РГР №4  
«Векторная модель текста»

Для каждого из языков работы №1 выбрать по 2 текста, выделив в них первых пять предложений.

Задание 1. На основании этих предложений получить векторную модель каждого текста на основе модели TF\*IDF.

**Алгоритм взвешивания признаков по TF-IDF**

1. Применить нормализацию текста (стемминг или лемматизацию), выделить базовые элементы
2. Построить частотный словарь для всех 𝑤
3. Проредить слова по частоте
4. Для каждого документа найти TF\*IDF.
5. Сформировать для каждого текста мультимножество TF\*IDF.

Задание 2. Сформировать векторную модель TF\*IDF для каждого текста.

Полученные модели сравнить на сходство тем с использованием косинусной меры подобия.

ПРИМЕР ФОРМИРОВАНИЯ МАТРИЦЫ ЧАСТОТ МОДЕЛИ TF\*IDF

Модель TF\*IDF

Для понижения значимости слов, которые встречаются почти во всех документах, вводят инверсную частоту термина IDF (inverse document frequency) — это логарифм отношения числа всех документов D к числу документов d, содержащих некоторое слово.

IDF = *lg (D/d)*

Параметр TF (term frequency) — это отношение числа раз kj , которое некоторое слово встретилось в документе, к общему числу слов в документе п. Нормализация длиной документа нужна для того, чтобы уравнять в правах короткие и длинные документы.

*TF = ki/n*

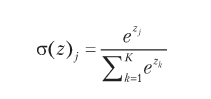
Коэффициент TF\*IDF равен произведению TF и IDF, при этом TF играет роль повышающего множителя, IDF — понижающего. Тогда весовыми параметрами векторной модели некоторого документа можно принять коэффициенты TF\*IDF входящих в него слов.

где nt - количество вхождений слова t в тексте *d*,   
 nк - общее количество слов в этом тексте,  
 *Di* - количество текстов из коллекции *D*, в которых встречается t.

Для того чтобы веса находились в интервале (0, 1), а векторы документов имели равную длину, значения TF\*IDF обычно нормализуются с использованием многопеременной логистической функции.

Эта функция преобразует вектор размерности в вектор той же размерности, где каждая координата полученного вектора представлена вещественным числом в интервале [0,1] и сумма координат равна 1.

Для каждого *z* из *K* выходных узлов выходное значение 𝜎 многомерной логистической функции можно вычислить с помощью нормированной экспоненциальной функции. Координаты 𝜎 вычисляются следующим образом:



Например, пусть коллекция состоит из 3 предложений.

1. *Мама мыла мылом Машу.*
2. *Мама мыла, мыла раму.*
3. *В магазине купила мама мыло.*

Вид словаря тогда будет следующим:

|  |  |  |  |
| --- | --- | --- | --- |
| Слово | Всего | Встретилось в предложении (d) | IDF |
| Мама | 3 | 3 | 0 |
| мыть | 3 | 2 | 0,18 |
| мыло | 2 | 2 | 0,18 |
| Маша | 1 | 1 | 0,47 |
| рама | 1 | 1 | 0,47 |
| магазин | 1 | 1 | 0,47 |
| купить | 1 | 1 | 0,47 |

Вид векторов:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 1 предложение | | | 2 предложение | | | 3 предложение | | |
| Слово | TF | TF\*IDF | Слово | TF | TF\*IDF | Слово | TF | TF\*IDF |
| Маша | 0,25 | 0,12 | рама | 0,25 | 0,11 | магазин | 0,25 | 0,12 |
| мыло | 0,25 | 0,05 | мыть | 0,5 | 0,09 | купить | 0,25 | 0,12 |
| мыть | 0,25 | 0,05 | мама | 0,25 | 0 | мыло | 0,25 | 0,05 |
| мама | 0,25 | 0 |  |  |  | мама | 0,25 | 0 |

Слово «мама» вообще можно не учитывать в векторном представлении. Так как оно встречается во всех предложениях коллекции, его значение TF\*IDF всегда будет равно нулю.

Заметим, что все слова примера приведены к нормальной форме (лемматизируются).

Таким образом, данную коллекцию документов можно описать матрицей частот

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Номер слова  Номер предложения | 1  (мыть) | 2  (мыло) | 3  (Маша) | 4  (рама) | 5  (магазин) | 6  (купить) |
| 1 | 0,05 | 0,05 | 0,12 |  |  |  |
| 2 | 0,09 |  |  | 0,11 |  |  |
| 3 |  | 0,05 |  |  | 0,12 | 0,12 |

Дальше обычно частоты нормируются так, чтобы сумма квадратов по строке (т. е. по документу) равнялась единице. Тогда окончательно получаем представление документов в виде векторов:

d1 = (0,36; 0,36; 0,86; 0; 0; 0);

d2 = (0,63; 0; 0; 0,77; 0; 0);

d3 = (0; 0,28; 0; 0; 0,68; 0,68).