Задание на семинар

Построить графики распределения в спам и не спам множествах следующих признаков:

- Количество слов на странице
- Количество слов в заголовке страниц (слова в теге <html><head><title>Some text</title>)
- Средняя длинна слова
- Количество слов в анкорах ссылок <html><body><a> some text

Пример шаблона дан в файле: *Atispam-Statistics.ipynb*

Домашнее задание

Цель

Разработать классификатор спама.

Описание

Для обучения предоставляется файл ./data/train-set-ru-b64-utf-8.txt.

Формат файла

Поля разделенные табуляциями:

0 - идентификатор документа

1 - метка класса 0 - не спам, 1 - спам

2 - урл документа

3 - документ в кодировке base64

Последовательность

- Из документов извлекаем признаки,
- Обучаем модель
- Присылаем модель и файл, который будет использовать обученную модель (*antispam_classifier.py*)

Для проверки качества модели используется скрипт *antispam_test.py*. Вызов скрипта

python antispam_test.py <имя файла с данными>

На выходе получаем результаты:

1. Точность, полнота и F1 мера для спама
2. Точность, полнота и F1 мера для не спама
3. Результат

Результат:

Минимальный порог прохождения - это 0.9 по F1 мере. Различие между F1 мерами для разных классов не должно превышать 10%.

Проверка осуществляется по тестовому дата сету.

Балы:

+2 – за сделанное задание семинара. Задание засчитывается если прислано до завершения семинара. Всего заданий 2 по количеству семинаров по фильтрации контента

+6 – за прохождение порога 0.9 F1 меры

+ 1 – бал за каждые 0.01 в F1 мере

Максимальный порог по F1 мере 0.96 по балам 15