N-грамма (N-gram)

> Синонимы: N-грамма
> Разделы: Алгоритмы

N-грамма — последовательность звуков, слогов, букв или слов из N элементов. При анализе текста чаще всего рассматривается по отношению к словам. Применяется для предугадывания пропущенных в тексте слов, выявления плагиата, автоматического определения языка, исправления орфографических ошибок, распознания речи, извлечения знаний из текста при интеллектуальном анализе данных.

Наиболее популярные разновидности N-грамм:

- биграмма — 2 элемента в последовательности;
- триграмма — 3 элемента в последовательности;
- четыреграмма — 4 элемента в последовательности.

При большем количестве последовательностей название выглядит как N-грамма, где N — количество элементов в последовательности. Чем больше элементов в последовательности, тем ближе N-грамма к исходному тексту. При составлении правил отбора слов, как правило, союзы, предлоги и знаки припинания отбрасываются.

Результат разбиения на N-граммы представлен ниже на примере предложения «Белая береза под моим окном принакрылась снегом, точно серебром».

Разбиение на биграммы:

1. Белая береза
2. береза моим
3. моим окном
4. окном принакрылась
5. принакрылась снегом
6. снегом точно
7. точно серебром

Разбиение на триграммы:

1. Белая береза моим
2. береза моим окном
3. моим окном принакрылась
4. окном принакрылась снегом
5. снегом точно серебром

Главная особенность N-грамм — возможность угадывания последнего слова в последовательности на основе вероятностных моделей. Это делает их особо популярными в системах интеллектуального поиска. В частности, данная особенность используется у большинства поисковых систем (подсказка следующего слова при вводе ключевой фразы в строку запроса реализована именно этим способом).

N-граммы являются более совершенным подходом при анализе текстовых данных, чем «мешок слов», т.к. позволяют производить количественную оценку не одного слова, а словосочетаний из N слов.