

# Vector Space Model

## Lecture 2: Sept 13, 2013

CS886-2: Natural Language Understanding  
University of Waterloo

CS886-2 Lecture Slides (c) 2013 P. Poupart

1

## Document Representation

- Bag-of-word model
  - Ignore order of words
  - Treat each word as a feature
- **Vector space model**
  - Document: vector of weights (one weight per word feature)
  - Often sufficient for topic modeling and information retrieval

CS886-2 Lecture Slides (c) 2013 P. Poupart

2

## Vector Space Model Example

- Weights: term frequencies (tf)

CS886-2 Lecture Slides (c) 2013 P. Poupart

3

## Information Retrieval

- Find document most relevant to a query
- Query types:
  - Set of keywords
  - Question (natural text)
  - Document
- Idea:
  - Represent query as a vector of word features
  - Rank documents based on distance measure between the query's vector and the vector of each document

CS886-2 Lecture Slides (c) 2013 P. Poupart

4

## Distance Measures

- Notation:

$v_q = (w_{1,q}, w_{2,q}, \dots, w_{n,q})$  : query vector

$v_d = (w_{1,d}, w_{2,d}, \dots, w_{n,d})$ : document vector

- Distance measures:

–  $L_p$  norms:  $\|v_d - v_q\|_p$

– **Angle cosine:**  $\frac{\sum_{i=1}^n w_{i,q} \times w_{i,d}}{\sqrt{\sum_{i=1}^n w_{i,q}^2} \times \sqrt{\sum_{i=1}^n w_{i,d}^2}}$

CS886-2 Lecture Slides (c) 2013 P. Poupart

5

## Cosine Illustration

- Picture

- Cosine values:

*cosine* = 1:

*cosine* = 0:

CS886-2 Lecture Slides (c) 2013 P. Poupart

6

## Two Problems

- Some words are meaningless
  - E.g., a, the, of, with, etc.
- Words with slightly different suffixes are considered different
  - E.g., computer vs computers, drive vs driver, eat vs eaten

CS886-2 Lecture Slides (c) 2013 P. Poupart

7

## Some Solutions

- Remove “stop” words
  - Mostly “function” words that do not carry any meaning
  - Several common lists available on the web
  - E.g., a, the, of, with, etc.
- Stemming: truncate words to their stem
  - Computer, computers, computing →
  - Eat, eaten →

CS886-2 Lecture Slides (c) 2013 P. Poupart

8

## Porter Stemmer

- Series of rules:
  - ATIONAL → ATE e.g., relational →
  - ING → ε e.g., motoring →
  - SSES → SS e.g., grasses →

CS886-2 Lecture Slides (c) 2013 P. Poupart

9

## Better weights

- Idea: combine term frequency (tf) with inverse document frequency (idf)
- Terminology:
  - $K$ : total # of documents
  - $k_i$ : # of documents that contain term  $i$
- **Inverse document frequency (idf)**

$$idf_i = \log\left(\frac{K}{k_i}\right)$$
- **Better weights (tf-idf):**  $w_i = tf_{i,d} \times idf_i$

CS886-2 Lecture Slides (c) 2013 P. Poupart

10

### Term Weighting with TF-IDF

Document source: Old Bailey Online t18100110-41

Term Weighting:  None  Term Frequency  Raw Document Frequency  Inverse Document Frequency: TF-IDF

Stopwords greyed out

charles bailey was indicted for feloniously stealing on the 29th of december last dressed deer skins value 20 to the property of samuel savage and richard savage richard savage as a leather seller 63 chiswell street partner name samuel savage few days previous to the 29th of december looked out seventy skins for an order these skins being of a bad colour directed them to be brimstoned to make them of equal colour pale on the 29th in the afternoon I saw them all smooth to a horse a few hours afterwards they appeared very much tumbled and one was thrown into the yard dirtied caused them to be brought in the warehouse and counted there was one gone out foreman went to worship street and brought armstrong vickrey they searched out found this skin in the prisoner breeches and the other skin was found in the workshop carter as foreman samuel richard savage seventy skins I saw with me savage looking them out I took them out of the stove and counted them on the horse and on friday counted them three times since there were no more than sixty eight instead seventy went to worship street brought me armstrong vickrey with me they waited till the men left work and when they came then they were searched out on the prisoner and skin was found john armstrong went to this gentleman house after the men came down vickrey searching minute vickrey called in I received this skin from him it was taken out of the prisoner breeches I have not in ever since john vickrey a few were seen armstrong yes while I was searching number one saw the prisoner very uneasy and the breeches were unbuttoned I put my hand in and took some skin from him he said he could not tell how it came there the property produced and identified the prisoner said nothing to his defence called four witnesses who gave him a good character guilty aged 27 confined six months in the house of correction and fined a second mistress jury before us recorder