

Лабораторная работа 7

Тема: Работа с естественным языком. Предобработка и классификация текста.

Задача состоит в предобработке и классификации текста. Классификация текста как правило заключается в:

- ✓ определении темы статьи или новости
- ✓ определении тональности (положительный или отрицательный отзыв).

Примерный алгоритм:

1. Найдите датасет, подходящий для классификации текстов. Это может быть классификация для определения тональности отзывов на товары или фильмы, или классификация для определения темы сообщения.
2. Проведите предобработку текста: токенизацию, лемматизацию, удаление спецсимволов, стоп-слов и т.д. (библиотеки NLTK или SpaCy).
3. Переведите текст в числовую форму для дальнейшего обучения (методы Bag of words, TF – IDF). **! Эмбендинги не используйте, эта тема будет рассмотрена в следующей лабораторной работе, поэтому пока остановимся на простых методах преобразования текста в числовую форму.**
4. Обучите 2-3 модели известных вам классификаторов (например, Naïve Bayes часто используют для текстов, RandomForest, LogisticRegression или иные).
5. Выберите наилучшую модель, используя свои знания метрик оценки качества моделей.
6. Получите новые текстовые данные соответствующие тематике вашего тренировочного датасета (отзывы, сообщения) путем парсинга сайтов. Проведите предобработку данных (см. п. 2).
7. Классифицируйте новые данные при помощи своей обученной модели.
8. **(по желанию)** Постройте облако слов для любого текста. Интересным может быть вариант построения облака слов из беседы в ТГ, в беседе можно выбрать сообщения конкретного человека (например, любимого человека, друга, мамы...), облако слов будет отражать наиболее частые слова, написанные этим человеком в диалоге с вами. Результаты могут быть достаточно интересными. Такое облако слов также можно оформить в виде любого контура и напечатать индивидуальную открытку. Для адекватного результата из текста нужно удалить стоп-слова.

Вопросы:

1. Что означает термин лемматизация.
2. Метод Bag of words.
3. Метод TF – IDF. Метрики TF и IDF.