Competition 1

Рысьмятова Анастасия

ВМК МГУ 317 группа

08.04.2015

Содержание

Для описания объявления

- применен стеммер
- удалены стоп-слова
- испозованы данные из файла Location_Tree
- посчитан tf-idf

Использованные параметры для VW

- -b 27 используется для указания функции хэширования (все признаки хэшируются, а сами хэши принимают значения от 0 до 2^{b-1})
- -q параметр который указывает, что мы также хотим добавить в модель парные признаки
- –ngram индикаторы того, что данные два слова встретились рядом
- –affix +4t,+4q,+4h,+4c,-4t,-4c генерирует префиксы и суффиксы признаков

Различные идеи

- посчитать среднюю зарплату для каждого слова
- обучать различные модели из sklearn на спарс матрице из tf-idf

vw -d F.csv –passes 400 –quantile_tau 0.44 –affix +4t,+4q,+4h,+4c,-4t,-4c -c -f model.vw –loss_function quantile -l 2900 –l1 0.00000026 –initial_t 0.01 –power_t 0.515 -q jt -q pt -q qt -q jh -q pq -q pc -q qq -q hh -q ht -q pp -q jj -q cc -q jp -q jy -q cq -q hc -q hq -q pi –ngram F2 –ngram t2 -b 27
vw -d F_test.csv -i model.vw -t -p pred.txt