Кафедра АД, «Автоматическая обработка текстов»

Задание 2
Кластеризация писем

Описание

В этом задании вам предстоит взять открытый датасет с электронными письмами Хиллари Клинтон и поэкспериментировать с построением более-менее интерпретируемых кластеров.

Организационные вопросы

Для сдачи задания выложите IPython/Jupiter notebook с кодом на github или nbviewer, и пришлите на почту xead@yandex-team.ru письмо со ссылкой на код. Тема письма должна иметь вид [АД Тексты 2016] Фамилия Имя – Задание 2 – кластеризация писем.

Задание

a. Загрузите датасет с kaggle: https://www.kaggle.com/kaggle/hillary-clinton-emails

b. Изучите, из чего состоит датасет.

c. Предобработайте тексты как сочтете правильным для первых экспериментов. Опишите, как вы его предобрабатываете, и почему так в блокноте в markdown ячейке

d. Выясните, какие биграммы чаще всего встречаются в датасете

e. Попробуйте выделить коллокации из двух слов по PMI с помощью nltk (примеры можно найти по ссылке: http://www.nltk.org/howto/collocations.html)

f. Выполните любую несложную кластеризацию писем (не тратьте на этот шаг много времени)

g. Придумайте, как визуализировать содержание кластеров. Например, можно выводить самые частые слова из каждого кластера (но, вероятно, это не самая удачная идея). Визуализируйте ту кластеризацию, которая у вас уже получилась.

h. Поработайте с признаками и методом кластеризации так, чтобы

кластеры выглядели наиболее интерпретируемыми.

i. Придумайте, как оценить интерпретируемость кластеров с помощью асессоров (какие вопросы задавать, как подсчитать качество на основе ответов). Для эксперимента воспользуйтесь кем-то из однокурсников в качестве асессора, и оцените интерпретируемость вашей кластеризации. Имейте ввиду, что такая оценка разумеется не статзначима и по-хорошему нужно привлекать более одного асессора, но протестировать придуманный вами способ оценки до какой-то степени так можно. Опишите ваш способ оценить интерпретируемость кластеризации и результаты в markdown ячейке в вашем ipython notebook.