Проект по курсу “Анализ неструктурированных данных. Часть 2. **70 из 100 баллов**”

23 октября 2017 г.

1 Тематический анализ

В прошлой части работы вы выкачали сообщения HackerNews за некоторый период, относящиеся к некоторому выбранному популярному бренду. Полученные данные являются сырыми, для них неизвестны ни метки классов, ни теги, ни что-либо подобное. В таких ситуациях применяются различные методы обучения «без учителя», т.е. методы кластеризации, направленные на выявление закономерностей в данных только по их структуре и статистической информации. В ситуации с текстами большую помощь могут оказать тематические модели, векторные представления слов и документов.

1. Установите и научитесь запускать один из предложенных на семинаре пакетов для тематического моделирования.

2. Выкачайте больше данных: по возможности увеличьте вашу выборку до 4-10K текстов (мы специально даём вам больше времени на задание). Имейте ввиду, что качество получаемых результатов существенно зависит от объёма выборки. Если что-то работает плохо, и вы уверены, что выполнили всё верно, попробуйте докачать больше данных.

3. Подготовьте ваши данные для моделирования (лемматизация, удаление стоп-слов, приведение к нижнему регистру, фильтрация по частотам, выделение коллокаций). Подробно опишите проделанную работу и параметры полученной коллекции.

4. Постройте простую модель PLSA или LDA с разным числом тем (от 10 до 200). Оцените темы по их наиболее вероятным словам (10-

20 штук). Насколько они интерпретируемы? Связаны ли с вашим брендом?

5. Постройте сжатые векторные представления слов, используя word2vec или Glove. Кластеризуйте полученное множество векторов произвольным алгоритмом кластеризации.

6. Визуализируйте полученные кластеры векторов word2vec, а также матрицу «слова-темы», полученную с помощью тематического моделирования. Выбор инструментов построения модели и визуализации свободный. Требуется, чтобы визуализация была наглядной, давала какую-то информацию о данных и была, по возможности, красивой. Рекомендуется смотреть в сторону t-SNE и библиотек визуализации графов и тематических моделей (самый простой вариант — LDAvis). Пример визуализации матрицы «слова-темы» можно найти здесь. Визуализации должны быть представлены в отчёте вместе с подробным описанием того, как они были получены и какие выводы были сделаны на их основании.

7. По совокупности результатов проделанной работы, сделайте выводы о том, какие сущности более всего обсуждаются по тематике выбранного бренда, а так же то, в каком аспекте проходит это обсуждение (пример хорошего вывода: «При обсуждении Apple часто говорят об iPhone, обсуждаются такие-то достоинства и такие-то недостатки»). Очень важно, чтобы каждый этап проделанной работы был подробно отражён в отчёте и снабжён дельными комментариями.

2 Правила игры

Ниже приведен список правил и условий:

1. Дедлайн по заданию: **23:59 26.11.2017**. **Дедлайн строгий!** Это означает, что задание, сданное после дедлайна, оценивается нулём баллов. При этом сдать задание всё равно будет нужно, без этого сдача курса становится невозможной.

2. Задание выполняется в группе. В случае использования какого-либо стороннего источника информации обязательно дайте на него ссылку (поскольку другие тоже могут на него наткнуться). Плагиат наказывается нулём баллов за задание и предвзятым отношением семинаристов в будущем.

3. Все пункты задания **обязательны к выполнению**, мы требуем присылать только полностью выполненные работы. Неполностью выполненное задание будет оцениваться каким-то количеством баллов на усмотрение проверяющих, но оно будет заведомо ниже полного балла.

4. При возниконовении проблем с выполнением задания обращайтесь с вопросами к преподавателям. Поэтому **настоятельно рекомендуется** выполнять задание заранее, оставив запас времени на всевозможные технические проблемы. Если вы начали читать условие в последний вечер и не успели из-за проблем с установкой какой-либо библиотеки — это ваши проблемы (см. п. 1), мы предупреждали.

5. Результат выполнения задания — это **отчёт** в формате `html` на основе Jupyter Notebook. **Отчёт должен быть написан по все правилам**, указанным на вики, небрежно его выполнение существенно отразится на итоговой оценке. Весь код из отчёта должен быть воспроизводимым, если для этого нужны какие-то дополнительные действия, установленные модули и т.п. — всё это должно быть прописано в тексте в явном виде.

6. Готовое задание нужно прислать на почту курса с темой «Фамилия Имя Проект Часть 2 (Фамилия семинариста)». Письма с иными темами рискуют остаться непроверенными. Просьба выполненное задание присылать единожды.