Мешок слов (Bag of words)

Разделы: Алгоритмы

Мешок слов — представление текста в виде массива, состоящего из отдельных слов и количества их использования. Применяется при анализе естественного языка и в составе алгоритмов компьютерного зрения.

Пример представления текста из двух предложений с помощью мешка слов представлен в таблице ниже.

Предложение	нужна	помощь	с	активация	программа	кл
Нужна помощь с активацией программы	1	1	1	1	1	
Нужен ключ для активации программы	1	0	0	1	1	
Итого	2	1	1	2	2	

Результатом представления является словарь в виде уникальных слов и их количества по предложениям и всему тексту в целом.

Недостаток мешка слов заключается в том, что с увеличением объема анализируемого текста происходит рост размерности массива. Каждое уникальное слово добавляет новый столбец. Это создает дополнительную сложность при анализе.

К тому же, для данного способа представления не важен порядок слов в тексте и грамматические языковые особенности. Это вносит свои ограничения на использование. К примеру, мешок слов будет сложно применим для ряда задач, в которых важен контекст и сочетание слов.

Тем не менее, без мешка слов практически невозможно рассчитать основные описательные статистики и проанализировать текст. В некоторых задачах и вовсе основную ценность имеет не количество слов в тексте, а образуемый словарь, содержащий все уникальные значения.

При анализе текста мешок слов является подготовительной стадией для подсчета метрики «tf-idf» и осуществляется после проведения операций токенизации и лемматизации.