# Document Representation 

Jaime Arguello<br>INLS 509: Information Retrieval<br>jarguell@email.unc.edu

## Document Representation



## Document Representation

## - How should this document be represented?



## Main page

Contents
Featured content
Current events
Random article
Donate to Wikipedia

- Interaction

Help
About Wikipedia
Community portal
Recent changes
Contact Wikipedia

- Toolbox
- Print/export
- Languages

Deutsch
Español
Bahasa Indonesia
Article
Discussion
Read Edit View history

## Gerard Salton

From Wikipedia, the free encyclopedia

Gerard Salton (8 March 1927 in Nuremberg - 28 August 1995), also known as Gerry Salton, was a Professor of Computer Science at Cornell University. Salton was perhaps the leading computer scientist working in the field of information retrieval during his time. His group at Cornell developed the SMART Information Retrieval System, which he initiated when he was at Harvard.

Salton was born Gerhard Anton Sahlmann on March 8, 1927 in Nuremberg, Germany. He received a Bachelor's (1950) and Master's (1952) degree in mathematics from Brooklyn College, and a Ph.D. from Harvard in Applied Mathematics in 1958, the last of Howard Aiken's doctoral students, and taught there until 1965, when he joined Cornell University and co-founded its department of Computer Science.
Salton was perhaps most well known for developing the now widely used Vector Space Model for Information Retrieval ${ }^{[1]}$. In this model, both documents and queries are represented as vectors of term counts, and the similarity between a document and a query is given by the cosine between the term vector and the document vector. In this paper, he also introduced TF-IDF, or term-frequency-inverse-document frequency, a model in which the score of a term in the a document is the ratio of the number of terms in that document divided by the frequency of the number of documents in which that term occurs. (The concept of inverse document frequency, a measure of specificity, had been introduced in 1972 by Karen Sparck-Jones ${ }^{[2]}$.) Later in life, he became interested in automatic text summarization and analysis ${ }^{[3]}$, as well as automatic hypertext generation ${ }^{[4]}$. He published over 150 research articles and 5 books during his life.

Salton was editor-in-chief of the Communications of the ACM and the Journal of the ACM, and chaired SIGIR. He was an associate editor of the ACM Transactions on Information Systems. He was an ACM Fellow (elected 1995), received an Award of Merit from the American Society for Information Science (1989), and was the first recipient of the SIGIR Award for outstanding contributions to study of information retrieval (1983) -- now called the Gerard Salton Award.

## Elements of a Document Representation

- Document attributes (metadata)
- source, publication date, language, length, etc.
- Controlled vocabulary index terms
- Free-text index terms
- terms selected from the document text itself
- may also include text from outside the document (e.g., anchor text)
- lots of room for creativity!


## Elements of a Document Representation

ticle Discussion
Read Edit View history
a Log in / create accoun

From Wkipedia, the tree encyclope
erard Salton (8 March 1927 in Nuremberg - 28 August 1995), also known as Gerry Salton, was a Professor of Intormation rettieval during his time. His group at Correll developedt the SMART Intormation Retrieval System, which he initiated when he was at Havard.
Satoon was borm Germard Anton Sahlmann on March 8, 1927 in Nuremberg, Germany. He received a Bachelor's (1950)
and Master's (1952) degree in mathematics trom Brookyn College, and a Ph.D. from Havard in Applied Mathematios in and Master's (1952) degree in mathematicics trom Brook kn College, and a Ph.D. from Harard in Applied Mathematics in co-tounded is department of Computer Science.
Salton was permaps most well known for developing the now widely used Vector Space Model for Intormation Retrieval. In this model, both documents and queries are represented as vectors of term counts, and the similarity between a document and a query is given by the cosine between the term vector and the document vector. In this paper, he also
introduced TF-IF, or term-ftequencery-iverse-document trequency a model in which the score of a term in the a Intoduced TF-IDF, or term-triequencry-inverse-document trequency, a model in which the score of a term in the a which that term occurs. (The concept tof inverse document trequency, a measure of specificty, had been introduced in 1972 by Karen Sparck.-Jonene ${ }^{\left[{ }^{[2}\right]}$. LLater in ilie, he became interested in automatic text summarization and analysis ${ }^{(3)]}$, a well as automatic hypertext generation ${ }^{46}$. He published over 150 research articles and 5 books during his life. Salton was editor-in-Chief of the Communications of the ACM and the Journal of the ACM, and chaired SIIIIR. He was an
associate edito of the ACM Transactions on Intormation Systems. He was an ACM Fellow elected 1995), recived an Award of Merit trom the American Society for Intormation Science (1989), and was the first reciiient of the SIIGIR Award for outstanding contributions to study of intormation rettieval (1983) - now called the Gerard Salton Award.

## controlledvocabulary index terms

Categories: 1927 births I 1995 deaths I American computer scientists I Computer pioneers I Harvard University alumni I Harvard University faculty I Cornell University faculty I Fellows of the Association for Computing Machinery I Guggenheim Fellows

## Elements of a Document Representation

Aricle Discussion
Read Edit View history
Gerard Sat
Gerard Salton
rom Wikipedia, the tree encyclopedia
Serard Salton (8 March 1927 in Nuremberg - 28 August 1995), also known as Gerry Satton, was a Protessor of Computer Science at Cornell Univerity. Salton was perhaps the leading computer scientist working in the field of
intormation retrieval durina his time. His srouv at Correll develooned the SMART Intormation Retrieval Svstem. which he
nititated when he wa
Salton was bor Ger nd Master's (1952)
958, the last of Hov 958, the last of Hov co-tounded its depar Sation was permaps
Satoo was permaps
nthis model, both d
cocument and a que
Introduced TF-IDF, 0
which that term occu
1972 by Karen Sparn
well as automatichy
well as automatic hy
Salton was editorin-
Award of Merit from
Tor outstanding contt Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia

- Interaction

Help
About Wikipedia
Community portal
Recent changes
Contact Wikipedia
p Toolbox

- Print/export
- Languages

Français
Русский
a Log in / create accoun

## anchor text (nearby terms?)

Log in / create accoun

Article Discussion

## Amit Singhal

Read View source View history Search
Q

From Wikipedia, the free encyclopedia
Amit Singhal is a software engineer at Google Inc., a Google Fellow, and the head of Google's core ranking team. ${ }^{[1]}$

```
Contents [hide]
1 Education
2 Career
3 References
4 External links
```

Education

Born in Jhansi, a city in the state of Uttar Pradesh, India, ${ }^{[2]}$ Amit received a Bachelor of Engineering degree in computer science from IIT Roorkee in $1989 .{ }^{[3]}$ He continued his computer science education in the United States and received an M.S. degree from University of Minnesota Duluth in 1991. ${ }^{[4]}$ He writes about UMD:
"UMD was the turning point in my life. Studying Information Retrieval with Don Crouch and then Don recommending that I move to Cornell to study with Gerard Salton, is the main reason behind my success today. Don gave me the love for search, I have just followed my passion ever since. ${ }^{[4]}$ -Amit Singhal

[^0]
## Text Processing

gerard salton 8 march 1978 in nuremberg 28 august 1995 also know as gerry salton was professor of computer science at cornell university salton was perhaps the leading computer scientist working in the field of information retrieval during his time his group at cornell developed the smart information retrieval system which he initiated when he was at harvard

- Our goal is to describe content using content
- After mark-up removal, down-casing, and tokenization, what we have is a sequence of terms
- What are the most descriptive words?

|  | Term-Frequencies <br> top 20 |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
| rank | term | freq. | rank | term | freq. |
| 1 | the | 34 | 11 | as | 9 |
| 2 | of | 29 | 12 | he | 9 |
| 3 | a | 20 | 13 | vector | 8 |
| 4 | in | 20 | 14 | an | 8 |
| 5 | and | 19 | 15 | s | 7 |
| 6 | salton | 18 | 16 | term | 7 |
| 7 | model | 15 | 17 | for | 7 |
| 8 | was | 12 | 18 | automatic | 7 |
| 9 | information | 11 | 19 | paper | 6 |
| 10 | retrieval | 10 | 20 | gerard | 6 |


|  | Term-Frequencies <br> top 20 |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: |
| rank | term | freq. | rank | term | freq. |
| I | the | 34 | 11 | as | 9 |
| 2 | of | 29 | 12 | he | 9 |
| 3 | a | 20 | 13 | vector | 8 |
| 4 | in | 20 | 14 | an | 8 |
| 5 | and | 19 | 15 | s | 7 |
| 6 | salton | 18 | 16 | term | 7 |
| 7 | model | 15 | 17 | for | 7 |
| 8 | was | 12 | 18 | automatic | 7 |
| 9 | information | 11 | 19 | paper | 6 |
| 10 | retrieval | 10 | 20 | gerard | 6 |

## IMDB Corpus <br> term-frequencies

| rank | term | frequency | rank | term | frequency |
| :---: | :---: | :---: | :---: | :---: | :---: |
| I | the | 1586358 | $I I$ | year | 25015 I |
| 2 | a | 854437 | 12 | he | 242508 |
| 3 | and | 822091 | 13 | movie | 24155 I |
| 4 | to | 804137 | 14 | her | 240448 |
| 5 | of | 657059 | 15 | artist | 236286 |
| 6 | in | 472059 | 16 | character | 234754 |
| 7 | is | 395968 | 17 | cast | 234202 |
| 8 | i | 390282 | 18 | plot | 234189 |
| 9 | his | 328877 | 19 | for | 207319 |
| 10 | with | 253153 | 20 | that | 197723 |

## Stopwords

- A stopword is a term that is discarded from the document representation
- Typically the same set of stopwords is used in processing all documents in the collection
- Stopwords are typically function words: determiners (a, the), prepositions (on, above), conjunctions (and, but)
- May also be corpus-specific: "plot" in the IMDB corpus
- Assumption: stopwords are unimportant because they are frequent in every document


## Lemur Stopword List first 60 (sorted alphabetically)

| a | all | amongst | anywhere | become | besides |
| :---: | :---: | :---: | :---: | :---: | :---: |
| about | almost | an | apart | becomes | between |
| above | alone | and | are | becoming | beyond |
| according | along | another | around | been | both |
| across | already | any | as | before | but |
| after | also | anybody | at | beforehand | by |
| afterwards | although | anyhow | av | behind | can |
| again | always | anyone | be | being | can |
| against | am | anything | became | below | cannot |
| albeit | among | anyway | because | beside | canst |

## Term-Frequencies after stopword removal

| rank | term | freq. | rank | term | freq. |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | salton | 18 | 11 | paper | 6 |
| 2 | model | 15 | 12 | document | 6 |
| 3 | information | 11 | 13 | acm | 6 |
| 4 | retrieval | 10 | 14 | 1975 | 4 |
| 5 | vector | 8 | 15 | frequency | 4 |
| 6 | s | 7 | 16 | science | 4 |
| 7 | term | 7 | 17 | cornell | 4 |
| 8 | automatic | 7 | 18 | award | 3 |
| 9 | gerard | 6 | 19 | 0 | 3 |
| 10 | space | 6 | 20 | 8 | 3 |

## Creating a Stopword List

- Sort vocabulary based on frequency in the corpus
- Examine the most frequent words
- Examine a query-log to see which frequent terms may be important
- $38 \%$ of unique AOL queries contain at least one stopword (may or may not be important - more later)
- $0.025 \%$ of unique AOL queries are nothing but stopwords
- longest: i want something else to get me through this (third eye blind lyrics)


## Trends in Stopword Removal

- The earliest systems used stopword lists of 200-300 terms
- To improve efficiency and effectiveness
- Very frequent terms were problematic for early retrieval models (e.g, OR operations in ranked boolean)
- Web search engines generally do not remove stopwords
- The latest trend is to index stopwords and (possibly) ignore them at query-time if they seem unimportant
- Newer retrieval models are better at handling very frequent terms (later lecture)


## Document Representation



## AOL Query-Log Examples stopword removal

wrong lyrics
am i wrong lyrics
i was wrong lyrics
wrong again lyrics
where did i go wrong lyrics wrong lyrics
got me wrong lyrics
what went wrong lyrics
buy house
who will buy my house
buy a house
buy my house
buy house
we buy house
how to buy a house

## change

be the change you want in others
how can i change me
change
where is my change
i want my change
never change
calculate bmi
calculate bmi
calculate my bmi
how to calculate your bmi
how to calculate bmi

## Morphological Analysis

## Morphology

- the study and description of word formation (as inflection, derivation, and compounding) in language


## Merriam-Webster Dictionary

## Morphology

- Inflectional morphology: changes to a word that encode its grammatical usage (e.g., tense, number, person)
- say vs. said, cat vs. cats, see vs. sees
- Derivational morphology: changes to a word to make a new word with related meaning
- organize, organization, organizational
- Compounding: combining words to form new ones
- shipwreck, outbound, beefsteak
- more common in other languages (e.g., german)
- lebensversicherungsgesellschaftangestellter


## Morphological Analysis

 in information retrieval- Basic question: words occur in different forms. Do we want to treat different forms as different index terms?
- Conflation: treating different (inflectional and derivational) variants as the same index term


## Morphological Analysis

 in information retrieval- Conflation: treating different (inflectional and derivational) variants as the same index term

| image | images | imaging | imag* (root form) |
| :---: | :---: | :---: | :---: |
| $d f=6$ | df=4 | $d f=3$ | $d f=6$ |
| I, 4 | I, 4 | I, 4 | I, 12 |
| 10, 1 | 10, 5 | 10, 5 | 10,11 |
| 15, 2 | 16, I | 16, I | 15,2 |
| 16, 1 | 68, I |  | 16, 3 |
| 33, 5 |  |  | 33, 5 |
| 68, 7 |  |  | 68, 8 |

## Morphological Analysis in information retrieval

- What are we trying to achieve by conflating morphological variants?
- Goal: help the system ignore unimportant variations of language usage


## Morphological Analysis in information retrieval

Guide to Computer Troubleshooting and Repair - PC ...
www.daileyint.com/hmdpc/manual.htm - Cached
PC's are actually much easier to repair these days than in the early 90 's when I wrote my original guide for technicians I was training. The number of discrete ...

Online Computer Training Courses - For all beginners and experts ... www.beyourownit.com/ - Cached
Want to learn about computers $\$$ You've found the right placel On thic website, you'll be able to find everything from simple computer repair articles and computer ..

Repairing basic computer hard ware problem (system disk failure ... www.instructables.com/.../Repairing-basic-computer-hard-ware-and... - Cached May 10, 2008 - THIS GUIDE IS NOT YET FINISHED, I WILL ADD MORE INFORMATION WHEN I GET A CHANCE.If you need any help with fixing a computer ...

## Computer-Repair Technicians

www.collegeboard.com ) ... ) Majors \& Careers Central, Profiles - Cached
Computer-repair technicians maintain and repa $\mathbf{r}$ computers, scanners, printers, monitors, and other computer equipment. Learn more about this career at ...

Computer repair NYC | Laptop repair ny | PC repair NY
ifixny.com/ - Cached
Computer repair ny. Data recovery nyc. We offers a full range of computer fix and technical support with free diagnostics and estimates, also iPhone BlackBerry ...

## Morphological Analysis in information retrieval

- The query "computer repairs" will match all combinations of:

computer computers computing<br>computation<br>computational<br>and

repair<br>repairs<br>repaired<br>repairing<br>repairable

## Morphological Analysis in information retrieval

- In English, conflating morphological variants is usually done using a stemmer
- Stemming: automatic suffix-stripping
- English word variations occur at the end of a word
- root/stem + suffix
- repair + s/ed/ing/able
- A stemmer conflates different variations by reducing them to a common root/stem


## Morphological Analysis

 in information retrieval- In some cases, whatever is left after suffix-stripping is not even a word (e.g., comput)
- Is this a problem?

computer<br>computers<br>computing<br>computation<br>computational

$\because:$
> repair
> repairs
> repaired
> repairing
> repairable
> $\because$

## Morphological Analysis



## Morphological Analysis

 the porter stemmer (porter '80)- A long list of rules that are applied in sequence
- apply the rule that removes the longest suffix
- check to see that the stem is likely to be a root (replac+ement vs. c+ement)
- Fast, effective, and, therefore, very popular


## Martin Porter's Home Page

No doubt you came here out of idle curiosity from the Porter Stemming Algorithm page. Before you hastily return, you are welcome to look at the following.

This (jerkily) spinning can is the work of Philip Holmes Esquire, ingenious graphic designer and inventor of visual puns. I could never have thought up anything so clever. (Apologies to the Dr Pepper people!)


## Morphological Analysis the porter stemmer (porter '80)

## - Example step (1 of 5)

## Step 1a:

- Replace sses by ss (e.g., stresses $\rightarrow$ stress).
- Delete $s$ if the preceding word part contains a vowel not immediately before the $s$ (e.g., gaps $\rightarrow$ gap but gas $\rightarrow$ gas).
- Replace $\boldsymbol{i e d}$ or $\boldsymbol{i e s}$ by $\boldsymbol{i}$ if preceded by more than one letter, otherwise by $\boldsymbol{i e}$ (e.g., ties $\rightarrow$ tie, cries $\rightarrow$ cri).
- If suffix is $\boldsymbol{u s}$ or $\boldsymbol{s} \boldsymbol{s}$ do nothing (e.g., stress $\rightarrow$ stress).

Step 1b:

- Replace eed, eedly by ee if it is in the part of the word after the first non-vowel following a vowel (e.g., agreed $\rightarrow$ agree, feed $\rightarrow$ feed).
- Delete ed, edly, ing, ingly if the preceding word part contains a vowel, and then if the word ends in $\boldsymbol{a} \boldsymbol{t}, \boldsymbol{b} \boldsymbol{l}$, or $\boldsymbol{i z}$ add $\boldsymbol{e}$ (e.g., fished $\rightarrow$ fish, pirating $\rightarrow$ pirate), or if the word ends with a double letter that is not $\boldsymbol{l l}, \boldsymbol{s s}$ or $\boldsymbol{z} \boldsymbol{z}$, remove the last letter (e.g., falling $\rightarrow$ fall, dripping $\rightarrow$ drip), or if the word is short, add $\boldsymbol{e}$ (e.g., hoping $\rightarrow$ hope).
- Whew!


## Morphological Analysis the porter stemmer (porter '80)

## - Original Text

gerard salton 8 march 1978 in nuremberg 28 august 1995 also know as gerry salton was professor of computer science at cornell university salton was perhaps the leading computer scientist working in the field of information retrieval during his time his group at cornell developed the smart information retrieval system which he initiated when he was at harvard

- Stemmed Text
gerard salton 8 march 1978 in nuremberg 28 august 1995 also know as gerri salton wa professor of comput scienc at cornel univers salton wa perhap the lead comput scientist work in the field of inform retriev dure hi time hi group at cornel develop the smart inform retriev system which he initi when he wa at harvard


## Morphological Analysis the porter stemmer (porter '80)

- false positives: two words conflated to the same root when they shouldn't have been

organization/organ<br>generalization/generic<br>numerical/numerous<br>policy/police<br>university/universe<br>addition/additive<br>negligible/negligent<br>execute/executive<br>past/paste<br>ignore/ignorant<br>special/specialized<br>head/heading

## Morphological Analysis the porter stemmer (porter '80)

- false negatives: two words not conflated to the same root word when they should have been

european/europe<br>cylinder/cylindrical<br>matrices/matrix<br>urgency/urgent<br>create/creation<br>analysis/analyses<br>useful/usefully<br>noise/noisy<br>decompose/decomposition<br>sparse/sparsity<br>resolve/resolution<br>triangle/triangular

## AOL Query-log Examples stemmed queries

russian translat
russian translations russian translator russian translation russian translate

## secret

secret
secretions
secrets
secretion
stock for sale
stockings for sale stocking for sale stocks for sale
smokei mountain nation park smokey mountains national park smokey mountain national park smokey mountains national parks
cat fenc
cat fencing
cat fences
cat fence
strawberri plant strawberry planting strawberry plants
strawberries planting

# AOL Query-log Examples stopped + stemmed queries 

bui comput
buy a computer
buying a computer
we buy computers
how to buy a computer buying computers
rid raccoon
get rid of raccoons
how to get rid of raccoons
how to get rid of a raccoon
what to use to get rid of raccoons how do i get rid of a raccoon

## auto repair

auto repairables
how to auto repairs
auto repair do it yourself do it yourself auto repair auto repair .com do it yourself auto repairs auto repair
water diet
the water diet
the all water diet
water and diet
water diet
water diets

## AOL Query-log Examples stopped + stemmed queries

planet orbit sun
why is there only one planet in each orbit around the sun why do the planets orbit the sun
planets that orbit the sun
univers
plant shade
plant shade
plants for shade
plants that do well in shade plants that like shade plants shade planting in the shade
universalism
universism
other universe
university
our universe
across the universe the universe within universities

## Morphological Analysis evaluation results

- Stemming
- English: 0-5\% improvements
- Finnish: 30\% improvement
- Spanish: $10 \%$ improvement
- Compound Splitting
- German: $15 \%$ improvements
- Swedish: $25 \%$ improvement


## Morphology Across Languages European Parliament Corpus

- Number of unique terms (remember, these are translations of the same text):
- English: 150,725
- Spanish: 213,486
- Portuguese: 219,121
- Danish: 367,282
- Finnish: 709,049
- German: 401,929


## To Stem or Not To Stem



## To Stem or Not To Stem



## Big Picture

- Text-processing requires making decisions about what to store in the index
- Two big decisions: stopword-removal and stemming
- My own recommendation (take it, leave it, question it)
- remove stopwords only if you have to (don't have enough disk-space)
- off-load the job to query-processing (removing stopwords from the query)
- stem depending on the importance of recall and the size of the collection


## What about homonyms?

(words that are spelled the same, but have different meaning)

## Words often have multiple senses

- bank (noun)

1. the rising ground bordering a lake, river, or sea
2. a mound, pile, or ridge above the surrounding level
3. a steep slope (as in "bank of a hill")
4. an establishment for the custody, loan, exchange, and issue of money
5. a supply of something held in reserve
6. the lateral inward tilt of a vehicle (as an airplane) when turning

## Word Sense Disambiguation

- Given a word in a particular context, automatically predict its correct sense from a finite set (bank 1-6)?
"I stopped by the bank to deposit some cash."

An establishment for the custody, loan, and exchange of money
"I stopped by the food bank to donate some food."

## A supply of something held in reserve

- An active area of research since the 1950's
- How would you do this?


## Word Sense Disambiguation

- Predict the sense whose definition contains terms that co-occur often with those in the surrounding context
"I stopped by the bank to deposit some cash."

An establishment for the custody, loan, and exchange of money

| mutual information from IMDB corpus | money debt dollars | raise money money | $\begin{aligned} & 2.686 \\ & 2.578 \\ & 2.567 \\ & \hline \end{aligned}$ |
| :---: | :---: | :---: | :---: |
|  | money | cash | 2.546 |
|  | buy | money | 2.471 |
|  | money | gambling | 2.436 |
|  | money | pay | 2.427 |
|  | money | bank | 2.387 |
|  | insurance | money | 2.117 |
|  | money | paid | 2.018 |

## Word Sense Disambiguation in information retrieval

1. Expand the indexed vocabulary so that each sense of a word is a different index term
2. Automatically predict the correct sense for each word in the collection (e.g, bank¹, bank ${ }^{2}$, ... , bank ${ }^{6}$ )

- lots of context (i.e., surrounding text)

3. Index the collection as usual
4. At query-time, predict the correct word sense in the query (e.g., "drive-through bank ${ }^{4}$ carrboro")

- more difficult, not much context

5. Retrieve documents as usual

## Word Sense Disambiguation in information retrieval

- Does it improve (average) retrieval effectiveness?


# Word Sense Disambiguation in information retrieval 

- Not much. Why not?


## Word Sense Disambiguation in information retrieval

- Not really a problem for long-queries (other query terms disambiguate the ambiguous ones)
- In theory, can improve performance for short queries
- However, these are precisely the queries for which disambiguation is the most difficult (not much context)


## Word Sense Disambiguation in information retrieval

- There is another reason. What is it?


## Word Sense Disambiguation in information retrieval

united bank<br>union bank california<br>union bank<br>tyra banks show<br>star bank republic bank<br>pnc bank<br>people bank

outer banks north carolina outer banks nc
online banking bank america national bank texas commerce bank
national bank south carolina national bank oneida national bank omaha national bank marin national bank alaska national bank merchants bank loans bank account hotels outer banks nc hotels outer banks guaranty bank freedom bank
farmers merchants bank

# Word Sense Disambiguation in information retrieval 

- Wait for it..., Wait for it...


## Word Sense Disambiguation in information retrieval

- Word senses also (more or less) follow Zipf's law: a few are very frequent and most a rare

united bank<br>union bank california<br>union bank<br>tyra banks show<br>star bank<br>republic bank<br>pnc bank<br>people bank<br>outer banks north carolina<br>outer banks nc<br>online banking bank america<br>national bank texas<br>commerce bank

national bank south carolina national bank oneida national bank omaha national bank marin national bank alaska national bank merchants bank loans bank account hotels outer banks nc hotels outer banks guaranty bank freedom bank farmers merchants bank

## Word Sense Disambiguation in information retrieval

$\left.\begin{array}{cc|rc}\begin{array}{c}\text { No. of } \\ \text { senses }\end{array} & \begin{array}{c}\text { Size } \\ \text { of }\end{array} & \begin{array}{l}\text { set }\end{array} & \begin{array}{l}\text { Most } \\ \text { sense }\end{array} \\ \hline \text { common }\end{array}\right)$

Table 10. Percentage of occurrences accounted for by the most common sense of a word. The figures in brackets (shown for comparison) is the percentage that would result if senses occurred $i$ equal amounts. Measurements made on the SEMCOR corpus.


[^0]:    Amit continued his studies at Cornell University in Ithaca, New York and received a Ph.D. degree in $1996 .{ }^{[4]}$ At Cornell Amit studied vith Gerard Salton, a poneer in the field of information retrieval, the academic discipline which forms the foundation of modern search. John Battelle, in his book "The Search" calls Gerard Salton "the father of digital search."

