

Закон ципфа

Закон Ципфа (Зипфа) — эмпирическая закономерность распределения частоты слов естественного языка: если все слова языка (или просто достаточно длинного текста) упорядочить по убыванию частоты их использования, то частота n -го слова в таком списке окажется приблизительно обратно пропорциональной его порядковому номеру n (так называемому рангу этого слова, см. шкала порядка). Например второе по используемости слово встречается примерно в два раза реже, чем первое, третье — в три раза реже, чем первое, и т. д.

TF-IDF и BM25

TF-IDF (от англ. TF — term frequency, IDF — inverse document frequency) — статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов или корпуса. Вес некоторого слова пропорционален количеству употребления этого слова в документе, и обратно пропорционален частоте употребления слова в других документах коллекции.

Объясним на примере. Если на странице 1000 слов, из которых 30 — слово “хомяк”, то получаем, что частота слова “хомяк” в документе равно 0,03 (30/1000). DF определяется как количество страниц, содержащих слово “хомяк” разделенное на количество всех страниц. То есть, если страниц, содержащих слово “хомяк”, 2 000 000, а всего страниц 5 000 000 000 (подразумевается количество страниц в интернете в целом), то $DF = 0,0004$ (2 000 000/5 000 000 000). Окончательный расчёт веса слова подразумевает деление TF на DF (или умножение на IDF). В нашем случае вес слова “хомяк” TF-IDF равен 75 (0,03/0,0004).

TF-IDF и BM25

BM25. Наиболее распространённую форму этой функции можно описать следующим способом:

Пусть дан запрос Q , содержащий слова $q_1, \dots, q_n$. Релевантность документа D запросу Q высчитывается следующим образом:

$$\text{score}(D, Q) = \sum_{i=1}^n \text{IDF}(q_i) \cdot \frac{f(q_i, D) \cdot (k_1 + 1)}{f(q_i, D) + k_1 \cdot (1 - b + b \cdot \frac{|D|}{\text{avgdl}})},$$

Немного теории поиска: волшебная формула ранжирования

где $f(q_i, D)$ есть частота слова (TF), q_i в документе D , $|D|$ есть длина документа (количество слов в нём), а avgdl — средняя длина документа в коллекции. k_1 и b — свободные коэффициенты, обычно их выбирают как $k_1 = 2.0$ и $b = 0.75$. $\text{IDF}(q_i)$ есть обратная документная частота (IDF) слова q_i .

Функция BM25 в её классическом виде оценивает “важность” страницы, исходя из встречаемости ключевых слов без учёта взаимоотношений между страницами (то есть ссылок). В настоящее время поисковыми системами используется модификация этой функции под названием BM25F. Она учитывает другие параметры, такие как заголовки, околоссылочный текст, количество ссылок и т.д. Каждому из этих параметров присваивается своя степень значимости, которая влияет на итоговый результат функции ранжирования.