**Проект NLP4.**

**Smart teeth**

Атнагулова Залина

Зубова Дарья

Наумов Александр

**Вступление.**

В качестве материалов для проекта были выбраны отзывы о ресторанах с сайта restoran.ru. Они были получены ручным способом. А затем данные были подвергнуты автоматической предобработке, чтобы привести их в нормальный вид.

**Подход. Mystem - WordNet**

**Шаг 1.**

Предобработка данных в парсере с помощью Mystem. Тексты были токенизированы и лемматизированы.

**Шаг 2.**

Лемматизированные токены обрабатывались и сравнивались с исходным списком оценочных слов из seed.txt по степени близости с помощью русского wordnet и следующих метрик.

Использовались метрики:

* Path Similarity
* Leacock-Chodorow Similarity

**Шаг 3.**

Основываясь на полученных коэффициентах близости, список оценочных слов был пополнен теми словами, которые имели наиболее схожие значения. Пороговым значением мы приняли коэффициент <1.

**Шаг 4.**

Далее было проведено сравнение полученных результатов с золотым стандартом. Мы считали кол-во слов из gold standard попавших в seed.txt и отбирали вручную слова, которые, по нашему мнению, должны были попасть в наш список, но не попали.

**Подход. Частотность по корпусам**

**Шаг 1.**

Мы берем отзывы в качестве материала и разбиваем их на несколько корпусов.

**Шаг 2.**

Производится лемматизация в Mystem, фильтруются токены и считается частотность в каждом тексте.

**Шаг 3.**

Производится сравнение самых частотных форм и, возможно, вручную отбираются наиболее подходящие варианты.