Отчёт за 23 марта.

Сухочев Александр

Реализовал метод Disriminant Adaptive Nearest Neighbors, основываясь на вот этой статье https://web.stanford.edu/~hastie/Papers/dann_IEEE.pdf и осознал, что этот метод пока нам не подходит, так как его использование подразумевает перемножение матриц с размерностью, равной количеству признаков. Текущее количество признаков, которое я имею не используя никакой предобработки кроме mystem, составляет 15000-16000 в зависимости от разбиения на обучающую и тренировочную выборки. С перемножением матриц таких объёмов мой компьютер не справляется, не говоря уже о том, что это просто долго. Необходимо сократить количество признаков по крайней мере до 1000.

Одним из инструментов сокращения признаков, не учитывающих семантику проблемы, является PCA (http://www.miketipping.com/papers/met-mppca.pdf). Для быстрого отыскания собственных векторов ковариационной матрицы он использует сингулярное разложение, но даже этого не хватает, чтобы осилить наше количество признаков. Существует решение Incremental PCA (http://www.cs.toronto.edu/~dross/ivt/RossLimLinYang_ijcv.pdf), которое может fit-титься на последовательно подгружаемых данных, но использование его также даёт memory error. Сейчас пробую бороться с этим, нарезая данные на части и обучаясь на них по очереди, но это по-прежнему вызывает memory-error. Слишком маленькие части подавать ему на вход нельзя, так как он сможет выделить, как я понял, столько главных компонент, какова минимальная величина из количества признаков и обучающих примеров. Таким образом, подавая слишком маленькие части на вход, мы сможем получить слишком маленькое количество главных компонент.

План работы:

В данный момент работаю над решением этой проблемы и над сокращением числа базовых признаков с учётом семантики проблемы. Также готов заняться решением проблемы расширенного датасета, где присутствуют запросы, не касающиеся отелей.

Владислав просил подсчитать 90%-ый квантиль времени отклика на один запрос, он составляет 0.066131 секунд.