**Отчет по проекту**

**Проект выполнили:**

* Настя Каприелова
* Настя Кромина
* Петрова Даша
* Андрей Зырянов

1. **Описание корпусов:**

Корпус, с помощью которого осуществлялось расширение оценочных списков слов, - корпус отзывов по ресторанам и его baseline-разметка аспектов + тональности. Объем корпуса 5000 текстов. Мы не анализировали весь корпус, так как лемматизация всех текстов корпуса значительно снижает скорость выполнения программы.

Также в программе использовалась обученная модель c векторами слов araneum\_none\_fasttextcbow\_300\_5\_2018 для определения косинусной близости слов.

1. **Немного о методах и признаках:**

В функции detect мы для каждого слова из предложения смотрим его косинусную близость со словами из списков seed. Если число близостей, которые > 0.3, превышает 75% от сида, добавляем слово в кандидаты для пополнения сида. Если в одном предложении встречается слово с аспектом и оценочное слово, пара становится кандидатом для списка оценочных слов. Все кандидаты далее проверяются путем сравнения с коллокациями adj+noun, которые мы получили из нашего корпуса. Если кандидат на добавление встречается в коллокациях, мы смотрим на существительное, которое с ним употребляется. Это существительное с помощью косинусной близости сопоставляется со словами из списков слов, отнесенных к “еде” и “сервису” (списки придуманы нами). Таким образом, мы проверяем, стоит ли добавлять кандидата к тому или иному списку.

Для придания финальной выдаче требуемого формата мы используем разметку UdPipe.

1. **Результаты тестирования + квантитативный и качественный анализ ошибок:**

При использовании выданного сида было много мусора из-за большого количества разнообразной лексики. Мы не могли это ограничить, и поэтому использовали придуманный нами сид поменьше с ярко выраженной оценкой. Пример такого сида с положительной оценкой: ['отличный','вкусный','сочный','удобный'].

При выделении оценочных слов и их оценки были замечены некоторые ошибки: для таких характеристик сервиса, как навязчивый, курящий и некурящий, были назначены оценки 1, 1, 0, соответственно. Что на деле характеризует сервис с совсем другой стороны. Но есть вероятность того, что для определенных контекстов “некурящий” действительно может быть недостатком, в то время как “курящий” - плюс.