**Срок сдачи: 01 октября 2018**

**3адание 1. Подготовка датасета**

Подготовить данные для выполнения задания 2 и последующих заданий:

1. На произвольном русскоязычном сайте новостей или агрегаторе (напр., Яндекс.Новости) найти новость о событии (лучше - международного масштаба), освещавшемся в СМИ в течение нескольких дней - стихийном бедствии (землетрясении, цунами, наводнении, лесных пожарах и т.п.).
2. На основе найденной статьи, выявить цепочку из новостей, которая будет включать не менее 4-5 стадий развития стихийного бедствия по их материалам, например:

i) факт наступления события,

ii) реакция спасательных служб (эвакуация, развертывание мобильных госпиталей, проведение спасательных операций, ...)

iii) факт продолжения стихийного бедствия / дополнительные разрушения (повторные толчки, распространение пожара, оползни вследствие непрекращающихся ливней, ….)

iv) протекание спасательных работ

v) ликвидация последствий стихийного бедствия

1. Скачать по одному новостному тексту на каждую стадию, сделать очистку от html, добавить метаданные (см.ниже).

**Формат:**

event\_id

subevent\_id\_1

subevent\_id\_1.txt (plain text, utf-8)

subevent\_id\_1\_metadata.json

subevent\_id\_1.html (source file)

subevent\_id\_i+1

subevent\_id\_i+1.txt (plain text, utf-8)

subevent\_id\_i+1\_metadata.json

subevent\_id\_i+1.html (source file)

….

subevent\_id\_n

subevent\_id\_n.txt (plain text, utf-8)

subevent\_id\_n\_metadata.json

subevent\_id\_n.html (source file)

Структура файла subevent\_id\_i\_metadata.json:

{

"URI": " ",

"title": " ",

"publishing\_date": " ",

"event\_date": " ",

"previous\_subevent\_uri": " ",

"next\_subevent\_uri": " ",

"topic": " "

}

**Задание 2.** **Подготовка данных для лемматизации.**

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.

ru\_ar\_cut.txt (выборка из веб-корпуса русского языка Aranea (RuTenTen).

**Задание 3. Сегментация. Предобработка.**

Ознакомиться с примером работ: text.cleaning.py, vanilla\_tokenizer\_mistakes.txt

Руководство и библиотека:

https://www.nltk.org

https://www.nltk.org/book/ch03.html

Удалить знаки пунктуации из текста, полученного в результате выполнения задания 2, используя библиотеку nltk, text.cleaning.py и регулярные выражения.

Написать регулярное выражение для различения дефиса (не удалять) и тире (удалять) , например: “Москва - столица” (тире), “диван-кровать”, “кто-нибудь” (дефис).

**Задание 3. Лемматизация.**

1. Запустить морфоанализатор mystem (ссылка на документацию:<https://tech.yandex.ru/mystem/doc/>) на текстовых файлах, собранных в задании 3.
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 5.** Поcчитать 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 (4-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

В отчет: первые 30 n-грамм