ЮУрГУ
1943
ВШЭКН

JINR
MLIT

ЛАБОРАТОРИЯ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ
имени М.Г. Мещерякова

НИЯУ МИФИ

Рекомендательная система подбора научных публикаций

Научные руководители:
доцент кафедры СП, к.т.н.,
М.В. Сухов
сотрудник ЛИТ ОИЯИ Антонов Е.В.

Автор:
студент ЮУрГУ (НИУ)
Д.А. Подрядова

Дубна – 2024

Актуальность

В настоящее время информации и исследований становится настолько много, что технологии все больше проникают в различные сферы жизни человека, начиная от автоматизации делопроизводства, заканчивая постановкой диагнозов рекомендательными системами.

В данной работе рассматривается прототип рекомендательной системы для определения близких по смыслу документов на основе научных интересов ученых. Подобного рода системы позволяют исследователям экономить время для отбора подходящих статей.

Цель и задачи

Цель:

разработать прототип рекомендательной системы подбора научных публикаций

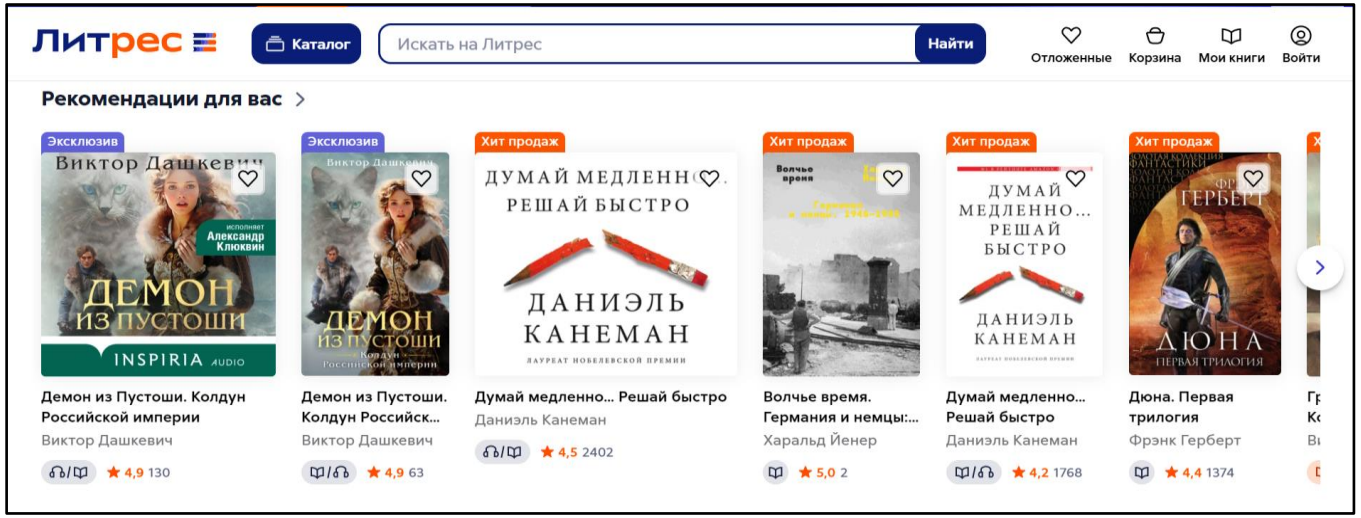
Задачи:

1. Изучить теоретическую часть методов обработки естественного языка.
2. Провести обзор аналогов рекомендательных систем.
3. Провести предобработку данных для обучения и тестирования.
4. Спроектировать и реализовать прототип рекомендательной системы подбора научных публикаций.
5. Провести тестирование реализованного прототипа.

Обзор аналогов

Используемые технологии:

- **Матричная факторизация** — это метод, который используется для определения двух (или более) матриц, так что при их умножении получается исходная матрица.



Рекомендательная система сервиса «Литрес»

- **Alternating least squares** (ALS) — модель, которая изучает бинарную цель взаимодействия каждого пользователя с каждым объектом, но взвешивает каждое бинарное взаимодействие по доверительному значению уверенности в этом взаимодействии пользователя/объекта.

Предобработка набора данных для обучения


```
{
    "title": "Study of the Effect of Pulsed-Periodic Electric
Properties of Liquid-Crystal Waveguide",
    "published": {
        "journal": "Physics of Wave Phenomena",
        "publisher": "Allerton Press",
        "volume": "26",
        "issue": "2",
        "pub_place": null,
        "page": {
            "start": "116",
            "end": "123"
        },
        "eISSN": "1934-807X",
        "ISSN": "1541-308X",
        "year": 2018,
        "month": 4,
        "day": null,
        "doi": "10.3103/s1541308x18020061",
        "submission_info": "Received January 19, 2018"
    },
    "authors": [
        {
            "name": "A. A Egorov",
            "email": "yegorov@kapella.gpi.ru",
            "ids": {},
            "aff_keys": [
                1
            ]
```

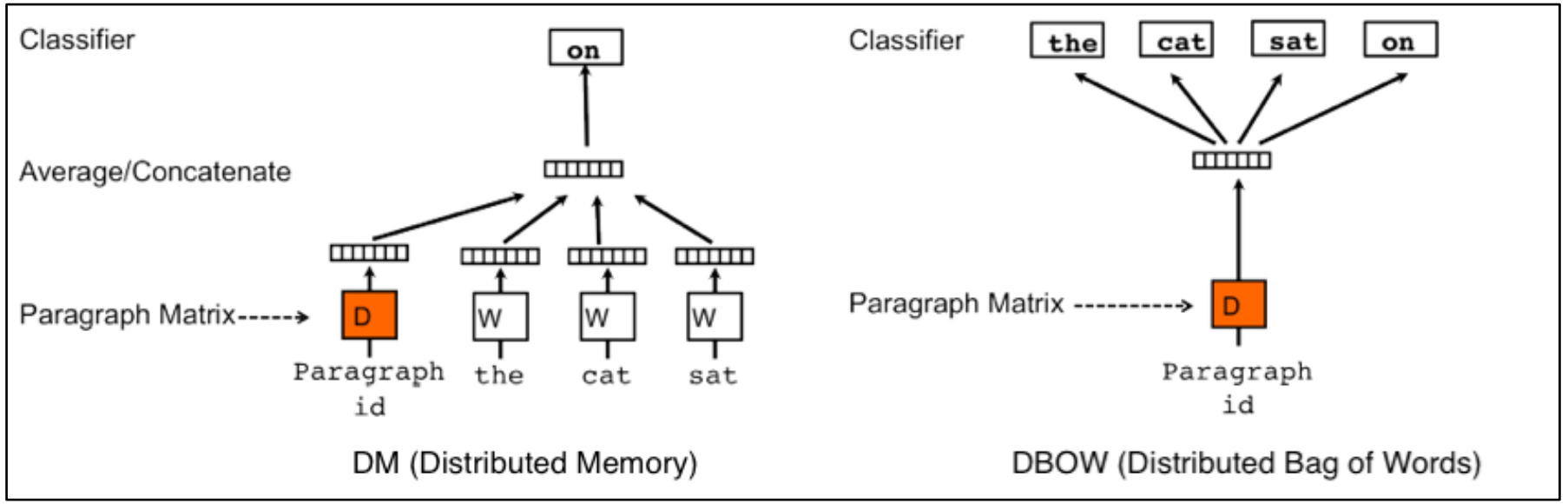

Пример «сырого» файла для набора данных

Обучающий набор	Тестовый набор
113 статей	225 статей

Этапы предобработки набора данных:

1. Преобразование JSON файла в список словарей.
2. Очистка словарей от лишних пунктов.
3. Преобразование глав в единый текст.
4. Преобразование списка словарей в удобную структуру – Датафрейм.
5. Удаление специальных и отдельно стоящих символов в тексте.
6. Конвертация слов к нижнему регистру.
7. Удаление всех слов длиной меньше 3
8. Лемматизация текста.

Используемая технология



Визуальное представление алгоритмов работы doc2vec

Doc2Vec — это подход на основе нейронных сетей, который изучает распределенное представление документов. Это метод обучения без учителя, который отображает каждый документ в вектор фиксированной длины в многомерном пространстве. Векторы изучаются таким образом, что аналогичные документы сопоставляются с близлежащими точками векторного пространства.

Средства реализации и характеристики

- **Обучение моделей:** Google Colaboratory
- **Характеристики:** Intel(R) Xeon(R) CPU @ 2.30GHz, ОЗУ 12 ГБ, GPU NVIDIA Tesla T4 24Гб, ОС Ubuntu 18.04
- **Язык программирования:** Python 3.11
- **Основные библиотеки:** NumPy 1.18.5, Pandas 1.1.4, Scipy, NLTK.tokenize, Gensim, Doc2Vec, SentenceTransformer

Тестирование модели

Айриян Александр Сержикович

№	Название статьи	Да/Нет	Оценка (от 0 до 10)	Близость doc2vec	Место в топе doc2vec	Близость tranformer all-MiniLM-L6-v2	Место в топе tranformer all-MiniLM-L6-v2
14	Distributed intelligence on the Edge-to-Cloud Continuum: A systematic literature review	да	8	0,7283	1	**0,419701**	**27**
29	Rosetta: A container-centric science platform for resource-intensive, interactive data analysis	да	8	0,7282	2	**0,338929**	**66**
21	Knowledge transfer at CERN	да	7	0,6805	11	**0,352984**	**59**
3	A neural system dynamics modeling platform and its applications in randomized controlled trial data analysis	да	6	**0,3765**	**192**	**0,118027**	**196**
22	Long term stability studies in the presence of crab cavities and high order multipoles in the CERN super proton synchrotron and high luminosity large hadron collider	да	6	0,7163	4	**0,172082**	**162**
23	Matter-antimatter gigaelectron volt gamma ray laser rocket propulsion	да	6	0,6653	15	**0,139775**	**183**
28	Renewable and Sustainable Energy Reviews	да	6	0,6564	23	**0,424388**	**23**
8	Cosmological inflation and meta-empirical theory assessment	нет	5	**0,7029**	**6**	**0,271726**	**103**
9	Data portability effects on data-driven innovation of online platforms: Analyzing Spotify	нет	5	0,5181	105	**0,31006**	**80**
13	Distributed denial of service attack prediction: Challenges, open issues and opportunities	нет	5	0,6684	14	**0,282905**	**96**
19	Future Generation Computer Systems	нет	5	0,6666	15	0,5000783	8
26	Nonlinear Analysis: Real World Applications	нет	5	**0,7267**	**3**	**0,136723**	**184**
31	Towards Green Big Data at CERN	нет	5	**0,6917**	**9**	0,3752077	46
1	3D point cloud data processing with machine learning for construction and infrastructure applications: A comprehensive review	нет	4	**0,6953**	**7**	0,4582466	15
18	Fusion Engineering and Design	нет	4	**0,644**	**28**	**0,275645**	**99**
20	IoT data analytic algorithms on edge-cloud infrastructure: A review	нет	4	**0,6457**	**27**	0,3741151	49
27	Point cloud generation for critical transportation infrastructure through Bézier curve	нет	4	**0,6775**	**13**	0,3274606	71
12	Deterministic simulation of the static neutronic characteristics for the lead core of VENUS-II facility	нет	3	**0,6903**	**11**	0,3579992	57

Обучающая выборка	
Айриян А.С.	23 статьи
Зрелов П.В.	43 статьи
Петросян А.Ш.	47 статей
Тестовая выборка	
225 статей	
Кол-во статей на оценку	
Айриян А.С.	32
Зрелов П.В.	32
Петросян А.Ш.	45

$$cos(\theta) = \frac{A \cdot B}{|A| \cdot |B|}.$$

Основные результаты

1. Изучена теоретическая часть методов обработки естественного языка.
2. Проведен обзор аналогов рекомендательных систем.
3. Проведена предобработка данных для обучения и тестирования.
4. Спроектирован и реализован прототип рекомендательной системы подбора научных публикаций.
5. Проведено тестирование реализованного прототипа.

Дальнейшее направление разработки

- Увеличение количества экспертов для оценки системы
- Повторное обучение модели с учетом экспертных оценок
- Изучение влияния колаборационных статей на общую оценку
- Корректировка пороговых значений по итогам расширенной экспертизы
- Реализация в качестве программного компонента, оценки возможности внедрения в рамках проекта «Цифровой ОИЯИ»

Тестирование модели на данных Зрелова Петра Валентиновича

№	Название статьи	DOI	Да/Нет	Оценка (от 0 до 10)	**Близость doc2vec**	**Место в топе doc2vec**	**Близость tranformer all-MiniLM-L6-v2**	**Место в топе tranformer all-MiniLM-L6-v2**
10	Distributed intelligence on	https://do	да	10	0,6563	21	0,631584704	7
15	IoT data analytic algorithm	https://do	да	10	0,6743	14	0,626565576	8
17	Long term stability studies	https://do	да	10	0,7315	1	**0,375577778**	**71**
18	Microprocessors and Micr	https://do	да	10	0,708	5	**0,498110354**	**30**
19	ML interpretability: Simple	https://do	да	10	0,6588	20	**0,354659975**	**84**
22	Neutron flux monitoring sy	https://do	да	10	0,6476	25	**0,399957806**	**60**
4	An Empirical Evaluation of	https://do	да	8	**0,2529**	**220**	**0,589244127**	**12**
26	PyApprox: A software pack	https://do	да	8	0,6702	16	**0,367841899**	**74**
27	Quantifying the value of us	https://do	да	8	**0,4424**	**161**	**0,396601915**	**62**
7	Blockchain model for envir	https://do	да	6	0,6906	9	**0,383065611**	**68**
24	Potentials of ionic liquids t	https://do	да	6	0,6783	12	**0,095771313**	**207**
25	Prenatal and childhood lea	https://do	да	6	0,6961	8	**0,02537067**	**218**
28	Renewable and Sustainable	https://do	да	6	0,6513	23	0,452118099	46
30	Sensor fault analysis of aer	https://do	да	6	0,7007	6	0,404205322	58
31	Supply chain hybrid simula	https://do	да	6	**0,3705**	**198**	0,513175905	23
2	A deep learning technique	https://do	да	5	0,5516	87	**0,342463613**	**88**
14	Heterodox underdetermin	https://do	нет	5	**0,6987**	**7**	**0,154450938**	**172**
23	Point cloud generation for	https://do	нет	5	**0,7191**	**4**	**0,300370395**	**104**
16	Journal Pre-proofs Ultrathi	https://do	нет	4	0,5098	125	0,277999878	115
3	Alloy core composition eff	https://do	нет	3	**0,631**	**31**	0,262162656	126
6	Biological functions are ca	https://do	нет	3	**0,5807**	**64**	0,193630472	153
29	Scaling procedures in clima	https://do	нет	3	**0,6624**	**18**	**0,126184165**	**189**
32	The dominant genera of ni	https://do	нет	3	**0,478**	**146**	0,268072397	119
12	European Journal of Opera	https://do	нет	2	**0,6876**	**10**	**0,435206771**	**49**
21	Navigation and star identif	https://do	нет	2	0,6772	13	0,164617151	163
1	3D point cloud data proces	https://do	нет	0	0,6851	11	0,416895688	54

Тестирование модели на данных Петросяна Артёма Шмавоновича

№	Названи е статьи	DOI	Да/Нет	Оценка (от 0 до 10)	Близость doc2vec	Место в топе doc2vec	Близость tranformer all-MiniLM-L6-v2	Место в топе tranformer all-MiniLM-L6-v2
41	Rosetta: A	https://do	да	7	0,8002	13	**0,415800542**	**6**
45	Towards C	https://do	да	3	**0,7668**	**25**	0,368207842	12
13	Continuou	https://do	да	2	**0,7726**	**24**	0,130279988	109
18	Distribute	https://do	да	2	**0,8207**	**11**	0,123065293	117
19	Distribute	https://do	да	2	**0,8075**	**16**	0,140154123	103
42	Strategies	https://do	да	2	**0,7523**	**37**	0,080565646	151
1	A break in	https://do	да	1	**0,7848**	**23**	0,021516021	203
4	Accountab	https://do	да	1	**0,7426**	**46**	0,117580183	121
10	Blockchair	https://do	да	1	**0,7754**	**28**	0,134696871	105
5	Active lear	https://do	да	0	0,7489	43	0,184560597	61
9	Bacterial c	https://do	да	0	**0,3762**	**209**	**0,029102355**	**198**
11	Cement ar	https://do	да	0	**0,4786**	**205**	**0,130148828**	**110**
27	Future Ge	https://do	да	0	**0,7795**	**31**	0,253691286	35
15	CRSExtrac	https://do	нет	2	**0,7267**	**66**	0,142229095	102
16	Data & Kn	https://do	нет	2	**0,5714**	**166**	0,21044457	49
17	Data conv	https://do	нет	2	**0,7536**	**46**	0,114289463	123
26	Federatior	https://do	нет	2	**0,7629**	**43**	0,145443708	97
36	Neural net	https://do	нет	2	**0,7345**	**66**	0,096616909	137
25	Facilitating	https://do	нет	1	**0,7417**	**59**	0,171973795	71
44	Text minin	https://do	нет	1	**0,7239**	**74**	0,055168103	185
2	A holistic f	https://do	нет	0	**0,6735**	**111**	0,184064269	62
3	A prospec	https://do	нет	0	**0,7324**	**72**	0,085899107	145
6	Analysis o	https://do	нет	0	**0,6141**	**151**	0,122716211	118
7	Animism a	https://do	нет	0	**0,7012**	**96**	0,030834688	197
8	Approachi	https://do	нет	0	**0,6708**	**118**	0,075845338	159