**Вопрос**. Насколько просто или сложно получать упоминаемые 500 тысяч абстрактов? Является ли это отдельной задачей?

**Ответ.** 500 тысяч абстрактов упомянутые в статье были взяты из готового кластера данных с сайта Kaggle. А для получения абстрактов для других тематик необходимо будет найти или разработать скрапер. Я нашел скрапер который собирает данные из arxiv {абстракт, титул, авторов, дату создания, категорию} Это библиотека в которой можно указать категорию, например cs:ai и собрать данные за указанный период, например с 2022 до 2024.

За 42 минуты работы кода извлек данные 200 тысяч статей по тематике Computer Science. Далее остается только по ключевым словам в абстрактах найти интересующие нас темы.

**Первое**. Тематика статьи изменится на следующую: *Использование тематического моделирования для исследование поля научных публикаций (БПЛА)*

**Реализация:** Собрать кластер данных, применить модели тематического моделирования и оценить динамику роста для научных публикации (БПЛА) и также определить различные кластеры для данной тематики.

**Второе.** Статьи по конкретным направлениям с анализом ключевых слов и динамики публикаций

**Реализация:** Как я понял вы предлагаете выделить наиболее важные ключевых слова для топиков и сравнить динамику роста определенных топиков со всеми тематиками. Я могу с помощью скрапера собрать данные сначала по необходимым тематикам, а после по всем тематикам в целом.

**Вопросы.**

За какой промежуток времени стоит брать данные публикаций?

Будет ли достаточно данных только с arxiv и только абстрактов?

**Scraper for papers**

Links:

[github.com/Carouge/arXiv-parsing](https://github.com/Carouge/arXiv-parsing/tree/master)

[pip arxivabscraper](https://pypi.org/project/arxivabscraper/)

[Mahdisadjadi/arxivscraper: v0.0.2](https://zenodo.org/records/889854)