Мини-отчет по прошедшему контесту

Паша Коваленко

Машинное обучение | ММП ВМК

Текст и код

Согласно моим наблюдениям, в большинстве случаев листинг в вопросе выделен табами или пробелами относительно текста.

Тело каждого вопроса разбивается на текст и код.

Текст приведен к нижнему регистру и разбит на слова.
Слово — последовательность из букв и цифр.

Простые фичи

- Длина тела вопроса
- Длина заголовка
- Длина текста
- Длина кода
- Число строк кода
- Число абзацев в тексте
- Число слов в тексте
- Число тегов
- Репутация в момент написания вопроса
- Число ответов автора к моменту написания вопроса
- Время с момента регистрации до момента написания поста
- Число вопросительных знаков в тексте
- Число заглавных букв в тексте
- И еще несколько неинтересных

Word2Vec

Обучен на текстах из обучающей выборки оригинального контеста со стандартными параметрами (100 признаков).

Для каждого текста брались по каждой компоненте вектора:

- Среднее
- Стандартное отклонение
- Минимум
- Максимум

В итоге текст превращается в вещественный вектор длины 400.

Vowpal Wabbit

Для обучения использовал только Vowpal Wabbit.

Если вдруг кому-то интересны параметры:

- Функция потерь: логистическая
- Learning rate: 0.15
- 26-битные хеши

Ну давай уже к делу

В итоге Vowpal Wabbit обучался на:

1. Нормализованных "простых фичах"
2. Индикаторах простых фичей — "key=value"
3. Заголовке, разбитом на слова + 3-skip-2-gramm'ы
4. Тегах, преобразованных в слова + 3-skip-2-gramm'ы
5. Тексте (без кода), разбитом на слова
6. Word2Vec от текста без кода (тот самый вектор длины 400)
7. Word2Vec от заголовка
8. Word2Vec от тегов
9. Word2Vec от первого тега

Ну давай уже к делу

Что из этого в итоге взлетело:

Числа справа — уменьшение AUC-ROC при удалении этой "фичи"

- Word2Vec — 0.008
  - От текста — 0.0015
  - От заголовка — 0.003
  - От тегов — 0.0015
  - От первого тега — 0.002
- Логистическая функция потерь — 0.019 (!!!)
- 26-битные хеши — 0.005
- Нормализация фич — 0.003
- Индикаторы фич — 0.015 (!!!)
- Текст, разбитый на слова — 0.012
- n-skip-k-gramm'ы для заголовка и тегов — 0.002
- Все неочевидные фичи в совокупности — 0.0025
- Разбиение тегов на слова — 0.001

Спасибо за внимание!