1. Сегментация и токенизация текста

Токенизация - разбиение текста на осмысленные элементы (слова, фразы, символы), называемые токенами.

Примеры простой токенизации:  
re.split(r' ', raw)   
re.split(r'[ \t\n]+', raw)

re.split(r'\W+', raw)

Необходимо разработать токенизатор на основе регулярных выражений. При токенизации должен корректно обрабатываться простой текст, содержащий слова и знаки пунктуации, а также:  
 1. телефонные номера, например: +7-901-000-00-00, 8(918)3213412 и т.д.

2. адрес, например: г. Санкт-Петербург, ул. Советская, д. 1294124, кв. 1

3. эмотиконы, например: привет:) как твои дела?

4. математические формулы, например: a = b\*c = (c+d)^2

Для тестирования результатов необходимо использовать текст объёмом не менее 50 словоупотреблений. Текст должен содержать как минимум одно из предложенных выше усложнений.

1. Стемминг - это процесс нахождения основы слова для заданного исходного слова. Основа слова не обязательно совпадает с морфологическим корнем слова.

Необходимо выполнить стемминг, используя одну из существующих библиотек, например, nltk [SnowballStemmer](https://www.nltk.org/howto/stem.html) (python).

1. Лемматизация - процесс приведения словоформы к лемме — её нормальной (словарной) форме.

Необходимо выполнить лемматизацию, используя одну из существующих библиотек, например pymorphy2 (python).

Проверить результаты лемматизации. Для лемматизации и проверки можно использовать существующие датасеты, например [word2lemma](https://github.com/Koziev/NLP_Datasets/blob/master/Lemmas/Data/word2lemma.7z) или [dictionary](https://drive.google.com/file/d/0B19FB37aGawaUjBWYlIwRzFCNk0/view?usp=sharing). В отчете привести примеры неправильных случаев лемматизации (не менее 3 примеров) и случаев омонимии (порождено несколько гипотез для одной словоформы), объяснить полученные результаты.

1. Сформировать для текста из 1 пункта задания аннотацию в следующем формате.

файл: <фамилия\_студента>.tsv (например, Ivanov.tsv)

содержимое файла:  
<token\_1> <stem\_1> <lemma\_1>

<token\_2> <stem\_2> <lemma\_2>

…..

<token\_N> <stem\_N> <lemma\_N>

Пример:

Клиника клиник клиника

располагается располага располагаться

в в в

центре центр центр

столицы столиц столица

1. Результаты необходимо выкладывать в свою папку на google.drive или выкладывать свой репозиторий на github. Если будете выкладывать в репозиторий, то необходимо предоставить ссылку на него в папке на google.drive.