

Text Analysis with Python

Arieda Muço

Central European University

Spring 2019

Simple Example

You have 2 documents:

1. Blue House
2. Red House

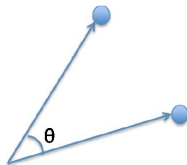
A document represented as a vector of word counts is called a "Bag of Words"

- "Blue House" \rightarrow (red,blue,house) \rightarrow (0,1,1)
- "Red House" \rightarrow (red,blue,house) \rightarrow (1,0,1)

Cosine Similarity

You can use cosine similarity on the vectors made to determine similarity:

$$\text{sim}(A, B) = \cos(\Theta) = \frac{A \cdot B}{\|A\| \|B\|}$$



Term Frequency and Inverse Document Frequency

- Improve on Bag of Words by adjusting word counts based on their frequency in corpus (the group of all the documents)
- Use Term Frequency - Inverse Document Frequency (TF-IDF)

TF-IDF

- Term Frequency - Importance of the term within that document
 - ▶ $TF(d,t)$ = Number of occurrences of term t in document d
- Inverse Document Frequency - Importance of the term in the corpus.
 - ▶ $IDF(t) = \log(D/t)$ where
 - ★ D = total number of documents
 - ★ t = number of documents with the term

TF-IDF term x in document y

$$TF_{x,y} \times \log\left(\frac{N}{DF_x}\right)$$

- $TF_{x,y}$ = frequency of x in y
- DF_x = number of documents containing x
- N total number of documents