Срок сдачи: 19 марта 2019

Задание 1. Сегментация “зашумленного” текста. Предобработка.

1. Ознакомиться с примерами: text.cleaning.py, vanilla_tokenizer_mistakes.txt
Руководство и библиотека:
https://www.nltk.org
https://www.nltk.org/book/ch03.html

2. Выбрать текст для сегментации, требования к тексту: не менее 100 словоупотреблений, должен содержать элементы, которые делают невозможной элементарную токенизацию по пробелам и знакам препинания, например: а) форумы и социальные сети (эмотиконы, гиперссылки, элементы разметки), б) технические тексты (единицы измерения, формулы, технические характеристики), в) химические и фармакологические тексты (названия химических соединений, условные обозначения), г) математические тексты (формулы, операторы, ошибки кодировки, разметка Tex), и т.д.

3. Разработать и обосновать принципы, по которым будет проводиться сегментация.

4. Написать регулярные выражения для правильной сегментации выбранного текста.

Удалить знаки пунктуации из текста, полученного в результате выполнения задания 2.

Задание 2. Подготовка данных для извлечения n-грамм.

Необходимо сгенерировать из исходного файла ru_ar_cut.txt (https://drive.google.com/open?id=0B4TmAgcGLMriMGpWS29yTloyVTg) файл с предложениями (1 предложение в строке (без тэгов)), который будет использоваться для дальнейшего тестирования алгоритмов извлечения n-грамм.

ru_ar_cut.txt (выборка из веб-корпуса русского языка Aranea (RuTenTen).

Задание 3. Лемматизация.

1. Запустить морфоанализатор mystem (ссылка на документацию:https://tech.yandex.ru/mystem/doc/) на текстовых файлах, собранных в задании 2.
2. Проверить результаты лемматизации. В отчете привести примеры неправильных случаев лемматизации (не менее 5 примеров) и случаев омонимии (порождено несколько гипотез для одной словоформы), объяснить полученные результаты.

Задание 4. Посчитать 3-граммы для корпуса аранеа.

1. На основе заданий 2-4 сформировать данные для расчета n-грамм по словоформам / лексемам (в зависимости от доставшегося варианта). Т.е. в

варианте со словоформами можно использовать файл полученный в задании 2, а в случае с лексемами можно получить файл: 1) используя mystem для исходного файла со словоформами (3-е задание), или 2) сгенерировать файл с предложениями на основе файла ru_ar_cut.txt, но брать значения именно лексем (2-й столбец).

2. Реализовать алгоритм расчета меры ассоциации (в зависимости от доставшегося варианта)
3. Проверить результаты с помощью библиотеки NLTK (скрипт)

Вариант:

1. словоформы, MI
2. словоформы, t-score
3. словоформы, log-likelihood
4. лексемы, MI
5. лексемы, t-score
6. лексемы, log-likelihood

На защите лабораторной продемострировать работу алгоритмов (собственный, nltk), показать список из top-30 3-грамм, объяснить различия в результатах.