Text Analysis and Visualization with Python

Arieda Muço

Central European University

Simple Example

You have 2 documents:

1. Blue House
2. Red House

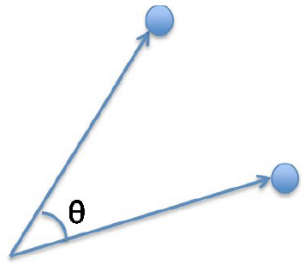
A document represented as a vector of word counts is called a "Bag of Words"

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

Cosine Similarity

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

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



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

TF-IDF term x in document y

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

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