**Тема дипломного проекта**: Разработка модуля векторизации научных статей.

Добрый день, Равиль Ильгизович.

**Задача 1.**

Я нашел датасет в нем не размечены топики. Но я могу например при обучении модели BERTopic указать необходимое мне количество топиков а это 25, если брать такое же количество как в статье, но я не знаю получится ли сделать топики такими же какие были при экспериментах в статье. И попробовать решить задачу сравнения результатов BERTopic, Top2vec и новый подход CFMf, который в статье указан.

**Задача 2.**

Или же начать решать задачу ‘Степень сходства между топиками в тематическом моделировании’

В документе более подробно расписал задачи.

Можете подсказать, какую задачу лучше решить первой.

Датасеты будут использоваться разные, т.к я уже смог использовать 550 тысяч абстрактов из научных статей для второй задачи и осталось только понять как правильно определить схожесть между топиками и найти способ как учитывать дату публикации.

[Top2vec](https://github.com/ddangelov/Top2Vec)

[BERTopic](https://github.com/MaartenGr/BERTopic)

[dataset](https://github.com/CRC-PSV/PhilSci_Topic_modeling_of_non_parallel_multilingual_corpora/tree/v1.0.0) https://zenodo.org/records/6484582

# Датасет нашел, *он не содержит топики*.

Полнотекстовое содержание 16 917 статей, опубликованных в 8 основных журналах по философии науки с 1931 по 2017 год

Датасет в папке data [ссылка](https://github.com/CRC-PSV/PhilSci_Topic_modeling_of_non_parallel_multilingual_corpora/tree/v1.0.0)

Имеет следующую структуру



**Journal\_id:** Идентификатор журнала.

**Title:** Заголовок статьи.

**Author:** Автор(ы) статьи, представлены в скобках с именем и фамилией.

**Year:** Год публикации.

**Volume:** Том журнала.

**Issue:** Номер выпуска журнала.

**Page\_range**: Диапазон страниц статьи.

**Citation:** Цитирование статьи, включая автора, год, заголовок и детали издания.

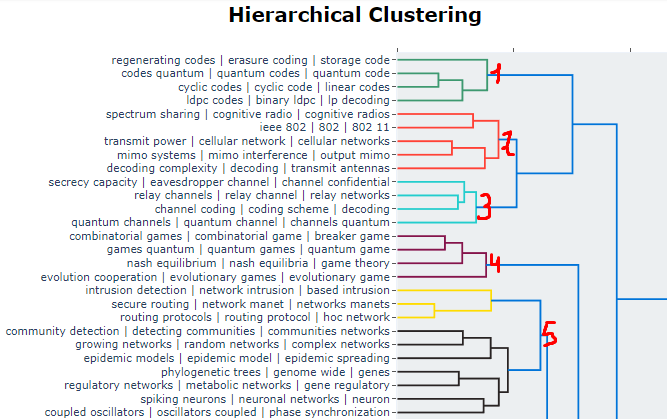
**Article\_ID:** Уникальный идентификатор статьи.

# Сравнить BERTopic и новый подход *CFMf* из статьи и также, Top2Vec и LDA.

# Степень сходства между топиками в тематическом моделировании

Учитывая временную шкалу, можно применить следующий подход:

1. **Группировка по Топикам:** Сначала группируем документы по топикам. Каждый топик содержит набор документов, связанных общей темой.
2. **Анализ Временных Рамок:** Затем анализируем временные рамки каждого топика. Например, в одном топике большинство документов опубликовано в 2008 году, в другом — в 2016.
3. **Определение Сходства:** Чтобы определить сходство между топиками, можно использовать метрики, учитывающие содержание документов (например, схожесть ключевых слов), а также их временные характеристики.
4. **Учёт Даты Публикации:** При определении сходства учитываем дату публикации, чтобы понять, как темы эволюционировали со временем.



# Применить полиномиальную для топиков с датой публикации за год

Top2vec hierarchical