



Школа информатики физики и  
технологий

Прикладной анализ данных и  
искусственный интеллект

Санкт-Петербург 2025

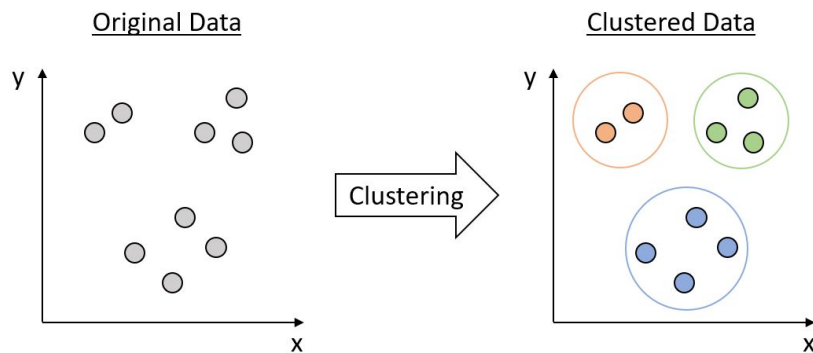
# Применение алгоритмов кластеризации для анализа больших текстовых медицинских данных

Соколовский Степан Павлович (3 курс, ПАДИИ)  
Научный руководитель: преподаватель Доронькин  
Максим Вячеславович



## Введение в область и проблематика

- Исследуемая область: кластеризация — метод машинного обучения, разбивающий неразмеченные данные на группы, внутри которых наблюдается сходство объектов между собой
  - Проблема: в современной медицине наблюдается экспоненциальный рост объёмов медицинских данных, которые можно и необходимо использовать для улучшения качества медицинской помощи
- Идея: использовать алгоритмы кластеризации для автоматического структурирования и использования медицинских документов





## Цель и задачи

**Цель** — обучить модели машинного обучения для структурирования и выявления значимых паттернов в больших массивах текстовых медицинских данных, а также разработать интерфейс для их использования

### **Задачи:**

- Изучить специфику медицинских данных и сформулировать требования к их обработке.
- Собрать и очистить текстовый корпус для обучения модели и экспериментов.
  - Провести эксперименты по кластеризации текстовых данных из корпуса и отобрать модели, верно кластеризирующие документы по их смыслу.
- Среди отобранных моделей выявить лучшую и провести её верификацию на тестовых данных.
- Сформулировать рекомендации по практическому использованию полученной модели



## Обзор аналогичных работ

### Статья “**A Lexical Approach for Text Categorization of Medical Documents**”[1] —

является наиболее близкой к моей по смыслу. В ней авторы предлагают свой способ кластеризации медицинских документов, учитывающий их специфику.

Однако, её авторы использовали для обучения и оценки качества модели данные 1991 года, которые потеряли актуальность.

Также в статье “**Using phrases and document metadata to improve topic modeling of clinical reports**”[2], хотя и иными методами, но преследуется схожая цель: моделирование тематик в медицинских документах. Однако решение, представленное в этой статье — не панацея, так как количество тематик ограничено и для работы такой модели требуются очень большие вычислительные ресурсы.

[1] R. Jindal, S. Taneja, “A Lexical Approach for Text Categorization of Medical Documents,” *Procedia Comput. Sci.*, vol. 46, pp. 314–320, 2015.

[2] W. Speier et al., “Using phrases and document metadata to improve topic modeling of clinical reports,” *J. Biomed. Inform.*, vol. 61, pp. 260–266, 2016.

## Порядок структурирования документов

Работа с входным корпусом текстов:

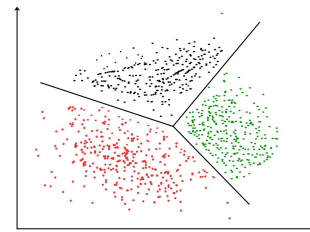
- построить векторные представления каждого текста (эмбединги)  
выбранным методом векторизации
  - применить к ним выбранный алгоритм кластеризации

Оценка адекватности модели:

- сравнить облака слов полученных кластеров между собой
- оценить взаиморасположение спроецированных кластеров на плоскости



| Words/<br>Documents | going | to   | today | i    | am   | it   | is   | rain |
|---------------------|-------|------|-------|------|------|------|------|------|
| Document 1          | 0     | 0.07 | 0.07  | 0    | 0    | 0.17 | 0.17 | 0.17 |
| Document 2          | 0     | 0    | 0.07  | 0.07 | 0.07 | 0    | 0    | 0    |
| Document 3          | 0     | 0.05 | 0     | 0.05 | 0.05 | 0    | 0    | 0    |



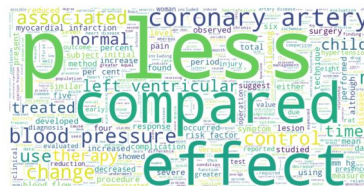


## Промежуточный результат

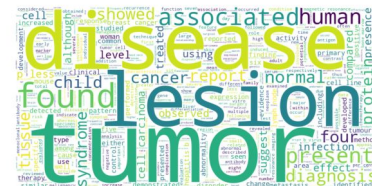
На данный момент были проведены эксперименты по применению различных методов векторизации текстов и алгоритмов кластеризации. Наилучшим образом себя проявила комбинация Word2Vec + k-Means.

В дальнейшем планируется продолжить исследования на новом корпусе текстов с возможными корректировками в оценке адекватности модели.

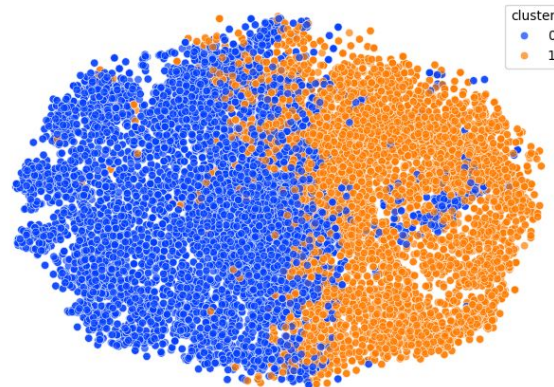
Облако для типа 0



Облако для типа 1



Разбиение текстов на 2 типов (проекция на плоскость)





## Оценка результатов

Качество разбиения документов на кластеры обученной моделью будет измеряться с помощью следующих способов:

- документы, лежащие в одном кластере, должны иметь схожую смысловую нагрузку
- документы из разных кластеров должны сильно отличаться по виду содержащейся в них информации
- количественные метрики, такие как Silhouette Score, Calinski-Harabasz Score, Davies-Bouldin Index, должны быть приемлимыми



## Планируемые улучшения

Отметим, что по достижении поставленной цели проект можно продолжить развивать в следующих направлениях:

- Расширение текстового корпуса и его более углубленная предобработка
- Использование иерархической кластеризации для построения древовидной структуры документов, включающей типы и подтипы таковых
- Предоставление возможности дообучения модели на новых данных без полного перезапуска обучения на старых данных





Спасибо за внимание!

Контакты:  
[spsokolovskii@edu.hse.ru](mailto:spsokolovskii@edu.hse.ru)