Tutorial 1

Information Retrieval and Text Mining

10/13/2020

Topics to be covered

- What is a corpus/vocabulary/bag-of-words?
- What are stopwords?
- What is Vector Space Model?
- VSM (TF based model).
- What is the problem with TF based model and its solution?
- TF-IDF based model.
- Some important Python data-structures.
- File Handling in Python.
- Google Colab
- Jupyter Notebook Sample Codes

What is a corpus?

- **The collection of all the possible words that occur in our set of documents.**
- E.g. All the words in your English dictionary can be your corpus.

What are stopwords?

- **Stopwords** usually refers to the most common words in a language.
- **Stopwords** are the words in any language which does not add much meaning to a sentence.
- Thus they can safely be ignored without sacrificing the meaning of the sentence.

contd.

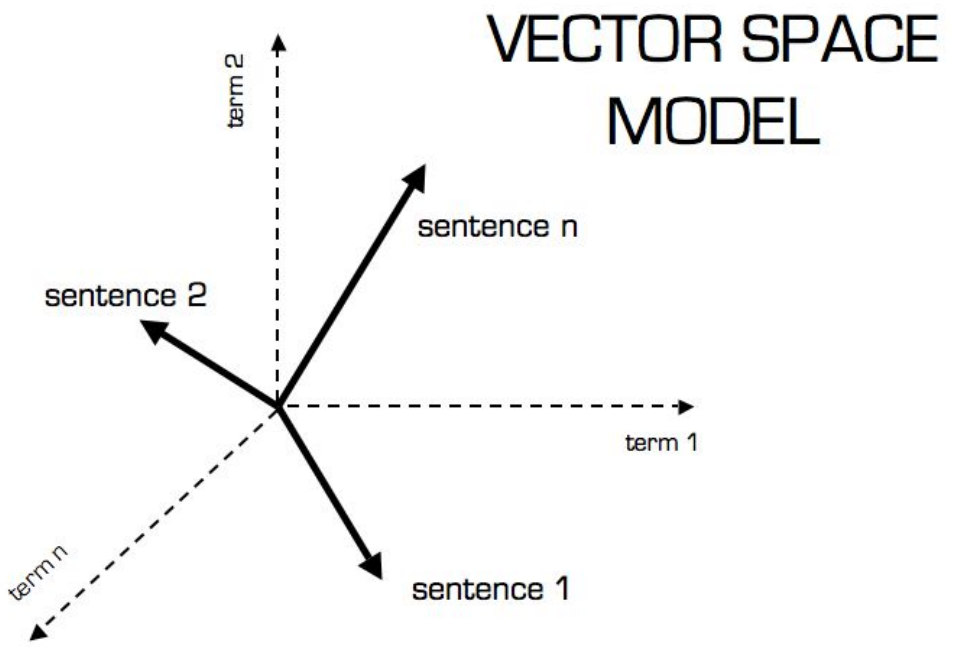
- ["i", "me", "my", "myself", "we", "our", "ours", "ourselves", "you", "your", "yours", "yourself", "yourselves", "he", "him", "his", "himself", "she", "her", "hers", "herself", "it", "its", "itself", "they", "them", "their", "theirs", "themselves", "what", "which", "who", "whom", "this", "that", "these", "those", "am", "is", "are", "was", "were", "be", "been", "being", "have", "has", "had", "having", "do", "does", "did", "doing", "a", "an", "the", "and", "but", "if", "or", "because", "as", "until", "while", "of", "at", "by", "for", "with", "about", "against", "between", "into", "through", "during", "before", "after", "above", "below", "to", "from", "up", "down", "in", "out", "on", "off", "over", "under", "again", "further", "then", "once", "here", "there", "when", "where", "why", "how", "all", "any", "both", "each", "few", "more", "most", "other", "some", "such", "no", "nor", "not", "only", "own", "same", "so", "than", "too", "very", "s", "t", "can", "will", "just", "don", "should", "now"]

contd.

- Most common python library to remove stopwords
  * **NLTK Library**
  * SpaCy Library
  * Gensim Library
  * Custom stop words

What is Vector Space Model?

- VSM is one of the ways by which we can represent our documents and query in the form of vectors.
- These vectors can then be used to address different NLP problems.



VSM (TF based model)

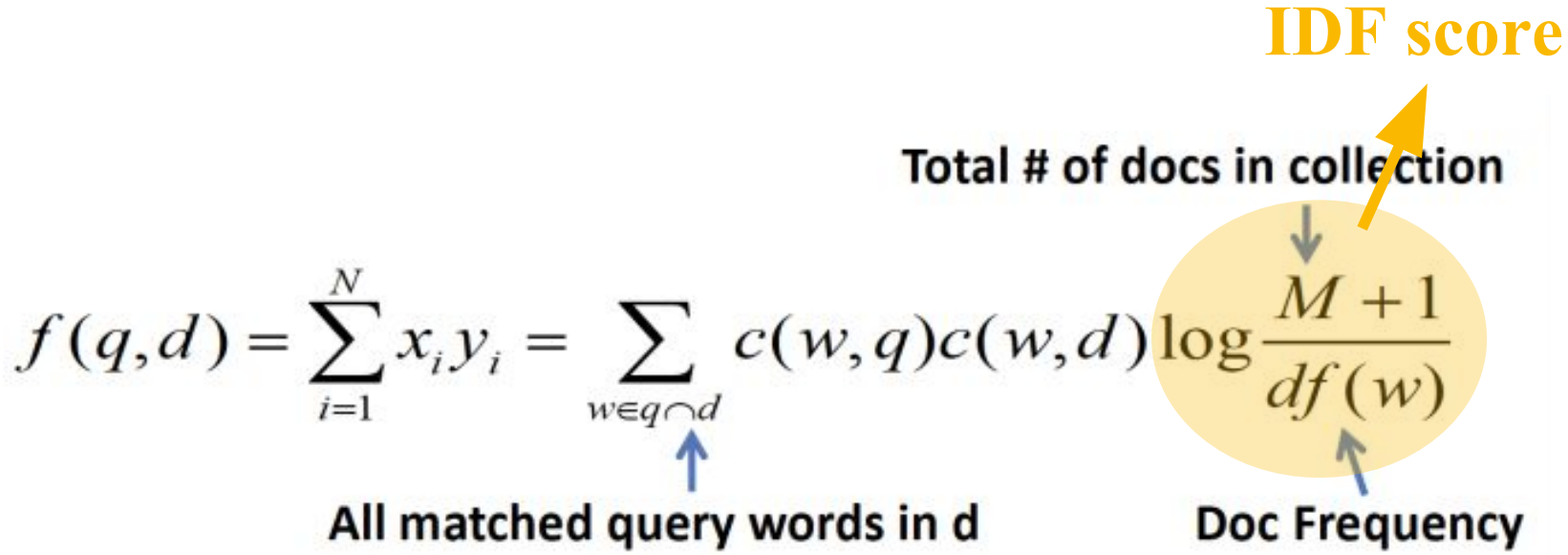
$\boldsymbol{q} = (x1,.....,xn)$
$\boldsymbol{d} = (y1,.....,yn)$
$\boldsymbol{x_i}$ = count of word $\mathbf{w}_i$ in query.
$\boldsymbol{y_i}$ = count of word $\mathbf{w}_i$ in document.

$$Sim(q,d) = \sum_{i=1}^{N} x_i\, y_i$$

What is the problem with TF based model and its solution?

???

VSM (TF-IDF based model)



Some important Python data-structures.

- **Lists**
  https://www.w3schools.com/python/python_lists.asp
- **Dictionaries**
  https://www.w3schools.com/python/python_dictionaries.asp
- **Tuples**
  https://www.w3schools.com/python/python_tuples.asp
- **Libraries you guys should explore**
  - Numpy
  - Pandas

File Handling in Python.

- https://www.w3schools.com/python/python_file_handling.asp

Google Colab

- Colaboratory is a **Google** research project created to help disseminate machine learning education and research. It's a Jupyter notebook environment that requires no setup to use and runs entirely in the cloud.
- **Why Colab?**

You get FREE GPU and TPU!!!

- **With some limitation of course**

- You can also mount your google drive.

Jupyter Notebook Sample Codes