Lecture 12:

Text mining

Outline

- Reminder: text data
- Similarity between documents
- Particularities of document clustering

Reminder

Extract features from a document (text file – unstructured data)

Bag-of-words approach):

“In document classification, a bag of words is a sparse vector of occurrence counts of words; that is, a sparse histogram over the vocabulary. In computer vision, a bag of visual words is a vector of occurrence counts of a vocabulary of local image features.”

Reminder

Extract features from a document – text file: bag-of-words approach

a) Remove the stop words

"In document classification, a bag of words is a sparse vector of occurrence counts of words; that is, a sparse histogram over the vocabulary. In computer vision, a bag of visual words is a vector of occurrence counts of a vocabulary of local image features."

"document classification bag words sparse vector occurrence counts words sparse histogram vocabulary computer vision bag visual words vector occurrence counts vocabulary local image features."

Reminder

Extract features from a document – text file: bag-of-words approach

b) Ignore inflections (reduce the words to their stems) – stemming (Porter algorithm)

“document classification bag words sparse vector occurrence counts words sparse histogram vocabulariy computer vision bag visual words vector occurrence counts vocabulariy local image features”

[http://textanalysisonline.com/nltk-porter-stemmer]

“document classif bag word spars vector occurr count word spars histogram vocabulari comput vision bag visual word vector occurr count vocabulari local imag featur”

Reminder

Extract features from a document – text file: bag-of-words approach

c) Compute the frequencies

“document classif bag word spars vector occurr count word spars histogram vocabulari comput vision bag visual word vector occurr count vocabulari local imag featur”

The extracted features:

(bag,2), (classif,1), (comput,1), (count,2), (document,1), (featur,1), (histogram,1), (imag,1), (local,1), (occurr,2), (spars,2), (vector,2), (vision,1), (visual,1), (vocabulari,2), (word,3)