

## Практические задания

1. Существующие приложения Natural Language Processing (NLP).
2. Сегментация слов. Проблемы токенизации слов в тексте.
3. Сегментация предложений в тексте. Бинарный классификатор выделения предложений.
4. Нормализация и лемматизация слов.
5. Использование регулярных выражений для выделения токенов.
6. Создайте текстовый файл. Разбейте текст на предложения.
7. Разделите на токены второе и третье предложения.
8. Создайте словарь уникальных токенов текста.
9. Скачайте любой понравившийся текст с сайта <http://www.gutenberg.org/catalog/> в файл text.txt. Все токены файла text.txt запишите маленькими буквами, оставив только буквенные токены, создайте словарь уникальных токенов, отсортируйте его. Запишите код с использованием технологии «List Comprehensions» (включение списка).
10. Рассмотрите различные варианты токенизации приведенного ниже текста. Опишите их особенности, по-разному определяя токены. В последнем варианте выделите только буквенные слова: " 'Kinto by Mozilla - An open source Parse alternative >> <https://github.com/Kinto/kinto/> #python #parse OOOOOOO!!! @kindl!"
11. Используя приведенный ниже список твитов, выделите с помощью функции **regexp\_tokenize()** в первом твите ( и последнем твите) токены с хэштегами, т. е. начинающиеся с символа # или @ (подобно #Data или @text): ["#True\_store In the ZONE: The Official Music Video, Jut Now @example", "Katie #UTE @BOTtersnike. By far your bestalbum since Won't Go Quietly @example back to your best", "@realDonaldTrump FLORIDA – it is imperative that you heed the directions of your State and Local Officials. Please be prepared, be careful and be SAFE! #HurricaneMichael ready gov", "@NWSTallahassee 8am Intermediate Advisory from @NHC\_Atlantic upgrades #HurricaneMichael into a category 2 hurricane."]

12. Используя экземпляр класса **TweetTokenizer**, выведите на экран токены всех твитов. Запишите код с использованием технологии «List Comprehensions» (включение списка).

13. Используя функцию **regex\_tokenize()** выделите из строки "That U.S.A. and USA poster-print costs \$12.40 at New-York... \$145.9%" (или любой другой подобной строки текста) токены, представляющие обычные слова, аббревиатуры, слова с дефисом, валюту (вещественное число с точкой, начинающееся со знака \$ и возможно имеющее знак % в конце числа) и многоточие (...) одновременно.