**Лабораторная работа 1: Морфологический анализ. Анализ по показателям TF-IDF**

[https://colab.research.google.com/drive/1-7ak8rHZCvNr57XJ\_d\_1gwK1s5JdIUf7#scrollTo=wSE-m6Er\_4w7](https://colab.research.google.com/drive/1-7ak8rHZCvNr57XJ_d_1gwK1s5JdIUf7%23scrollTo=wSE-m6Er_4w7)

Цель задания: Научиться анализировать тексты с использованием показателей TF-IDF, проводить морфологический анализ, оценивать лексическую близость текстов, выявлять ключевые слова и темы, а также формировать отчет о проделанной работе.

▎Шаги выполнения задания:

1. Выбор документов:
   * Найдите два текста (статьи, исследования, отчеты) на схожую тематику (например, психология, социология, технологии и т.д.).
2. Предобработка текстов:
   * Удалите лишние символы (знаки препинания, специальные символы).
   * Приведите все слова к нижнему регистру.
   * Удалите стоп-слова (распространенные слова, не несущие смысловой нагрузки).
3. Морфологический анализ:
   * Используйте библиотеку для морфологического анализа (например, pymorphy2 или natasha в Python).
   * Проведите анализ текста, выделив части речи и их грамматические характеристики (род, число, падеж и т.д.).
4. Создание таблицы:
   * Составьте таблицу с результатами морфологического анализа, где будут указаны:
     + Слово
     + Часть речи
     + Грамматические характеристики (род, число, падеж и т.д.)
5. Анализ результатов:
   * Определите, какие части речи преобладают в тексте.
   * Выделите ключевые слова и фразы.
   * Обсудите, как морфологические характеристики слов влияют на смысл текста.
6. Расчет TF-IDF:
   * Используйте библиотеку для обработки текста (например, scikit-learn в Python) для расчета значений TF-IDF для каждого слова в обоих документах.
   * Сохраните результаты в виде таблицы или списка.
7. Анализ значений TF-IDF:
   * Выделите слова с наивысшими значениями TF-IDF в каждом документе.
   * Определите, какие темы или ключевые идеи они представляют.
   * Сравните ключевые слова между двумя документами и сделайте выводы о различиях и сходствах.
8. Расчет косинусного расстояния:
   * Используйте значения TF-IDF для расчета косинусного расстояния между двумя документами.
   * Объясните, что означает полученное значение и как оно отражает схожесть документов.
9. Написание отчета:
   * Включите в отчет следующие разделы:
     + Введение: Опишите цель работы и выбранные документы.
     + Методология: Опишите процесс предобработки текстов.
     + Результаты: Приложите результаты каждого этапа анализа для каждого документа.
     + TF-IDF: Обсудите ключевые слова, темы и их значимость для каждого документа.
     + Косинусное расстояние: Представьте результаты расчета косинусного расстояния и его интерпретацию.
     + Морфологический анализ: Представьте результаты морфологического анализа и его интерпретацию.
     + Выводы: Подведите итоги анализа, отметьте основные различия и сходства.
10. Приложение формы отчета:
    * Создайте шаблон для отчета (можно в формате Word или Google Docs), который включает все вышеупомянутые разделы.
    * Приложите образцы результатов.

▎Формат отчета:

* Объем: 3-5 страниц.
* Шрифт: Times New Roman, размер 12.
* Межстрочный интервал: 1.5.
* Ссылки на использованные источники.

Удачи в выполнении задания!