Отчёт по проекту.

Для работы был выбран корпус из 19 тысяч рецензий на рестораны. К ним в качестве обучающих данных были предоставлены 2 файла со словами из тем Food и Service, в каждом из которых было по 20 слов. Каждому из этих слов был дан лейбл тональности Для эмбеддинга слов я выбрал word2vec, который я обучал на данных словах, для получения большего правдоподобия. Потом я разбил методом LabelSpreading словарь на две темы Food и Service. Я это сделал перед определением тональностей, так как у каждой темы может быть своя семантика, поэтому необходимо разделять тональные слова каждой темы. Потом взяв оба корпуса, я начал искать ближайшие вектора слов к уже размеченным, получив таким образом по 100 слов каждой тональности, но количество может быть увеличено, для увеличения точности. К сожалению не получилось сделать поиск по биграммам и именным группам и сделать более точное разбиение на общего словаря корпуса на классы, из-за чего точность может сильно падать, так как связи между словами учитываются, но выражены они не явно.