UNIVERSITY OF PORTSMOUTH

Intelligent Data and Text Analytics

Text Mining

Part 1

Text Mining

- What is text mining?
- Examples of existing applications using text mining
- Text representation – Vector space model
- Natural Language processing

What is "Text Mining"?

- "*Text mining, also referred to as* ***text data mining****, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text.*" - wikipedia
- *"Another way to view text data mining is as a process of* ***exploratory*** *data analysis that leads to* ***heretofore unknown*** *information, or to answers for questions for which the answer is not currently known."* - Hearst, 1999

Knowledge discovery from text data

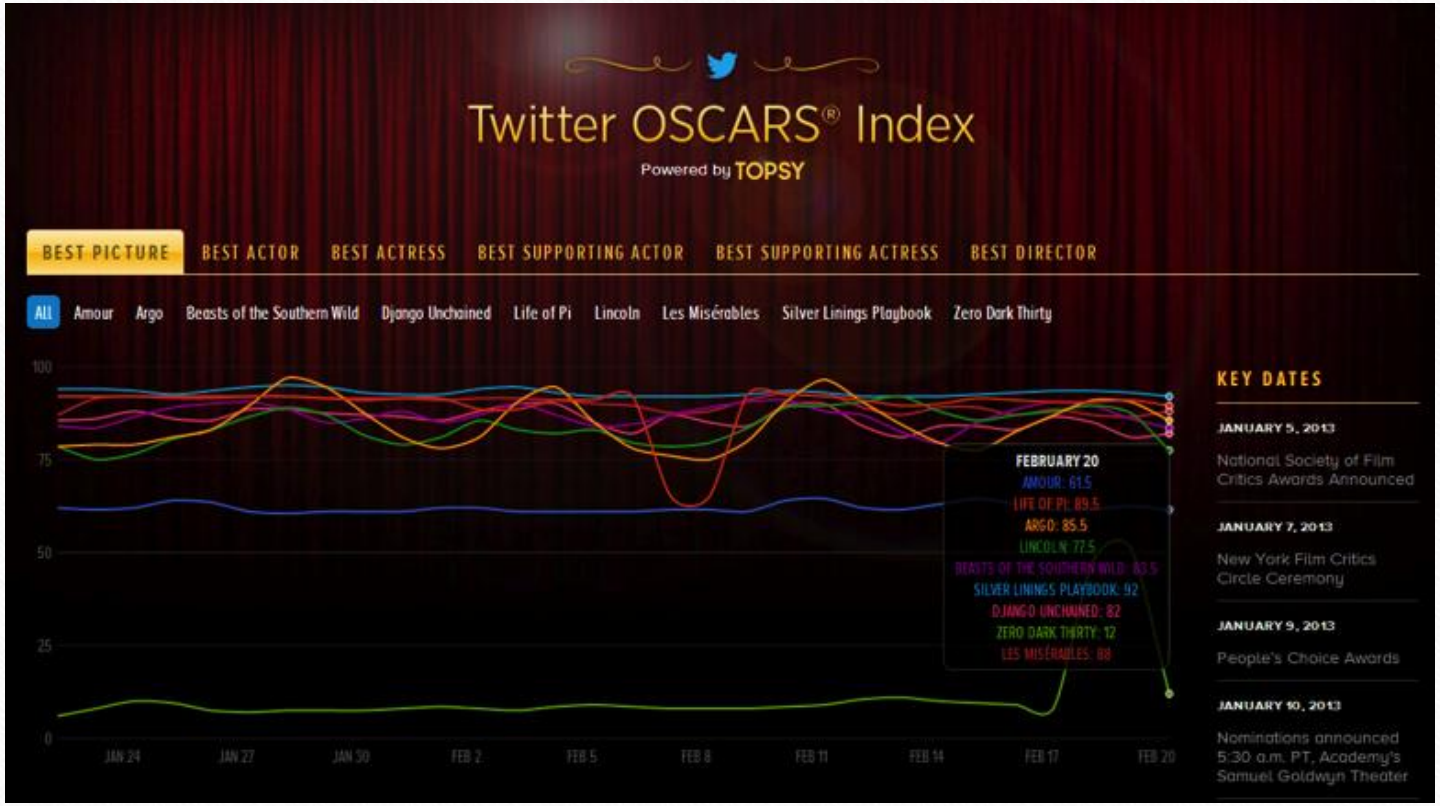
- IBM's Watson wins at Jeopardy! - 2011

What is inside Watson?

- *"Watson had access to 200 million pages of structured and unstructured content consuming four terabytes of disk storage including the full text of Wikipedia" – PC World*
- *"The sources of information for Watson include encyclopedias, dictionaries, thesauri, newswire articles, and literary works. Watson also used databases, taxonomies, and ontologies. Specifically, DBPedia, WordNet, and Yago were used." – AI Magazine*

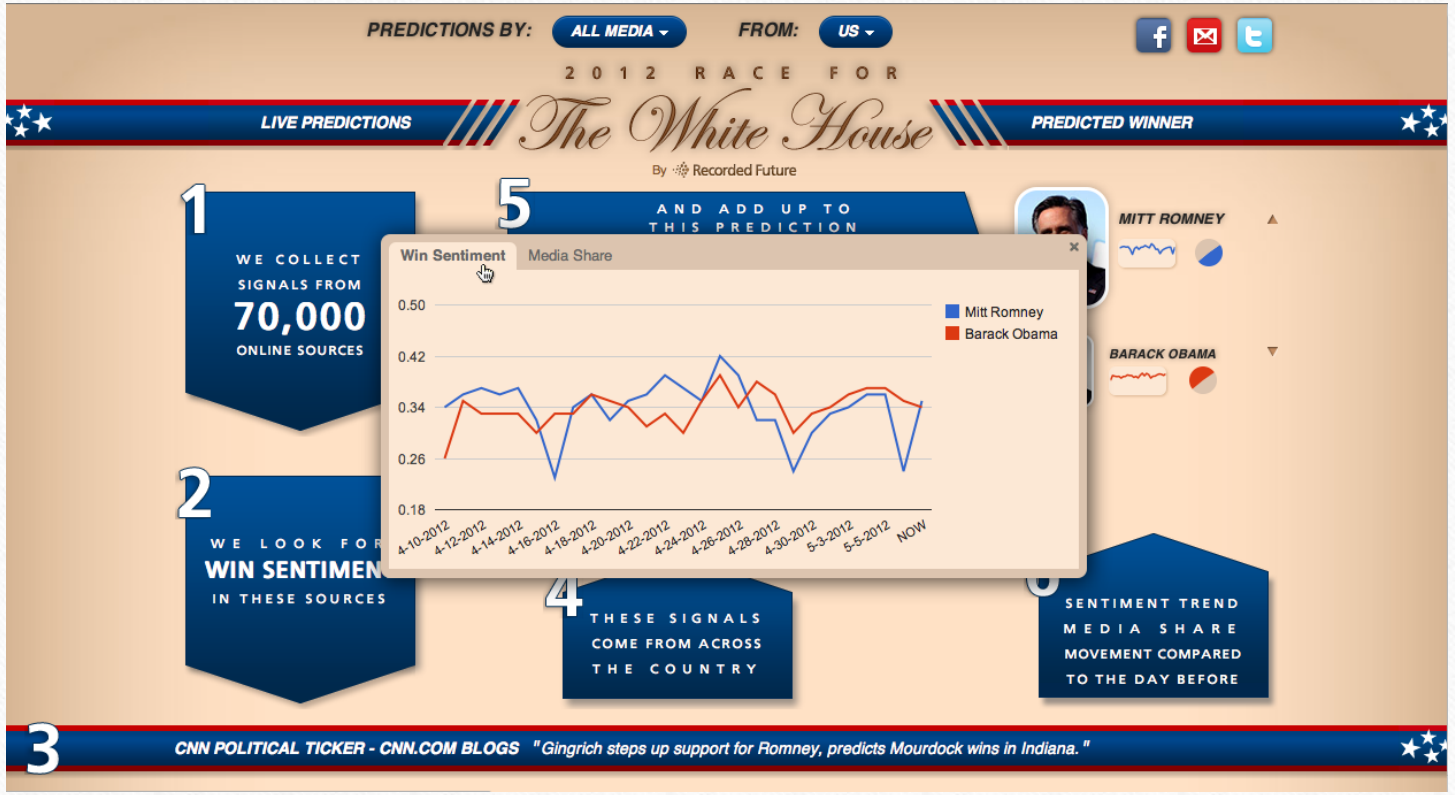
Text mining around us

- Sentiment analysis



Text mining around us

- Sentiment analysis

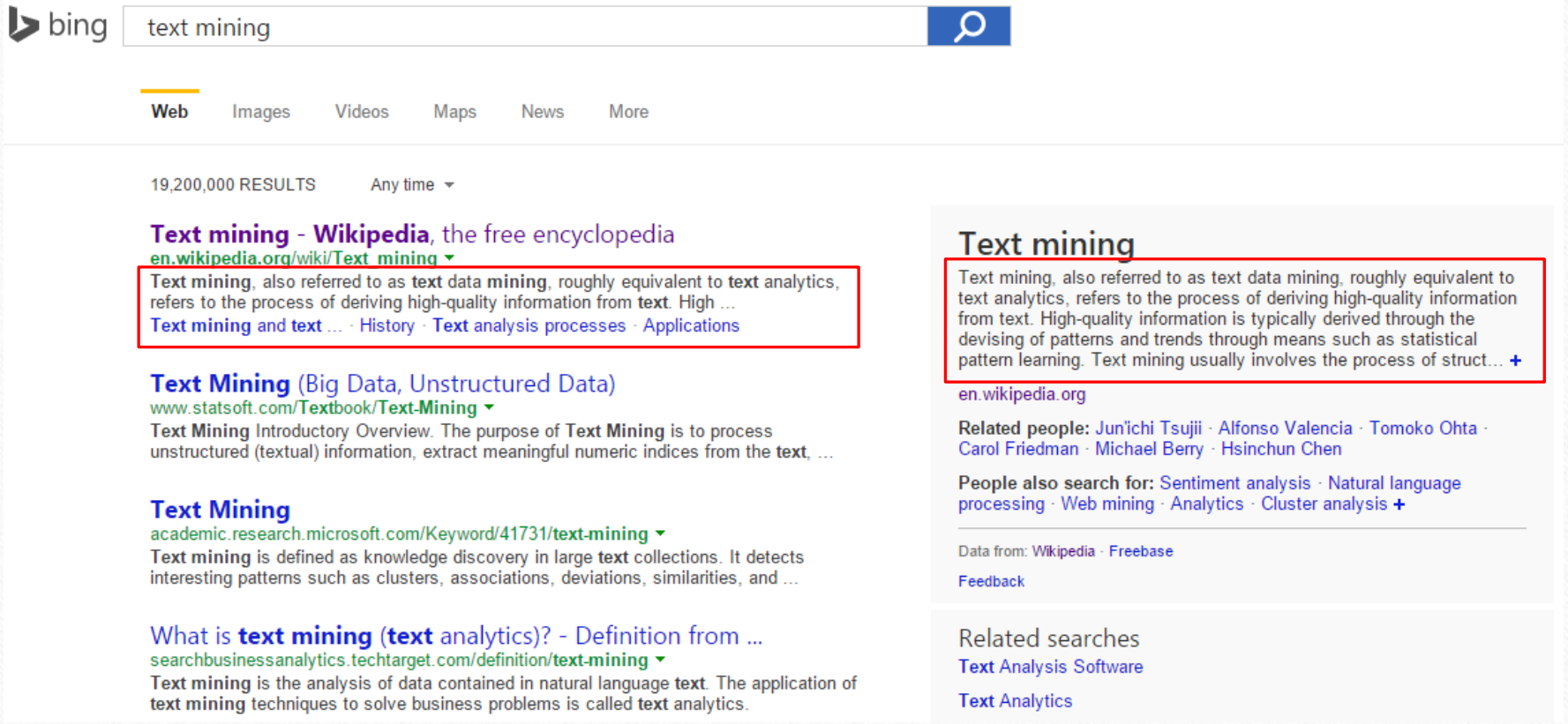


Text mining around us

- Document summarization

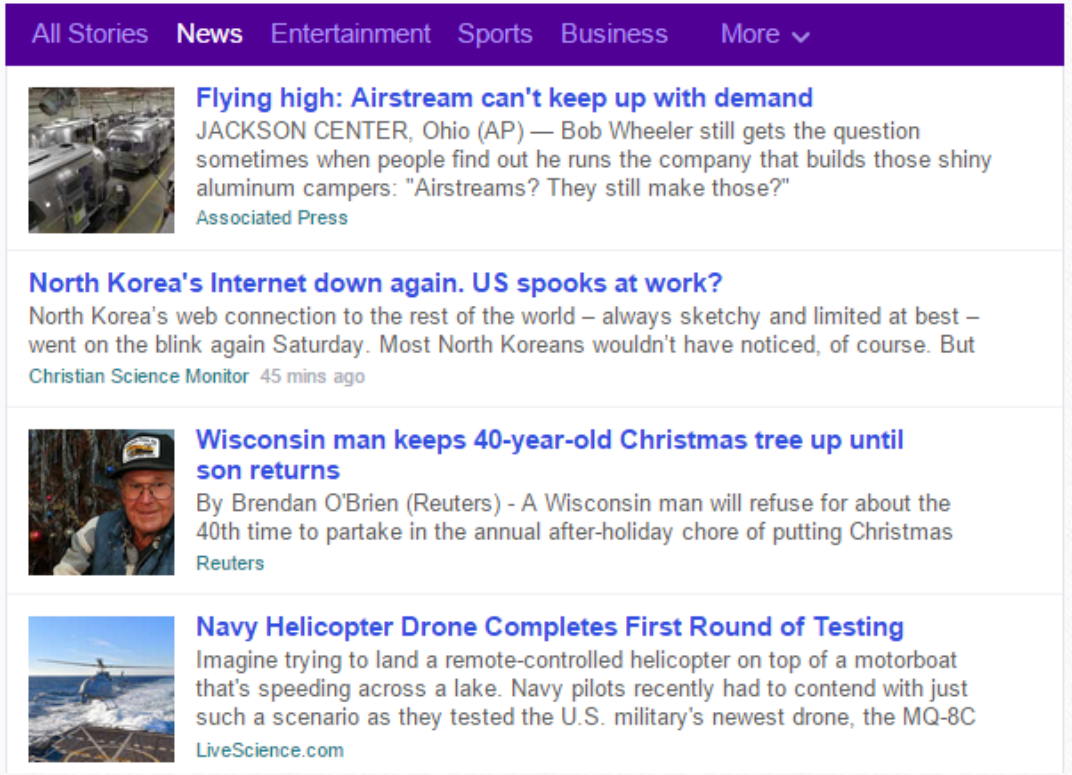
Text mining around us

- Document summarization



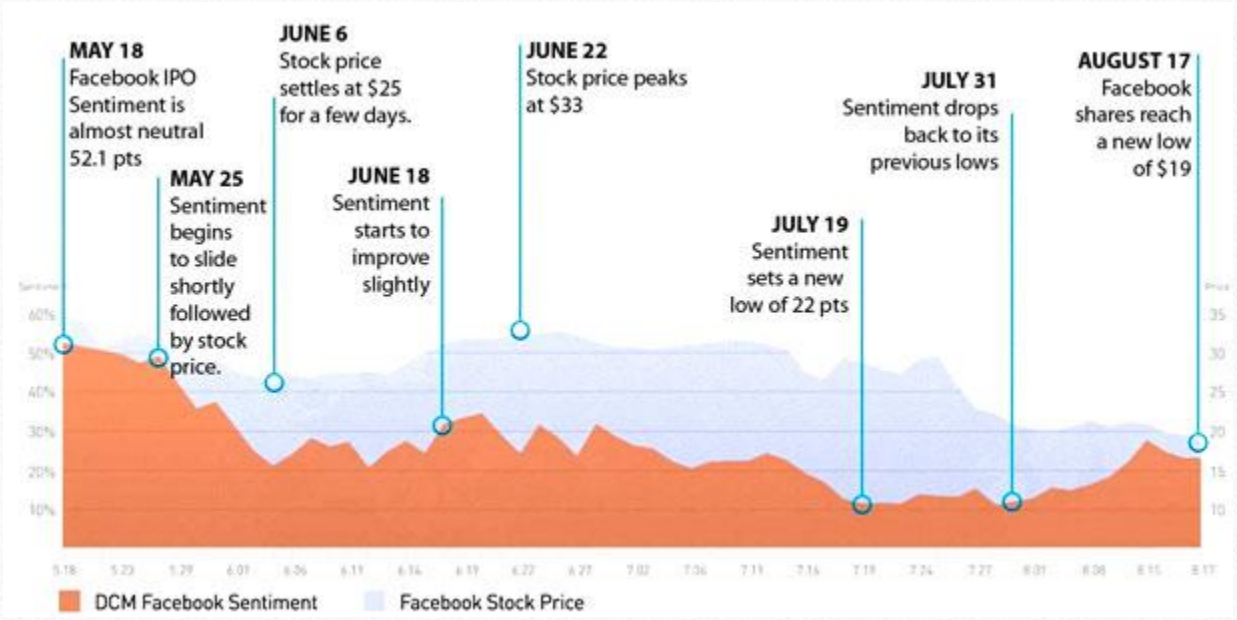
Text mining around us

- News recommendation



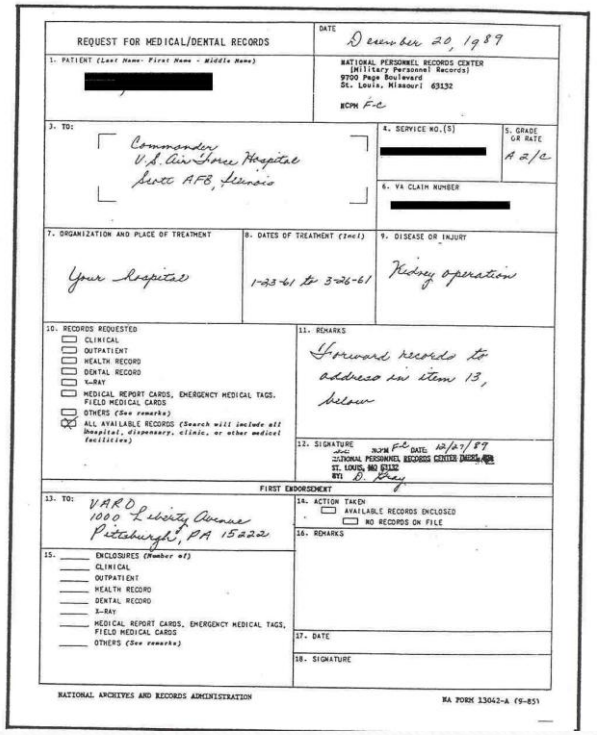
Text mining around us

- Text analytics in financial services



Text mining around us

- Text analytics in healthcare



REQUEST FOR MEDICAL/DENTAL RECORDS

DATE: December 20, 1989

1. PATIENT (Last Name- First Name - Middle Name): [redacted]

NATIONAL PERSONNEL RECORDS CENTER (Military Personnel Records) 9700 Page Boulevard St. Louis, Missouri 63132

NCPM F-C

3. TO: Commander U.S. Air Force Hospital Scott AFB, Illinois

4. SERVICE NO.(S): [redacted]

5. GRADE OR RATE: A 2/c

6. VA CLAIM NUMBER: [redacted]

7. ORGANIZATION AND PLACE OF TREATMENT: Your Hospital

8. DATES OF TREATMENT (Inclusive): 1-23-61 to 3-26-61

9. DISEASE OR INJURY: Kidney operation

10. RECORDS REQUESTED
- ☐ CLINICAL
- ☐ OUTPATIENT
- ☐ HEALTH RECORD
- ☐ DENTAL RECORD
- ☐ X-RAY
- ☐ MEDICAL REPORT CARDS, EMERGENCY MEDICAL TAGS, FIELD MEDICAL CARDS
- ☐ OTHERS (See remarks)
- ☒ ALL AVAILABLE RECORDS (Search will include all Hospital, dispensary, clinic, or other medical facilities)

11. REMARKS: Forward records to address in item 13, below

12. SIGNATURE: NCPM F-C DATE: 12/27/89 NATIONAL PERSONNEL RECORDS CENTER (MPR) 9700 PAGE BLVD. ST. LOUIS, MO 63132 BY: D. Gray

FIRST ENDORSEMENT

13. TO: VARO 1000 Liberty Avenue Pittsburgh, PA 15222

14. ACTION TAKEN: ☐ AVAILABLE RECORDS ENCLOSED ☐ NO RECORDS ON FILE

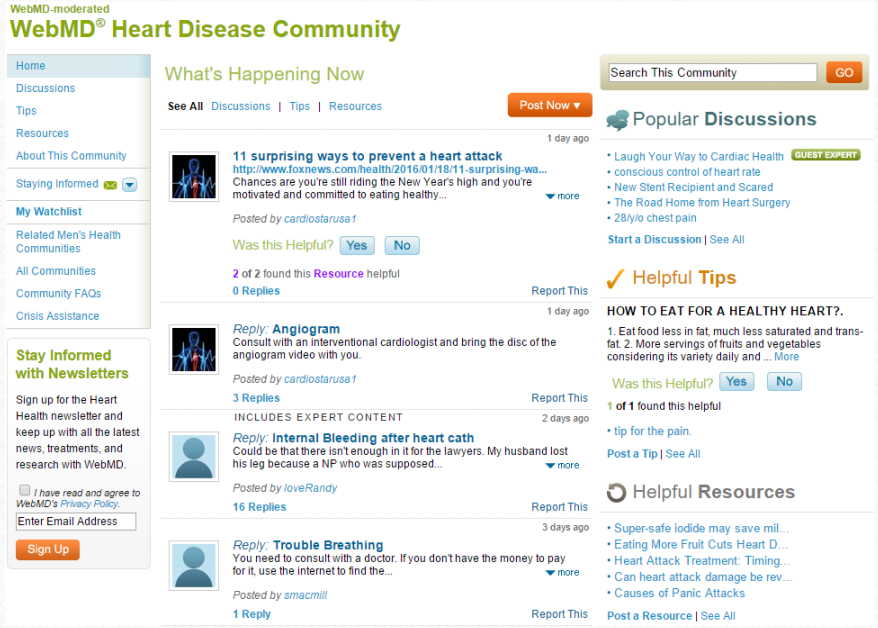
15. ENCLOSURES (Number of)
- ____ CLINICAL
- ____ OUTPATIENT
- ____ HEALTH RECORD
- ____ DENTAL RECORD
- ____ X-RAY
- ____ MEDICAL REPORT CARDS, EMERGENCY MEDICAL TAGS, FIELD MEDICAL CARDS
- ____ OTHERS (See remarks)

16. REMARKS

17. DATE

18. SIGNATURE

NATIONAL ARCHIVES AND RECORDS ADMINISTRATION — NA FORM 13042-A (9-85)



WebMD-moderated

WebMD® Heart Disease Community

Home | Discussions | Tips | Resources | About This Community | Staying Informed | My Watchlist | Related Men's Health Communities | All Communities | Community FAQs | Crisis Assistance

Stay Informed with Newsletters

Sign up for the Heart Health newsletter and keep up with all the latest news, treatments, and research with WebMD.

I have read and agree to WebMD's Privacy Policy.

Enter Email Address

Sign Up

What's Happening Now

See All | Discussions | Tips | Resources

Post Now

1 day ago

11 surprising ways to prevent a heart attack
http://www.foxnews.com/health/2016/01/18/11-surprising-wa...
Chances are you're still riding the New Year's high and you're motivated and committed to eating healthy... more

Posted by cardiostarusa1

Was this Helpful? Yes No

2 of 2 found this Resource helpful

0 Replies | Report This

1 day ago

Reply: Angiogram
Consult with an interventional cardiologist and bring the disc of the angiogram video with you.

Posted by cardiostarusa1

3 Replies | Report This

INCLUDES EXPERT CONTENT — 2 days ago

Reply: Internal Bleeding after heart cath
Could be that there isn't enough in it for the lawyers. My husband lost his leg because a NP who was supposed... more

Posted by loveRandy

16 Replies | Report This

3 days ago

Reply: Trouble Breathing
You need to consult with a doctor. If you don't have the money to pay for it, use the internet to find the... more

Posted by smacmill

1 Reply | Report This

Search This Community GO

Popular Discussions

- Laugh Your Way to Cardiac Health — GUEST EXPERT
- conscious control of heart rate
- New Stent Recipient and Scared
- The Road Home from Heart Surgery
- 28/y/o chest pain

Start a Discussion | See All

Helpful Tips

HOW TO EAT FOR A HEALTHY HEART?.

1. Eat food less in fat, much less saturated and trans-fat. 2. More servings of fruits and vegetables considering its variety daily and ... More

Was this Helpful? Yes No

1 of 1 found this helpful

- tip for the pain.

Post a Tip | See All

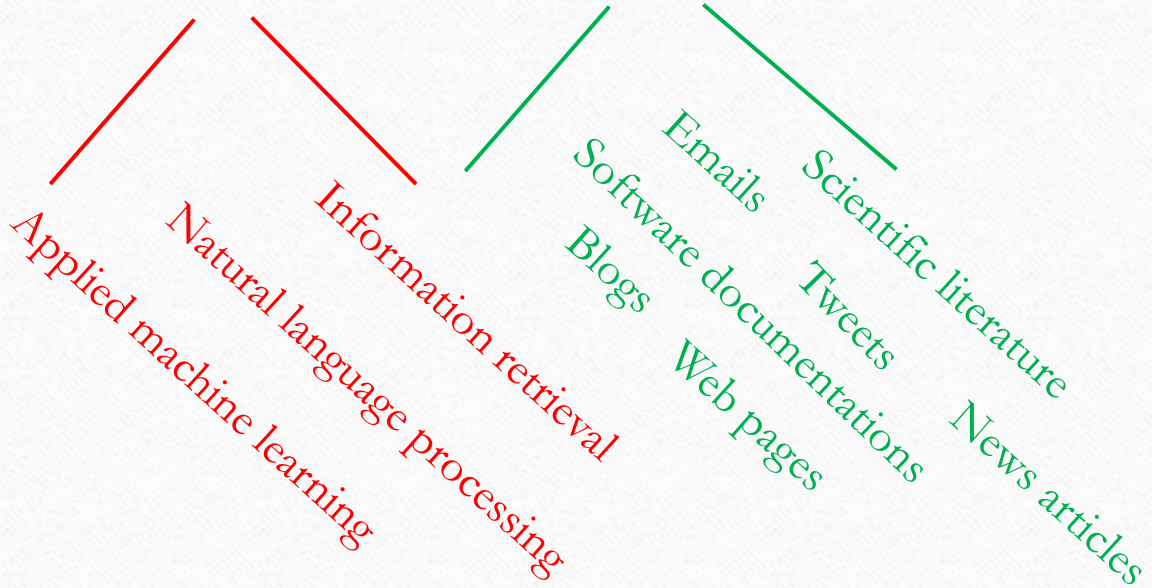
Helpful Resources

- Super-safe iodide may save mil...
- Eating More Fruit Cuts Heart D...
- Heart Attack Treatment: Timing...
- Can heart attack damage be rev...
- Causes of Panic Attacks

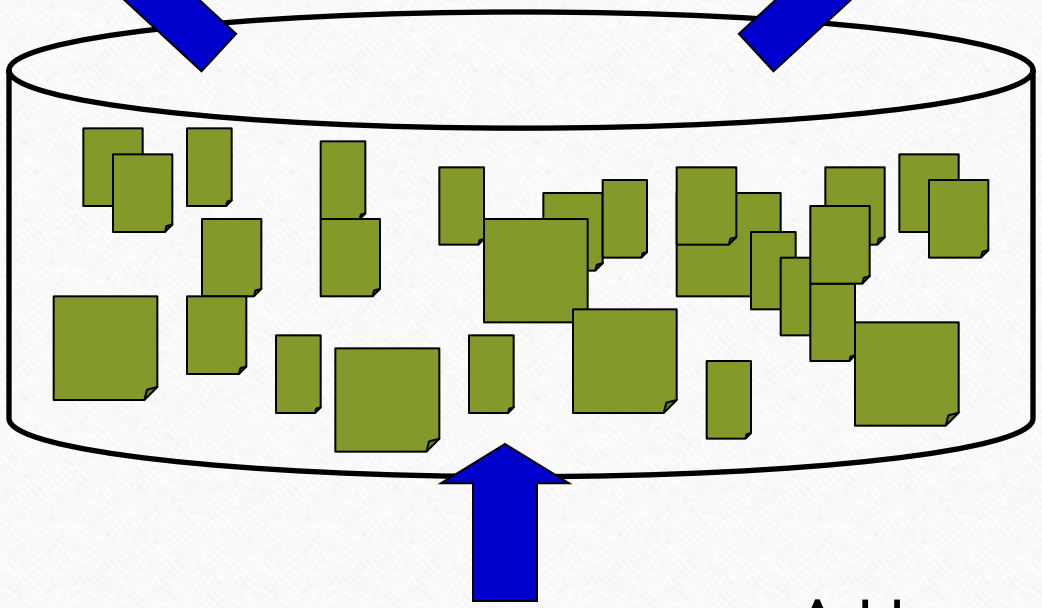
Post a Resource | See All

How to perform text mining?

- As computer scientists, we view it as
  - Text Mining = Data Mining + Text Data



Text mining in general



Access

Serve for IR applications

Sub-area of DM research

Mining

Filter information

Discover knowledge

Based on NLP/ML techniques

Organization

Add Structure/Annotations

Challenges in text mining

- Data collection is “free text”
  - Data is not well-organized
    - Semi-structured or unstructured
  - Natural language text contains ambiguities on many levels
    - Lexical, syntactic, semantic, and pragmatic
  - Learning techniques for processing text typically need annotated training examples
    - Expensive to acquire at scale
- What to mine?

Vector Space Model

How to represent a document

- Represent by a string?
  - No semantic meaning
- Represent by a list of sentences?
  - Sentence is just like a short document (recursive definition)

University of Virginia

From Wikipedia, the free encyclopedia

The **University of Virginia** (**UVA** or **U.Va.**), often referred to as simply **Virginia**, is a public research university in Charlottesville, Virginia. UVA is known for its historic foundations, student-run honor code, and secret societies.

Its initial Board of Visitors included U.S. Presidents Thomas Jefferson, James Madison, and James Monroe. President Monroe was the sitting President of the United States at the time of the founding; Jefferson and Madison were the first two rectors. UVA was established in 1819, with its Academical Village and original courses of study conceived and designed entirely by Jefferson. UNESCO designated it a World Heritage Site in 1987, an honor shared with nearby Monticello.[4]

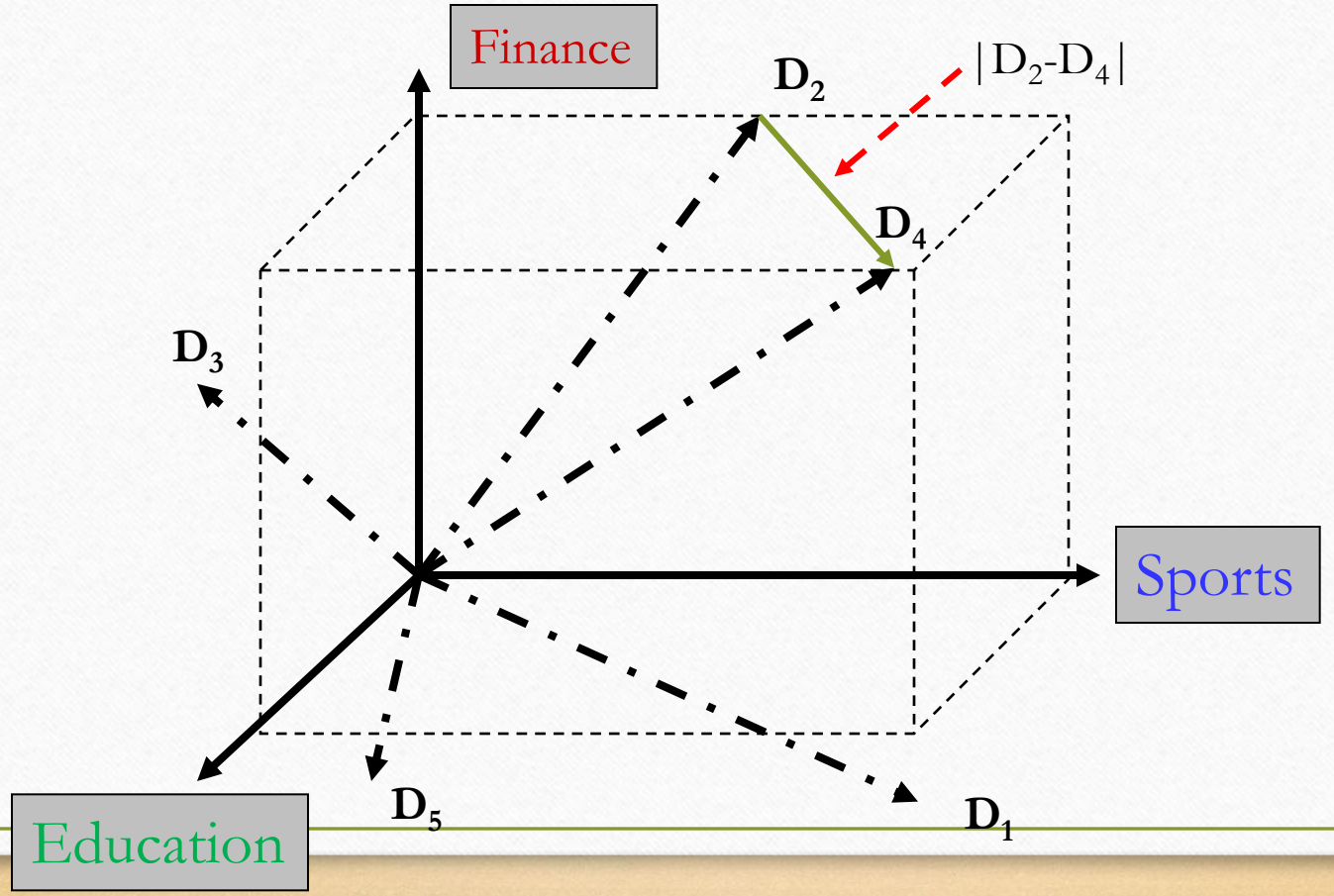
The first university of the American South elected to the Association of American Universities in 1904, UVA is classified as *Very High Research Activity* in the Carnegie Classification. The university is affiliated with 7 Nobel Laureates, and has produced 7 NASA astronauts, 7 Marshall Scholars, 4 Churchill Scholars, 29 Truman Scholars, and 50 Rhodes Scholars, the most of any state-affiliated institution in the U.S.[5][6][7] Supported in part by the Commonwealth, it receives far more funding from private sources than public, and its students come from all 50 states and 147 countries.[2][8][9] It also operates a small liberal arts branch campus in the far southwestern corner of the state.

Vector space model

- Represent documents by concept vectors
  - Each concept defines one dimension
  - *k* concepts define a high-dimensional space
  - Element of vector corresponds to concept weight
    - E.g., $d=(x_1,\ldots,x_k)$, x_i is "importance" of concept i in d
- Distance between the vectors in this concept space
  - Relationship among documents

An illustration of VS model

- All documents are projected into this concept space



What the VS model doesn't say

- How to define/select the "basic concept"
  - Concepts are assumed to be orthogonal (/independent)
- How to assign weights
  - Weights indicate how well the concept characterizes the document
- How to define the distance metric

Recap: vector space model

- Represent documents by concept vectors
  - Each concept defines one dimension
  - *k* concepts define a high-dimensional space
  - Element of vector corresponds to concept weight
    - E.g., $d=(x_1,\ldots,x_k)$, x_i is "importance" of concept i in d
- Distance between the vectors in this concept space
  - Relationship among documents

What is a good "Basic Concept"?

- Orthogonal
  - Linearly independent basis vectors
    - "Non-overlapping" in meaning
  - No ambiguity
- Weights can be assigned automatically and accurately
- Existing solutions
  - Terms or N-grams, a.k.a., Bag-of-Words

Bag-of-Words representation

- Term as the basis for vector space
  - Doc1: Text mining is to identify useful information.
  - Doc2: Useful information is mined from text.
  - Doc3: Apple is delicious.

	text	information	identify	mining	mined	is	useful	to	from	apple	delicious
Doc1	1	1	1	1	0	1	1	1	0	0	0
Doc2	1	1	0	0	1	1	1	0	1	0	0
Doc3	0	0	0	0	0	1	0	0	0	1	1

Tokenization

- Break a stream of text into meaningful units
  - Tokens: words, phrases, symbols
    - **Input:** It's not straight-forward to perform so-called "tokenization."
    - **Output(1):** 'It's', 'not', 'straight-forward', 'to', 'perform', 'so-called', '"tokenization."'
    - **Output(2):** 'It', "'", 's', 'not', 'straight', '-', 'forward, 'to', 'perform', 'so', '-', 'called', "“", 'tokenization', '.', "”"
  - Definition depends on language, corpus, or even context

Tokenization

- Solutions
  - Regular expressions
    - [\w]+: so-called -> 'so', 'called'
    - [\S]+: It's -> 'It's' instead of 'It', ''s'
  - Statistical methods
    - Explore rich features to decide where the boundary of a word is
      - Apache OpenNLP (http://opennlp.apache.org/)
      - Stanford NLP Parser (http://nlp.stanford.edu/software/lex-parser.shtml)
    - Online Demo
      - Stanford (http://nlp.stanford.edu:8080/parser/index.jsp)
      - UIUC (http://cogcomp.cs.illinois.edu/curator/demo/index.html)

Bag-of-Words representation

	text	information	identify	mining	mined	is	useful	to	from	apple	delicious
Doc1	1	1	1	1	0	1	1	1	0	0	0
Doc2	1	1	0	0	1	1	1	0	1	0	0
Doc3	0	0	0	0	0	1	0	0	0	1	1

- Assumption
  - Words are independent from each other
- Pros
  - Simple
- Cons
  - Basic vectors are clearly not linearly independent!
  - Grammar and order are missing
- ***The most frequently used document representation***
  - ***Image, speech, gene sequence***

Bag-of-Words with N-grams

- N-grams: a contiguous sequence of N tokens from a given piece of text
  - E.g., '*Text mining is to identify useful information.*'
  - Bigrams: '*text_mining*', '*mining_is*', '*is_to*', '*to_identify*', '*identify_useful*', '*useful_information*', '*information_*.'
- Pros: capture local dependency and order
- Cons: a purely statistical view, increase the vocabulary size $O(V^N)$

Automatic document representation

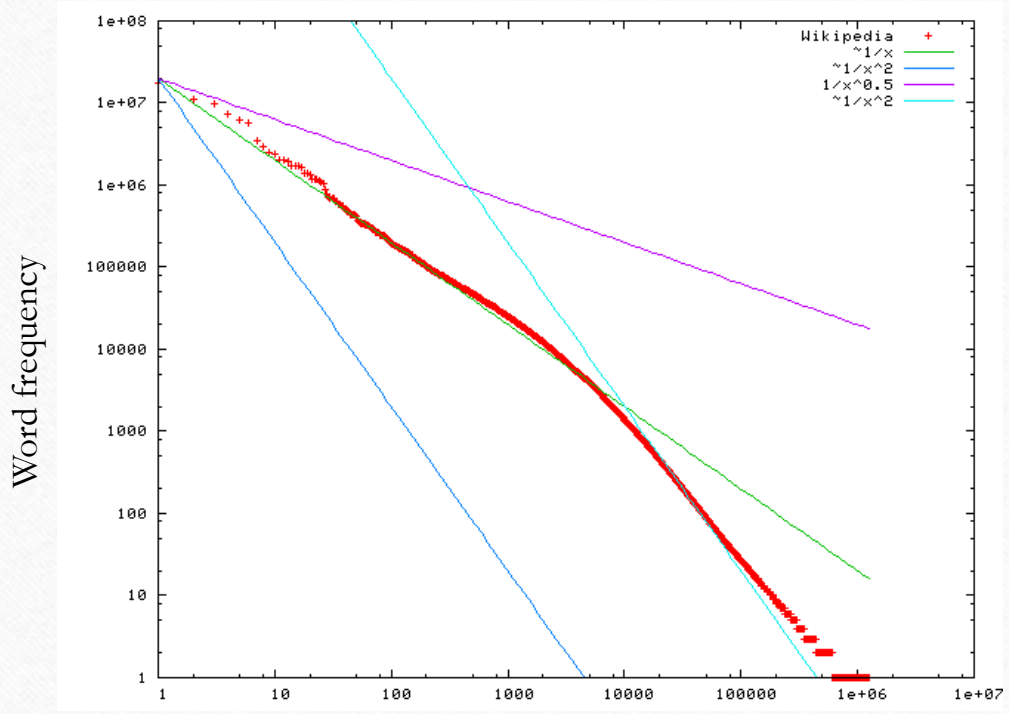
- Represent a document with all the occurring words
  - Pros
    - Preserve all information in the text (hopefully)
    - Fully automatic
  - Cons
    - Vocabulary gap: cars v.s., car, talk v.s., talking
    - Large storage: N-grams needs $O(V^N)$
  - Solution
    - Construct controlled vocabulary

A statistical property of language

- Zipf's law
  - Frequency of any word is inversely proportional to its rank in the frequency table
  - Formally
    - $f(k; s, N) = \frac{1/k^s}{\sum_{n=1}^{N} 1/n^s}$

    where k is rank of the word; N is the vocabulary size; s is language-specific parameter
  - Simply: $f(k; s, N) \propto 1/k^s$

A statistical property of language



A plot of word frequency in Wikipedia (Nov 27, 2006)

*In the Brown Corpus of American English text, the word "the" is the most frequently occurring word, and by itself accounts for nearly **7%** of all word occurrences; the second-place word "of" accounts for slightly over **3.5%** of words.*

Zipf's law tells us

- Head words take large portion of occurrences, but they are semantically meaningless
  - E.g., the, a, an, we, do, to
- Tail words take major portion of vocabulary, but they rarely occur in documents
  - E.g., dextrosinistral
- The rest is most representative
  - To be included in the controlled vocabulary

Automatic document representation

Remove non-informative words

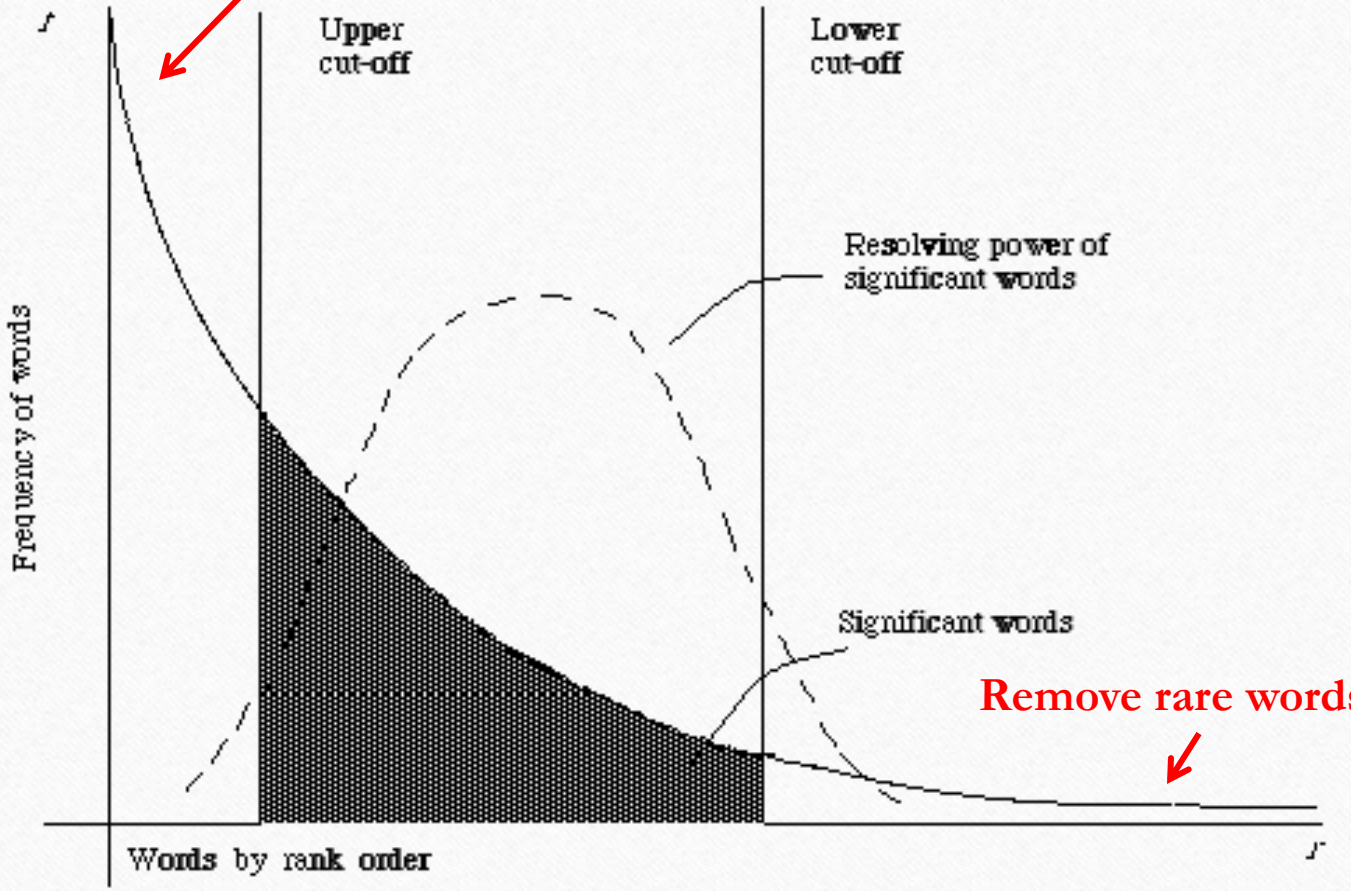


Figure 2.1. A plot of the hyperbolic curve relating f, the frequency of occurrence and r, the rank order (Adapted from Schultz[44] page 120)

Normalization

- Convert different forms of a word to a normalized form in the vocabulary
  - U.S.A. -> USA, St. Louis -> Saint Louis
- Solution
  - Rule-based
    - Delete periods and hyphens
    - All in lower cases
  - Dictionary-based
    - Construct equivalent class
      - Car -> "automobile, vehicle"
      - Mobile phone -> "cellphone"

Stemming

- Reduce inflected or derived words to their root form
  - Plurals, adverbs, inflected word forms
    - E.g., ladies -> lady, referring -> refer, forgotten -> forget
  - Bridge the vocabulary gap
  - Solutions (for English)
    - Porter stemmer: patterns of vowel-consonant sequence
    - Krovetz stemmer: morphological rules
  - Risk: lose precise meaning of the word
    - E.g., lay -> lie (a false statement? or be in a horizontal position?)

Stopwords

- Useless words for document analysis
  - Not all words are informative
  - Remove such words to reduce vocabulary size
  - No universal definition
  - Risk: break the original meaning and structure of text
    - E.g., this is not a good option -> option

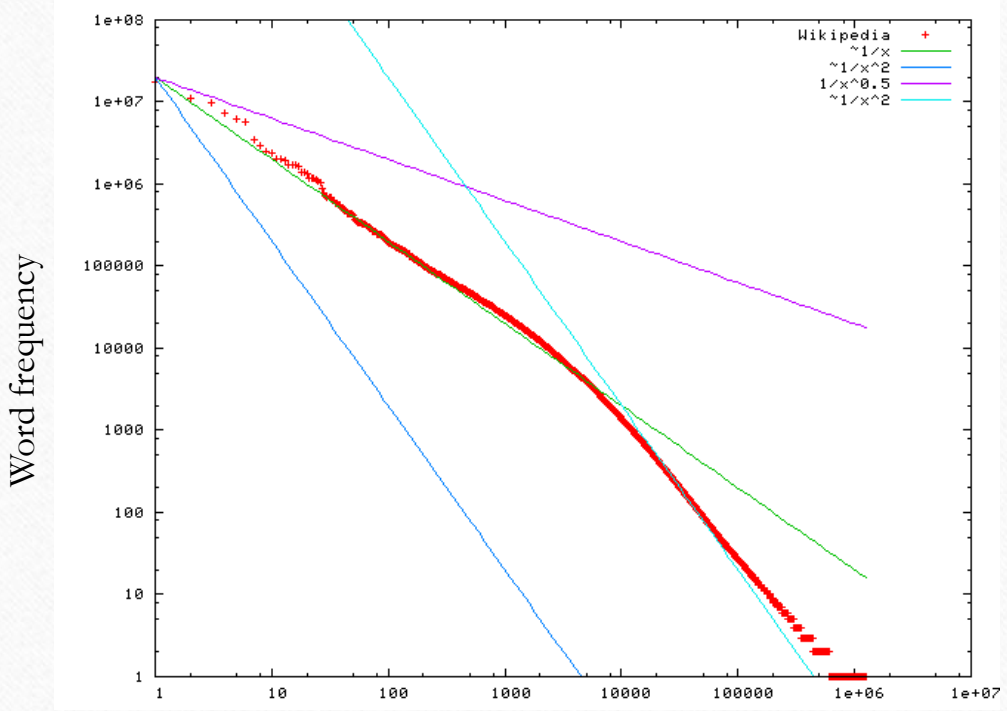
      to be or not to be -> null

Stopwords

Nouns	Verbs	Adjectives	Prepositions	Others
1. time	1. be	1. good	1. to	1. the
2. person	2. have	2. new	2. of	2. and
3. year	3. do	3. first	3. in	3. a
4. way	4. say	4. last	4. for	4. that
5. day	5. get	5. long	5. on	5. I
6. thing	6. make	6. great	6. with	6. it
7. man	7. go	7. little	7. at	7. not
8. world	8. know	8. own	8. by	8. he
9. life	9. take	9. other	9. from	9. as
10. hand	10. see	10. old	10. up	10. you
11. part	11. come	11. right	11. about	11. this
12. child	12. think	12. big	12. into	12. but
13. eye	13. look	13. high	13. over	13. his
14. woman	14. want	14. different	14. after	14. they
15. place	15. give	15. small	15. beneath	15. her
16. work	16. use	16. large	16. under	16. she
17. week	17. find	17. next	17. above	17. or
18. case	18. tell	18. early		18. an
19. point	19. ask	19. young		19. will
20. government	20. work	20. important		20. my
21. company	21. seem	21. few		21. one
22. number	22. feel	22. public		22. all
23. group	23. try	23. bad		23. would
24. problem	24. leave	24. same		24. there
25. fact	25. call	25. able		25. their

The OEC: Facts about the language

Recap: a statistical property of language



A plot of word frequency in Wikipedia (Nov 27, 2006)

Constructing a VSM representation

D1: *'Text mining is to identify useful information.'*

1. **Tokenization:**

   D1: *'Text', 'mining', 'is', 'to', 'identify', 'useful', 'information', '.'*

2. **Stemming/normalization:**

   D1: *'text', 'mine', 'is', 'to', 'identify', 'use', 'inform', '.'*

3. **N-gram construction:**

   D1: *'text-mine', 'mine-is', 'is-to', 'to-identify', 'identify-use', 'use-inform', 'inform-.'*

4. **Stopword/controlled vocabulary filtering:**

   D1: *'text-mine', 'to-identify', 'identify-use', 'use-inform'*

Terms	Documents													
	M1	M2	M3	M4	M5	M6	M7	M8	M9	M10	M11	M12	M13	M14
abnormalities	0	0	0	0	0	0	0	1	0	1	0	0	0	0
age	1	0	0	0	0	0	0	0	0	0	0	1	0	0
behavior	0	0	0	0	1	1	0	0	0	0	0	0	0	0
blood	0	0	0	0	0	0	0	1	0	0	1	0	0	0
close	0	0	0	0	0	0	1	0	0	0	1	0	0	0
culture	1	1	0	0	0	0	0	1	1	0	0	0	0	0
depressed	1	0	1	1	1	0	0	0	0	0	0	0	0	0
discharge	1	1	0	0	0	1	0	0	0	0	0	0	0	0
disease	0	0	0	0	0	0	0	0	1	0	1	0	0	0
fast	0	0	0	0	0	0	0	0	0	1	0	1	1	1
generation	0	0	0	0	0	0	0	0	1	0	0	0	1	0
oestrogen	0	0	1	1	0	0	0	0	0	0	0	0	0	0
patients	1	1	0	1	0	0	0	1	0	0	0	0	0	0
pressure	0	0	0	0	0	0	0	0	0	0	1	0	0	1
rats	0	0	0	0	0	0	0	0	0	0	0	0	1	1
respect	0	0	0	0	0	0	0	1	0	0	0	1	0	0
rise	0	0	0	1	0	0	0	0	0	0	0	0	0	1
study	1	0	1	0	0	0	0	0	1	0	0	0	0	0

Documents in a vector space!

How to assign weights?

- Important!
- Why?
  - Corpus-wise: some terms carry more information about the document content
  - Document-wise: not all terms are equally important
- How?
  - Two basic heuristics
    - TF (Term Frequency) = Within-doc-frequency
    - IDF (Inverse Document Frequency)

Term frequency

- Idea: a term is more important if it occurs more frequently in a document
- TF Formulas
  - Let $c(t, d)$ be the frequency count of term t in doc d
  - Raw TF: $tf(t, d) = c(t, d)$

Which two documents are more similar to each other?

Doc A: 'good',10
Doc B: 'good',2
Doc C: 'good',3

TF normalization

- Two views of document length
  - A doc is long because it is verbose
  - A doc is long because it has more content
- Raw TF is inaccurate
  - Document length variation
  - "Repeated occurrences" are less informative than the "first occurrence"
  - Information about semantic does not increase proportionally with number of term occurrence
- Generally penalize long document, but avoid over-penalizing
  - Pivoted length normalization

TF normalization

- Sub-linear TF scaling
  - $tf(t,d) = \begin{cases} 1 + \log c(t,d), if\ c(t,d) > 0 \\ 0, otherwise \end{cases}$

TF normalization

- Maximum TF scaling
  - $tf(t,d) = \alpha + (1-\alpha)\frac{c(t,d)}{\max_t c(t,d)}$, if $c(t,d) > 0$
  - Normalize by the most frequent word in this doc

Document frequency

- Idea: a term is more discriminative if it occurs only in fewer documents

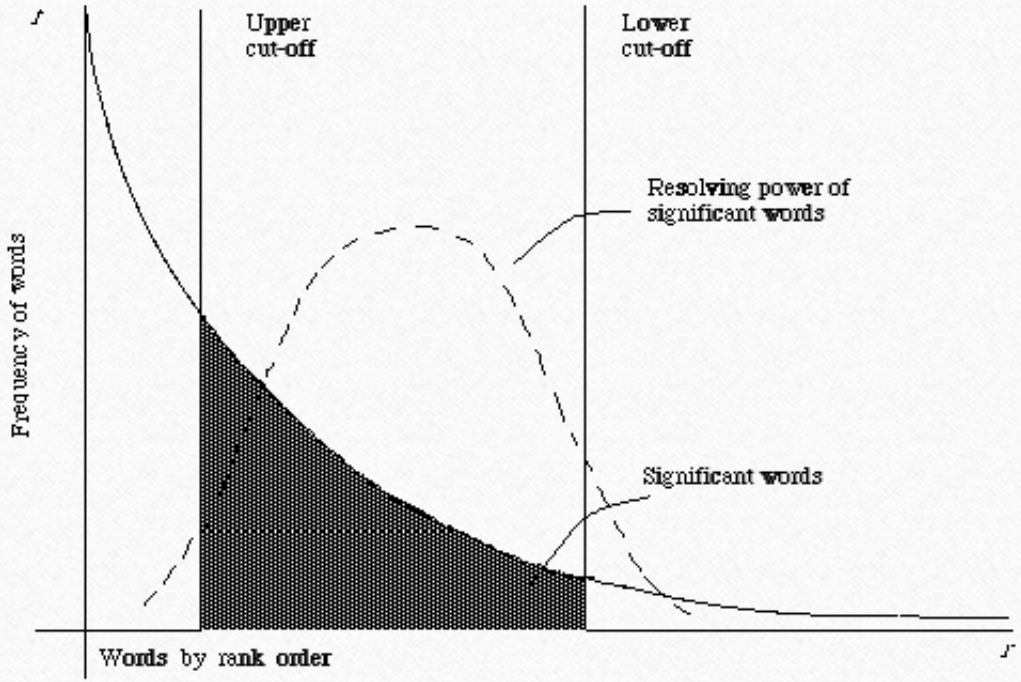


Figure 2.1. A plot of the hyperbolic curve relating f, the frequency of occurrence and r, the rank order (Adapted from Schultz[44] page 120)

Inverse document frequency

- Solution
  - Assign higher weights to the rare terms
  - Formula
    - $IDF(t) = 1 + \log(\frac{N}{df(t)})$
  - A corpus-specific property
    - Independent of a single document

Non-linear scaling

Total number of docs in collection

Number of docs containing term t

Why document frequency

- How about total term frequency?
  - $ttf(t) = \sum_d c(t, d)$

Table 1. Example total term frequency v.s. document frequency in Reuters-RCV1 collection.

Word	ttf	df
try	10422	8760
insurance	10440	3997

- Cannot recognize words frequently occurring in a subset of documents

TF-IDF weighting

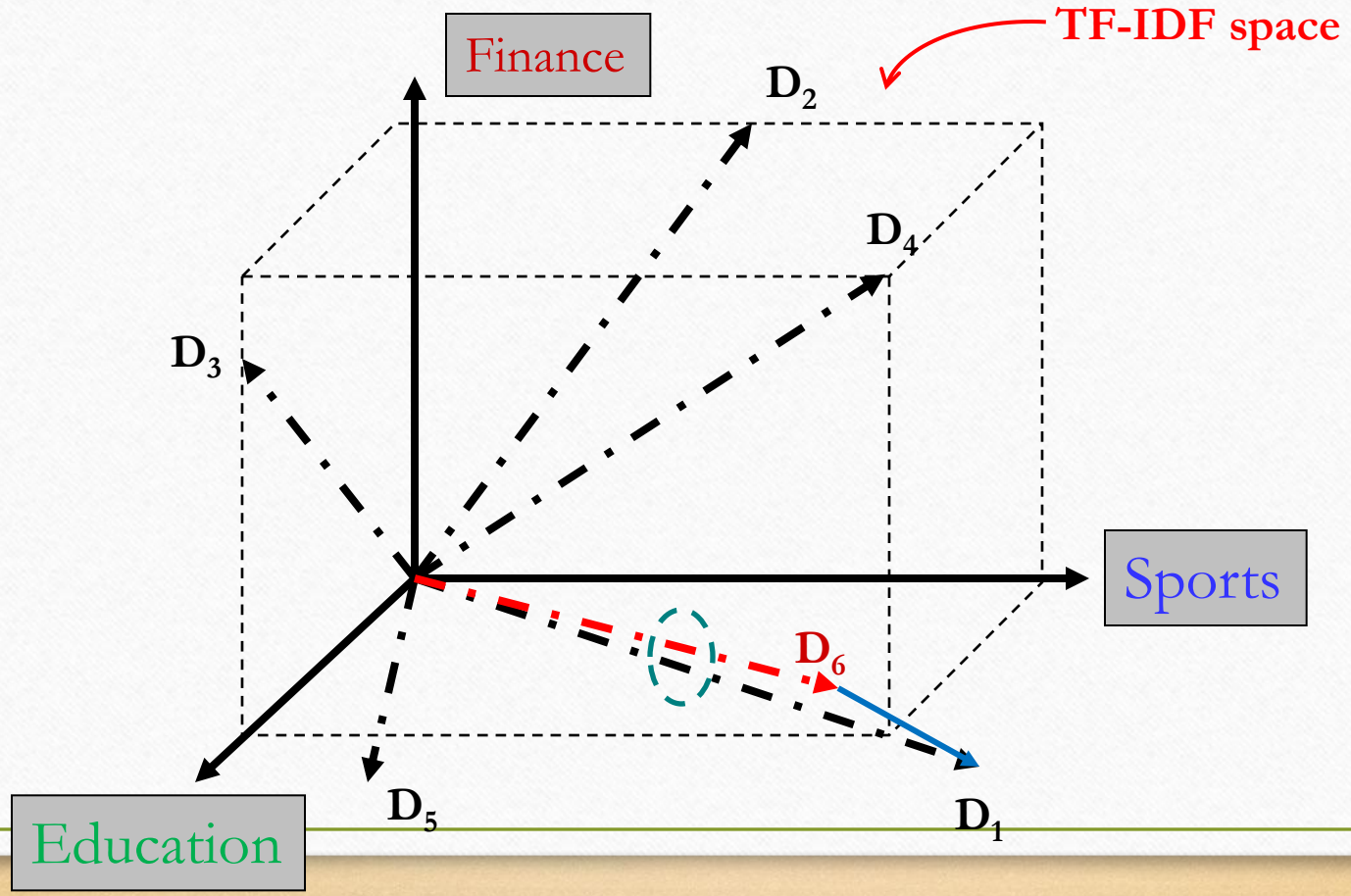
- Combining TF and IDF
  - Common in doc → high tf → high weight
  - Rare in collection→ high idf→ high weight
  - $w(t,d) = TF(t,d) \times IDF(t)$
- Most well-known document representation schema in IR! (G Salton et al. 1983)

"Salton was perhaps the leading computer scientist working in the field of information retrieval during his time." - wikipedia

Gerard Salton Award – highest achievement award in IR

How to define a good similarity metric?

- Euclidean distance?

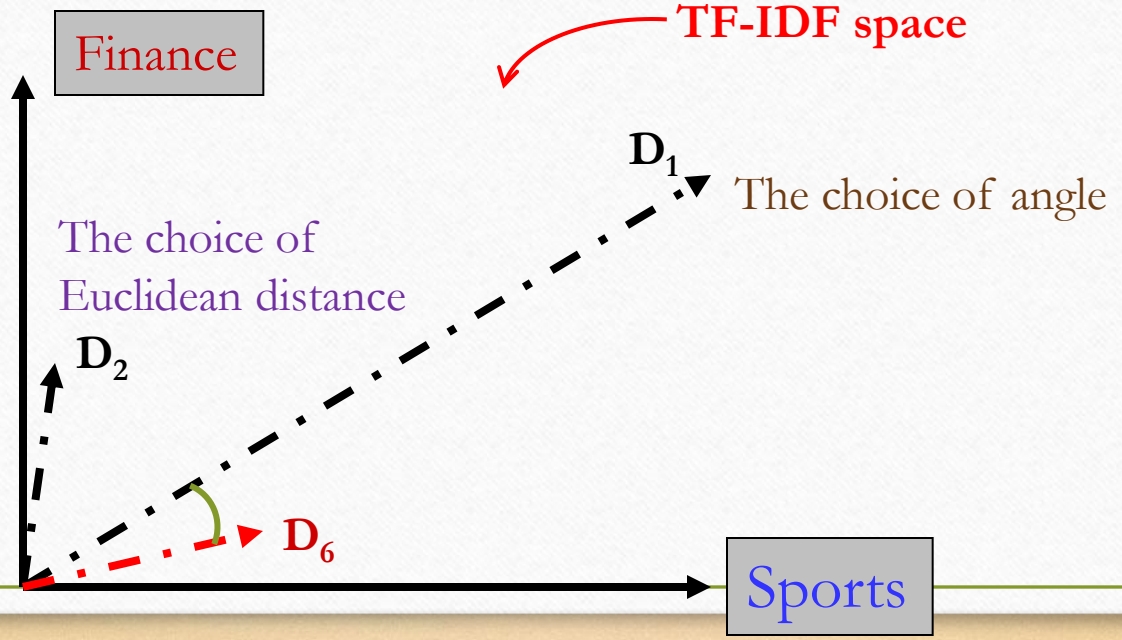


How to define a good similarity metric?

- Euclidean distance
  - $dist(d_i, d_j) = \sqrt{\sum_{t \in V} [tf(t, d_i) idf(t) - tf(t, d_j) idf(t)]^2}$
  - Longer documents will be penalized by the extra words
  - We care more about how these two vectors are overlapped

From distance to angle

- Angle: how vectors are overlapped
  - Cosine similarity – projection of one vector onto another



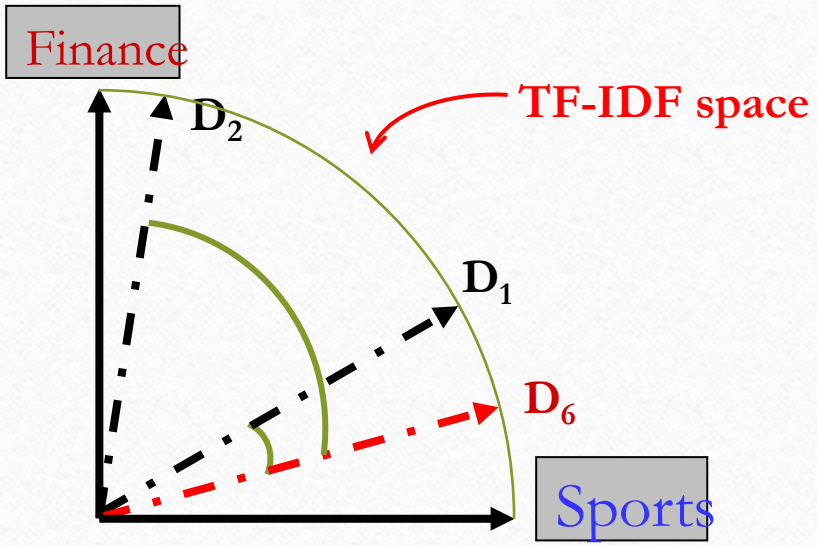
Cosine similarity

- Angle between two vectors
  - $cosine(d_i, d_j) = \frac{V_{d_i}^T V_{d_j}}{|V_{d_i}|_2 \times |V_{d_j}|_2}$

TF-IDF vector

Unit vector

  - Documents are normalized by length



Advantages of VS model

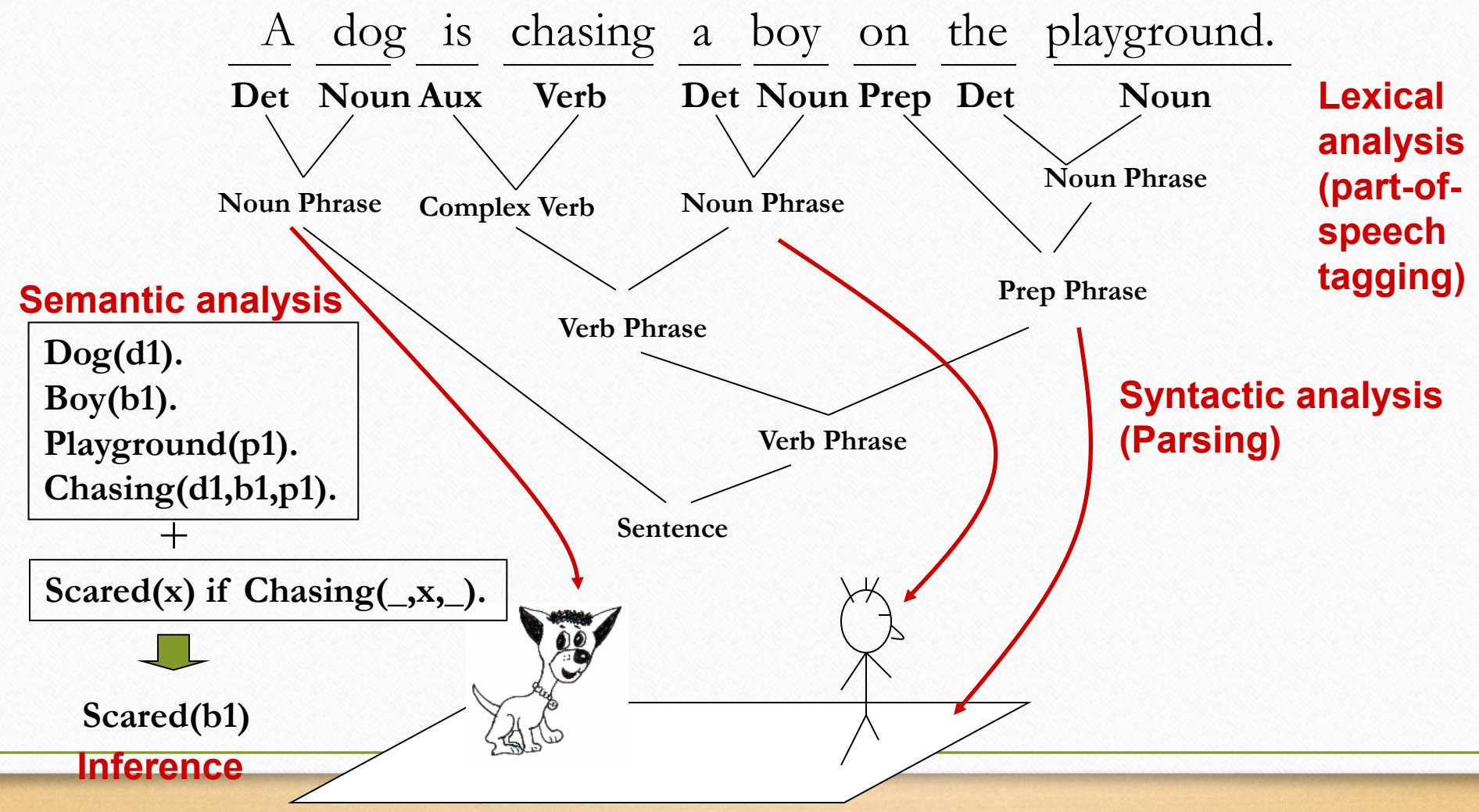
- Empirically effective!
- Intuitive
- Easy to implement
- Well-studied/mostly evaluated
- Warning: many variants of TF-IDF!

Disadvantages of VS model

- Assume term independence
- Lack of "predictive adequacy"
  - Arbitrary term weighting
  - Arbitrary similarity measure
- Lots of parameter tuning!

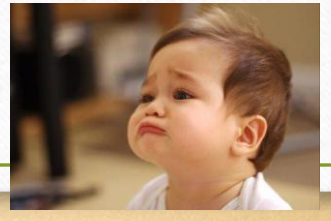
Natural Language Processing

An example of NLP



- *Automatically answer our emails*
- *Translate languages accurately*
- *Help us manage, summarize, and aggregate information*
- *Use speech as a UI (when needed)*
- *Talk to us / listen to us*

If we can do this for all the sentences in all languages, then …



- *Automatically answer our emails*
- *Translate languages accurately*
- *Help us manage, summarize, and aggregate information*
- *Use speech as a UI (when needed)*
- *Talk to us / listen to us*

If we can do this for all the sentences in all languages, then …

BAD NEWS:

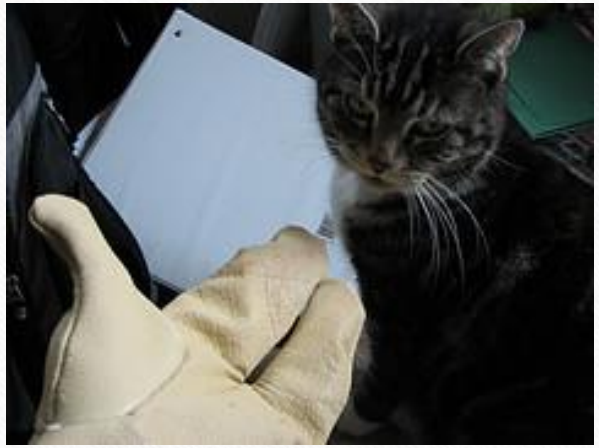
- **Unfortunately, we cannot right now.**
- **General NLP = "Complete AI"**

NLP is difficult!!!!!!!

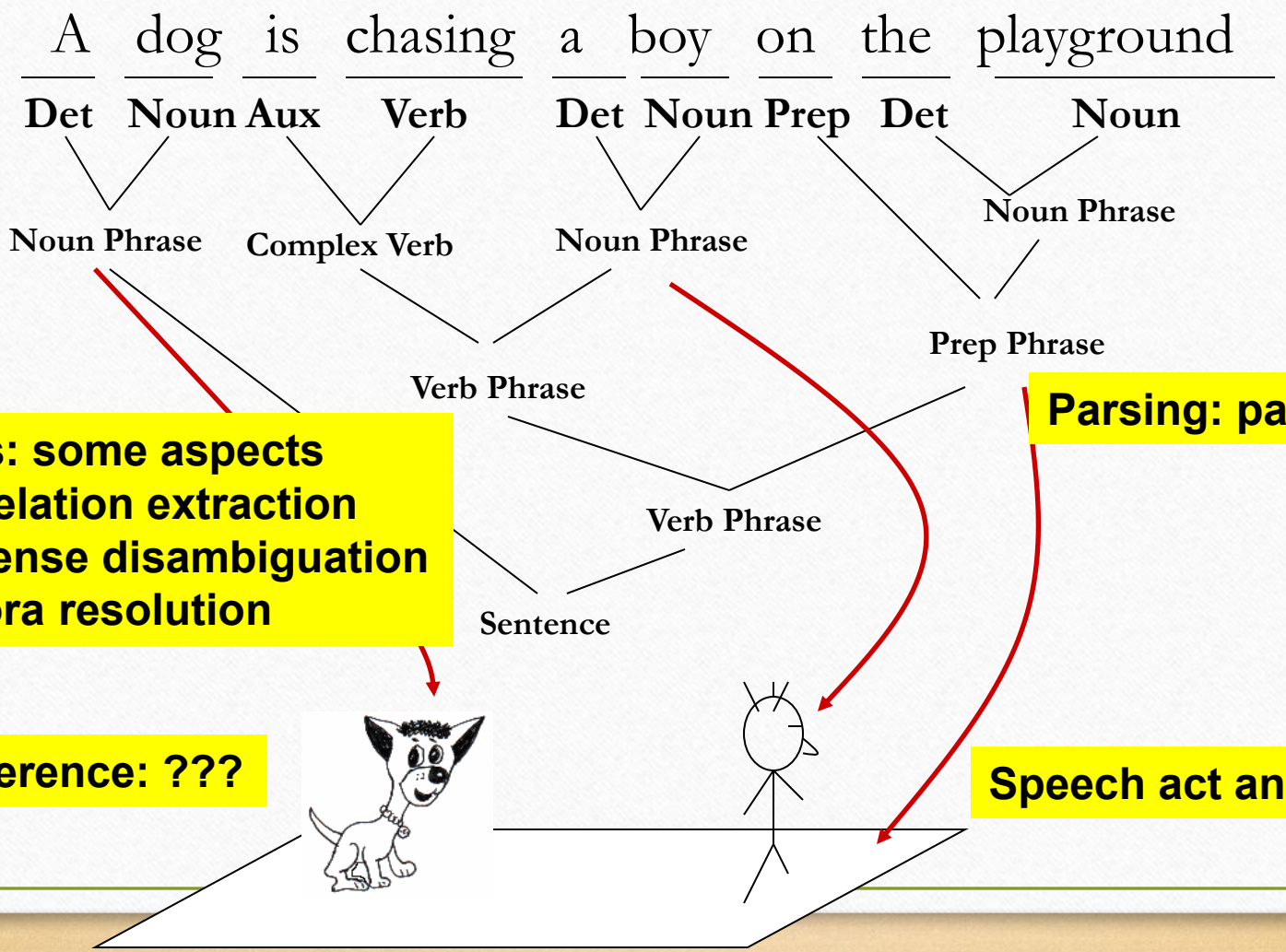
- Natural language is designed to make human communication efficient. Therefore,
  - We omit a lot of "common sense" knowledge, which we assume the hearer/reader possesses
  - We keep a lot of ambiguities, which we assume the hearer/reader knows how to resolve
- This makes EVERY step in NLP hard
  - Ambiguity is a "killer"!
  - Common sense reasoning is pre-required

An example of ambiguity

- Get the cat with the gloves.



The state of the art



POS Tagging: 97%

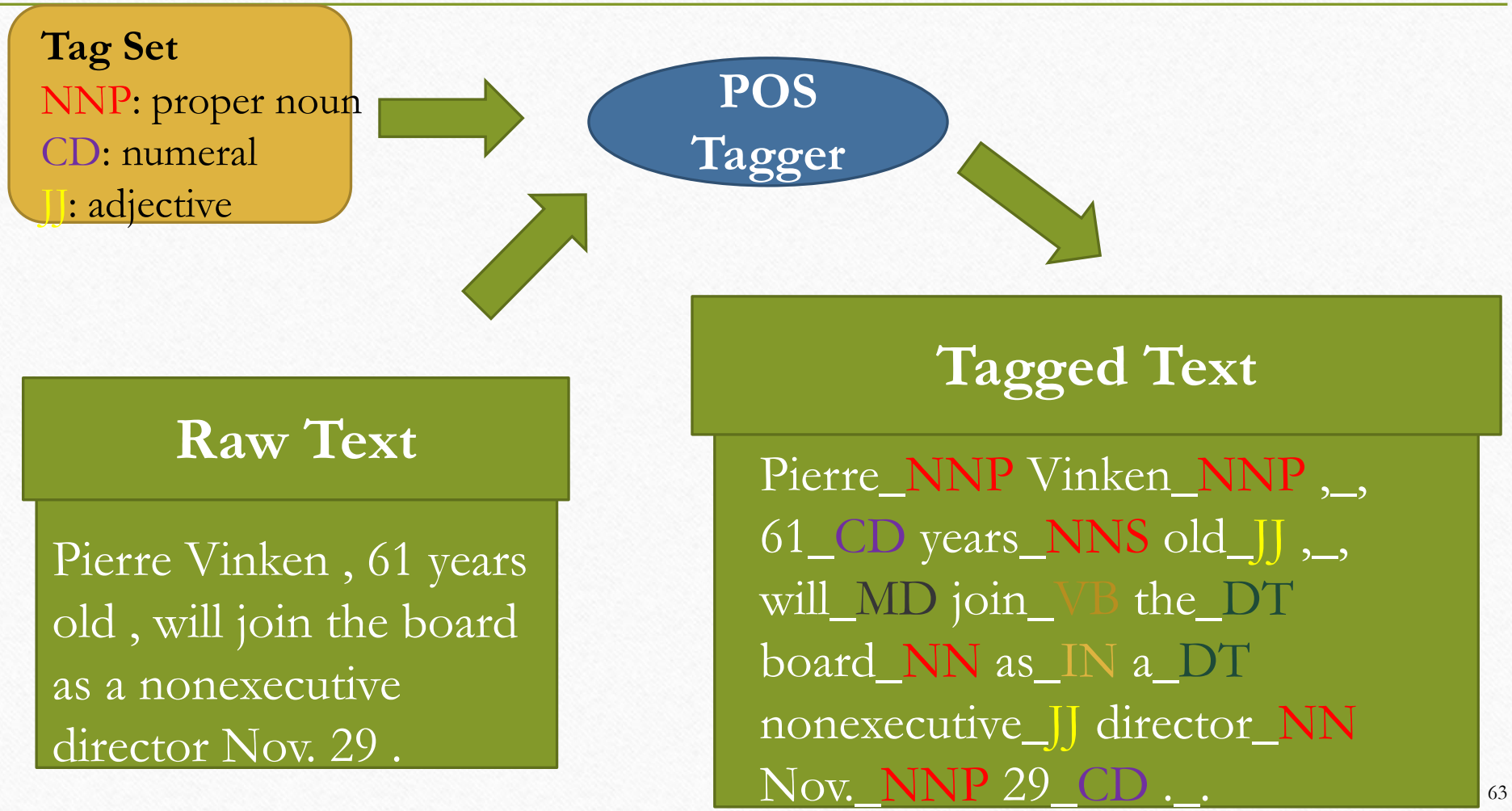
Parsing: partial >90%

Semantics: some aspects
- **Entity/relation extraction**
- **Word sense disambiguation**
- **Anaphora resolution**

Inference: ???

Speech act analysis: ???

What is POS tagging



Why POS tagging?

- POS tagging is a prerequisite for further NLP analysis
  - Machine translation
    - The meaning of a particular word depends on its POS tag
      - Verb: "They mean well but lack tact."
      - Noun: "The mean is 20." (mathematical mean)
      - Adjective: "Mean look."
  - Sentiment analysis
    - Adjectives are the major opinion holders
      - Good v.s. Bad, Excellent v.s. Terrible

Why POS tagging?

- POS tagging is a prerequisite for further NLP analysis
  - Syntax parsing
    - Basic unit for parsing
  - Information extraction
    - Indication of names, relations

Why POS tagging?

- Used for lemmatizing, an alternative to stemming:
  - *Stemming* usually refers to a crude heuristic process that chops off the ends of words in the hope of achieving this goal correctly most of the time, and often includes the removal of derivational affixes.
  - *Lemmatization* usually refers to doing things properly with the use of a vocabulary and morphological analysis of words, normally aiming to remove inflectional endings only and to return the base or dictionary form of a word, which is known as the *lemma*.
  - Example: the token *saw* (i.e. either a noun representing a tool for cutting or the past tense of the verb see)
    - Stemming returns: *s*
    - Lemmatization returns: *see* or *saw (depending if verb or noun)*

Challenges in POS tagging

- Words often have more than one POS tag
  - The back door (adjective)
  - On my back (noun)
  - Promised to back the bill (verb)
- Simple solution with dictionary look-up does not work in practice
  - One needs to determine the POS tag for a particular instance of a word from its context

Define a tagset

- We have to agree on a standard inventory of word classes
  - Taggers are trained on a labeled corpora
  - The tagset needs to capture semantically or syntactically important distinctions that can easily be made by trained human annotators

Word classes

- Open classes
  - Nouns, verbs, adjectives, adverbs
- Closed classes
  - Auxiliaries and modal verbs
  - Prepositions, Conjunctions
  - Pronouns, Determiners
  - Particles, Numerals

Public tagsets in NLP

- Brown corpus - Francis and Kucera 1961
  - 500 samples, distributed across 15 genres in rough proportion to the amount published in 1961 in each of those genres
  - 87 tags
- Penn Treebank - Marcus et al. 1993
  - Hand-annotated corpus of Wall Street Journal, 1M words
  - 45 tags, a simplified version of Brown tag set
  - Standard for English now
    - Most statistical POS taggers are trained on this Tagset

Is POS tagging a solved problem?

- Baseline
  - Tag every word with its most frequent tag
  - Tag unknown words as nouns
  - Accuracy
    - Word level: 90%
    - Sentence level
      - Average English sentence length 14.3 words
      - $0.9^{14.3} = 22\%$

Accuracy of State-of-the-art POS Tagger

- *Word level: 97%*
- *Sentence level:* $0.97^{14.3} = 65\%$

Public POS taggers

- Brill's tagger
  - http://www.cs.jhu.edu/~brill/
- TnT tagger
  - http://www.coli.uni-saarland.de/~thorsten/tnt/
- Stanford tagger
  - http://nlp.stanford.edu/software/tagger.shtml
- SVMTool
  - http://www.lsi.upc.es/~nlp/SVMTool/
- GENIA tagger
  - http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/
- More complete list at
  - http://www-nlp.stanford.edu/links/statnlp.html#Taggers

Pre-processing in practice

Pre-processings

- Remove punctuation
- Remove stopwords and short words
- Remove numbers
- POS tagging
- Stemming / Lemmatizing
- Case converter
- Create bag of words
- Filter terms with low frequency

Acknowledgments

- Slides have been compiled from several sources:
- Hongning Wang, Lecture slides on Text Mining, University of Virginia, USA
- Jiawei Han, Micheline Kamber, and Jian Pei, University of Illinois at Urbana-Champaign & Simon Fraser University (Data Mining: Concepts and Techniques 3rd ed.)