Токенизация (Tokenization)

> Разделы: Алгоритмы

Токенизация — процесс разделения текста на составляющие (их называют «токенами»). Чаще всего текст разделяется на токены по словам или предложениям. В русском и английском языках в качестве индикаторов, по которым осуществляется разделение, используются:

- для слов — пробел;
- предложений — точка, восклицательный и вопросительный знаки.

Использование тех или иных индикаторов для разделения текста зависит от особенностей языка, на котором написан текст.

Токенизация по предложениям преимущественно используется в таком направлении, как машинный перевод, когда важен контекст целого предложения, а не отдельных слов. Другое направление — синтез человеческой речи, в рамках которого происходит воспроизведение (озвучивание) заданного текста.

Токенизация по словам имеет массовое применение. На ее основе работает большинство автокорректоров орфографии, а также может осуществляться разметка текста, частным случаем которой является классификация (категоризация) текста.

Часто при токенизации по словам производят дополнительную разметку еще и по части речи для каждого токена. После осуществления подобной операции у каждого слова появляется дополнительный признак в виде части речи, к которой оно относится.

В некоторых случаях используется более редкий вид токенизации — по группе слов. Он является частным случаем токенизации по словам. При подобном подходе важно не одно слово, а сочетание определенных пар (например, глагол + существительное).

Токенизация является важным обязательным процессом при применении алгоритмов машинного обучения и нейронных сетей для анализа текста.