# Vector Space Model
# Lecture 2: Sept 13, 2013

CS886-2: Natural Language Understanding

University of Waterloo



## 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

# Vector Space Model Example

- Weights: term frequencies (tf)



# 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

# Distance Measures

- Notation:
  $v_q = (w_{1,q}, w_{2,q}, \ldots, w_{n,q})$ : query vector
  $v_d = (w_{1,d}, w_{2,d}, \ldots, w_{n,d})$: document vector
- Distance measures:
  – $L_p$ norms: $\left|\left|v_d - v_q\right|\right|_p$
  – Angle cosine: $\frac{\sum_{i=1}^{n} w_{i,q} \times w_{i,d}}{\sqrt{\sum_{i=1}^{n} w_{i,q}^2} \times \sqrt{\sum_{i=1}^{n} w_{i,d}^2}}$



# Cosine Illustration

- Picture
- Cosine values:
  $cosine = 1$:
  $cosine = 0$:

# 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



# 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 →

# Porter Stemmer

- Series of rules:

  ATIONAL → ATE  e.g., relational →

  ING → $\epsilon$  e.g., motoring →

  SSES → SS  e.g., grasses →



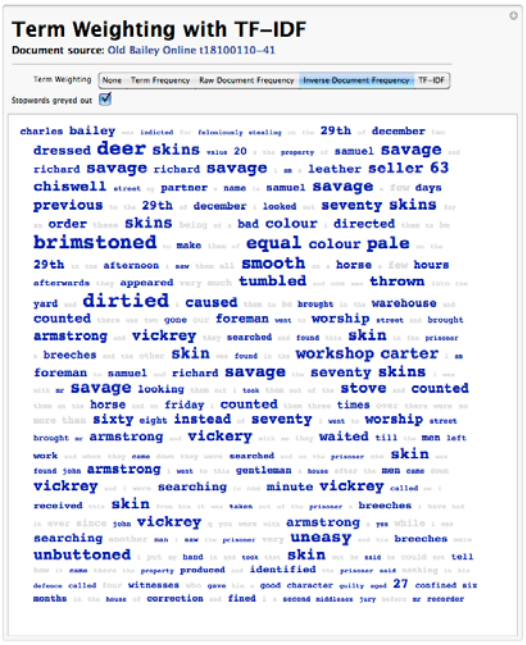
# Better weights

- 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): $w_i = tf_{i,d} \times idf_i$


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
29th
december
dressed deer skins
samuel savage
richard savage richard savage
leather seller 63
chiswell
partner
samuel savage
days
previous
29th
december
looked
seventy skins
order
skins
bad colour
directed
brimstoned
equal colour pale
29th
afternoon
smooth
horse
hours
appeared
tumbled
thrown
dirtied
caused
brought
warehouse
counted
foreman
worship
street
brought
armstrong
vickrey
searched
found
skin
prisoner
breeches
found
skin
workshop carter
foreman
samuel
richard savage
seventy skins
savage
looking
took
stove
counted
horse
friday
counted
times
sixty
eight
instead
seventy
worship
street
brought
armstrong
vickery
waited
till
men
left
work
came
searched
skin
found
john
armstrong
gentleman
house
men
came
vickrey
searching
minute
vickrey
called
received
skin
taken
prisoner
breeches
ever since
john
vickrey
armstrong
yes
while
searching
saw
prisoner
very
uneasy
breeches
unbuttoned
took
skin
said
could not
tell
came
property
produced
identified
prisoner
said
nothing
defence
called
four
witnesses
gave
good character
guilty
aged
27
confined
six
months
house
correction
fined
second
middlesex
jury
before
mr
recorder