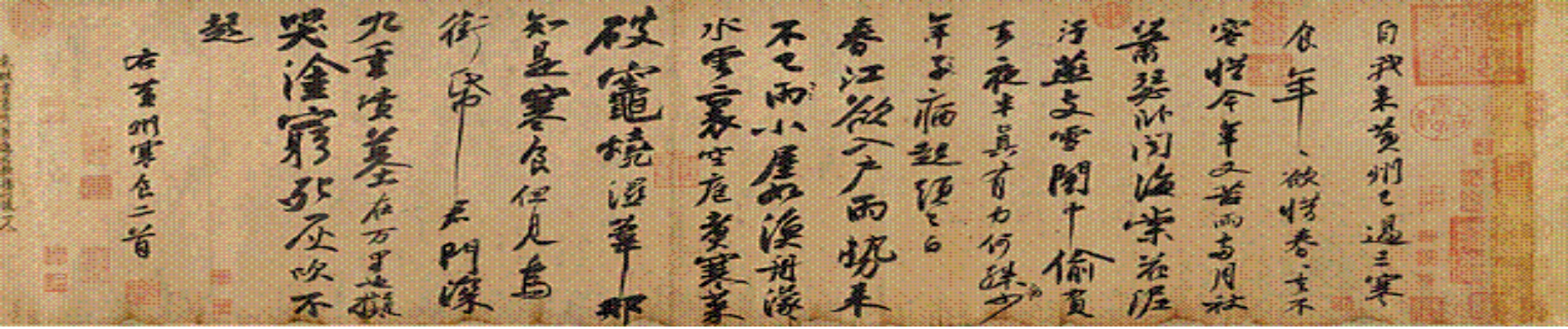


信息检索
Information Retrieval

教师：孙茂松

Tel:62781286

Email:sms@tsinghua.edu.cn

TA：**胡锦毅**

Email:hu-jy21@mails.tsinghua.edu.cn

郑重声明

- 此课件仅供选修清华大学计算机系本科生课《信息检索》(40240372)的学生个人学习使用，所以只允许学生将其下载、存贮在自己的电脑中未经孙茂松本人同意，任何人不得以任何方式扩散之（包括放到9#服务器上）。否则，由此可能引起的一切涉及知识产权的法律责任，概由该人负责。
- 此课件仅限孙茂松本人讲课使用。除孙茂松本人外，凡授课过程中，PTT文件显示此《郑重声明》之情形，即为侵权使用。

第三章 文本分析及自动标引 (Part 2)

3.2 Term的自动抽取及其加权

一个简单的自动标引过程

Step 1: Removal of high-frequency function words

Stop list: 250 common words in English (40-50% texts)

Step 2: stemming: suffixes, word stem form

analysis, analyze, analyzing, analyzer, analyzed, analysing

==> analy: enhancing recall

Step 3: Term weighting and producing of document vectors

3.2 Term的自动抽取及其加权

Caution: term deletion

(1) removal of some broad high-frequency terms may produce unwanted recall losses;

(2) removal of certain low-frequency terms reduces indexing exhaustivity and may result in reduced retrieval recall and precision.

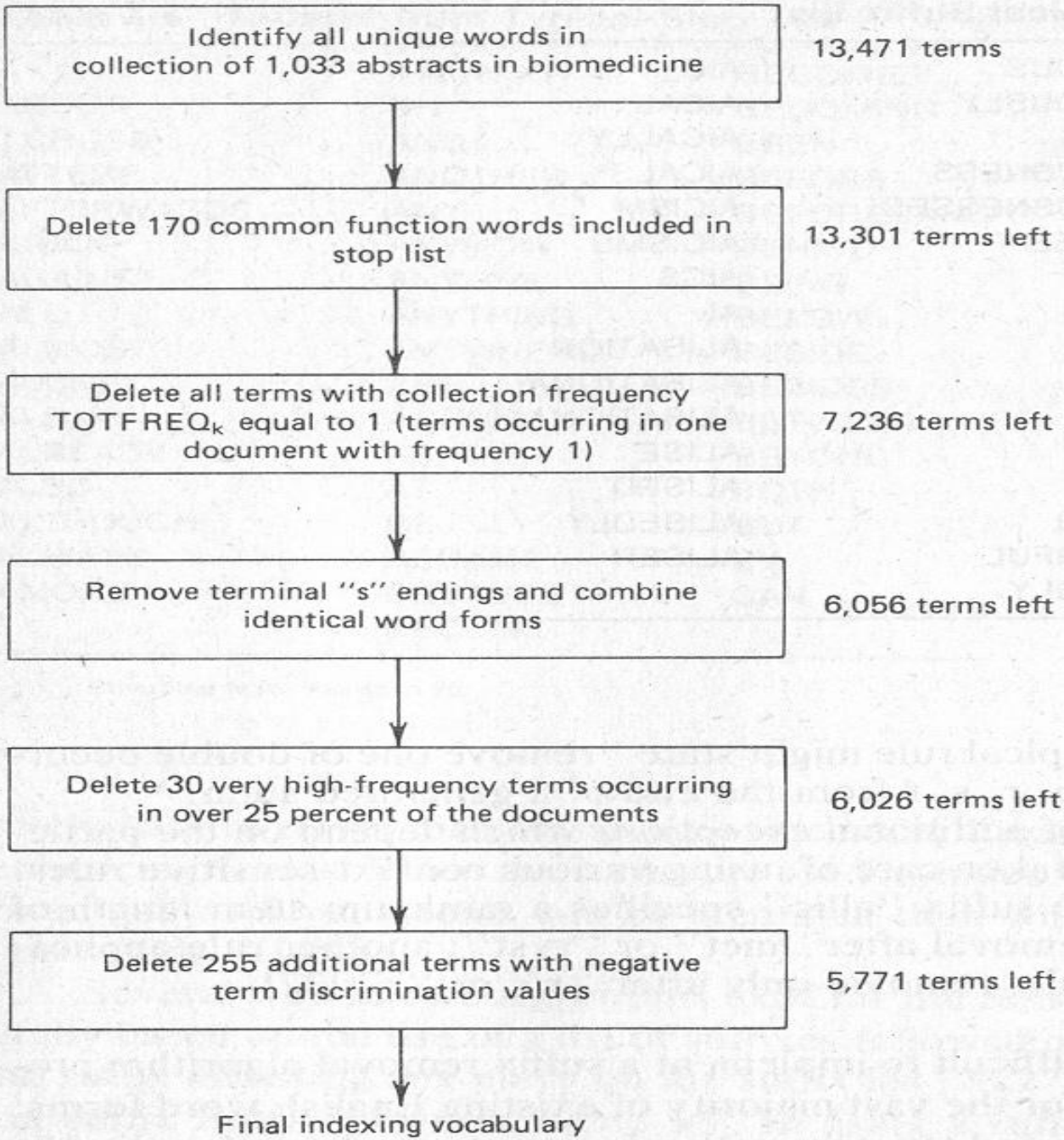


Figure 3-3 Typical term deletion algorithm (data for 1,033 documents in medicine).

3.2 Term的自动抽取及其加权

BM25 (https://www.elastic.co/cn/blog/practical-bm25-part-2-the-bm25-algorithm-and-its-variables)

$$\sum_{i}^{n} IDF(q_i) \frac{f(q_i, D) * (k1+1)}{f(q_i, D) + k1 * (1 - b + b * \frac{fieldLen}{avgFieldLen})}$$

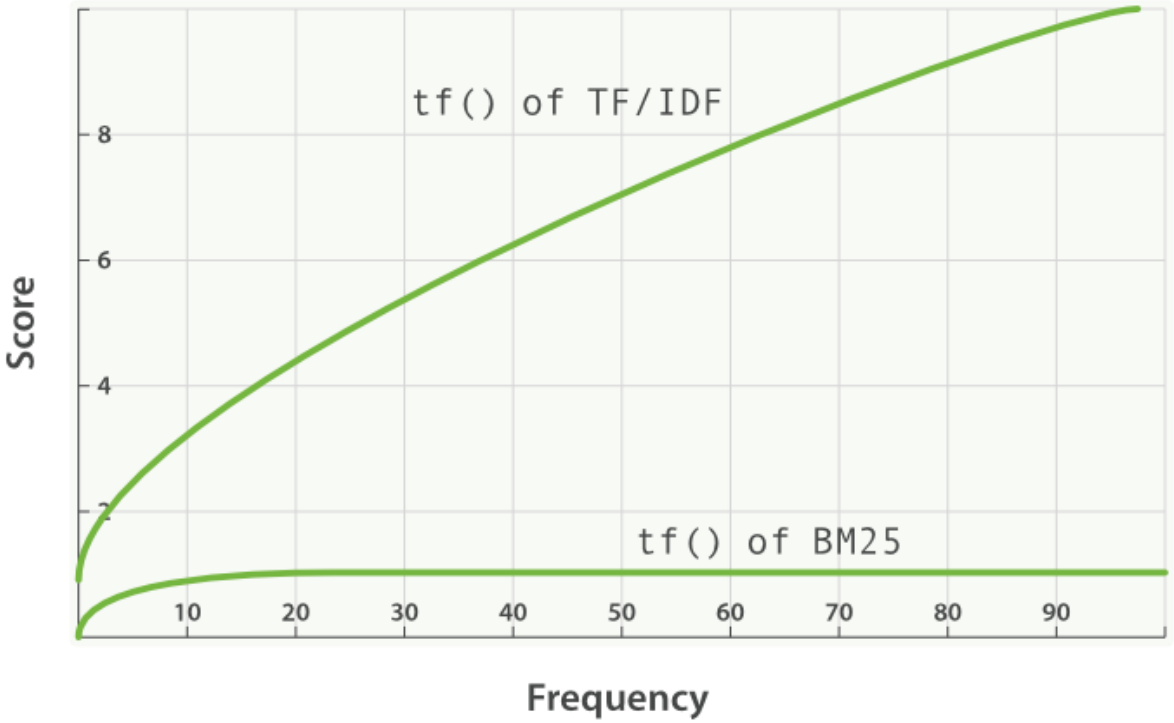
- IDF in Elasticsearch:

$$ln\left(1 + \frac{(docCount - f(q_i) + 0.5)}{f(q_i) + 0.5}\right)$$

- 文档长度的影响：By default, b has a value of 0.75 in Elasticsearch

3.2 Term的自动抽取及其加权

- TF in Elasticsearch:



* Term frequency saturation

* The curve of the impact of tf on the score grows slower and slower when tf() > k1. By default, k1 has a value of 1.2.

3.3 向量表示与相似度计算

Vector Space Model:

Developed in the SMART system (Salton,1970)

$$\begin{array}{c|cccc} & T_1 & T_2 & .... & T_t \\ D_1 & w_{11} & w_{21} & \cdots & w_{t1} \\ D_2 & w_{12} & w_{22} & \cdots & w_{t2} \\ \vdots & \vdots & \vdots & & \vdots \\ \vdots & \vdots & \vdots & & \vdots \\ D_n & w_{1n} & w_{2n} & \cdots & w_{tn} \end{array}$$

Binary → count → weight matrix

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth
Antony	5.25	3.18	0	0	0	0.35
Brutus	1.21	6.1	0	1	0	0
Caesar	8.59	2.54	0	1.51	0.25	0
Calpurnia	0	1.54	0	0	0	0
Cleopatra	2.85	0	0	0	0	0
mercy	1.51	0	1.9	0.12	5.25	0.88
worser	1.37	0	0.11	4.15	0.25	1.95

Each document is now represented by a real-valued vector of tf-idf weights $\in R^{|V|}$

- Independence assumption among terms

Documents as vectors

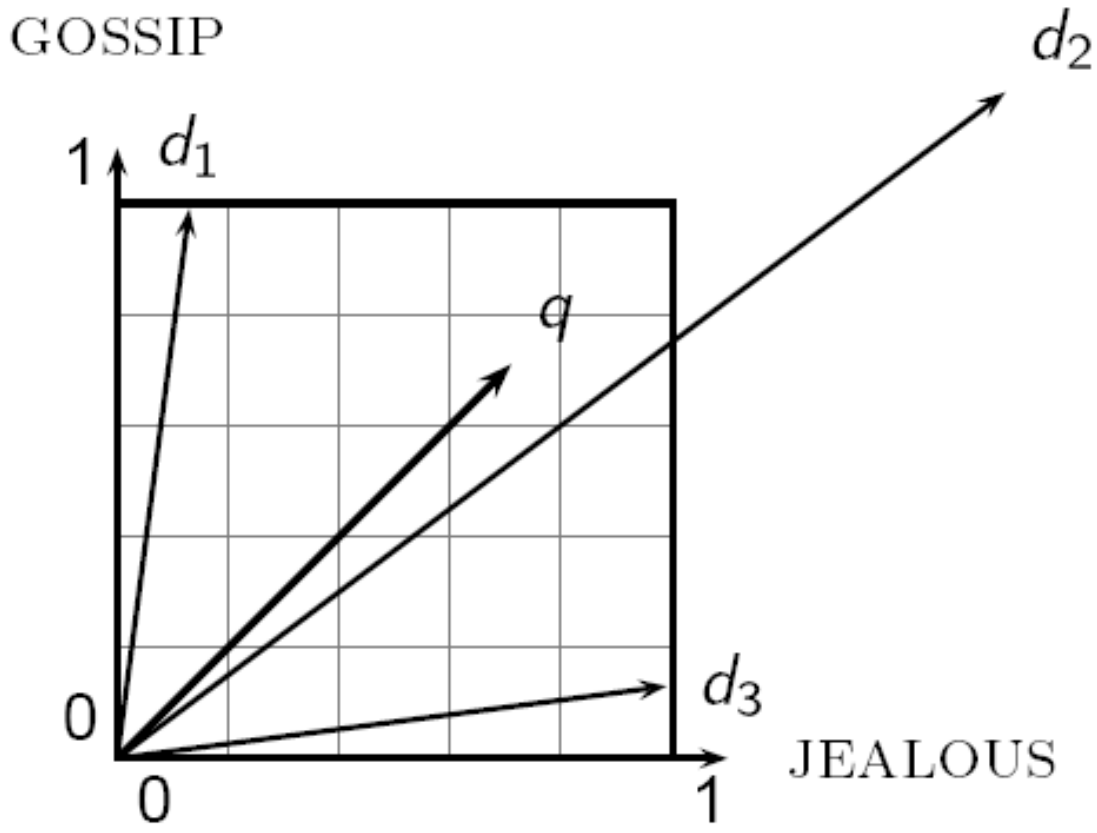
- So we have a |V|-dimensional vector space
- Terms are axes of the space
- Documents are points or vectors in this space
- Very high-dimensional: tens of millions of dimensions when you apply this to a web search engine
- These are very sparse vectors - most entries are zero.
- Vector representation doesn't consider the ordering of words in a document
- *John is quicker than Mary* and *Mary is quicker than John* have the same vectors
- This is called the bag of words model.

Formalizing vector space proximity

- First cut: distance between two points
  - (= distance between the end points of the two vectors)
- Euclidean distance?
- Euclidean distance is a bad idea . . .
- . . . because Euclidean distance is large for vectors of different lengths.

Why distance is a bad idea

The Euclidean distance between $\vec{q}$ and $\vec{d_2}$ is large even though the distribution of terms in the query $\vec{q}$ and the distribution of terms in the document $\vec{d_2}$ are very similar.



Use angle instead of distance

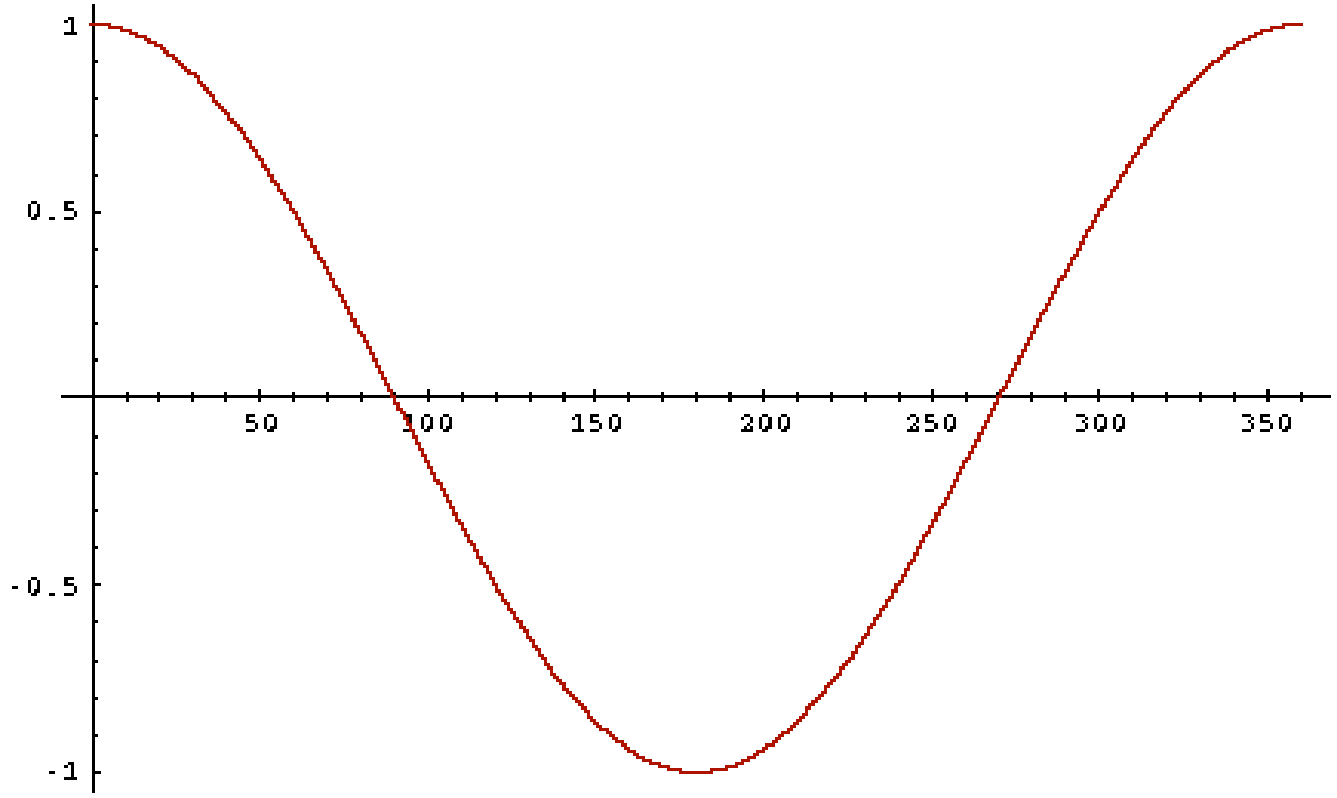
- Thought experiment: take a document *d* and append it to itself. Call this document *d*′.
- "Semantically" d and d′ have the same content
- The Euclidean distance between the two documents can be quite large
- The angle between the two documents is 0, corresponding to maximal similarity.

- Key idea: Rank documents according to angle with query.

From angles to cosines

- The following two notions are equivalent.
  - Rank documents in decreasing order of the angle between query and document
  - Rank documents in increasing order of cosine(query,document)
- Cosine is a monotonically decreasing function for the interval [0^o, 180^o]

From angles to cosines



- But how – *and why* – should we be computing cosines?

cosine(query,document)

Dot product

Unit vectors

$$\cos(\vec{q},\vec{d}) = \frac{\vec{q}\bullet\vec{d}}{|\vec{q}||\vec{d}|} = \frac{\vec{q}}{|\vec{q}|}\bullet\frac{\vec{d}}{|\vec{d}|} = \frac{\sum_{i=1}^{|V|} q_i d_i}{\sqrt{\sum_{i=1}^{|V|} q_i^2}\sqrt{\sum_{i=1}^{|V|} d_i^2}}$$

q_i is the tf-idf weight of term i in the query

d_i is the tf-idf weight of term i in the document

Length normalization

- A vector can be (length-) normalized by dividing each of its components by its length – for this we use the L_2 norm:

$$\|\vec{x}\|_2 = \sqrt{\sum_i x_i^2}$$

- Dividing a vector by its L_2 norm makes it a unit (length) vector (on surface of unit hypersphere)
- Effect on the two documents d and d′ (d appended to itself) from earlier slide: they have identical vectors after length-normalization.
  - Long and short documents now have comparable weights

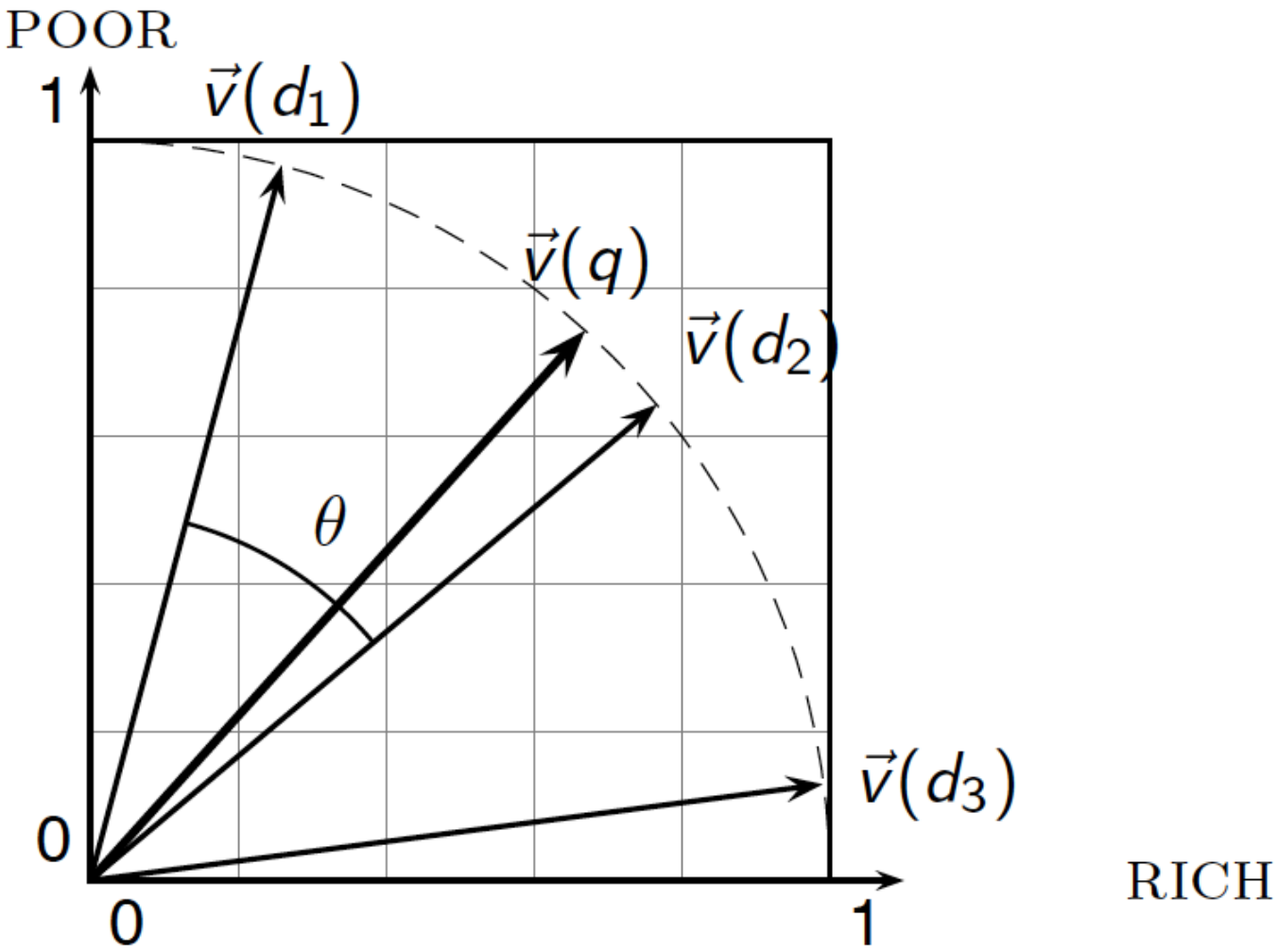
Cosine for length-normalized vectors

- For length-normalized vectors, cosine similarity is simply the dot product (or scalar product):

$$\cos(\vec{q},\vec{d}) = \vec{q} \bullet \vec{d} = \sum_{i=1}^{|V|} q_i d_i$$

for q, d length-normalized.

Cosine similarity illustrated



Cosine similarity amongst 3 documents

After length normalization

term	SaS	PaP	WH
affection	0.789	0.832	0.524
jealous	0.515	0.555	0.465
gossip	0.335	0	0.405
wuthering	0	0	0.588

cos(SaS,PaP) ≈

$0.789 \times 0.832 + 0.515 \times 0.555 + 0.335 \times 0.0 + 0.0 \times 0.0$

≈ 0.94

cos(SaS,WH) ≈ 0.79

cos(PaP,WH) ≈ 0.69

Naïve Implementation

Convert all documents in collection D to tf-idf weighted vectors, d_j, for keyword vocabulary V.

Convert query to a tf-idf-weighted vector q.

For each d_j in D do

    Compute score $s_j = \text{cosSim}(d_j, q)$

Sort documents by decreasing score.

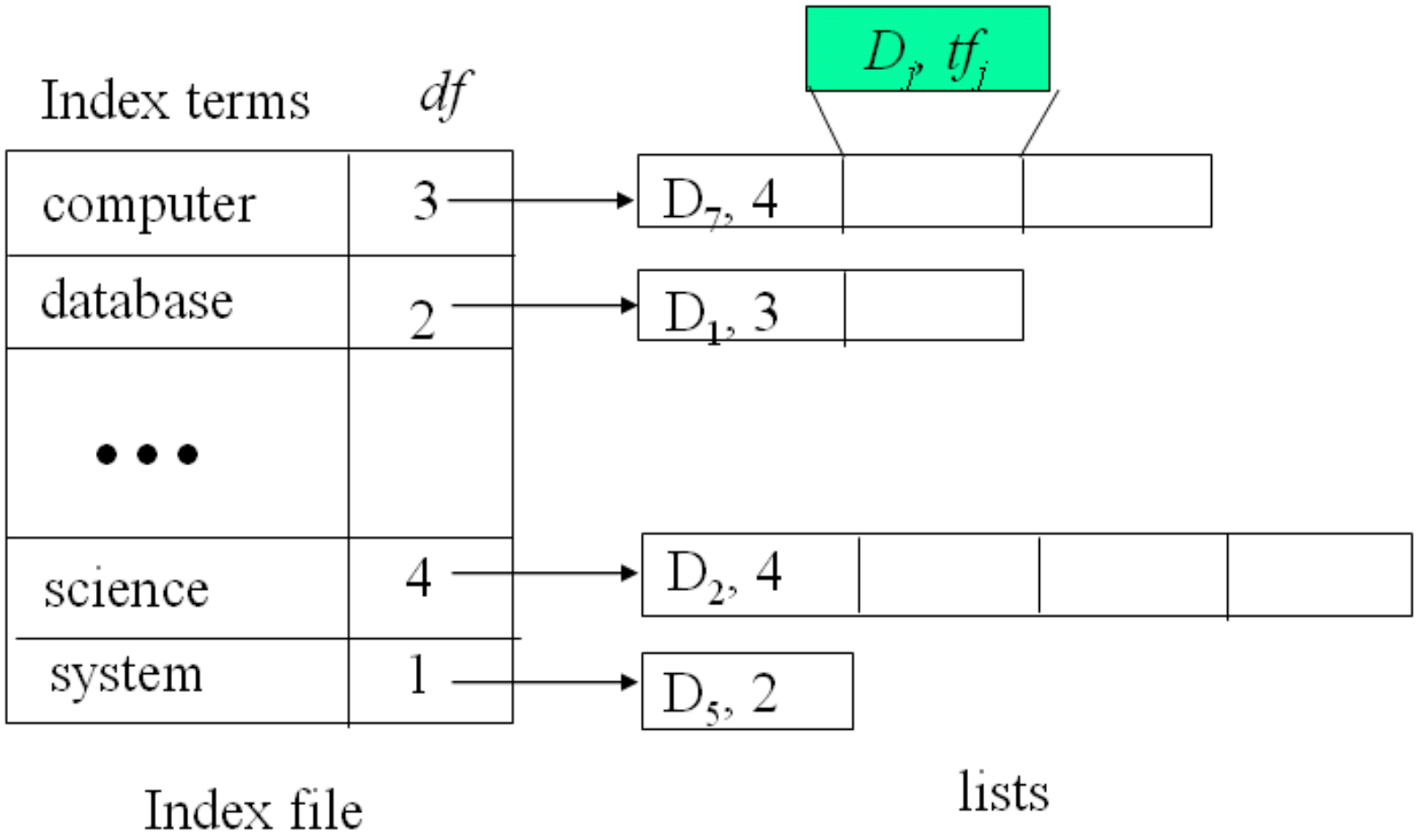
Present top ranked documents to the user.

Time complexity: $O(|V| \cdot |D|)$ Bad for large V & D !

$|V| = 10{,}000$; $|D| = 100{,}000$; $|V| \cdot |D| = 1{,}000{,}000{,}000$

- Faster?

3.3 向量表示与相似度计算



Computing cosine scores

```
CosineScore(q)
 1  float Scores[N] = 0
 2  float Length[N]
 3  for each query term t
 4  do calculate w_{t,q} and fetch postings list for t
 5      for each pair(d, tf_{t,d}) in postings list
 6      do Scores[d] += w_{t,d} × w_{t,q}
 7  Read the array Length
 8  for each d
 9  do Scores[d] = Scores[d]/Length[d]
10  return Top K components of Scores[]
```

3.4 Thesaurus及term自动关联

Very high-frequency terms?
Very low-frequency terms?

A term broadening step: stemming

Using associations between terms: a basic idea in improving the usefulness of index terms with questionable discrimination properties

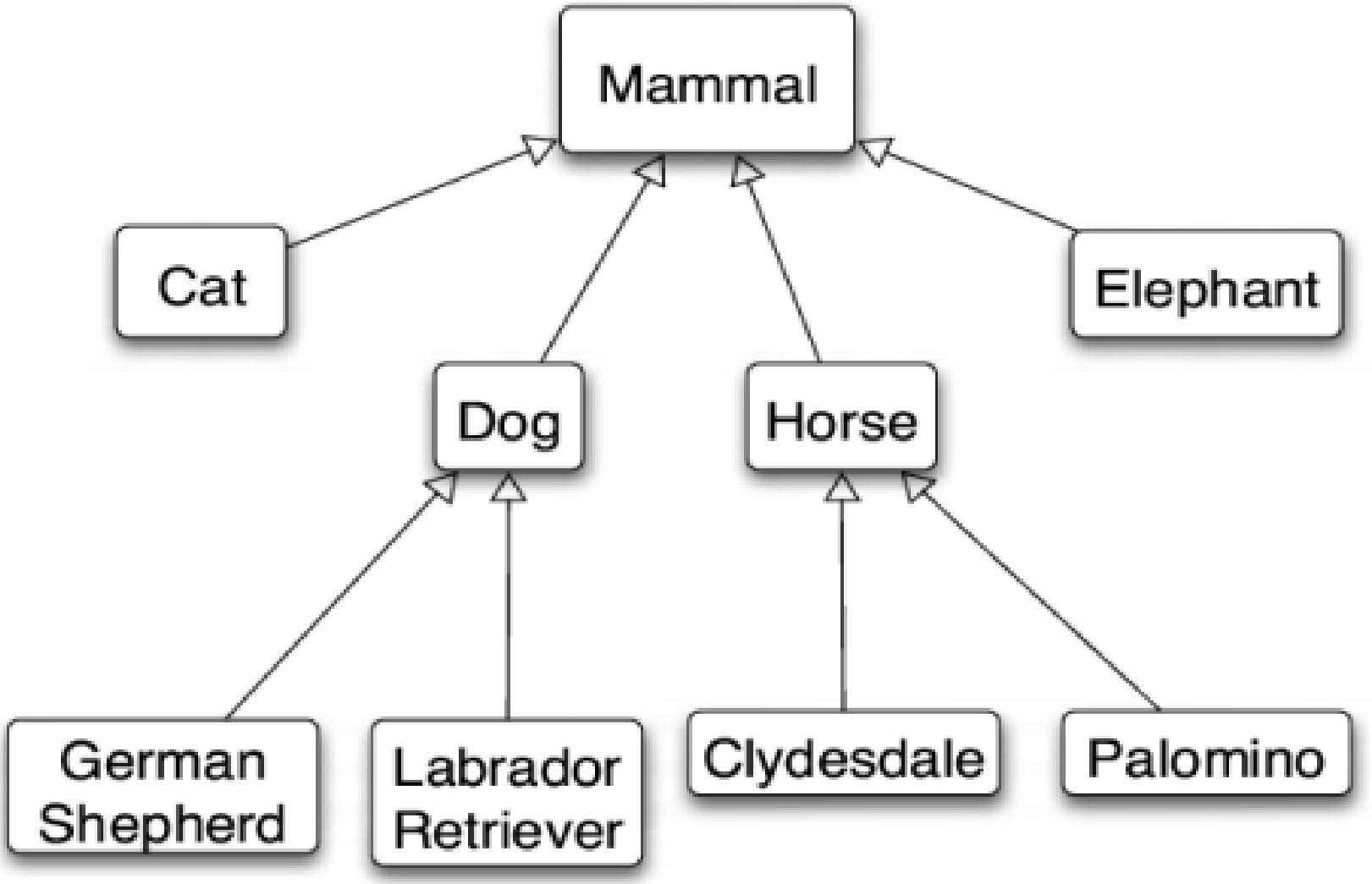
3.4 Thesaurus及term自动关联

Thesaurus

- thesaurus classes
  thesaurus category identifiers
  concept numbers
  car = automobile

- improving recall

3.4 Thesaurus及term自动关联



3.4 Thesaurus及term自动关联

- entity|实体
  - ├ thing|万物
    - … ├ physical|物质
      - … ├ animate|生物
        - … ├ AnimalHuman|动物
          - … ├ human|人
            - │ └ humanized|拟人
            - …
          - └ animal|兽
            - ├ beast|走兽
            - …

Thesaurus的树状层次结构

Change the scope of terms:
broader or narrower

"machine – computer – minicomputer"

3.4 Thesaurus及term自动关联

同义词、近义词（相似词）

医生 大夫

医生 护士

相关词

医生 病人

医生 手术刀

医生 医院

WordNet https://wordnet.princeton.edu/

What is WordNet
People
News
Use Wordnet Online
Download
Citing WordNet

What is WordNet?

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the creators of WordNet and do not necessarily reflect the views of any funding agency or Princeton University.

When writing a paper or producing a software application, tool, or interface based on WordNet, it is necessary to properly cite the source. Citation figures are critical to WordNet funding.

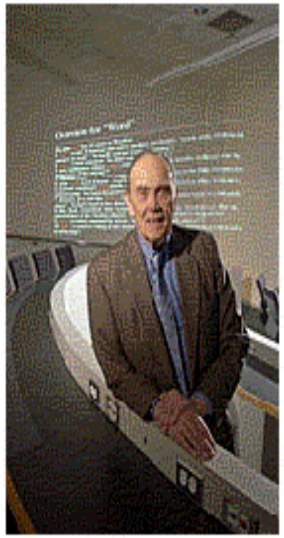
About WordNet

Note

Due to funding and staffing issues, we are no longer able to accept comment and suggestions.

We get numerous questions regarding topics that are addressed on

3.4 Thesaurus及term自动关联



George A. Miller

James S. McDonnell Distinguished University Professor of Psychology, Emeritus.

Princeton University
Department of Psychology
1-S-5 Green Hall
Princeton, NJ 08544

Phone: 609-258-5973 or 609-258-2972
Fax: 609-258-1113

Senior research psychologist and principal investigator of WordNet and Reader.

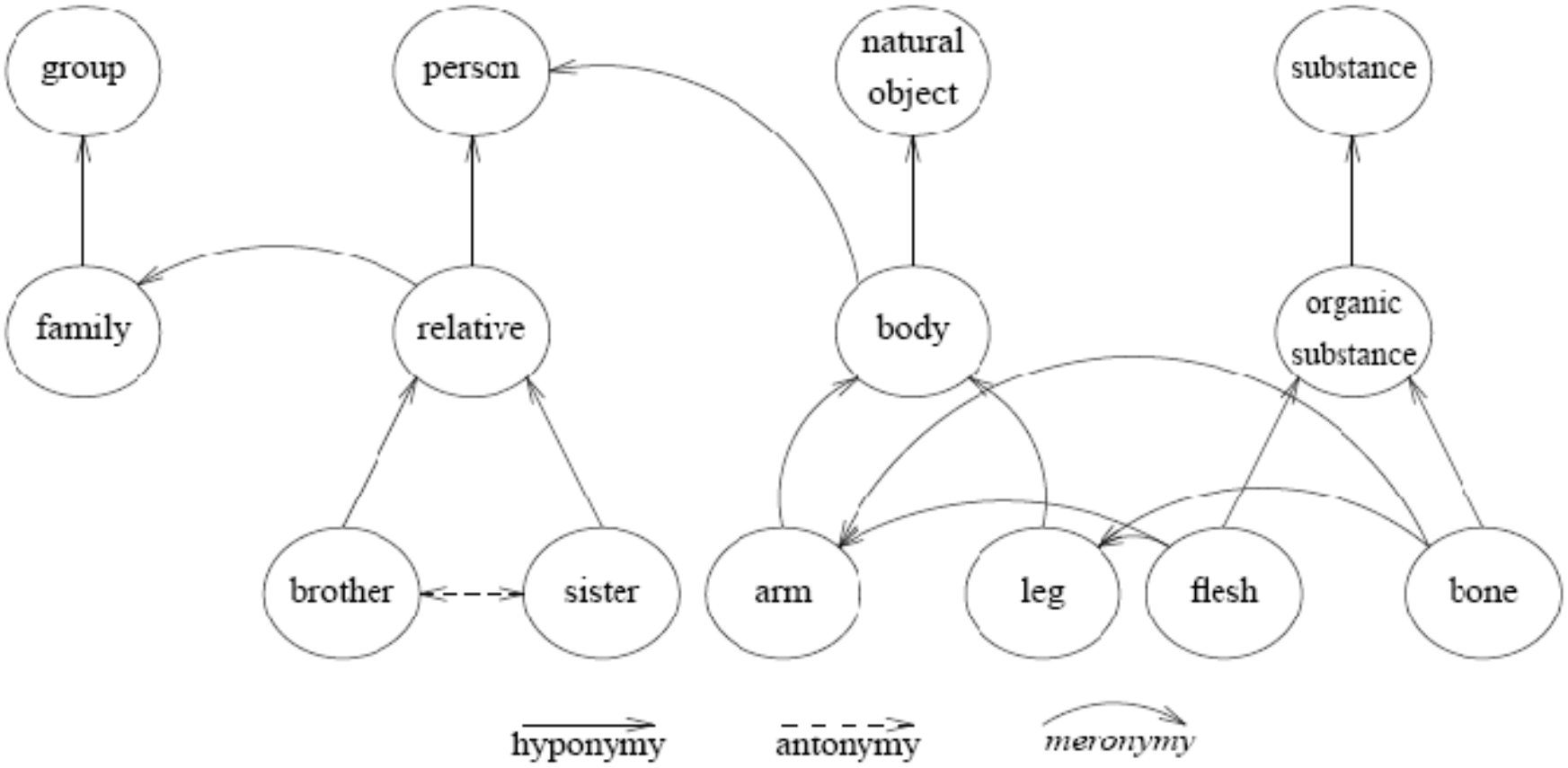
Author of The Magical Number Seven, Plus or Minus Two: Some Limits on our Capacity for Processing Information, (1956), *The Psychological Review.*

Recntly published in "TRENDS in Cognitive Sciences" is *The cognitive revolution: a historical perspective.*

http://psychclassics.yorku.ca/Miller/

3.4 Thesaurus及term自动关联

同义(synonymy)，反义(antonymy)，ISA（Hyponymy），PartOf（Meronymy）



Word to search for: bank Search WordNet

Display Options: (Select option to change) Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) **bank** (sloping land (especially the slope beside a body of water)) *"they pulled the canoe up on the bank"; "he sat on the bank of the river and watched the currents"*
- S: (n) depository financial institution, **bank**, banking concern, banking company (a financial institution that accepts deposits and channels the money into lending activities) *"he cashed a check at the bank"; "that bank holds the mortgage on my home"*
- S: (n) **bank** (a long ridge or pile) *"a huge bank of earth"*
- S: (n) **bank** (an arrangement of similar objects in a row or in tiers) *"he operated a bank of switches"*
- S: (n) **bank** (a supply or stock held in reserve for future use (especially in emergencies))
- S: (n) **bank** (the funds held by a gambling house or the dealer in some gambling games) *"he tried to break the bank at Monte Carlo"*
- S: (n) **bank**, cant, camber (a slope in the turn of a road or track; the outside is higher than the inside in order to reduce the effects of centrifugal force)
- S: (n) savings bank, coin bank, money box, **bank** (a container (usually with a slot in the top) for keeping money at home) *"the coin bank was empty"*
- S: (n) **bank**, bank building (a building in which the business of banking transacted) *"the bank is on the corner of Nassau and Witherspoon"*
- S: (n) **bank** (a flight maneuver; aircraft tips laterally about its longitudinal axis (especially in turning)) *"the plane went into a steep bank"*

Verb

- S: (v) **bank** (tip laterally) *"the pilot had to bank the aircraft"*
- S: (v) **bank** (enclose with a bank) *"bank roads"*
- S: (v) **bank** (do business with a bank or keep an account at a bank) *"Where do you bank in this town?"*

Word to search for: bank Search WordNet

Display Options: (Select option to change) Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) **bank** (sloping land (especially the slope beside a body of water)) *"they pulled the canoe up on the bank"; "he sat on the bank of the river and watched the currents"*
  - *direct hyponym* / *full hyponym*
    - S: (n) riverbank, riverside (the bank of a river)
    - S: (n) waterside (land bordering a body of water)
  - *direct hypernym* / ***inherited hypernym*** / *sister term*
    - S: (n) slope, incline, side (an elevated geological formation) *"he climbed the steep slope"; "the house was built on the side of a mountain"*
      - S: (n) geological formation, formation ((geology) the geological features of the earth)
        - S: (n) object, physical object (a tangible and visible entity; an entity that can cast a shadow) *"it was full of rackets, balls and other objects"*
          - S: (n) physical entity (an entity that has physical existence)
            - S: (n) entity (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))
  - *derivationally related form*
- S: (n) depository financial institution, **bank**, banking concern, banking company (a financial institution that accepts deposits and channels the money into lending

Word to search for: bank Search WordNet

Display Options: (Select option to change) Change

Key: "S:" = Show Synset (semantic) relations, "W:" = Show Word (lexical) relations

Display options for sense: (gloss) "an example sentence"

Noun

- S: (n) **bank** (sloping land (especially the slope beside a body of water)) *"they pulled the canoe up on the bank"; "he sat on the bank of the river and watched the currents"*
  - *direct hyponym* / *full hyponym*
    - S: (n) riverbank, riverside (the bank of a river)
    - S: (n) waterside (land bordering a body of water)
  - *direct hypernym* / ***inherited hypernym*** / *sister term*
    - S: (n) slope, incline, side (an elevated geological formation) *"he climbed the steep slope"; "the house was built on the side of a mountain"*
      - S: (n) geological formation, formation ((geology) the geological features of the earth)
        - S: (n) object, physical object (a tangible and visible entity; an entity that can cast a shadow) *"it was full of rackets, balls and other objects"*
          - S: (n) physical entity (an entity that has physical existence)
            - S: (n) entity (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))
  - *derivationally related form*
- S: (n) depository financial institution, **bank**, banking concern, banking company (a financial institution that accepts deposits and channels the money into lending

WordNet 3.0 database statistics

Number of words, synsets, and senses

POS	Unique Strings	Synsets	Total Word-Sense Pairs
Noun	117798	82115	146312
Verb	11529	13767	25047
Adjective	21479	18156	30002
Adverb	4481	3621	5580
Totals	155287	117659	206941

Polysemy information

POS	Monosemous Words and Senses	Polysemous Words	Polysemous Senses
Noun	101863	15935	44449
Verb	6277	5252	18770
Adjective	16503	4976	14399
Adverb	3748	733	1832
Totals	128391	26896	79450

POS	Average Polysemy Including Monosemous Words	Average Polysemy Excluding Monosemous Words
Noun	1.24	2.79
Verb	2.17	3.57
Adjective	1.40	2.71
Adverb	1.25	2.50

https://wordnetcode.princeton.edu/5papers.pdf

Five Papers on WordNet 前两篇（25页之前）必读，后三篇选读