Цель: реализация программного средства для рекомендаций научных статей

Задачи:

- Обзор существующих моделей нахождения векторных представлений документов
- Реализация программы по автоматическому сбору данных для обучения моделей NLP
- Обучение моделей на собранном датасете и выбор наиболее эффективной
- Реализация программного средства для рекомендаций научных статей на основе выбранной модели:
  - Создание API для взаимодействия с Google Drive
  - Создание базы данных на PostgreSql для протоколирования и сохранения результатов модели
  - Реализация программного модуля для поиска рекомендаций статей
  - Интеграция с существующим графическим интерфейсом

Полнотекстовый поиск

Документ 1	Документ 2	Схожесть
Документ_n	Документ_k	0.95
Документ_n	Документ_w	0.55

Практическая значимость:

специалист сокращает время на поиск научных статей, которые могут помочь в исследовании конкретной проблемы, темы.

Опробованные модели

TF-IDF – статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов или корпуса

Word2Vec и развивающие данный подход модели **Doc2Vec** и **FastText** – модели на основе искусственных нейронных сетей, предназначенные для получения векторных представлений слов на естественном языке

BERT – это предобученная на большой коллекции тестовых документов языковая модель, основанная на использовании трансформеров. BERT показал state-of-the-art результаты в ряде NLP задач

APTM – метод аддитивной регуляризации тематических моделей. Данный подход позволяет обучать тематические модели, используя регуляризаторы и модальности, учитывающие специфику предметной области

Исходные данные

С помощью реализованной программы с ресурса arxiv.org было собрано 3752 научных статьи по вычислительному материаловедению на английском языке.

Проведена подготовка статей для использования в моделях:

1. Удаление стоп-слов
2. Лемматизация
3. Токенизация

	pdf_key	title	author	journal	year	abstract	url	memory (Byte)
0	14113136v1.pdf	Shannon Entropy and Many-Electron Correlations...	['Luigi Delle Site']	International Journal of Quantum Chemistry	2014-11-12T10:59:24Z	In this paper I will discuss the overlap betwe...	https://export.arxiv.org/pdf/1411.3136v1	223318
1	191206192v2.pdf	Investigating solvent effects on the magnetic ...	['Loïc Halbert', 'Małgorzata Olejniczak', 'Val...	International Journal of Quantum Chemistry	2019-12-12T20:19:22Z	We investigate the ability of mechanical and e...	https://export.arxiv.org/pdf/1912.06192v2	16801744
2	condmat0507292v1.pdf	Interaction induced magnetic field asymmetry o...	['Markus Buttiker', 'David Sanchez']	International Journal of Quantum Chemistry	2005-07-13T08:50:28Z	We demonstrate that the nonlinear I-V characte...	https://export.arxiv.org/pdf/cond-mat/0507292v1	186070

Обучение моделей

Подзадача:

1. Удаление вхождений в документ химических элементов и их синонимов (Al, He, O2)
2. Разметка датасета на основе вхождений вышеперечисленных элементов
3. Обучение моделей NLP (18 моделей)
4. С помощью полученных векторных представлений обучить на размеченном датасете модель логистической регрессии (80% - обучающая выборка, 20% - тестовая, средний размер выборки – 500 документов)

Результаты моделирования

Хим. элемент / Модель	O2	He	Al
Word2Vec	0.55	0.59	0.72
Doc2Vec	0.69	0.71	0.81
FastText	0.68	0.7	0.85
TF-IDF	0.75	0.74	0.86
Bert	0.46	0.32	0.67
APTM	0.72	0.78	0.89

Метрика качества – F1-мера
Лучшая модель - APTM

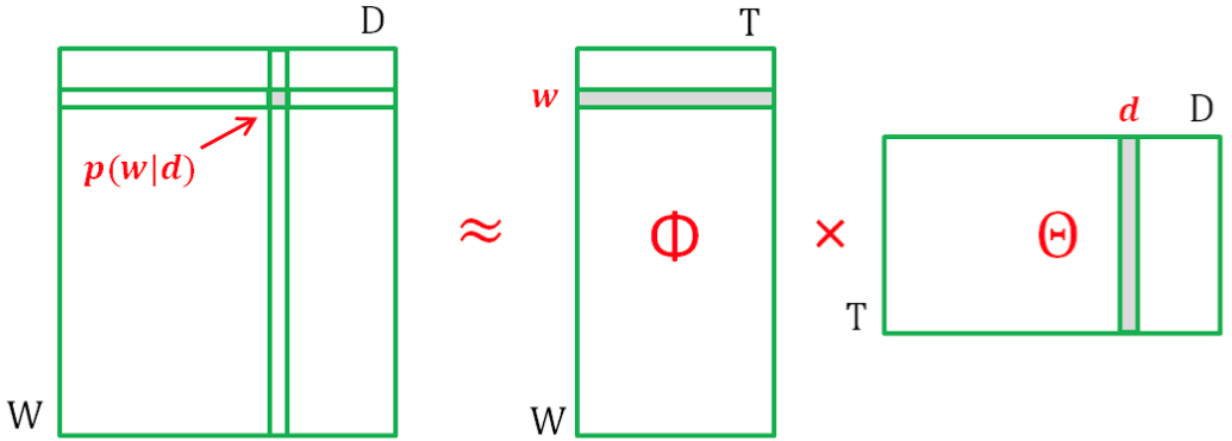
Тематическое моделирование

$p(w|t)$ – вероятность, с которой токен w встречается в теме t

$p(t|d)$ – вероятность, с которой тема t встречается в документе d

$t \in T, w \in W, d \in D$

$$\Phi = p(w|t)_{W \times T},$$
$$\Theta = p(t|d)_{T \times D},$$
$$F_{W \times D} \approx \Phi_{W \times T} \times \Theta_{T \times D}$$



Vorontsov, K., Ianina, A. Multimodal topic modeling for exploratory search in collective blog. In Intelligent Data Processing: Theory and Applications: Book of abstracts of the 11th International Conference (pp.186-187).

APTM

$$\mathrm{L}(\Phi,\Theta)=\sum_{d\in D}\sum_{w\in d}n_{dw}\ln\sum_{t\in T}\phi_{wt}\theta_{td}+\sum_{i}\tau_i R_i(\Phi,\Theta)\to\max_{\Phi,\Theta}$$

$$\sum_{w\in W}\phi_{wt}=1,\qquad \phi_{wt}\geq 0,\qquad \sum_{t\in T}\theta_{td}=1,\qquad \theta_{td}\geq 0,$$

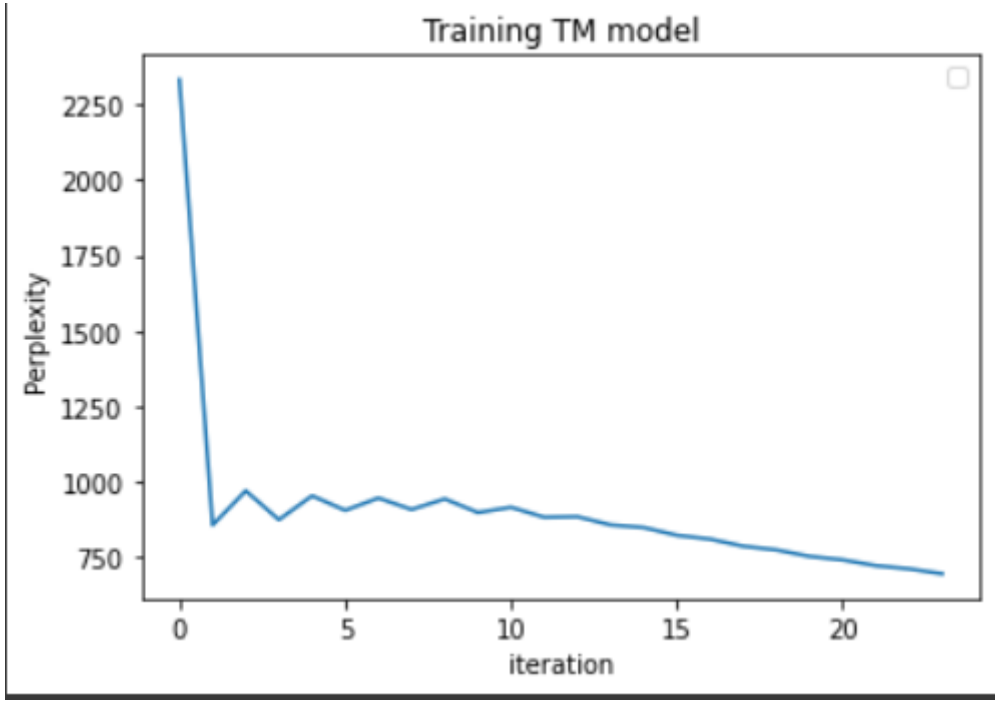
где $\tau_i \geq 0$ – коэффициент регуляризации.

EM-алгоритм

$$\begin{cases} p_{tdw} = p(t|d,w) = \underset{t\in T}{\text{norm}}(\phi_{wt}\theta_{td}) \\ \phi_{wt} = \underset{w\in W}{\text{norm}}\left(\sum_{d\in D} n_{dw}p_{tdw} + \phi_{wt}\frac{\partial R(\Phi,\Theta)}{\partial \phi_{wt}}\right) \\ \theta_{td} = \underset{w\in W}{\text{norm}}\left(\sum_{w\in W} n_{dw}p_{tdw} + \theta_{td}\frac{\partial R(\Phi,\Theta)}{\partial \theta_{td}}\right), \end{cases}$$

где $\underset{t\in T}{\text{norm}}(x_t) = \frac{\max\{x_t,0\}}{\sum_{s\in T}\max\{x_t,x_s\}}$ – операция нормировки вектора.

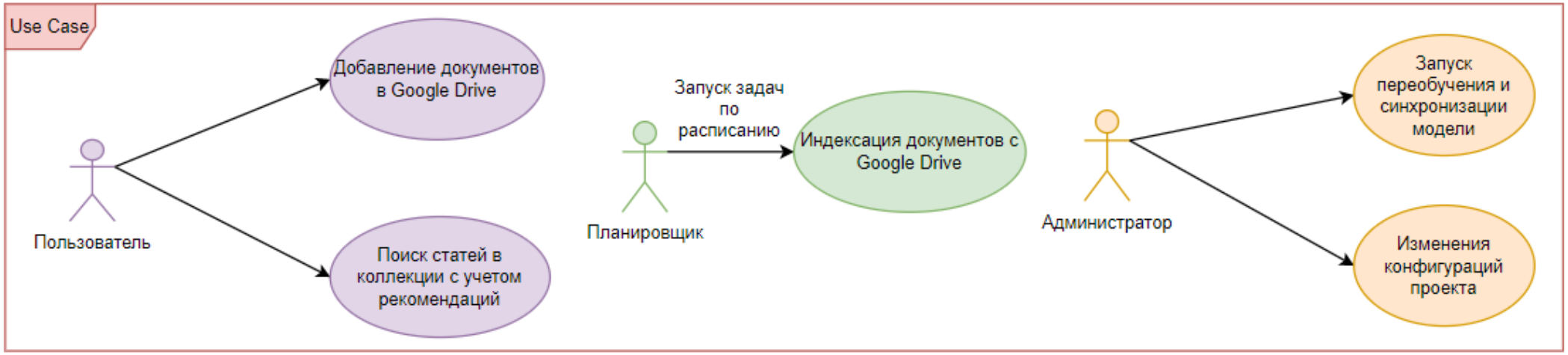
Результаты моделирования



Ход обучения:

- 1-ая стадия (5 итераций) – регуляризатор декоррелирования Φ (τ = 1.0e+5)
- 2-ая стадия (19 итераций) – регуляризатор разреживания Θ (τ = -1.5)

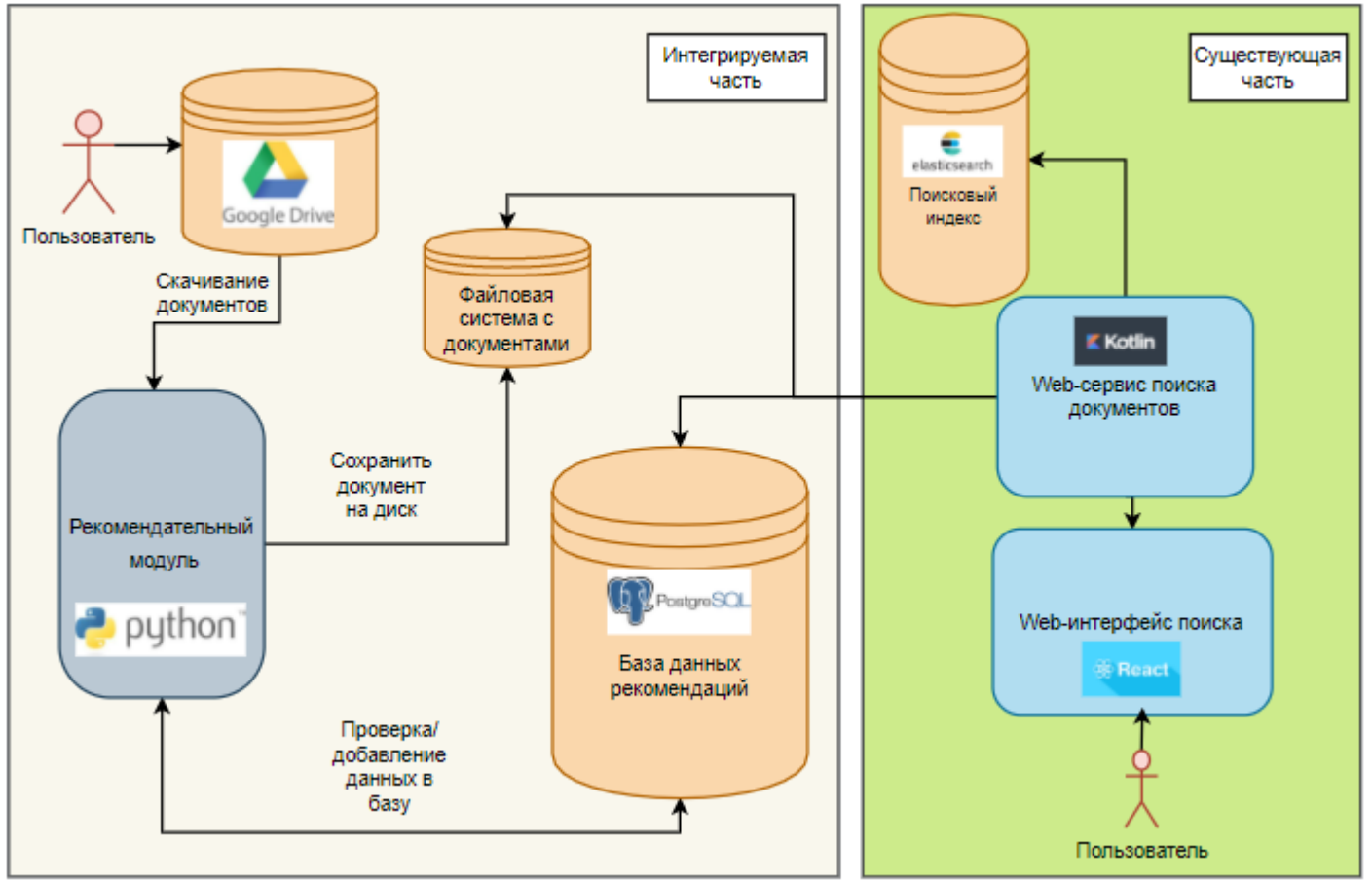
Сценарии использования



Аналогичный проект

- Проект Воронцова К.В. по тематическому поиску всей науки https://scisearch.ai
- Connected Papers – проект команды arxiv.org умного поиска по документу в коллекции, использующий работу с графами https://www.connectedpapers.com/about

Архитектура проекта



Поток данных

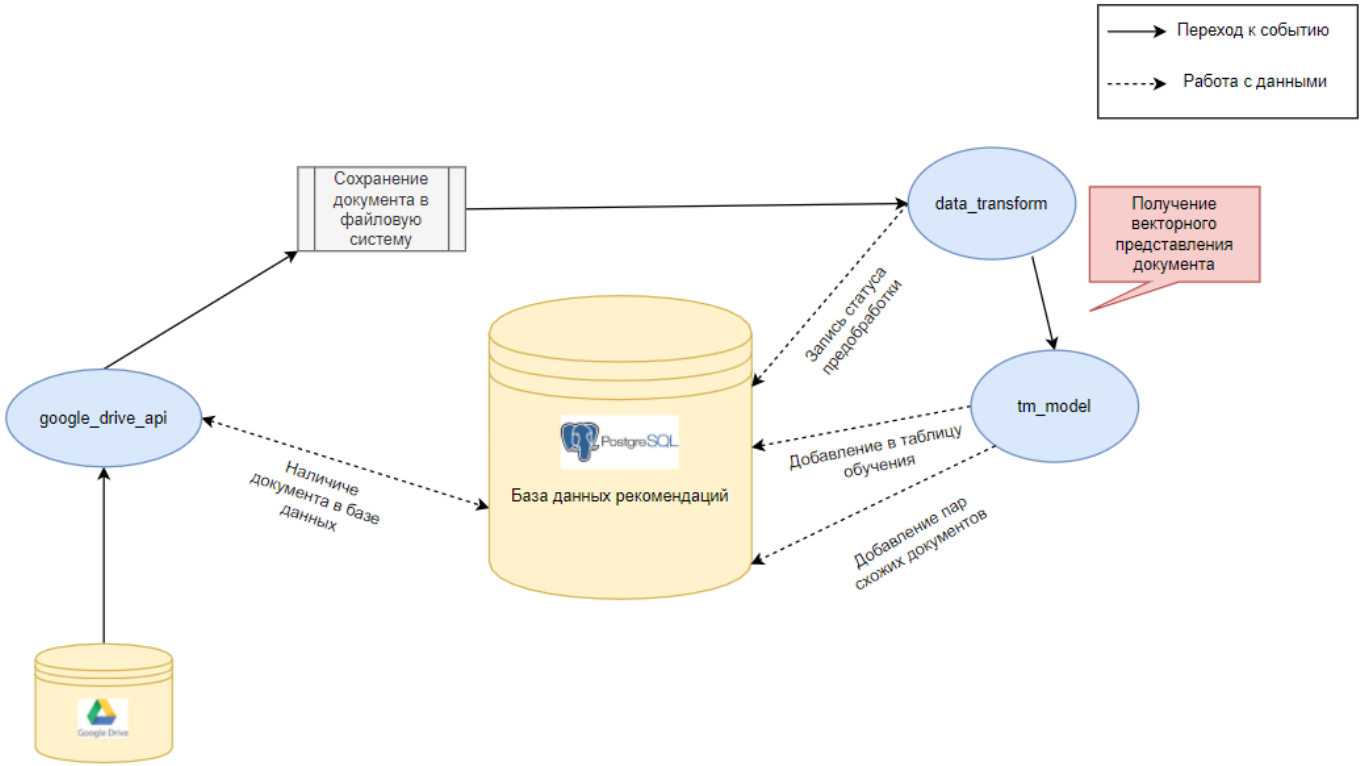
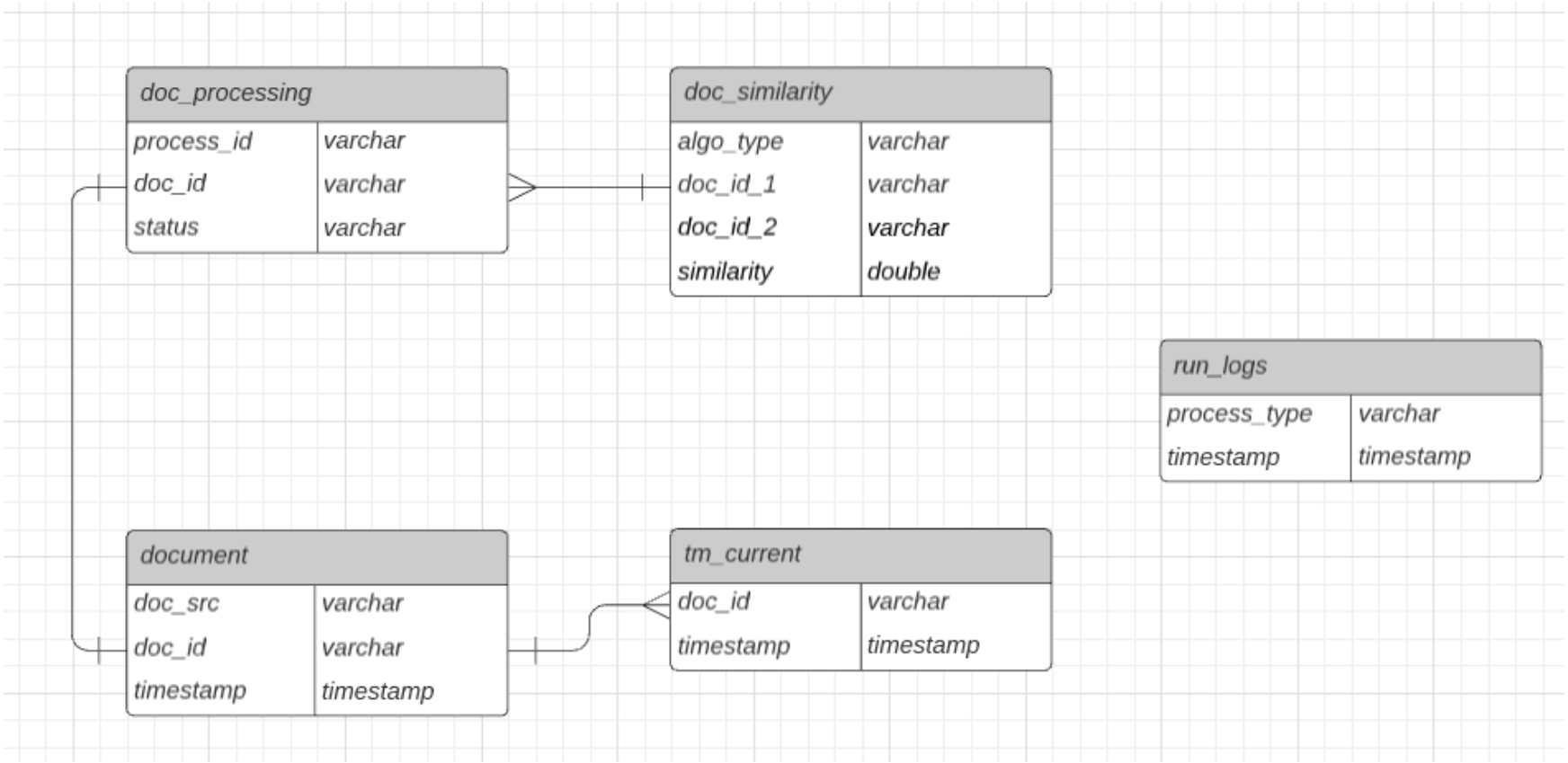


Схема базы данных



Результаты тематического моделирования для рекомендаций

Добавленные документы

Название	Владелец	Последнее изме...	Размер файла
ACB25_925.pdf	я	10 мая 2021 г.	2 МБ
condmat0506749v1.pdf	я	30 мар. 2021 г.	586 КБ
condmat0512535v1.pdf	я	30 мар. 2021 г.	414 КБ
Lattice+parameters+of+gallium+nitride.pdf	я	12 мая 2021 г.	55 КБ
pattsol9604003v1.pdf	я	7 мая 2021 г.	236 КБ
PhysRevB.66.115202.pdf	я	11 мая 2021 г.	257 КБ
qbio0701005v1.pdf	я	7 мая 2021 г.	268 КБ
quantph9803064v1.pdf	я	1 апр. 2021 г.	141 КБ

Близость документов

	algo_type character varying	doc_id_1 character varying	doc_id_2 character varying	similarity double precision
1	TM	Lattice+parameters+of...	151102305v1.pdf	0.600046790947024
2	TM	151102305v1.pdf	Lattice+parameters+of...	0.600046790947024
3	TM	Lattice+parameters+of...	condmat0507587v1.pdf	0.6007119810525342
4	TM	condmat0507587v1.pdf	Lattice+parameters+of...	0.6007119810525342
5	TM	Lattice+parameters+of...	07103088v1.pdf	0.6012732308911711
6	TM	07103088v1.pdf	Lattice+parameters+of...	0.6012732308911711
7	TM	Lattice+parameters+of...	condmat0601035v2.pdf	0.6014413961099255
8	TM	condmat0601035v2.pdf	Lattice+parameters+of...	0.6014413961099255
9	TM	Lattice+parameters+of...	07104542v2.pdf	0.601738375757237
10	TM	07104542v2.pdf	Lattice+parameters+of...	0.601738375757237

Результаты

- Собрана коллекция документов с помощью реализованного программного модуля
- Изучены и протестированы модели векторизации текстовых документов
- Реализован модуль для работы с моделями NLP
- Реализована база данных для сохранения документов и мониторинга процессов
- Реализован модуль по работе с Google Drive
- Реализованы стратегии и методы отказоустойчивости проекта
- Проводится интеграция с существующим программным комплексом
- По теме диссертации приняты к публикации тезисы в сборнике трудов Международной научно-технологической конференции студентов и молодых ученых «Молодежь. Инновации. Технологии»

Перспективы развития

- Реализация модулей в виде микросервисов
- Добавление возможности поиска по предложенным темам, полученных в результате обучения модели
- Добавление возможности удаления файлов из модели
- Интеграция других моделей векторизации текста