Citation network project

MADE'22 fall

Команда № 15
Тьютор: Игорь Иткин

Команда проекта:
Екатерина Потапова
Ирина Пугаева
Артем Ткаченко
Николай Шаманков
Евгений Шаров

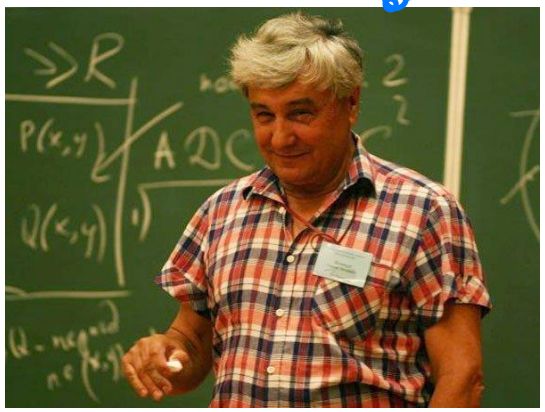
Научное сообщество

Помощь в поиске статей и соавторов

Что бы еще почитать?

Студент

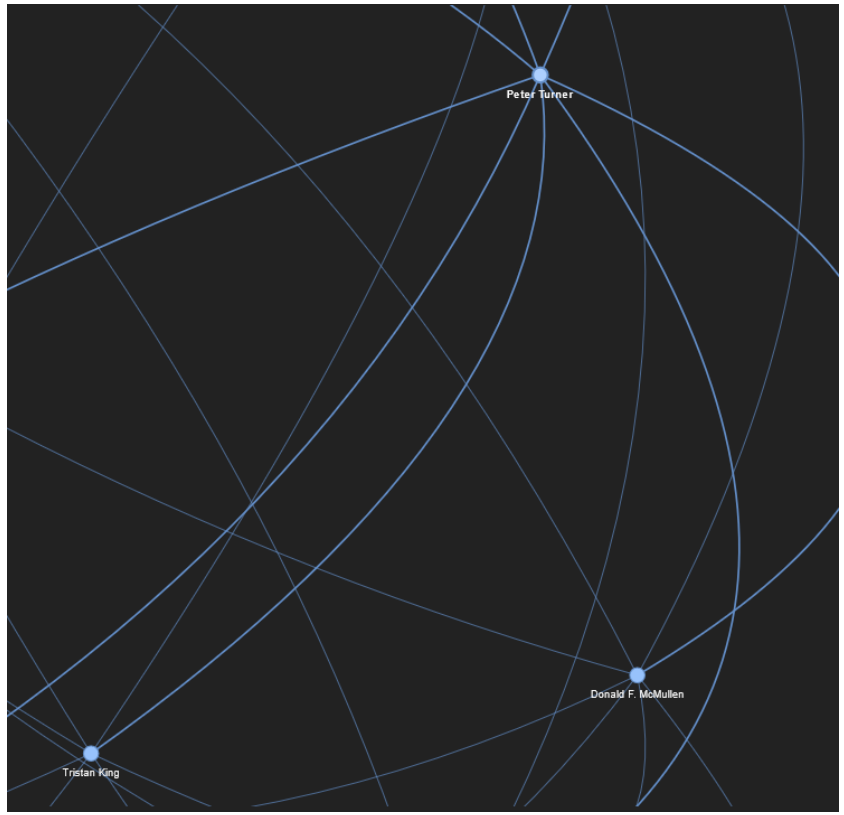
С кем бы написать статью?



Научный сотрудник

Артефакты

- EDA
- Кластеризация
- Рекомендации статей
- Репозиторий
- Веб-приложение



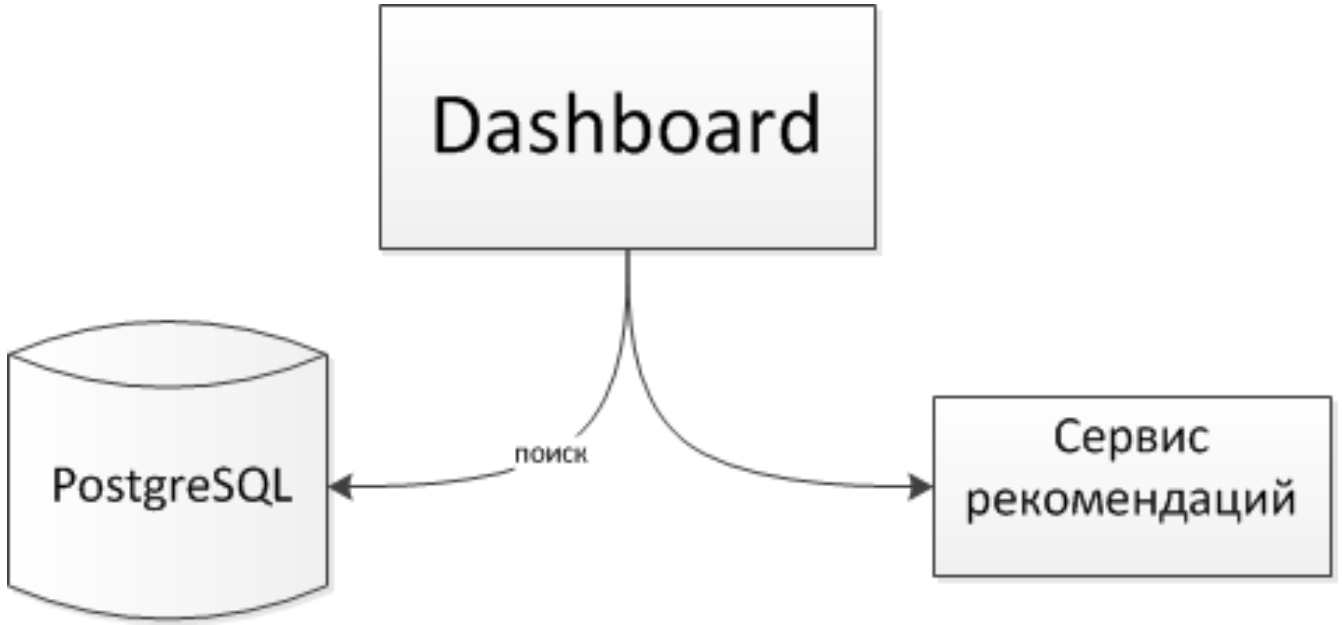
Технологии и архитектура

Технологии

- Python
- sklearn
- Uvicorn
- dash
- FastAPI
- Faiss
- pyvis.network, networkx, plotly
- git
- PostgreSQL

Модели и метрики

- sentence-transformer
- all-mpnet-base-v2



Демонстрация проекта

Citation network project

We will recommend some articles for you based on your search.

Write the title you are looking for:

Search: Gene

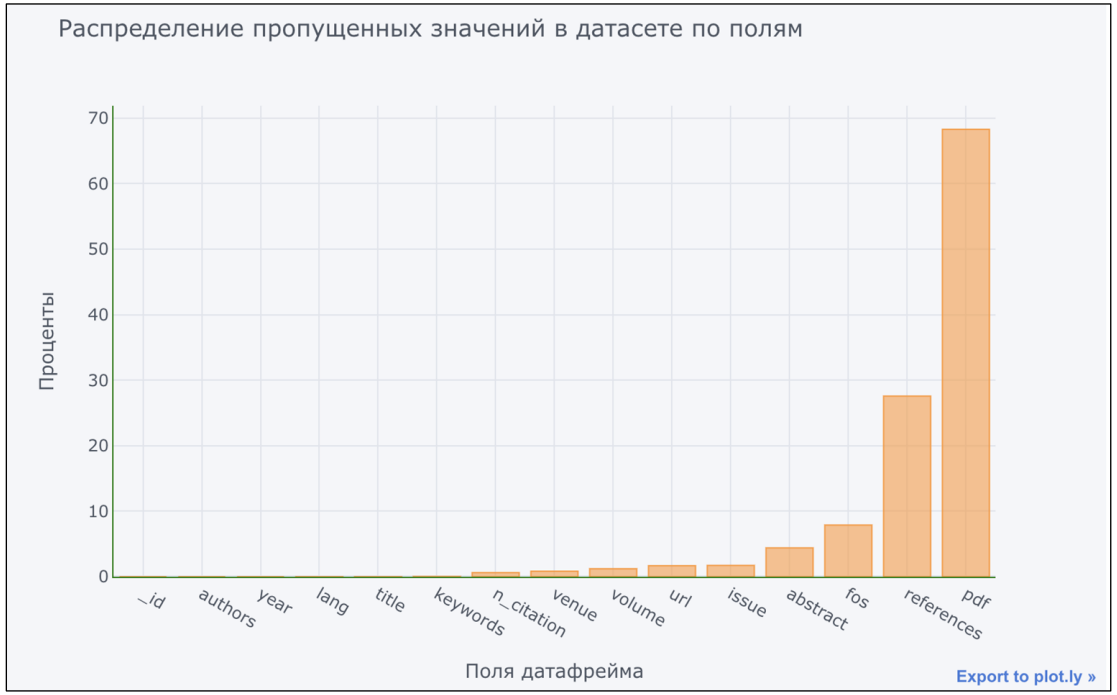
title	abstract	id	volume	year	n_citation	lang	index
Generalized Veronesean	We classify all embeddings θ: PG(n, q) → PG(d, q), with $$d \\geqslant \\tfrac{{n(n + 3)}} {2}$$, such that θ maps the set of points of each line to a set of coplanar points	53e99d65b7602d970261fbb9	31	2011	8	en	35
Genetic Algorithm		53e99d65b7602d970261f9de		2005	0	en	49

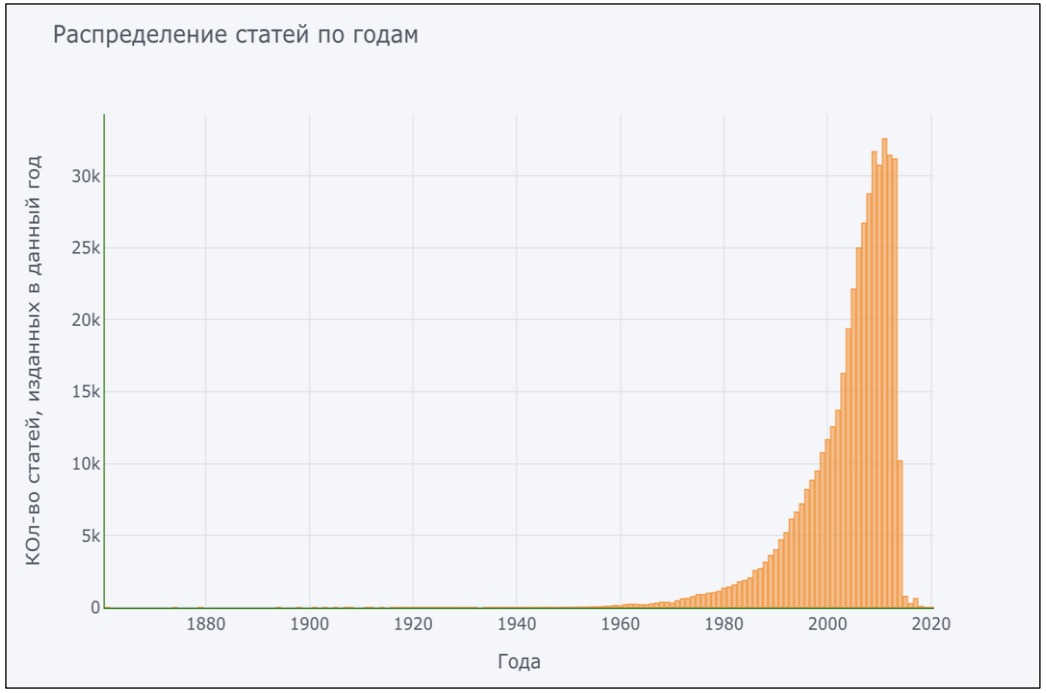
« ‹ 1 › »

These article might be interesting for you:

title	abstract	id	volume	year	n_citation	lang	index
Performance study of	This paper proposes five bandwidth reallocation algorithms (BRAs) that can be applied to the quantitative provisioning in a differentiated services (DS) network supporting	53e99d65b7602d970261f607	24	2001	4	en	1
Source level debugging of	We describe a method for providing source level debugging for programs that have been automatically parallelized for distributed memory, MIMD machines. We call this method a	53e99d65b7602d970261f610	26	1991	7	en	2

EDA. Выводы по данным. Очистка данных





Модель кластеризации. Используемые технологии

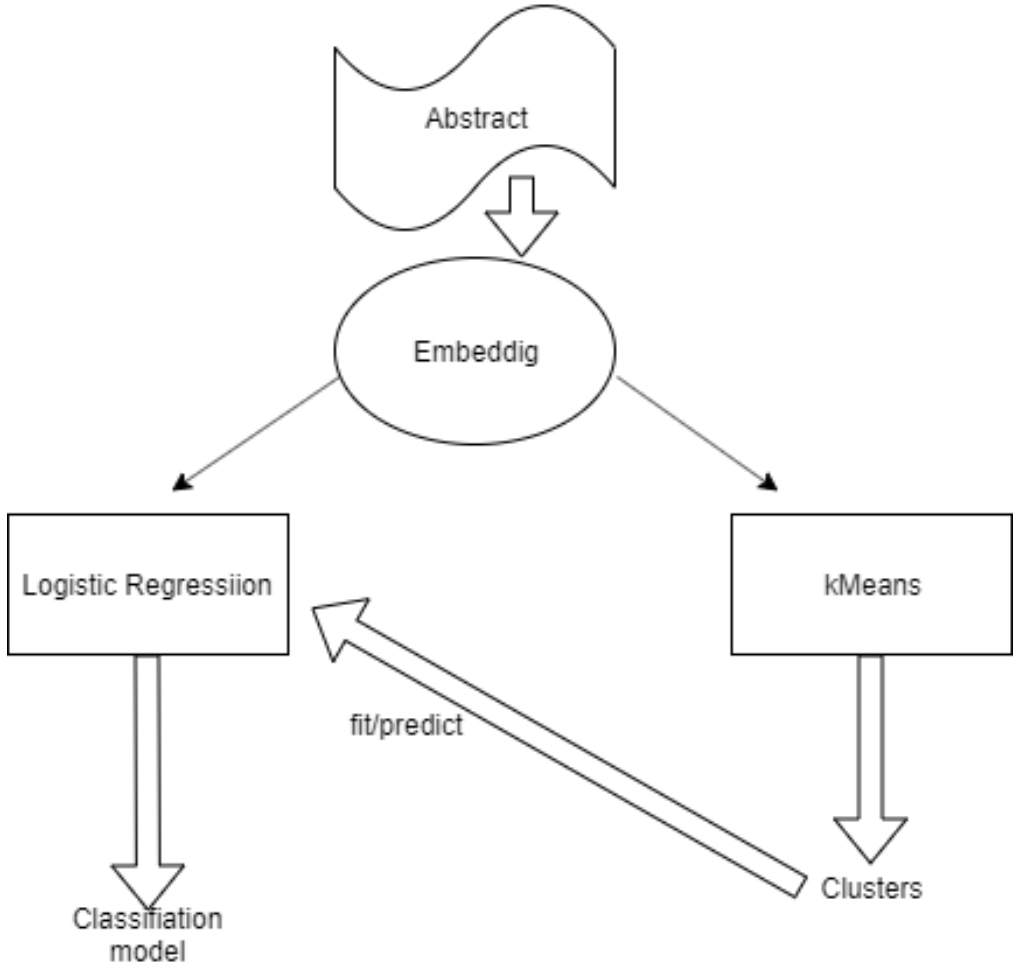
- all-MinilM-L6-v2 sentence-transformers model с относительно небольшим размером + высокой скоростью

Model Name	Performance Sentence Embeddings (14 Datasets)	Performance Semantic Search (6 Datasets)	Avg. Performance	Speed	Model Size
ll-mpnet-base-v2	69.57	57.02	63.30	2800	420 MB
nulti-qa-mpnet-base-dot-v1	66.76	57.60	62.18	2800	420 MB
ll-distilroberta-v1	68.73	50.94	59.84	4000	290 MB
ll-MiniLM-L12-v2	68.70	50.82	59.76	7500	120 MB
nulti-qa-distilbert-cos-v1	65.98	52.83	59.41	4000	250 MB
ll-MiniLM-L6-v2	68.06	49.54	58.80	14200	80 MB

- k-means + подбор оптимального количества кластеров по методу локтя - 25 кластеров

Модель Классификации

- Общая схема построения классификатора
- Используется тот же эмбеддер, что и на этапе
- Кластеризации
- В качестве финального решения была выбрана модель логистической регрессии, обладающей максимальной метрикой F1-score



Модель рекомендации статей. Используемые технологии, дизайн эксперимента

- all-MinilM-L6-v2 sentence-transformers model с относительно небольшим размером + высокой скоростью

Model Name	Performance Sentence Embeddings (14 Datasets)	Performance Semantic Search (6 Datasets)	Avg. Performance	Speed	Model Size
ll-mpnet-base-v2	69.57	57.02	63.30	2800	420 MB
nulti-qa-mpnet-base-dot-v1	66.76	57.60	62.18	2800	420 MB
ll-distilroberta-v1	68.73	50.94	59.84	4000	290 MB
ll-MiniLM-L12-v2	68.70	50.82	59.76	7500	120 MB
nulti-qa-distilbert-cos-v1	65.98	52.83	59.41	4000	250 MB
ll-MiniLM-L6-v2	68.06	49.54	58.80	14200	80 MB

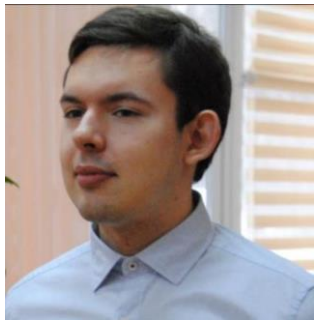
- FAISS

Facebook AI Research Similarity Search – разработка команды Facebook AI Research для быстрого поиска ближайших соседей и кластеризации в векторном пространстве. Высокая скорость поиска позволяет работать с очень большими данными – до нескольких миллиардов векторов.

Команда проекта

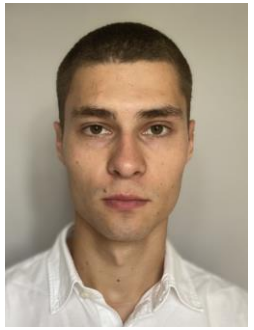
Ирина Пугаева

DS



Николай Шаманков

DS



Евгений Шаров

Дэшборд и микросервис для рекомендаций

Екатерина Потапова

Engeneering

Артем Ткаченко

Инфраструктура, база данных

Спасибо
за внимание!

Итоговый отчет по проекту

Команда № 15

Дэшборд описание

Дэшборд разработан с помощью библиотеки dash.
Данные о статьях выгружаются из базы данных.
Пользователям доступен поиск по названию статей.
На основании найденных статей пользователям рекомендуются другие статьи.

RecSys модель развернута в отдельном микросервисе, на которые отправляются запросы из дэшборда.

Реализовать авторизацию не получилось, поскольку нативная реализация dash_auth блокировала поиск по данным после 5-10 поисковых запросов.
Проблему устранить не получилось.

Дэшборд для просмотра статей и рекомендаций

Citation network project

We will recommend some articles for you based on your search.

Write the title you are looking for:

Search: Gene

title	abstract	id	volume	year	n_citation	lang	index
Generalized Veronesean	We classify all embeddings θ: PG(n, q) → PG(d, q), with $$d \\geqslant \\tfrac{{n(n + 3)}} {2}$$, such that θ maps the set of points of each line to a set of coplanar points	53e99d65b7602d970261fbb9	31	2011	8	en	35
Genetic Algorithm		53e99d65b7602d970261f9de		2005	0	en	49

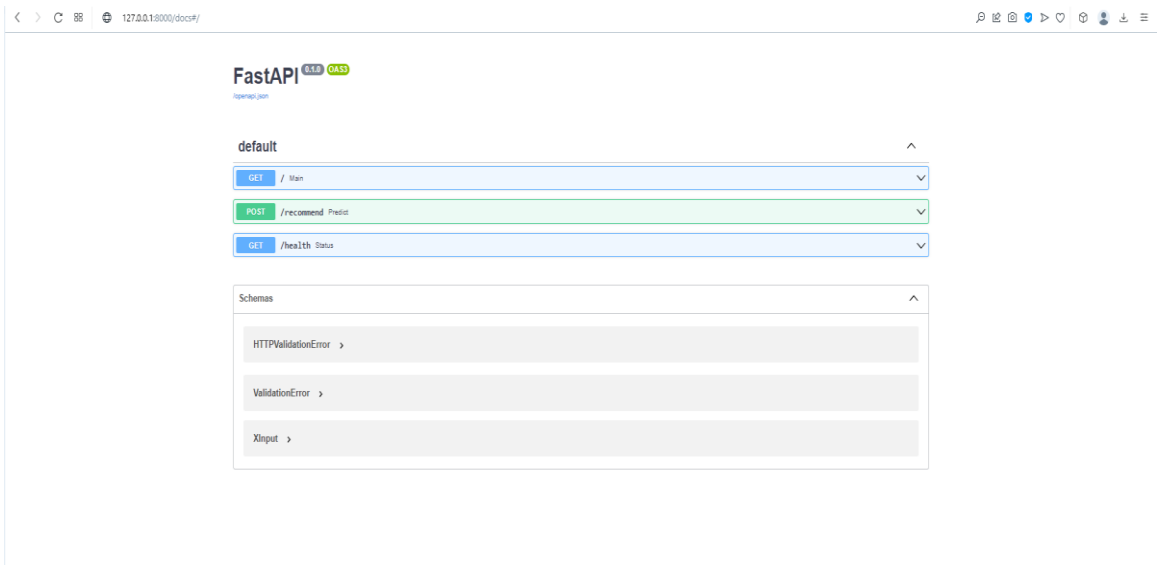
« ‹ 1 › »

These article might be interesting for you:

title	abstract	id	volume	year	n_citation	lang	index
Performance study of	This paper proposes five bandwidth reallocation algorithms (BRAs) that can be applied to the quantitative provisioning in a differentiated services (DS) network supporting	53e99d65b7602d970261f607	24	2001	4	en	1
Source level debugging of	We describe a method for providing source level debugging for programs that have been automatically parallelized for distributed memory, MIMD machines. We call this method a	53e99d65b7602d970261f610	26	1991	7	en	2

Модель рекомендации соавторов. Используемые технологии

- Для построения рекомендаций используется алгоритмический подход: рекомендуем тех соавторов, с которыми ранее автор имел наибольшее число публикаций
- Данное решение легко ложится на реальное поведение людей в научных коллективах
- Итоговый результат на текущий момент представлен в виде локального веб-сервиса, позволяюсь рекомендовать соавторов по id автора публикации

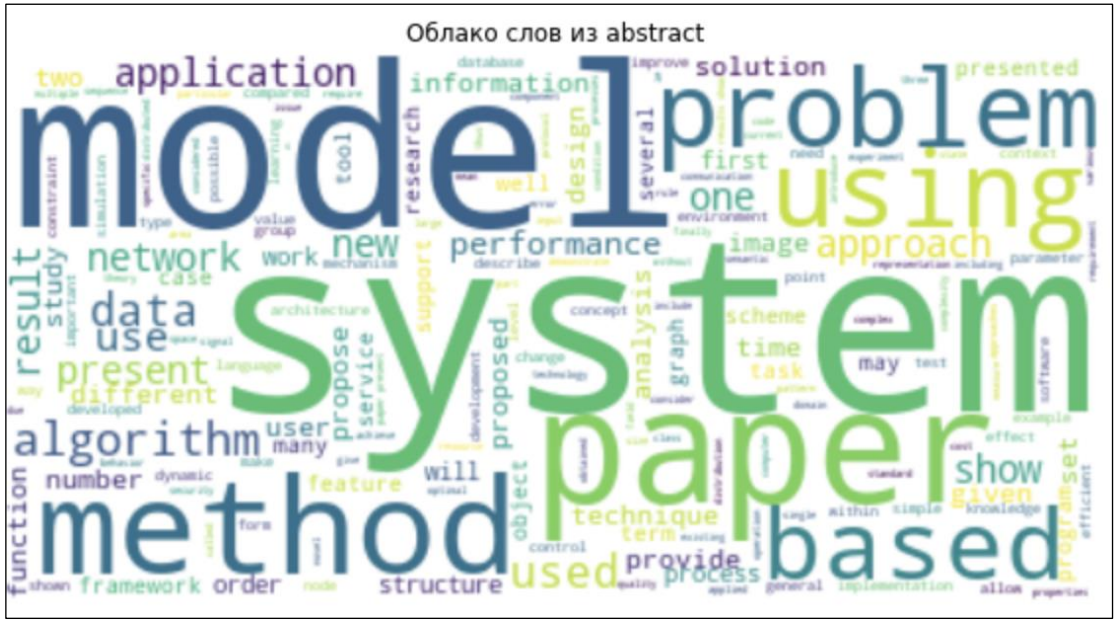


Микросервис для прогнозирования

Микросервис разработан с помощью fastapi.
При запуск загружает модель и данные из БД. Принимает на вход данные о статьях и возвращает id рекомендованных статей.

Сервер не поддерживает асинхронные вызовы, поскольку ожидает запросов только от дэшборда.

EDA. Выводы по данным. Очистка данных



EDA. Выводы по данным. Очистка данных

- Учёный имеет индекс h, если h из его N статей цитируются как минимум h раз каждая, в то время как оставшиеся (N — h) статей цитируются не более чем h раз каждая

- Мы написали функция для вычисления индекса Хирша и применили ее к массиву из числа цитирований каждой статьи автора

Инфраструктура

- Virtual private server (VPS) – гарантирует гибкость и доступность
- Centos 7
- 1 Gb RAM
- 20 Gb HDD
- СУБД PostgreSQL – свободное ПО (as in freedom), большое сообщество

- Ограничения
- Загружено примерно 40% от данных Citation Network Dataset (не хватило места на сервере)
- Не использовался docker (больше ручной работы при установке Python и прочего ПО, меньше гибкости)

DB Scheme

- Virtual private server (VPS) – гарантирует гибкость и доступность
- Centos 7
- 1 Gb RAM
- 20 Gb HDD
- СУБД PostgreSQL – свободное ПО (as in freedom), большое сообщество

- Ограничения
- Загружено примерно 40% от данных Citation Network Dataset (не хватило места на сервере)
- Не использовался docker (больше ручной работы при установке Python и прочего ПО, меньше гибкости)