Text

- Unstructured text (free text)
  - Exact keyword query
  - Syntactically similar keyword query
  - Semantically similar keyword query
- Semistructured text (SGML,XML)
- Structured text
  - Structural similarity query

Keyword based retrieval

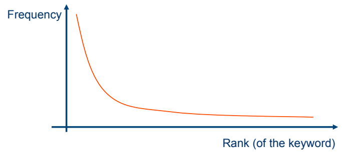
- Given a keyword, find all documents that contain the keyword
- Inverted indices when word boundaries are known
  - Use B-trees for exact retrieval
  - Use “trie”s and suffix trees for prefix based retrieval
- Need substring search when word boundaries are not known
  - exact: Bayer, Moore (BM) or Knuth,Morris,Pratt (KMP)
  - syntactic similarity: Wu,Manber

Text (as a collection of keywords) …..

- Each document is represented as a multi-set of keywords
  - Content words (terms)
  - Non-content words (stop words)
- Preprocessing
  - Stop word removal: eliminates stop words
  - Stemming: identifies roots of the terms
  - Phrasing: identifies compound terms

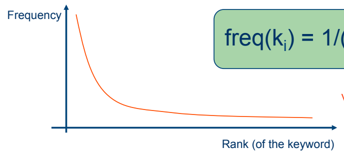
Zipfian Distribution

- The frequency of the k^{th} most frequent word in a collection is $(1/k)^{\ominus}$ times the most frequent word.



Zipfian Distribution

- The frequency of the k^{th} most frequent word in a collection is $(1/k)^{\ominus}$ times the most frequent word.

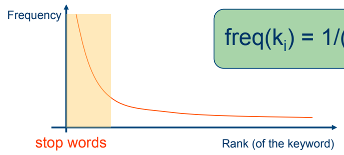


$$freq(k_i) = 1/(r_i \ln(1.78V))$$

Vocabulary size

Zipfian Distribution

- The frequency of the k^{th} most frequent word in a collection is $(1/k)^{\ominus}$ times the most frequent word.



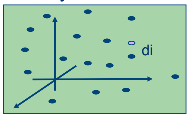
$$freq(k_i) = 1/(r_i \ln(1.78V))$$

Vector representation

- Given a set of keywords, each document is represented as a vector:

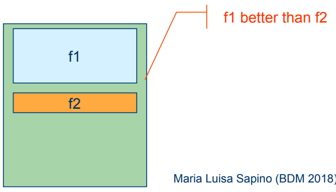
  di = <wi1,wi2,wi3,………,win>

  where
  - wij = 0, if the keyword does not occur in di
  - wij > 0, if the keyword occurs in di



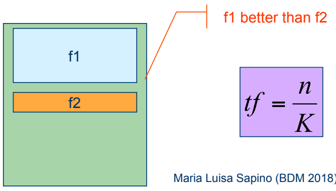
What are the weights????

- They need to capture how
  - good the term (feature) is in describing the content of the object



What are the weights????

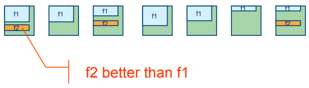
- They need to capture how
  - good the term (feature) is in describing the content of the object



$$tf = \frac{n}{K}$$

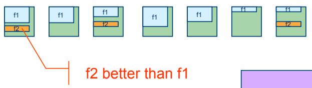
What are the weights????

- They need to capture how
  - differentiating the term (feature) is..



What are the weights????

- They need to capture how
  - differentiating the term (feature) is..



$$idf = \log(\frac{N}{m})$$

What are the weights????

- They need to capture how
  - good the term (feature) is in describing the content of the object
  - differentiating the term (feature) is..

$$tfidf = \frac{n}{K}\log(\frac{N}{m})$$

What are the weights????

- They need to capture how
  - good the term (feature) is in describing the content of the object
  - differentiating the term (feature) is..

Idf of the keyword

$$norm\_tfidf = \frac{n}{K} \frac{\log(\frac{N}{m})}{\max idf}$$

Experiment results suggest that

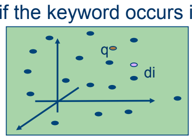
- Poor terms have high document frequency
- Good terms have low document frequency
  - Problem: may not be queried often enough to be useful
- Best terms have medium document frequency

How about query terms??

- Given a set of keywords, each query is also represented as a vector:

  q = <wq1,wq2,wq3,………,wqn>

  where
  - wqj = 0, if the keyword does not occur in dq
  - wqj > 0, if the keyword occurs in dq



How about query terms??

- They need to capture how
  - good the term (feature) is in describing the query
  - differentiating the term (feature) is..
  - Salton&Buckley suggests..

$$tfidf = \left(0.5 + 0.5 \frac{\frac{n}{K}}{\max\ freq} \right) \log\left(\frac{N}{m}\right)$$

Frequency of the term in the query

Maximum term frequency in the query

Are keywords independent??

Are keywords independent??

- Vector model assumes that they are..
- ..but are they really?

Are keywords independent??

- Syntactic similarity
  - Prefix relationship
    - “cat” vs. “catle”
  - Edit distance:
    - “table” vs. “cable”: 1 (replace “t” with “c”)
    - “table” vs. “bale”:   2 (delete “t”; swap “a” and “b”)

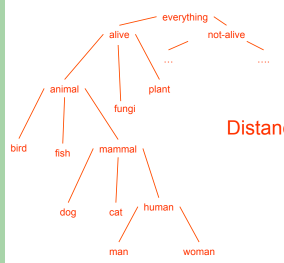
Semantic relationships

- Corpora-independent
  - Synonymy
    - Different but same meaning
  - Polysemy
    - More than one meaning
  - Hyponymy
    - K1 is an hyponym of K2 iff K1 is a K2
  - Hypernymy
    - K1 is an hypernim of K2 iff K2 is a K1
- Corpora-dependent
  - cooccurence

Semantic distance

- How dissimilar two terms are?
  - dist(man, woman)?
  - dist(man, child)?
  - dist(man, human)?

Semantic distance

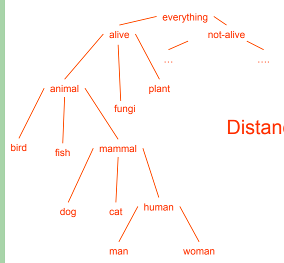


Class hierarchy

Distance = number of edges

Maria Luisa Sapino (BDM 2018)

Semantic distance



Class hierarchy

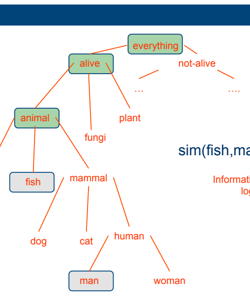
How about co-occurrence?

Distance = number of edges

How about frequency?

Maria Luisa Sapino (BDM 2018)

Semantic distance (P. Resnick)



Class hierarchy

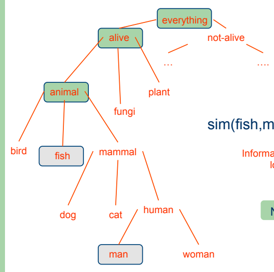
common ancestor

sim(fish,man) = max{inf_content(ca(man,fish))}

Information content
log(1/f)

Maria Luisa Sapino (BDM 2018)

Semantic distance (P. Resnick)



Class hierarchy

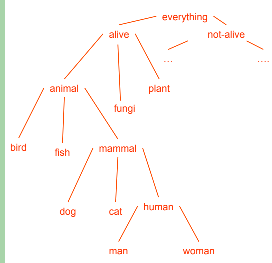
common ancestor

sim(fish,man) = max{inf_content(ca(man,fish))}

Information content
log(1/f)

Note: sim(fish,man) = sim(fish, bird)

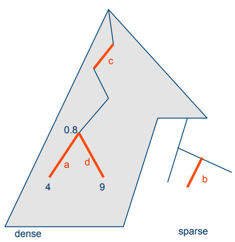
Semantic distance (Richardson et al.)



Class hierarchy

Weight each edge
- density of the hierarchy
- depth of the edge
- information content of each end

Semantic distance (Richardson et al.)



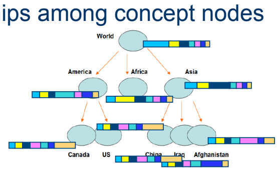
Class hierarchy

Weight each edge
- density of the hierarchy
- depth of the edge
- information content of each end

Δa < Δb (density)
Δa < Δc (depth)
Δa < Δd (ic)

Building the concept-vector space

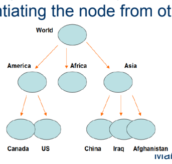
- Each concept is represented as a vector
- Concept vectors represent semantic relationships among concept nodes



Maria Luisa Sapino (BDM 2018)

Building the concept-vector space

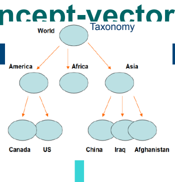
- CP/CV [CIKM06] process assigns a concept-vector to each concept node in the taxonomy:
  - A concept node clusters all its descendants nodes and essentially acts as a context for the descendant nodes
  - Descendants of a given node may also act as a context for the node, differentiating the node from others that are similarly labeled



Maria Luisa Sapino (BDM 2018)

Building the concept-vector space

- An example



Concept vectors $\vec{cv}_i$

	world	Asia	Africa	America	Afghanistan	Iraq	China	Canada	US
$\vec{cv}_{world}$	0.450	0.169	0.141	0.158	0.018	0.018	0.018	0.021	0.021
$\vec{cv}_{Asia}$	0.052	0.469	0.006	0.006	0.156	0.156	0.156	0.0003	0.0003
$\vec{cv}_{Africa}$	0.100	0.012	0.873	0.012	0.0006	0.0006	0.0006	0.0007	0.0007
$\vec{cv}_{America}$	0.057	0.007	0.007	0.520	0.0003	0.0003	0.0003	0.204	0.204
$\vec{cv}_{Afghanistan}$	0.004	0.100	0.0002	0.0002	0.872	0.012	0.012	0	0
$\vec{cv}_{Iraq}$	0.004	0.100	0.0002	0.0002	0.012	0.872	0.012	0	0
$\vec{cv}_{China}$	0.004	0.100	0.0002	0.0002	0.012	0.012	0.872	0	0
$\vec{cv}_{Canada}$	0.006	0.0003	0.0003	0.165	0	0	0	0.806	0.023
$\vec{cv}_{US}$	0.006	0.0003	0.0003	0.165	0	0	0	0.023	0.806

Maria Luisa Sapino (BDM 2018)

Term-to-term correlation

- Computes relationships between keywords given a corpus of documents
- Keyword connection matrix

Probability that keyword i and l occur in the same document

$$c_{i,l} = \frac{n_{i,l}}{n_i + n_l - n_{i,l}}$$

Probability that keyword i occurs in a document

Probability that keyword l occurs in a document

Vector model

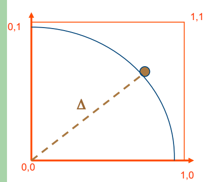
- Given a set of keywords, each document is represented as a vector:

  di = <wi1,wi2,wi3,………,win>

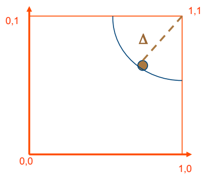
- We already discussed the salient features of this model…

Extended Boolean Model

- Salton, Fox, Wu(83)



k1 or k2 = Δ



k1 and k2 = √2 - Δ

Probabilistic Model

- Robertson&Jones(76)
  - Binary Independence Model
- Given a query and a document, estimate the probability that the user will find the document interesting
  - Assumption: there is an ideal set!! Can we estimate the properties of the ideal set.
  - We will come back to this model later..

Relevance feedback!!!!

Maria Luisa Sapino (BDM 2018)

Fuzzy Set Model

- Each query term defines a fuzzy set
- Each document has a degree of membership in this set
- Example: membership degree of document d_j in keyword k_i

$$\mu_{i,j} = 1 - \prod_{k_l \in d_j} (1 - c_{i,l})$$

- We will come back to this model later

Query processing!!!

Maria Luisa Sapino (BDM 2018)