Calculate TF-IDF

|  |
| --- |
| **A corpse** on chats on a computer model with **4 documents**: |
|  |
| Important “words” after ***pre-processing* (lower case, remove punctuation, remove accented characters, remove stopwords, lemmatization)**:  like, computer, model, XX15a, most, update, buy, yesterday |
| **TF (for each “word” in each document)**  e.g. “like” in Document 1:  no. of “like” = 1  total no. of words in document 1 = 6  TF for “like” in Document 1 = 1/6  It counts the relative frequency in the document.  *[Note: Sometimes, other forms of frequencies, e.g. count frequency (i.e. 1 for “like”) is used]*  **Higher relative frequency -> has higher importantance in the document**   |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | Document1 | Document 2 | Document 3 | Document 4 | | like | 1/6 | 0/4 | 0/5 | 1/5 | | computer | 1/6 | 0/4 | 0/5 | 0/5 | | model | 1/6 | 0/4 | 1/5 | 1/5 | | XX15a | … | … | … | … | | most |  |  |  |  | | update |  |  |  |  | | buy |  |  |  |  | | yesterday |  |  |  |  |   **IDF (for each “word” in the whole corpse)**  e.g. “like”  no. of documents in the corpse = 4  no. of documents containing “like” = 1  IDF for “like” = log(4/1)  It indicates how “rare” it occurs in a document.  **The more “rare”, its occurrence in a document may mean the word has more important meaning.**   |  |  | | --- | --- | | like | log(4/2) | | computer | log(4/1) | | model | log(4/3) | | XX15a | … | | most |  | | update |  | | buy |  | | yesterday |  |   **TF-IDF (for each “word”, each document)**  TF-IDF = TF\*IDF  e.g. “like” in Document 1  TF= 1/6  IDF= log(4/2)  TF-IDF = (1/6)\*log(4/2)  **Higher TF-IDF -> the “word” has higher importance in the document**   |  |  |  |  |  | | --- | --- | --- | --- | --- | |  | Document1 | Document 2 | Document 3 | Document 4 | | like | (1/6)\* log(4/2) | 0 | 0 | 1/5\* log(4/2) | | computer | 1/6\* log(4/1) | 0 | 0 | 0 | | model | 1/6\* log(4/3) | 0 | 1/5\* log(4/3) | 1/5\* log(4/3) | | XX15a | 1/6\*log(4/1) | … | … | … | | most | 0 |  |  |  | | update | 0 |  |  |  | | buy | 0 |  |  |  | | yesterday | 0 |  |  |  |   From the TF-IDF, the **most important** “word” in **document 1** are “**computer**” and “**XX15a**” |

[*Note:*

(1/6)\* log(4/2)=0.050

1/6\*log(4/1)=0.100 *(largest -> most important)*

1/6\* log(4/3)=0.021

1/6\*log(4/1)=0.100 *(largest -> most important*)]