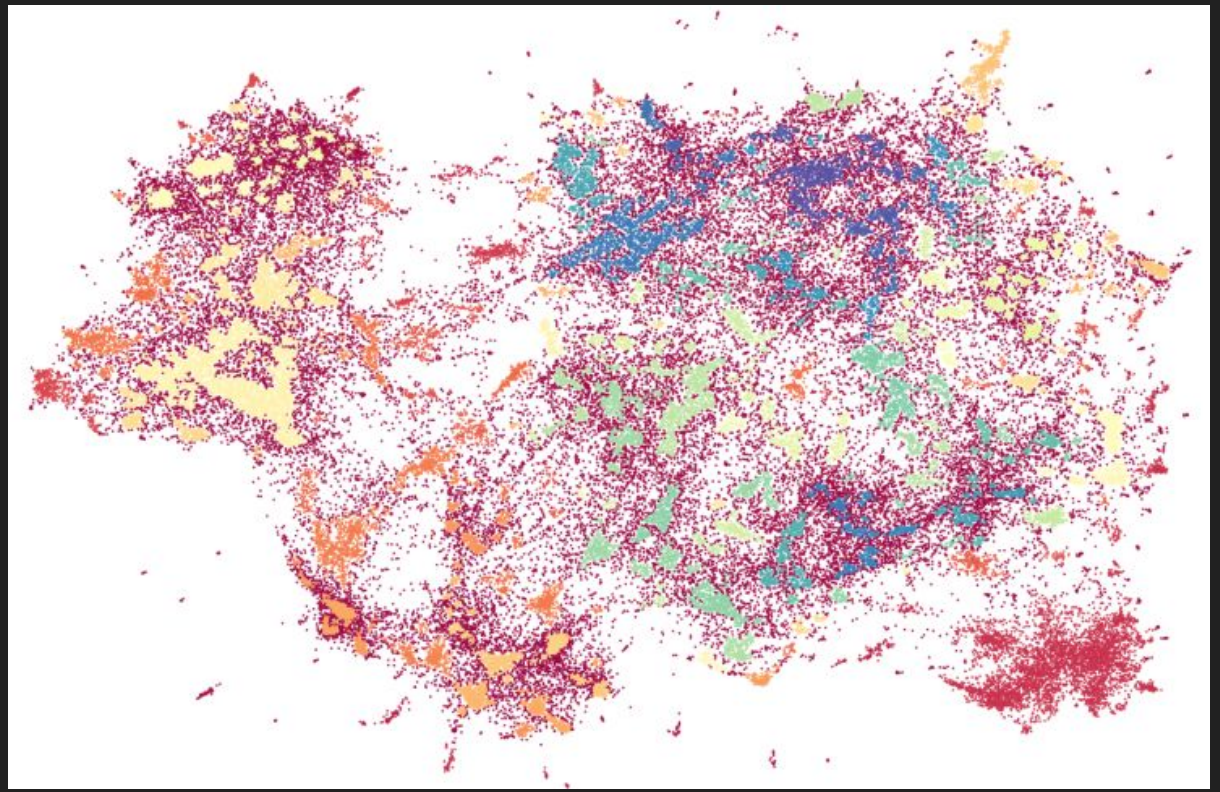
1. Lovercase
2. Digits removing
3. Tokenization
4. Lemmatization
5. Stop words removing

```
Эпидемиологи КНР установили новый вид коронавируса. Он
\ха0стал возбудителем вспышки пневмонии в\ха0городе
Ухань. Об\ха0этом в\ха0четверг сообщило Центральное
телевидение Китая.\n\n«В\ха0результате лабораторных
исследований был найден новый вид коронавируса, удалось
установить весь его геном. При помощи нуклеинового
анализа подтверждено 15 случаев заражения новым видом
коронавируса, у\ха0одного из\ха0инфицированных...
```

```
эпидемиолог кнр устанавливать новый вид коронавирус
становиться возбудитель вспышка пневмония город ухань
это четверг сообщать центральный телевидение китай
результат лабораторный исследование находить новый вид
коронавирус удаваться устанавливать весь ген помощь
нуклеиновый анализ подтверждать случай заражение новый
вид коронавирус инфицированный...
```

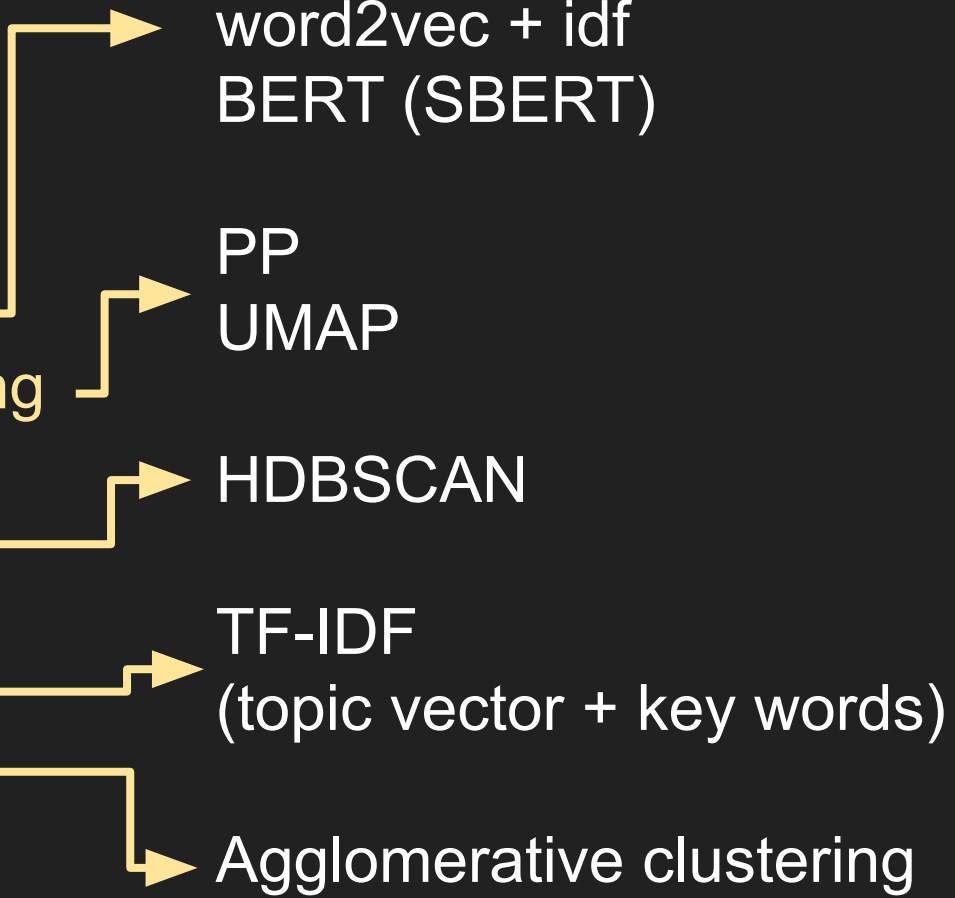
Top2vec

1. Doc2vec
2. UMAP
3. HDBSCAN
4. Topic vector
5. Key words

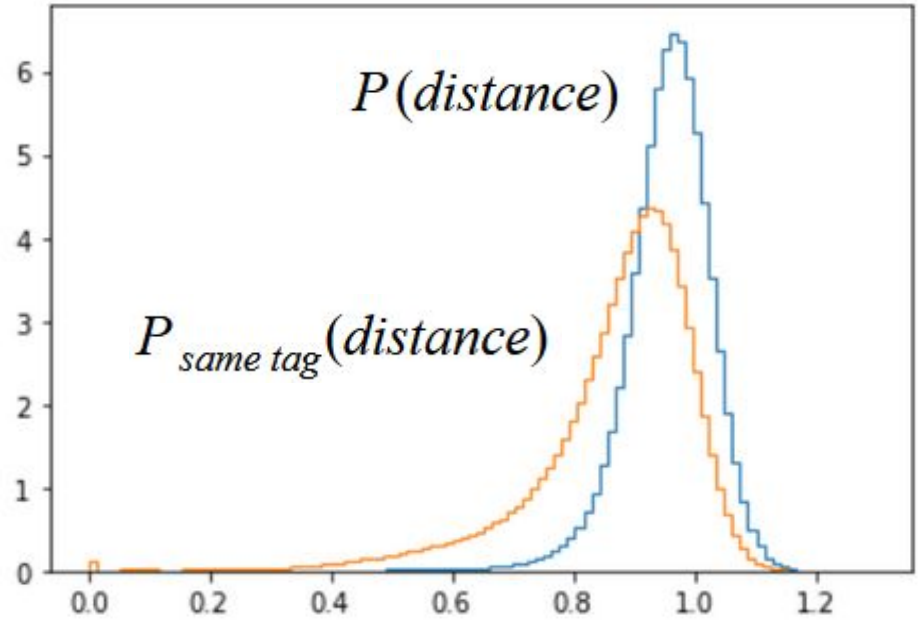


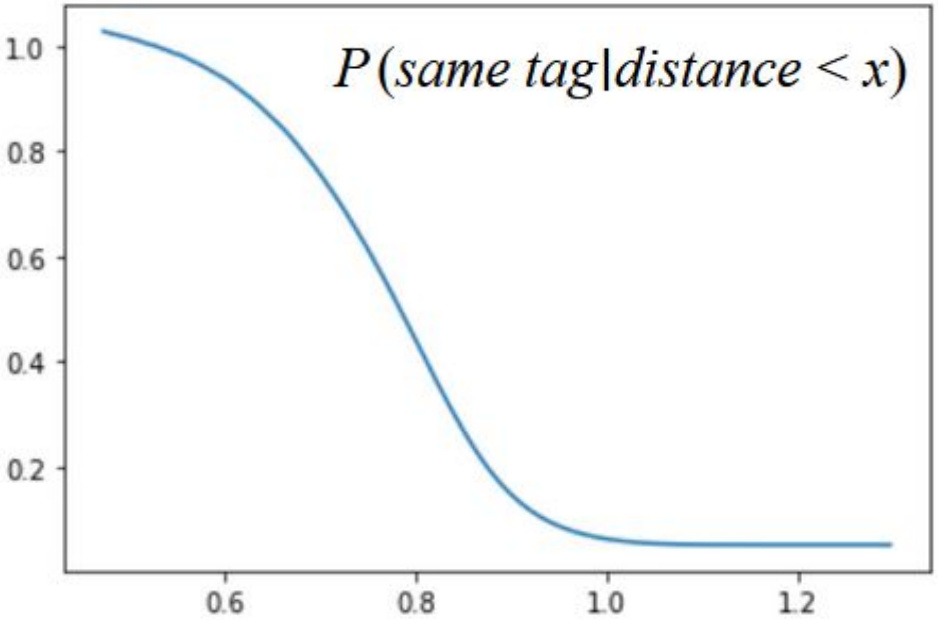
Pipeline

1. Text preprocessing
2. Embeddings
3. Embeddings postprocessing
4. Embeddings evaluation
5. Clustering
6. Clustering evaluation
7. Key words
8. Topic dendrogram



Embeddings evaluation





$$x = quantile_{P(distance)}(0.01)$$

Embeddings evaluation

[P]	Без обработки	PP	UMAP	PP+UMAP
doc2vec	0.670	0.647	0.679	0.624
BERT	0.453	0.599	0.432	0.522
word2vec+idf	0.588	0.690	0.574	0.596

Clustering evaluation

[precision - recall]	UMAP	PP+UMAP
doc2vec	0.747 - 0.026	0.692 - 0.019
BERT	0.699 - 0.017	0.680 - 0.027
word2vec+idf	0.792 - 0.032	0.766 - 0.030

$$precision = \frac{TP}{TP+FP}, \; recall = \frac{TP}{TP+FN}$$

Result pipeline

1. Text preprocessing
2. Embeddings - word2vec + idf
3. Embeddings postprocessing - PP / UMAP
4. Clustering - HDBSCAN
5. Key words - TF-IDF
6. Topic dendrogram - Agglomerative clustering

Технологии системы

Spark - распределённые вычисления

Delta Lake - формат хранения таблиц

Dagster - оркестрация пайплайна

Streamlit - визуализации

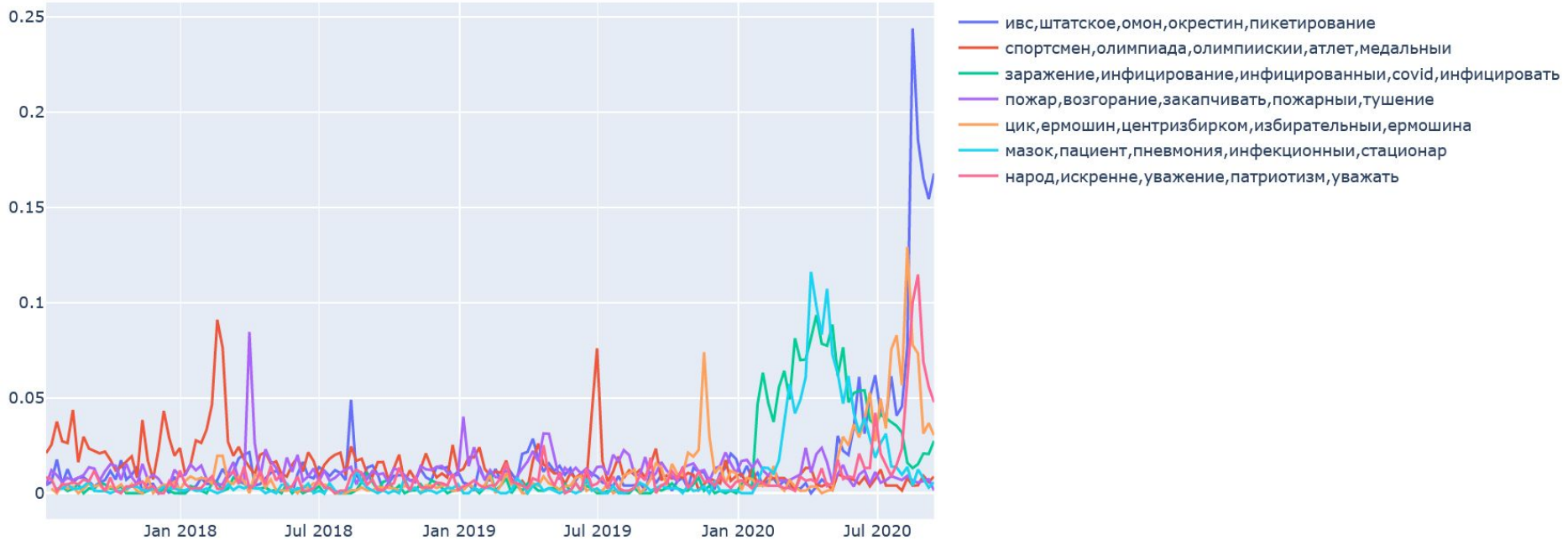
Result

- Разработан пайплайн тематического моделирования с использованием эмбеддингов
- Разработана система для применения пайплайна и визуализации.

Site screenshots

(https://share.streamlit.io/sinytim/topics)

Trends


0.25
0.2
0.15
0.1
0.05
0
Jan 2018
Jul 2018
Jan 2019
Jul 2019
Jan 2020
Jul 2020
ивс,штатское,омон,окрестин,пикетирование
спортсмен,олимпиада,олимпиискии,атлет,медальныи
заражение,инфицирование,инфицированныи,covid,инфицировать
пожар,возгорание,закапчивать,пожарныи,тушение
цик,ермошин,центризбирком,избирательныи,ермошина
мазок,пациент,пневмония,инфекционныи,стационар
народ,искренне,уважение,патриотизм,уважать

Latest articles:

[2021-05-08 13:36:00] **Евросоюз выходит из локдауна. В школу, в ресторан, в отпуск — уже можно?** (Евросоюз, Коронавирус, вакцинация, туризм)

[2021-05-07 12:20:00] **В Индии новый рекорд по суточной заболеваемости COVID-19 — больше 414 тысяч человек** (В мире, Коронавирус, эпидемии)

[2021-05-06 08:59:00] **В Индии за сутки выявили рекордное число заразившихся коронавирусом** (Коронавирус, эпидемии)

[2021-05-05 13:15:00] **В Индии установлен очередной антирекорд смертности от коронавируса** (В мире, Коронавирус, эпидемии)

The most similar articles to the selected one:

[2021-05-08 12:10:00] **Инфекционист — о поставках в Беларусь вакцины от Pfizer и BioNTech и реакциях на прививку от COVID-19** [здоровье, Коронавирус, медицина, вакцинация, Здравоохранение] [вакцина, вакцинация, прививка, доза, прививать] (distance: 0.000)

[2021-03-30 13:41:00] **Как отличаются вакцины, каким будет иммунитет. Инфекционист ответил на вопросы по поводу прививок от COVID-19** [здоровье, Коронавирус, медицина, вакцинация, Здравоохранение, наука] [вакцина, вакцинация, прививка, доза, прививать] (distance: 0.021)

[2020-12-29 17:27:00] **Что полезно знать о российской вакцине «Спутник V», которой начали прививать белорусов** [здоровье, Коронавирус, медицина, Беларусь - Россия, вакцинация, министерства, наука, ОРВИ, грипп, эпидемии] [вакцина, вакцинация, прививка, доза, прививать] (distance: 0.023)

Dendrogram


кредит, поручительство, мес, банк, залог
банк, visa, карта, byn, карточка
банк, обменник, абсолютбанк, валюта, рубль
валюта, byn, курсы, скачать, рубль
дтп, водитель, гаи, авария, tutbylive
дтп, пешеход, гаи, наезд, сбивать
дтп, водитель, неосторожность, повлечь, пдд
гаи, инспектор, водитель, пьяный, промилле
угнать, угон, автомобиль, милиция, дтп
смартфон, galaxy, iphone, pro, xiaomi
абонент, тариф, мтс, оператор, тарифный
скидка, tamby, акция, руб, tamby
скидка, диагностика, ремонт, замена, подвеска
hyundai, blue, creta, accent, сигарета
автомобиль, кроссовер, двигатель, машина, geely
renault, kaptur, stepway, duster, трейдин
nissan, драйвмоторс, xtrail, qashqai, автоцентр
lada, минсклад, серов, автоцентр, vesta
боровой, девелопмент, пирс, квартал, квартира
квартира, комплекс, челюскинец, маяк, квартал
realtby, квадратный, квартира, метр, эквивалент