Проект по курсу “Анализ неструктурированных данных. Часть 1. **30 из 100 баллов**”

26 сентября 2017 г.

1 Составление собственного корпуса

В этом задании вам предстоит самостоятельно составить собственный корпус технических новостей из блогов. В качестве источника данных используйте HackerNews.

HackerNews представляет собой доску ссылок, один пост – одна ссылка на внешний ресурс. Про каждый пост HackerNews известна дата его публикации. Посты могут быть прокомментированы.

1. Используя API HackerNews (http://news.ycombinator.com), получите посты за последние полгода, сохраните все комментарии, число лайков, даты публикации и, непосредственно, ссылки на внешние ресурсы.

2. Напишите краулер, который будет переходить по всем ссылкам на внешние ресурсы, полученные на предыдущем шаге. По каждой ссылке нужно будет получить текстовые сообщения, для этого используйте readability (https://github.com/buriy/python-readability).

3. Отфильтруйте все сообщения, относящиеся к какому-нибудь вами выбранному бренду (например, Uber, Apple, Google, Elon Musk). Выберите достаточно популярный бренд, чтобы не иметь проблем с объёмом выборки.

4. Ответьте на следующие вопросы. Ответы проиллюстрируйте вычислениями и диаграммами.

    - Сколько постов в собранном корпусе?

- Из каких источников собран корпус (подсказка: используя регулярки или urlparse (https://docs.python.org/3/library/urllib.parse.html) найдите в ссылках доменное имя)?
- Какие слова (не считая стоп-слов) встречаются чаще всего?
- Из всех собранных ссылок, сколько ведут на блоги и новостные издания, а сколько – на github и другие неновостные издания?

2 Правила игры

Ниже приведен список правил и условий:

1. Дедлайн по заданию: **23:59 10.10.2017**. **Дедлайн строгий!** Это означает, что задание, сданное после дедлайна, оценивается нулём баллов. При этом сдать задание всё равно будет нужно, поскольку от его выполнения напрямую зависит следующая часть проекта.

2. Задание выполняется в группе (2-4 человека). 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 **данные**, представленные в структурированном виде, и **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 «Фамилия1, Фамилия2, Фамилия3 Проект Часть 1 (Фамилия Семинариста)». Письма с иными темами рискуют остаться непроверенными. Просьба выполненное задание присылать единожды.