Московский государственный университет имени М.В.Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра суперкомпьютеров и квантовой информатики

Акопян Микаэла Тиграновна

Исследование возможностей предиктивного анализа публикационной активности

Отчет о научно-исследовательской работе

Группа 323
Научный руководитель:
к.т.н.
Григорьева Мария Александровна

Москва, 2024

Содержание

1 Введение

Физика высоких энергий (ФВЭ) – это область физики, изучающая фундаментальные взаимодействия и структуры материи при экстремально высоких энергиях. Она является одной из наиболее динамично развивающихся областей современной науки, требующей анализа огромных объемов информации. Конференции играют ключевую роль в научной коммуникации, объединяя исследователей для представления и обсуждения результатов. Важным аспектом таких мероприятий являются публикации, основанные на представленных материалах, которые отражают актуальность тем и уровень интереса научного сообщества.

InspireHEP — это цифровая библиотека в области ФВЭ с открытым доступом. С появлением подобных платформ стало возможным централизованное хранение и доступ к данным о конференциях, публикациях и связанных метаданных. Эти данные открывают новые перспективы для анализа и прогнозирования тенденций в научной деятельности, что особенно актуально в условиях увеличивающегося объема информации.

С InspireHEP удобно работать, используя существующий API InspireHEP, который позволяет получить информацию об имеющихся на платформе публикациях, авторах, конференциях, цитированиях, институтах и диссертациях. Задачей курсовой работы является изучение данных для выявления закономерностей в научной коммуникации, анализа популярности тем и оценки динамики развития областей знаний.

2 Полученные результаты и используемые методы

2.1 Извлечение информации

Был изучен API и извлечена информация о прошедших за последние пять лет конференциях и публикациях в них. О конференциях были получены: даты проведения, id в системе, краткие названия, публикации на каждой из них. О каждой публикации: названия, id в системе, авторы, ключевые слова, краткие описания(абстракты). Главным недостатком полученной информации является то, что ключевые слова, по которым можно было бы сравнительно просто классифицировать статьи, есть только у 70 процентов публикаций. Поэтому было принято решение работать с названиями статей и извлекать тему из них, так как название является обязательным атрибутом. Первым шагом в работе является кластеризация документов по темам. Был написан модуль на Python и запущен на небольшой тестовой выборке, состоящей из статей с одной конференции (около 1000 документов).

Подходящий метод выбирался среди Latent Dirichlet Allocation (LDA), Non-Negative Matrix Factorization (NMF) и Embedded Topic Model (ETM). LDA – популярный, классический метод, но его ограничение в учёте семантической близости слов стало существенным недостатком при работе с узкоспециализированными текстами. NMF неудобен в интерпретации результатов и требует большое число настроек. После проведённого анализа был выбран ETM, который сочетает вероятностный подход с использованием предобученных эмбеддингов слов. Это позволяет учитывать контекстные связи между словами, что особенно важно для анализа научных публикаций.

2.2 Описание используемого подхода

Embedded Topic Model (ETM) — это подход для тематического анализа текстов, который сочетает вероятностное моделирование тем с использованием векторных представлений слов.

2.3 Генеративная модель

ETM описывается следующим процессом:

- Каждая тема k представляется как вектор $\beta_k \in \mathbb{R}^L$, где L — размерность эмбеддингового пространства.
- Для документа d сэмплируется распределение тем $\theta_d \sim \mathrm{Dir}(\alpha)$, где α — гиперпараметр Дирихле.
- Для каждого слова w_{dn} в документе:
  1. Сэмплируется тема $z_{dn} \sim \mathrm{Categorical}(\theta_d)$.
  2. Генерируется слово w_{dn}:

$$w_{dn} \sim \mathrm{Multinomial}\left(\mathrm{softmax}(\beta_{z_{dn}} \cdot \mathbf{e}_w)\right),$$

  где $\mathbf{e}_w$ — эмбеддинг слова w, а $\beta_{z_{dn}}$ — вектор темы.

2.4 Основные обозначения

- β_k — вектор темы k в эмбеддинговом пространстве.
- $\mathbf{e}_w$ — эмбеддинг слова w.
- θ_d — распределение тем для документа d.
- α — гиперпараметр Дирихле.
- z_{dn} — тема для слова w_{dn}.
- w_{dn} — слово, сгенерированное из темы z_{dn}.
- softmax — функция, преобразующая вектор в распределение вероятностей.

2.5 Обучение модели

Обучение модели осуществляется с помощью алгоритма **Adam** (Adaptive Moment Estimation). Данный метод автоматически адаптирует скорость обучения (learning rate) для каждого параметра, используя скользящее среднее первого и второго моментов градиента. Благодаря этому Adam обеспечивает более быструю и стабильную сходимость по сравнению с классическим стохастическим градиентным спуском (SGD).

2.6 Используемые эмбеддинги

Для представления слов в векторном виде, $\mathbf{e}_w$, используются предобученные эмбеддинги из набора данных **GloVe (Global Vectors for Word Representation)**. Данные вектора отражают семантическую близость слов, что позволяет повысить качество тематического анализа за счёт более точного учёта лексических связей.

2.7 Основные этапы реализации

- **Предобработка текстов.** Документы(названия) преобразуются в матрицу «документ–слово» с использованием `CountVectorizer` (библиотека `scikit-learn`).
- **Загрузка эмбеддингов.** Предобученная модель GloVe конвертируется в формат word2vec, после чего извлекается эмбеддинг-матрица для включения семантических связей между словами.
- **Построение модели.** Определяются два линейных слоя:
  1. `beta`, отвечающий за распределение тем по словам (использует загруженные эмбеддинги);
  2. `theta`, позволяющий получать распределение тем для каждого документа.
- **Обучение.** Параметры `beta` и `theta` оптимизируются с помощью Adam путём минимизации отрицательного логарифма вероятности сгенерированных моделью слов.
- **Кластеризация.** После оценки тематических распределений для всех документов к ним применяется метод `KMeans`, что даёт возможность объединять документы на основе их тематической близости. KMeans был выбран для кластеризации из-за его простоты, высокой скорости работы и способности эффективно обрабатывать большие объемы данных.

2.8 Результаты

Результатом работы программы являются кластеры и топ-слова для каждого из них, а также информация о принадлежности каждого из документов к определённому кластеру, которая заносится в исходный csv - файл с полученными из API данными, позволяя

отследить корректность выполнения.

Число кластеров – это гиперпараметр, который можно регулировать. Важно отметить, что модель не даёт названия кластеров. Названия подбираются пользователем, основываясь на топ-словах, так что корректное именование тоже является важной задачей, требующей подробного рассмотрения.

Топ-слова	**Предложенное название кластера**	**Пример названия документа в кластре**	**Число документов в кластере**
beam, steel, processor, positrons, future, automated, criteria	Automation and Future Beam Systems	Application of Passive Wedge Absorbers for Improving the Performance of Precision-Science Experiments	217
design, cell, employing, beamlines, loop, chopper, matching	Design and Optimization	Research and Development of RF System for SC200 Cyclotron	245
vertical, beamlines, space, positrons, online, design, cesrta	Beamlines and Space Operations	Soft Chemical Polishing and Surface Analysis of Niobium Samples	349
beamlines, cylindrical, corrector, nanotube, design, nano, positrons	Advanced Beamline Technologies	Construction and Commissioning of the S-Band High-Gradient RF Laboratory at IFIC	189

Таблица 1: Результаты работы

3 Планы на следующий семестр

1. Полная обработка данных: Провести полный анализ текстовых данных для выявления тематик публикаций. Улучшить код, чтобы он корректно работал на больших данных. Использовать более мощные вычислительные ресурсы (GPU)
2. Анализ исторических данных: На основе собранных данных выявить закономерности в распределении публикаций по тематикам и конференциям. Подумать о том, как выбрать оптимальное число искомых кластеров и о том, как корректно назвать полученные группы.
3. Построение предиктивной модели: Разработать модель, которая на основе выявленных закономерностей и текущих данных сможет предсказать количество публикаций на конференциях в будущем.

Список литературы

[1] *Dieng, Adji B.* Topic Modeling in Embedding Spaces. — 2019. https://arxiv.org/abs/1907.04907.

[2] *Pennington, Jeffrey.* GloVe: Global Vectors for Word Representation / Jeffrey Pennington, Richard Socher, Christopher Manning // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) / Ed. by Alessandro Moschitti, Bo Pang, Walter Daelemans. — Doha, Qatar: Association for Computational Linguistics, 2014. — . — Pp. 1532–1543. https://aclanthology.org/D14-1162.

[3] *Kingma, Diederik P.* Adam: A Method for Stochastic Optimization. — 2017. https://arxiv.org/abs/1412.6980.

[4] *MacQueen, James.* Some methods for classification and analysis of multivariate observations / James MacQueen et al. // Proceedings of the fifth Berkeley symposium on mathematical statistics and probability / Oakland, CA, USA. — Vol. 1. — 1967. — Pp. 281–297.