Лемматизация (Lemmatisation)

Разделы: Алгоритмы

Лемматизация — это процесс приведения всех изменяемых форм слова к единому значению. Снижает вариативность одного и того же слова, что повышает качество анализа текста.

Алгоритм лемматизации основан на поиске наиболее подходящего варианта слова по словарю. При анализе текстовой информации обычно используются данные, полученные после процесса токенизации, подразумевающего разделение текста на отдельные слова или предложения. После сопоставления со словарем все словоформы одного слова заменяются на одно конкретное значение.

В языках со сложным словообразованием (например, русском) может потребоваться помимо стандартных словарей использовать дополнительные, учитывающие специфику речи. Отдельно к процессу лемматизации подключаются словари сленга, аббревиатуры и сокращений.

Пример для английского языка: слово «walk» будет являться единой формой для глаголов «walking», «walked», «walks» и непосредственно «walk», а слово «eat» для «ate», «eaten», «eating», «eat». Таким образом, все словоформы после применения алгоритма лемматизации примут одно единое значение.

Пример для русского языка: слово «отправлять» будет являться единой формой для глаголов «отправляю», «отправляешь», «отправляет», «отправляем», «отправляете», «отправляют», и, непосредственно, «отправлять».