

# Проект “Анализ цитируемости статей”

Выполнил: Лысенко Иван Егорович ММОВС23  
Куратор: Кофанова Мария

Высшая Школа Экономики

## Описание задачи

- **Цель работы:** предсказание текущих трендов в современных статьях по машинному обучению.
- **Задачи:**
  - Сбор данных и EDA
  - Анализ литературы по графовой кластеризации и NLP методам для кластеризации
  - Анализ графа цитирований с помощью простых методов графовой кластеризации (например, спектральной кластеризации)
  - Topic modeling
  - Кластеризация статей по контенту, например, используя эмбединги BERT [2]
  - Использование моделей, учитывающих как структуру графа, так и контент вершин (например, attributed graph clustering [6])
  - Анализ полученных результатов

- **Исходная выборка:** arXiv dataset [5]
- **Преобразования над исходной выборкой:**
  - Выбирались только статьи 2023 года
  - Выбирались только статьи с arXiv тегами “cs.CV”, “cs.LG”, “cs.CL”, “cs.AI”, “cs.NE” и “stat.ML”
  - С помощью Semantic Search [3] API для каждой статьи загружалась дополнительная метаинформация, в частности ее ссылки
  - Для каждой ссылки также в список статей добавлялись ее ссылки с помощью информации из Semantic Search [3] API
- **Итоговая выборка:** граф цитирований, где каждая вершина – статья (название, абстракт, год, Semantic Scholar ID), а ребра – связи по цитированию.

- Были невалидные вершины (например, с неверным Semantic Scholar ID). Они были удалены.
- 411988 вершин, из которых примерно 35000 – изначальные статьи 2023 года с arXiv
- 271 компонента связности, одна из которых размера 411706. Далее работа шла только с ней.
- Плотность большой компоненты связности  $0.00011196709822370937 \ll 1$  – граф сильно разреженный

## Эксперимент 1: спектральная кластеризация

- 30 собственных векторов Лапласиана большой компоненты связности, соответствующие 30 наименьшим собственным значениям
- k-means кластеризация с 8 кластерами. Размеры кластеров:

[654, 1, 1901, 3614, 1213, 3498, 1119, 23220]

- Облака слов по упрощенным с помощью nltk [1] текстам вида title + abstract, а также ручная проверка нескольких статей из каждого кластера не дают выделить темы кластеров

**Итог:** информации о связи вершин в графе недостаточно для адекватной кластеризации на темы.

## Эксперимент 1: спектральная кластеризация

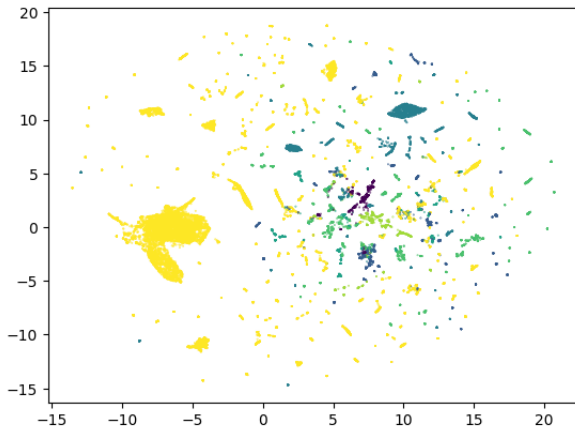


Рис.: UMAP [4] визуализация кластеров

## Эксперимент 2: topic modeling, LDA

- Тексты вида title + abstract были упрощены с помощью nltk [1]
- С помощью gensim создавался словарь токенов, встречающихся не менее, чем в 20 документах, но не более, чем в 50% документов
- Из словаря были убраны частые токены “algorithm”, “learning”, “training”, “data”
- На текстах с помощью словаря из 6715 токенов была обучена модель LDA с `n_topics=10`
- Итоговое распределение размеров тем:

[3367, 4206, 4100, 2094, 1932, 2182, 2830, 6291, 4969, 3492]

**Итог:** не все темы интерпретируемые, но есть довольно понятные.

## Эксперимент 2: topic modeling, HDP

- Препроцессинг такой же, как и у LDA
- Выбирать число тем вручную не нужно, большой плюс по сравнению с LDA
- Coherence scores гораздо ниже, чем у LDA

**Итог:** темы вышли очень плохие и непонятные.



`https://t.me/mlds_paper_analysis_bot`

- `/start` - старт бота
- `/help` - сообщение с описанием команд
- `/get_random_papers` - получить случайные статьи из кластера
- `/get_paper_stats` - получить статистику о статьях и кластерах
- `/rate_bot` - оценить бота
- `/get_avg_rating` - получить среднюю оценку бота

## Бот Telegram: /get\_paper\_stats

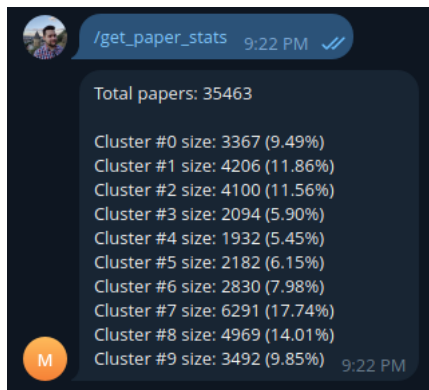


Рис.: Вывод команды `/get_paper_stats`

## Бот Telegram: /get\_random\_papers

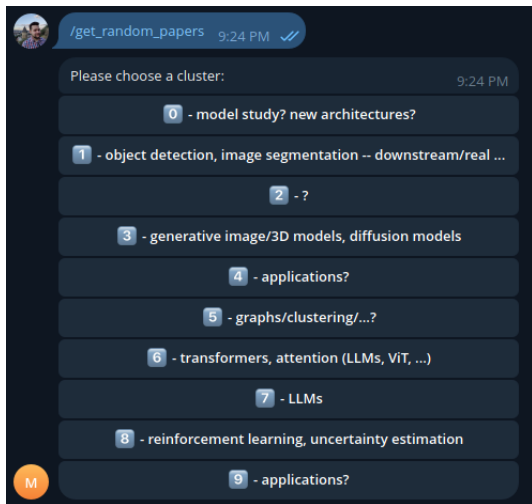


Рис.: Вывод команды `/get_random_papers`

# Бот Telegram: /get\_random\_papers

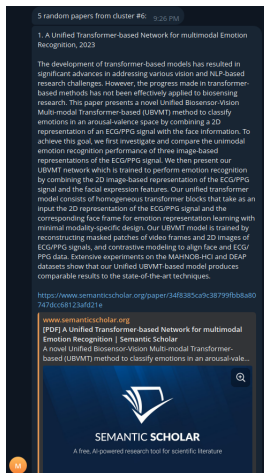


Рис.: Вывод команды /get\_random\_papers

## References I

- [1] Edward Loper Bird Steven. *Natural Language Processing with Python*. 2009.
- [2] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. [arXiv: 1810.04805 \[cs.CL\]](#).
- [3] Rodney Michael Kinney et al. "The Semantic Scholar Open Data Platform". In: *ArXiv abs/2301.10140* (2023). URL: <https://api.semanticscholar.org/CorpusID:256194545>.
- [4] Leland McInnes, John Healy, and James Melville. *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*. 2020. [arXiv: 1802.03426 \[stat.ML\]](#).

## References II

- [5] arXiv.org submitters. *arXiv Dataset*. 2024. DOI: 10.34740/KAGGLE/DSV/7352739. URL: <https://www.kaggle.com/dsv/7352739>.
- [6] Chun Wang et al. *Attributed Graph Clustering: A Deep Attentional Embedding Approach*. 2019. arXiv: 1906.06532 [cs.LG].