В рамках домашнего задания предлагается попробовать реализовать несколько “классических” методов для классификации текстов.

В рамках задания мы используем датасет для определения спама в смс сообщениях. Задание выполняется в предоставленном ноутбуке.

1. После загрузки датасета его необходимо предобработать. В качестве базового решения предлагается токенизация и приведение к нижнему регистру. Приветствуется эксперименты по удалению пунктуации, лемматизации и т.д.
2. Постройте словарь наиболее частых слов. Напишите функцию для преобразования текстов в вектор частот слов из вашего словаря. В рамках задания не используйте готовые реализации из sklearn.
3. Заполните пропуски в классе BinaryNaiveBayes для классификации с помощью наивного байесовского классификатора.
   1. self.p\_y — вероятности классов 0 и 1
   2. word\_counts\_positive и word\_counts\_negative — частоты слов для разных классов
   3. self.p\_x\_given\_positive и self.p\_x\_given\_negative — условные вероятности слов при том или ином классе
   4. score\_negative и score\_positive — вероятности классов 0 и 1 для входного текста. Вычисляется как произведение условных вероятностей токенов на входе и вероятностей классов. Используя логарифм вычисления можно ускорить.
4. Найдите слова, которые наиболее важны для того или иного класса. Для этого необходимо найти отношение условных частот слов.
5. Далее попробуем обучить логистическую регрессию, преобразовав текст в вектора различными способами:
   1. Вектора частот слов в текстах
   2. Вектора TF-IDF. Реализуйте TF-IDF самостоятельно
   3. Сумма векторов каждого слова из текста. Используйте библиотеку gensim
6. Посчитайте точность и нарисуйте ROC-AUC кривую для каждого подхода. Сравните результаты.
7. \* Попробуйте также другой датасет для определения тональности в рецензиях на фильмы.