



Цель: реализация программного средства для рекомендаций научных статей

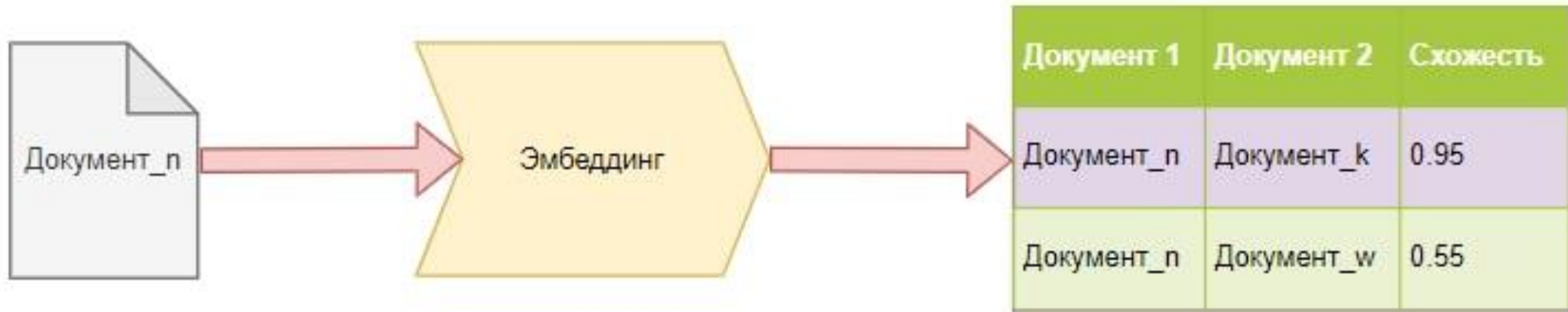
Задачи:

- 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

Поиск по документу



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 ...	['Loic Halbert', 'Malgorzata 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/condmat/0507292v1	186070



Обучение моделей

1. Разметка датасета на основе вхождений химических формул (Al, He, O2) и их синонимов
2. Удаление вхождений в документ вышеперечисленных формул
3. Обучение моделей NLP (18 моделей)
4. С помощью полученных векторных представлений решить задачу классификации с помощью логистической регрессии



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

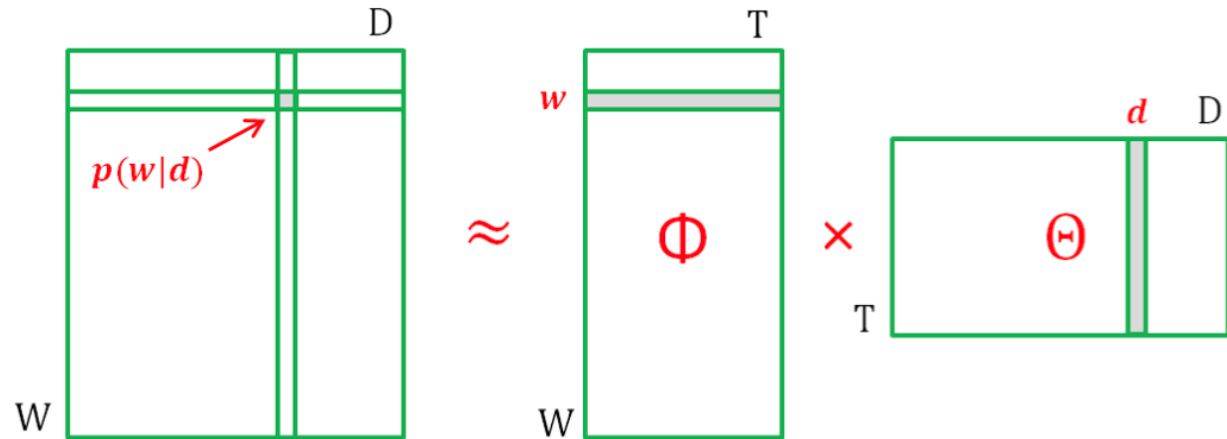
$p(t|d)$ – вероятность, с которой тема t встречается в документе d

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

$$\Phi = p(w|t)_{W \times T},$$

$$\Theta = p(t|d)_{T \times D},$$

$$F_{W \times D} \approx \Phi_{W \times T} \times \Theta_{T \times D}$$





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

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

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



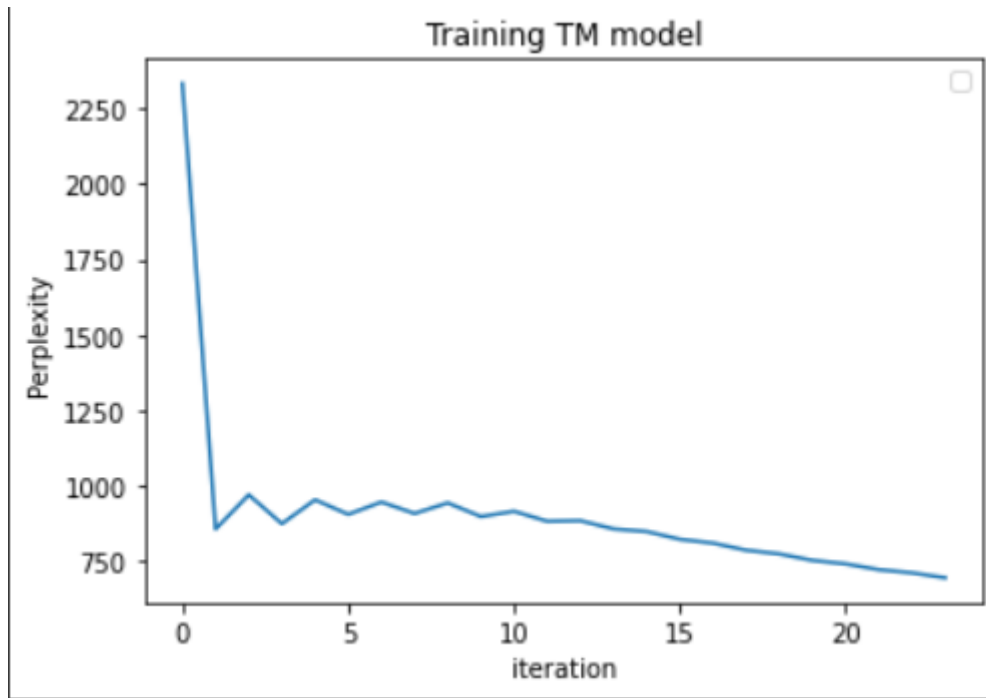
EM-алгоритм

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

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



Результаты моделирования

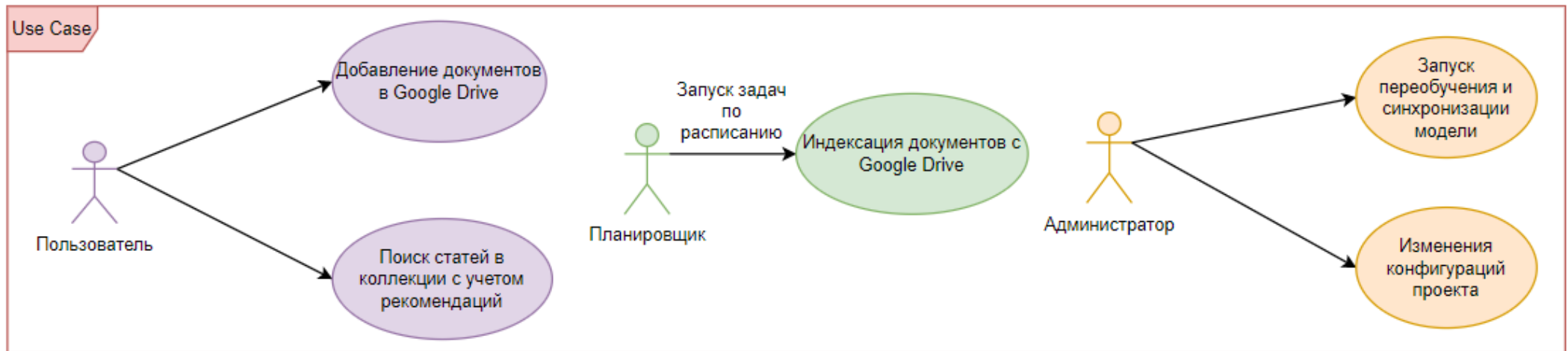


Ход обучения:

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



Сценарии использования





Аналогичные проекты

- 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

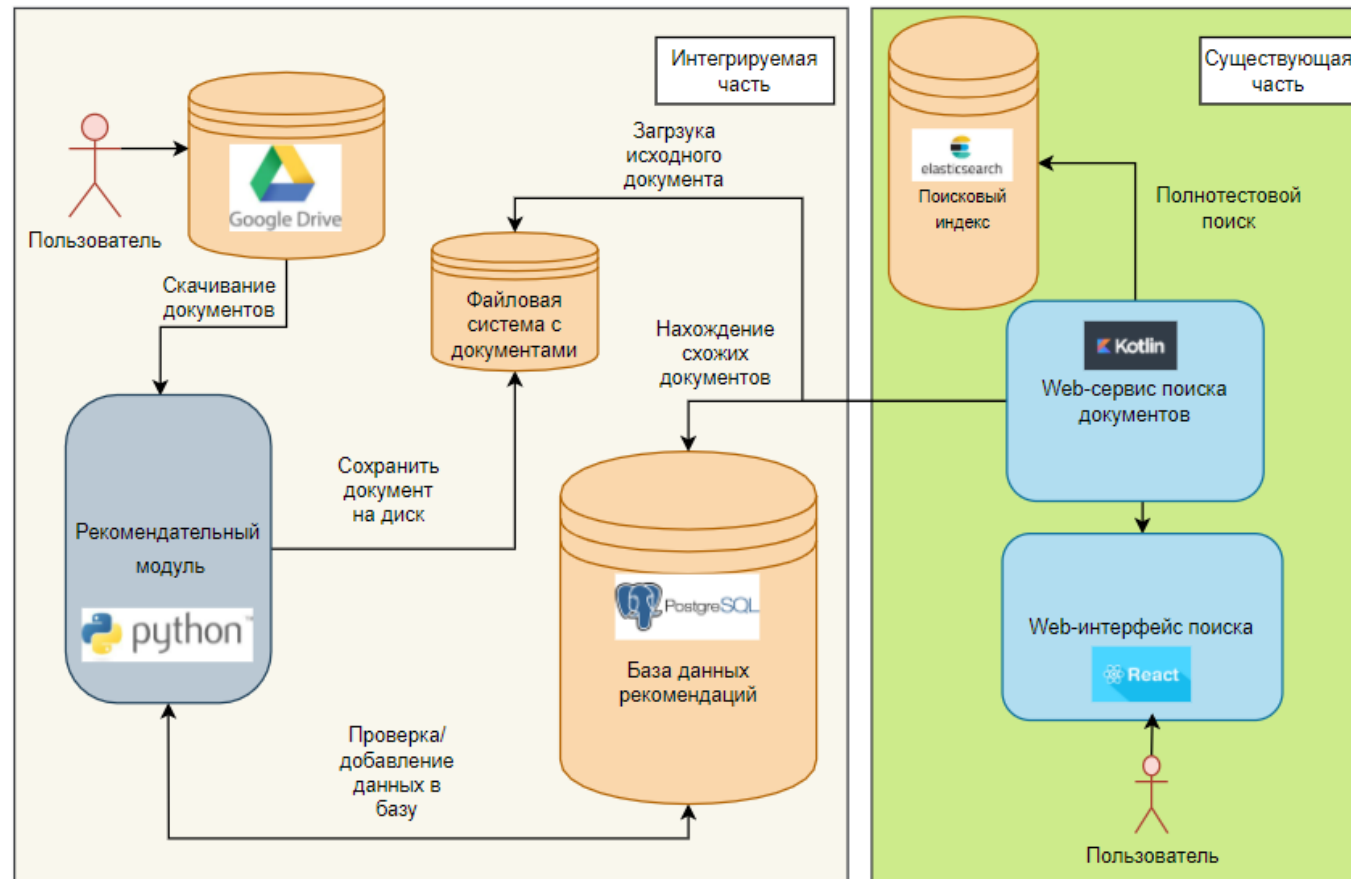
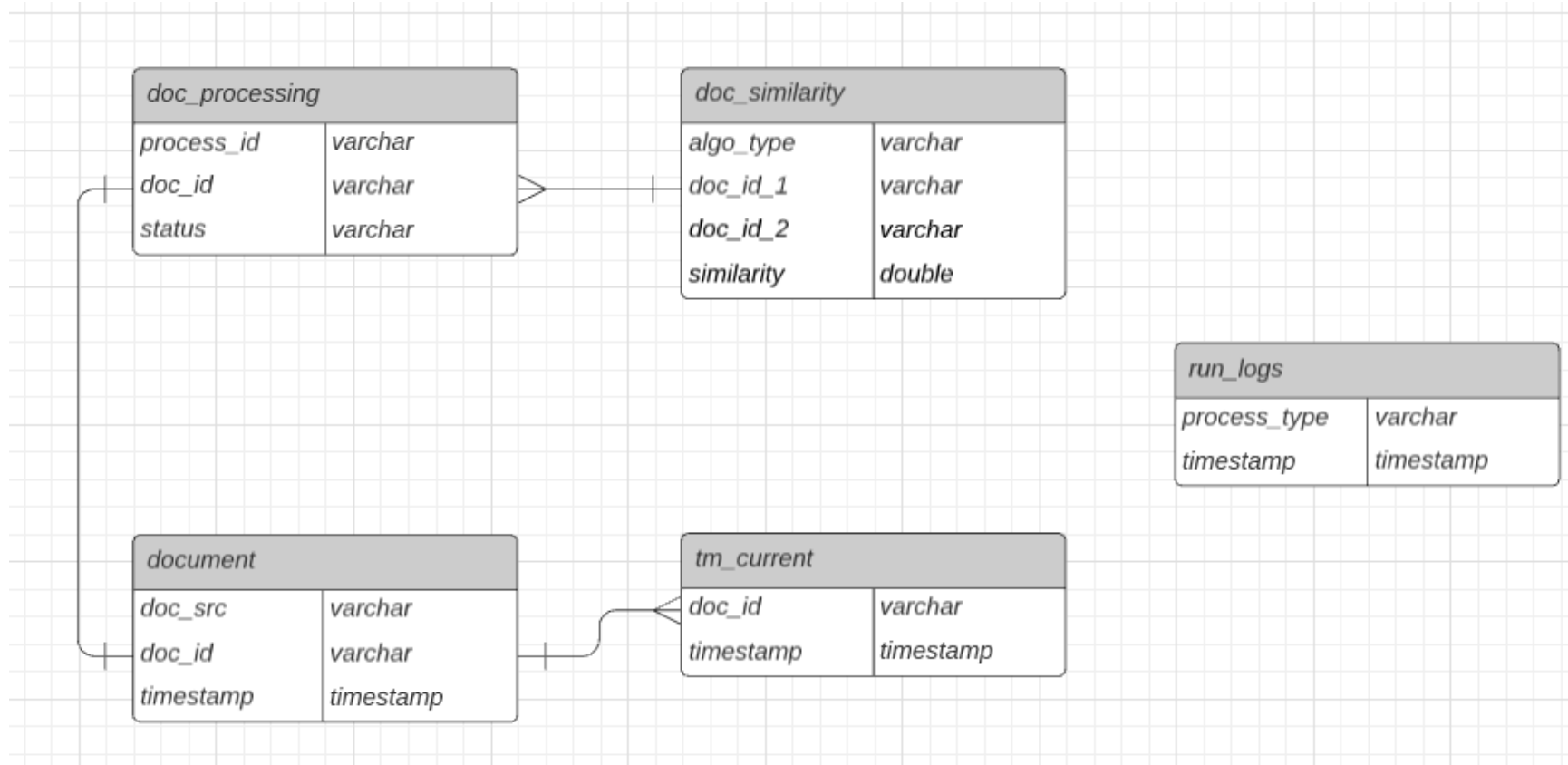




Схема базы данных





Результаты тематического моделирования для рекомендаций

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

Добавленные документы

	algo_type	doc_id_1	doc_id_2	similarity
	character varying	character varying	character varying	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

Близость документов



Отображение рекомендаций

Arxiv 2020

Химические элементы Химическая формула Кристаллическая система Тип атомного радиуса

Результатов: 190

[200311355v1.pdf](#)

Дата: Thu Nov 19 2020

arXiv:2003.11355v1 [physics.gen-ph] 19 Mar 2020 (2+1) dimensional **cosmological** models in $f(R, T)$ gravity with $\Lambda(R, T)$ Safiqul Islam a, Praveen Kumar b, G.S. Khadekar c and Tapas K Das d a,dHarish-Chandra research Institute, HBNI, Chhatnag Road, Jhansi, Allahabad-211019, India b,cDepartment of Mathematics, Rashtrasant Tukadoji Maharaj Nagpur University,

Nagpur 440033, India asafiqulislam@hri.res.in bpkumar6743@gmail.com cgkhadekar@yahoo.com dtapas@hri.res.in Abstract. We intend to study a new class of **cosmological** models in $f(R, T)$ modified theories of gravity, hence define the **cosmological** constant Λ as a function of the trace of the stress energy-momentum-tensor T and the Ricci scalar R , and name such a model $\Lambda(R, T)$ gravity where we have

specified a certain form of $\Lambda(R, T)$. $\Lambda(R, T)$ is also defined in the perfect fluid and dust case. Some physical and geometric properties of the model are also discussed. The pressure, density and energy conditions are studied both when Λ is a positive constant and when $\Lambda = \Lambda(t)$, i.e a function of **cosmological** time, t . We study behavior of some **cosmological** quantities such as Hubble

Похожие документы:

[hep-th/0407078v2.pdf](#)

[200601822v1.pdf](#)



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»