

Лабораторная работа #2: Language Modeling

В данной лабораторной работе вам предстоит закончить методы класса `LanguageModel`, являющегося реализацией n -gram'овой модели, которая предсказывает вероятность предложений. Модель обучается на предложениях из файла **train.tsv**. Файл **test.tsv** содержит пары предложений. Одно предложение настоящее, другое -- искусственно сгенерированно. Задача модели -- предсказывать где настоящее, а где сгенерированное. .

Файл **ngram.py** содержит класс *LanguageModel*, реализующий n -граммы. Они представлены в виде туплей целых чисел, счётчик n -грам реализован как словарь, ключи к которому являются n -граммами, а значениями -- количества их вхождений. Код почти закончен, ваши задачи:

1. По аналогии со счётчиком n -грам, реализовать счётчик $\{n-1\}$ -грам (контекстный)
2. Реализуйте формулу подсчёта вероятности предложения как сумму логарифма входящих в него n -грам
3. Реализуйте формулу подсчёта вероятности n -граммы, на основе количества её вхождений и вхождений её контекста

Для вашего удобства задания помечены как TODO в коде. После успешного выполнения **ngram.p**, собирается файл **submission.csv**. Вы можете проверить своё решение послав его на [ВОТ ЭТО](#) kaggle-соревнование. Если вы всё сделали правильно, ваш результат будет не менее 0.8.