ТОКЕНЕЗАЦИЯ В NLP

Вашкевич Е.К.

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Борискевич И.А. – канд. тех. наук

Обработкой естественного языка (NLP – Natural Language Processing) называется активно развивающаяся научная дисциплина, занимающаяся поиском смысла и обучением на основании текстовых данных. Токенизация – это процесс разбиения фразы, предложения, абзаца или всего текстового документа на более мелкие единицы, например, отдельные слова или термины. Каждое из этих меньших подразделений называется токенами. В статье проведен краткий обзор типов и средств токенизации.

Перед обработкой естественного языка нужно определить слова, которые составляют строку символов. В связи с этим токенизация является основным шагом для работы с **NLP**. Важность токенизации обусловлена тем, что значение текста можно легко интерпретировать, анализируя слова, присутствующие в тексте.

Пример токенизации:

"This is a cat." → ['This', 'is', 'a', cat '].

Токенизированную форму можно использовать для подсчета базовых статистик, например, количества слов в тексте или частоты слова, как необходимый шаг перед более сложными шагами обработки текста.

Существует несколько методов токенизации на языке программирования Python.

Базовым методом является токенизации (для «западных» языков, таких как русский или английский) с использованием функции Python split(). Данная функция возвращает список строк после разбиения заданной строки указанным разделителем. По умолчанию split() разбивает строку по пробелам.

Input:

text = """London of the capital of Great Britain."""

Output: [‘London’, ‘of’, ’the’, ‘capital’, ‘of’, ‘Great’, ‘Brittan.’]

Токенизация с использованием NLTK. NLTK, сокращение от Natural Language ToolKit, – это библиотека машинного обучения, написанная на Python для символьной и статистической обработки естественного языка.

NLTK содержит модуль tokenize(), который далее классифицируется на две подкатегории:

– Word tokenize: используется метод word_tokenize(), чтобы разбить предложение на токены или слова;

– Sentence tokenize: используется метод sent_tokenize(), чтобы разбить документ или абзац на предложения.

Пример работы подкатегории Word Tokenization:

Input:

from nltk.tokenize import word_tokenize

text = """London of the capital of Great Britain. It's one of the largest cities in the world."""

Output: [‘London’, ‘of’, ’the’, ‘capital’, ‘of’, ‘Great’, ‘Brittan’, ’.’, ‘It’, ’’’, ’s’, ‘one’, ‘of’, ‘the’, ‘largest’, ‘cities’, ‘in’, ‘the’, ‘world’, ’.’]

Все знаки препинания библиотека NLTK определяет, как отдельные токены.

Пример работы подкатегории Sentence Tokenization:

Input:

from nltk.tokenize import sent_tokenize

text = """London of the capital of Great Britain. It's one of the largest cities in the world."""

Output: ['London of the capital of Great Britain.', ' It's one of the largest cities in the world.']

Существуют ещё множество алгоритмов токенизации, однако библиотека NLTK – ведущая платформа для создания NLP-программ на Python. У нее достаточно легкие в использовании интерфейсы для многих языковых корпусов, имеются библиотеки для обработки текстов для классификации, токенизации и стемминга. А также это бесплатный опенсорсный проект.

Таким образом, токенизация является важным шагом в любом проекте по обработке и анализе текстов.

Список использованных источников:

1. Николенко С., Кадурин А., Архангельская Е. Глубокое обучение. – СПб.: Питер, 2018. – 480 с.: ил. – (Серия «Библиотека программиста»).

2. Боярский К. К. Введение в компьютерную лингвистику. Учебное пособие. – СПб: НИУ ИТМО, 2013. – 72 с.

3. Е.И. Большакова, Э.С. Клышинский, Д.В. Ландэ, А.А. Носков, О.В.Пескова, Е.В. Ягунова. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика. – М.:МИЭМ, 2011.-272с.