DSCS.pro

Санкт-Петербургский Федеральный исследовательский центр Российской академии наук

Санкт-Петербургский государственный университет

2025 год

Интеллектуальная система агрегации научной литературы

Олисеенко Валерий Дмитриевич

— Научный сотрудник лПИИ СПб ФИЦ РАН
— Исполнительный директор по исследованию данных в Сбере
— Старший преподаватель кафедры информатики СПбГУ

Вяхирев Иван Олегович

Куратор-студент, программа “Инженерия Искусственного Интеллекта” ИТМО

Плетка Даниил Ильич

Студент 4-го курса программы ПИИТ

Описание проекта

С ростом числа публикаций на arXiv исследователям становится все труднее отслеживать ключевые работы в своей области!

Основная проблема: Стандартный поиск по ключевым словам часто неэффективен для сложных и узкоспециализированных запросов.

Репозиторий

Leveraging LLMs for Persona-Based Visualization of Election Data

Swaroop Panda and Arun Kumar Sekar

Northumbria University

ABSTRACT

Visualizations are essential tools for disseminating information regarding elections and their outcomes, potentially influencing public perceptions. Personas, delineating distinctive segments within the populace, furnish a valuable framework for comprehending the nuanced perspectives, requisites, and behaviors of diverse voter demographics. In this work, we propose making visualizations tailored to these personas to make election information easier to understand and more relevant. Using data from UK parliamentary elections and new developments in Large Language Models (LLMs), we create personas that encompass the diverse demographics, technological preferences, voting tendencies, and information consumption patterns observed among voters.Subsequently, we elucidate how these personas can inform the design of visualizations through specific design criteria. We then provide illustrative examples of visualization prototypes based on these criteria and evaluate these prototypes using these personas and LLMs. We finally propose some actionable insights based upon the framework and the different design artifacts.

KEYWORDS

Election Data Visualization, User-Centered Visualization, User Personas, LLMs, Prototypes

1. Introduction

Electoral data and visualizations are prominently featured across various media publications [1, 2]. Political analysts examine with great scrutiny every poll, trend, and demographic shift in an effort to interpret the electorate's sentiment and forecast the electoral outcome. Media professionals construct narratives around these quantitative indicators, developing content that appeals to their readership and potentially influences public discourse [3]. Furthermore, data scientists and visualization designers extensively engage in transforming raw statistical information into sophisticated graphics and interactive presentations. These visualizations range from choropleth maps that illuminate regional voting patterns to dynamic charts depicting campaign expenditure over time, thereby providing a comprehensive framework through which voters may engage with the electoral process [4]. Social media platforms serve as significant conduits for the dissemination of these visualizations, facilitating widespread discussion. Visual content including infographics and data-driven publications permeate digital spaces, influencing online discourse and shaping public perception regarding candidates and pertinent issues [5]. In the contemporary digital environment, characterized by information abundance and diminished attention capacity, the significance of electoral data visualization in molding public opinion is substantial and cannot not be

Author Email: swaroop.panda@northumbria.ac.uk

Цель

Создать интеллектуальную систему, которая по запросу пользователя будет находить, анализировать и структурировать релевантные научные статьи.

Задачи

- Сбор данных
- Ранжирование
- Корректно использование числа цитирований при сортировке
- Суммаризация информации по статье

Основные технологии

Язык программирования:

PYTHON

Библиотеки:

ARXIV

GOOGLE-GENAI

SENTENCETRANSFORMER

PANDAS

Важная информация с arXiv

- Entry_id
- Published
- Title
- Summary
- Pdf_url

ПЕРИОД

Один год

Выбор модели

HUGGING FACE

ALLENAI-SPECTER

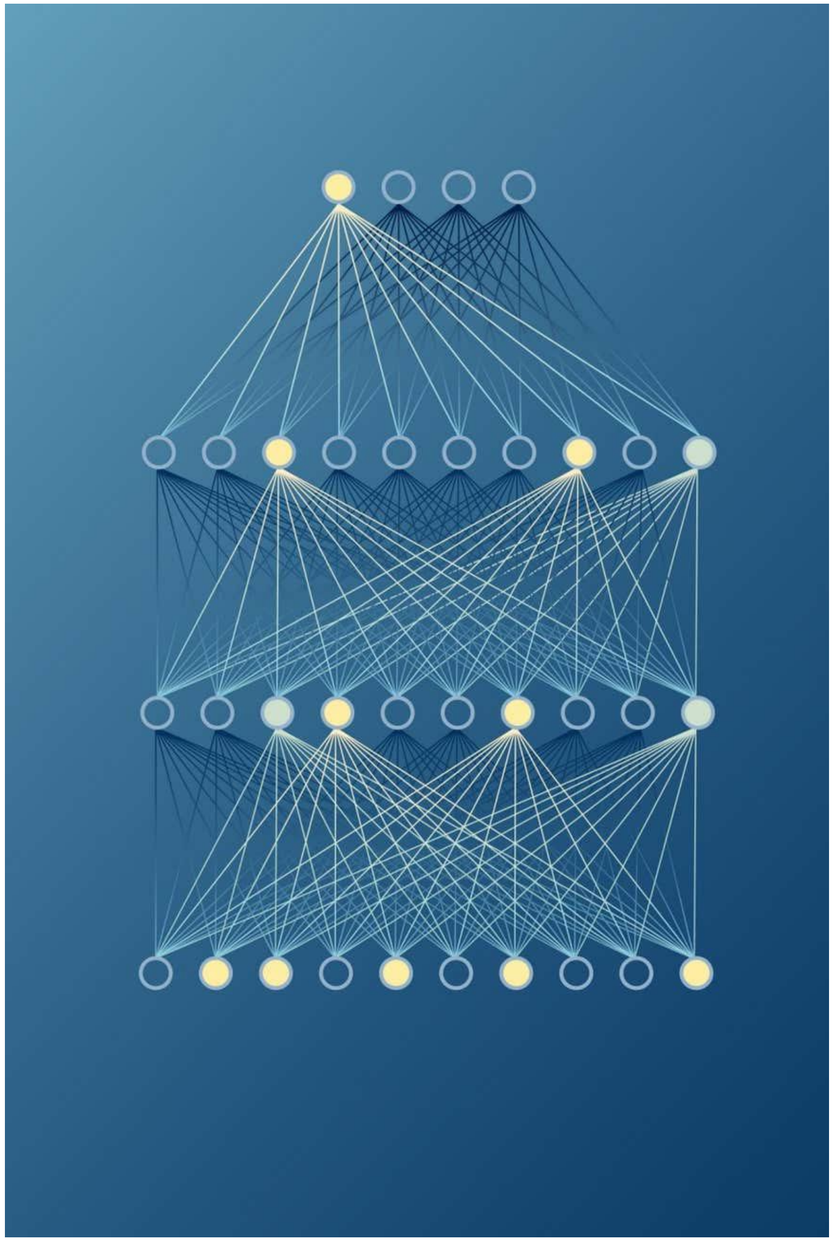
Специализированная модель для научных статей. Достаточно быстро работает.

ALL-MINILM-L6-V2

Маленькая и быстрая. Достигается хороший баланс между качеством и производительностью.

ALL-MPNET-BASE-V2

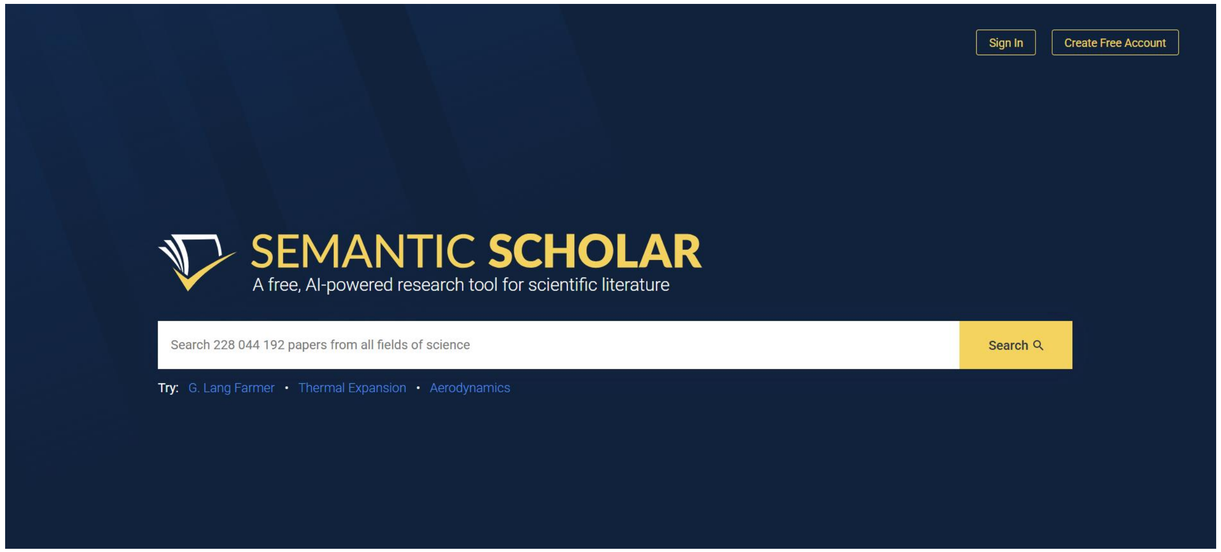
По сравнению с прошлыми моделями более точная, но работает медленнее из-за большего размера и вычислительной нагрузки.

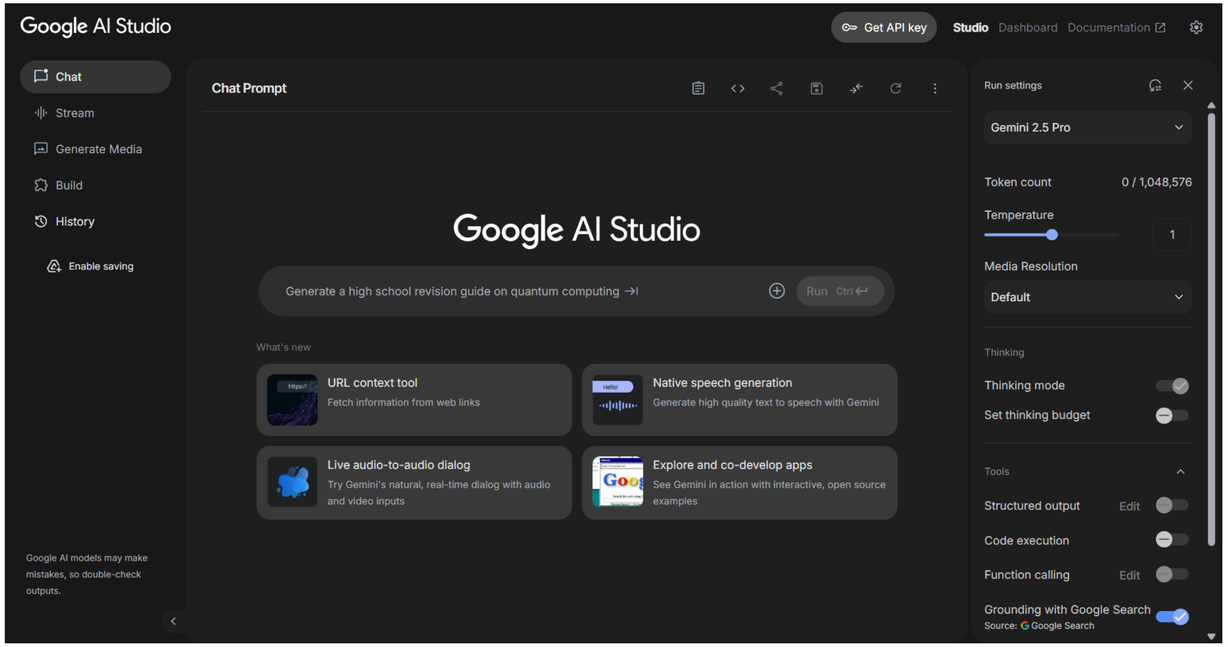


Дополнительные необходимые ресурсы

ПОЛУЧЕНИЕ ЧИСЛА ЦИТИРОВАНИЙ СТАТЬИ

- На arXiv нет информации о цитировании
- Решение – использование ресурса **Semantic Scholar**





ФОРМИРОВАНИЕ СОБСТВЕННОЙ SUMMARY

- Решение – использование ресурса **Google AI Studio**.

Как учитывать число цитирований?

РАСЧЕТНЫЕ ФОРМУЛЫ

$$citations = \frac{\text{citation\_num}}{\text{current\_year} - \text{published\_year} + 1}$$

$$\alpha \cdot (\text{similarity}) + \beta \cdot \log(citations)$$

Предсказание числа цитирований

Идея:

Научиться предсказывать, будет ли статья цитироваться в будущем

[PDF] Role of **AI** in education.

A Harry - Interdiciplinary Journal & Hummanity (INJURITY), 2023 - radensa.ru

… **Artificial intelligence** (**AI**) has the potential to revolutionize the … **AI** in education refers to the use of **artificial intelligence** … **AI** has the potential to revolutionize the way we learn and teach, …

☆ Сохранить Цитировать Цитируется: 352 Похожие статьи Все версии статьи (2) »

Nanotechnological applications in **medicine**

SD Caruthers, SA Wickline, GM Lanza - Current opinion in Biotechnology, 2007 - Elsevier

Nanotechnology-based tools and techniques are rapidly emerging in the fields of medical imaging and targeted drug delivery. Employing constructs such as dendrimers, liposomes, …

☆ Сохранить Цитировать Цитируется: 686 Похожие статьи Все версии статьи (7)

Основные метрики:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (Predicted_i - Actual_i)^2}{N}}$$

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \overline{y})^2}$$


```
{
"id":int1091
"authors":[...]8 items
"title":string"Preliminary Design of a Network Protocol Learning Tool Based on the Comprehension of High School Students: Design by an Empirical Study Using a Simple Mind Map"
"year":int2013
"n_citation":int1
"page_start":string"89"
"page_end":string"93"
"doc_type":string"Conference"
"publisher":string"Springer, Berlin, Heidelberg"
"volume":string""
"issue":string""
"doi":string"10.1007/978-3-642-39476-8_19"
"references":[...]2 items
"indexed_abstract":{...}2 items
"fos":[...]8 items
"venue":{...}3 items
}
```


Датасет

Демонстрация

https://drive.google.com/file/d/1srL0j5rcd64BlnGDZiW8x7Z8PZrvMs-H/view?usp=drive_link

РЕЗУЛЬТАТЫ

Осуществляется корректный сбор статей с arXiv по запросу пользователя

При ранжировании учитываются не только ключевые слова, но и число цитирований, которое в большой мере влияет на сортировку результата

Предоставляется альтернативная сводка summary на русском языке по тексту статьи

DSCS.pro

Санкт-Петербургский Федеральный исследовательский центр Российской академии наук

Санкт-Петербургский государственный университет

2025 год

Интеллектуальная система агрегации научной литературы

Олисеенко Валерий Дмитриевич

— Научный сотрудник лПИИ СПб ФИЦ РАН
— Исполнительный директор по исследованию данных в Сбере
— Старший преподаватель кафедры информатики СПбГУ

Вяхирев Иван Олегович

Куратор-студент, программа “Инженерия Искусственного Интеллекта” ИТМО

Плетка Даниил Ильич

Студент 4-го курса программы ПИИТ