Метрика TF-IDF (Term frequency–inverse document frequency)

Разделы: Метрики

TF-IDF — статистический показатель, применяемый для оценки важности слова в контексте категории, документа или коллекции документов. Используется при анализе текстовых данных.

Как правило, TF-IDF определяется для каждого слова. Чем выше значение данного показателя, тем значимее слово в контексте категории, документа, коллекции. При этом данный показатель также позволяет учесть и широкоупотребляемые слова, понизив их значимость в контексте объекта для анализа.

Формула для определения показателя имеет следующий вид:

$$TF - IDF = TF * IDF,$$

где TF — частота слова в конкретной категории/документе/коллекции (в зависимости от того, какие данные анализируются), IDF — обратная частота документа (популярность слова).

Частота слова в категории определяется по формуле:

$$TF = \frac{n_t}{\sum_{i=1}^{k} n_i},$$

где n_t количество отдельных слов в категории/документе/коллекции, $\sum_{i=1}^{k} n_i$ общее количество всех слов в категории/документе/коллекции.

Обратная частота документа (также часто называют инверсией частоты) определяется по формуле:

$$IDF = ln(\frac{n_c}{\sum_{j=1}^{m} n_j}),$$

где n_c — количество категорий/документов/коллекций всего, $\sum_{j=1}^{m} n_j$ — количество категорий/документов/коллекций в которых содержится интересующее слово.

Первый компонент формулы для вычисления TF-IDF фактически всегда не меняется. Метод расчета инвариации частоты может различаться в зависимости от специфики задачи, объема данных для анализа, количества категорий. При этом основной смысл показателя остается без изменений и он позволяет снизить «вес» широкоупотребляемых слов.

При анализе текстовых данных метрику TF-IDF лучше всего рассчитывать после проведения процессов токенизации, а также лемматизации или стемминга.