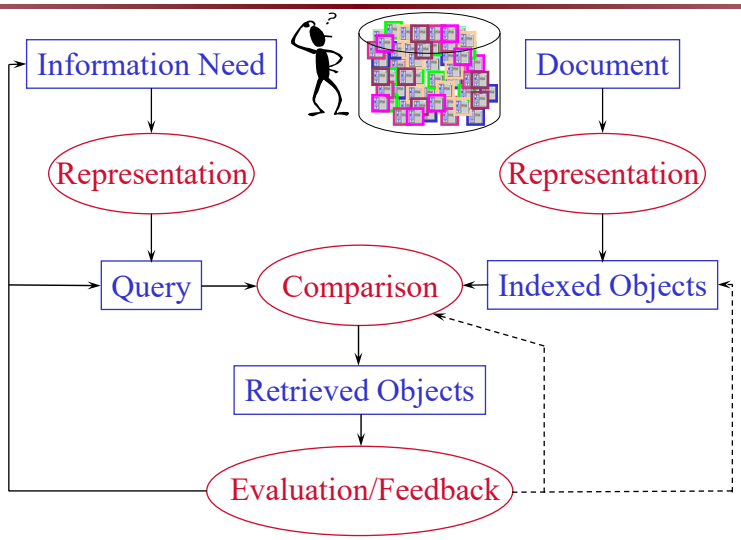
11-442 / 11-642 / 11-742:
Search Engines

Document Representation

Jamie Callan
Carnegie Mellon University
callan@cs.cmu.edu

1

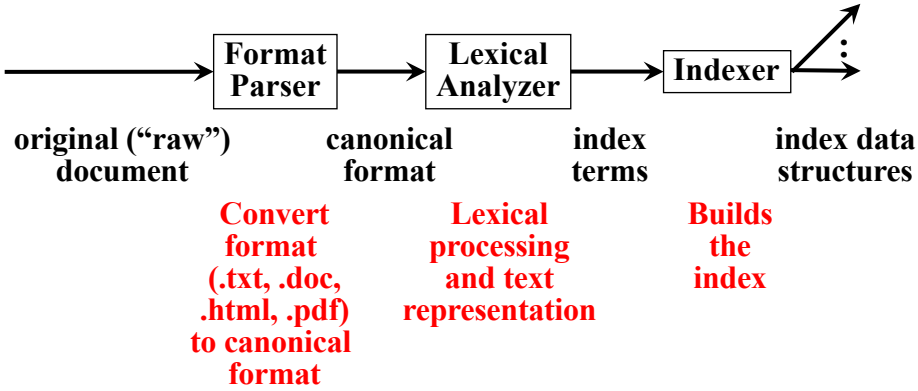
Overview of Information Retrieval Processes



2

2

Lexical Processing and Text Representation: Overview



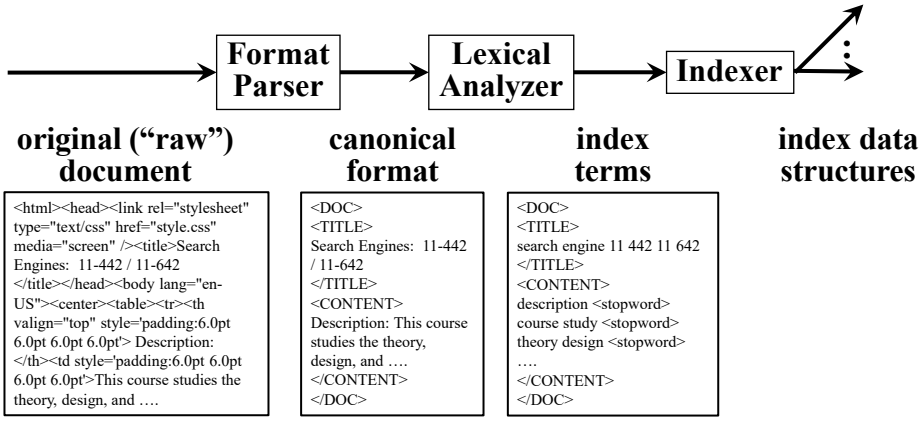
Task: (Quickly) convert document tokens into index terms

Issues: What makes a good index term?

3

3

Lexical Processing and Text Representation: Overview



Task: (Quickly) convert document tokens into index terms

Issues: What makes a good index term?

4

4

A Document is an Object That Contains Information

Metadata

- Typically <attribute, value> data
- E.g., date, author, price, language, …

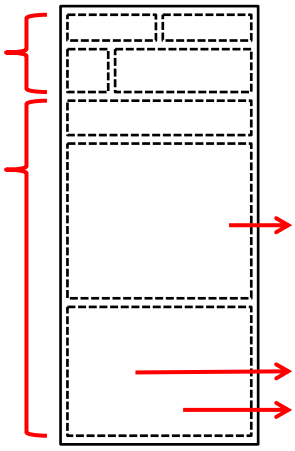
Content

- Maybe organized in fields, sections, elements
- Fields/elements may be related or unrelated
  - E.g., title, body (related)
  - E.g., complaint, payment history (unrelated)

Relations with other documents

- E.g., citations, hyperlinks (→)

A document



5

5

Document Attributes

We don't talk much about document attributes in this course, but they are an important component of search interfaces

6

6

How is the Information Content in a Document Represented?

There are two approaches to representing information content

- **Free-text or full-text index terms (this lecture)**
  - Terms selected from the text of the document
  - Terms selected from texts related to this document
  - Invented in modern times, but more familiar to many people
- **Controlled vocabulary index terms (next week)**
  - Terms selected from a well-defined classification scheme
  - Invented in ancient times

7

7

Free-Text Indexing

Main Idea: Select a few index term from the document

- **Note:** This is an uncontrolled vocabulary
- **Advantages:**
  - Index terms guaranteed to be a good match to document contents
  - No need to learn a (possibly complex) controlled vocabulary
  - Possibly easier to automate than controlled-vocabulary indexing
- **Disadvantage:**
  - Greater possibility of vocabulary-mismatch problems
    - » E.g., document says “automobile”, query says “car”

8

8

Free-Text & Full-Text Indexing

How should the terms be chosen?

- **Use selected terms from the document ("free-text indexing")**
  - Historically this was tried first
  - Usually done manually
  - Major issues: Which terms? Selected how?
    - » Essentially a feature selection problem
- **Use all terms from the document ("full text indexing")**
  - Avoids selection problems
  - Easy to automate
  - Major issue: The terms aren't equally useful
    - » Feature improvement, feature weighting, …

9

9

Free-Text & Full-Text Indexing

Free-text and full-text indexing are appealing

… but they are harder than they seem

- Words are very specific – are they really good index terms?
  - There are many ways to express the same concept
- What is a word, anyway?

Full-text indexing

- Transform (messy) language into reliable index terms

10

10

Full-Text Indexing: Overview

Basic lexical processing

- Tokens
- Case conversion
- Stopwords
- Morphological processing ("stemming")

Other representations

- Phrases, citations and inlink text, paths and urls

Multiple representations

11

11

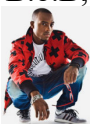
Tokens

The text stream is segmented into tokens

Typically, segment English text on whitespace and punctuation

It sounds easy, but …

- trade-in, quad-core, well-qualified, 12-month, all-star
- crowd-pleasing, family-friendly, CE 46–120, 747-400, …
- 802.11 b/g/n, cancel/extend, AT&T, O'Neill, …
- B.o.B, will.i.am, Too $hort



Usually this part of the system is carefully tuned heuristics

12

12

Lexical Processing

The text parser typically processes one token at a time

… looking and looking for a new camera to …

↑

Current token

Why?

- Lexical processing needs to be <u>really fast</u>, so it must be lightweight
  - You're touching every byte of a very big file
- Usually lightweight, local processing is sufficient
  - Deeper NLP hasn't provided much additional value (yet)

13

13

Lexical Processing

Search engines use
shallow language analysis and heuristics
to convert lexical tokens (usually words)
into index terms ('features')

This improves the ability to match queries to documents

- It ignores 'unimportant' differences in language usage

14

14

Lexical Processing

What should go into the index?

- **Are these useful?**
  - Stopwords
- **Are these the same concept?**
  - Morphological variants
- **Are these the same concept?**
  - Proper names
- **Are these the same concept?**
  - Case conversion

… looking and looking for a new camera to …

… any of these cameras …

… was set to buy the Sony …

… the new Best Buy in …

… a 3x Optical zoom …

… The optical zoom …

(topjimmy5150, Apr 21, 2003, Epinions.com)

15

15

Lexical Processing

Heuristic methods are used to map tokens to indexing terms

- Discard some tokens (“stopwords”)
  - E.g., “and”, “the”
- Normalize a token (e.g., case conversion)
  - E.g., “Optical” → “optical”
- Map a token to another token (“stemming”, “conflation”)
  - E.g., “images” → “image”
- **…**

This part of the system has a big effect on accuracy

- Often poor performance is due to a poor text representation

16

16

Full-Text Indexing

Let's generalize the full-text idea slightly

- **Select features or indexing terms from the document**
  - Maybe a feature is derived from words in the document
  - Maybe a feature is only related to words in the document
- **Maybe don't use every feature in the document**
  - "Feature selection"
- **Full-text indexing**
  - Document words / tokens → Index features / terms

17

17

Stopwords

Stopwords: Words that are discarded from a document representation

- Typically function words: a, an, and, as, for, in, of, the, to, …

Why remove stopwords?

- Reduces index size
  - Significantly!
- Can improve accuracy
  - Why?

Rank	Term	Frequency	Proportion
1	the	4,352,160	6.31%
2	of	2,134,125	3.09%
3	to	2,023,402	2.93%
4	a	1,811,373	2.63%
5	in	1,546,782	2.24%
6	and	1,507,140	2.18%
7	s	855,190	1.24%
8	that	787,792	1.14%
9	for	780,138	1.13%
10	is	605,988	0.88%
Total			**23.77%**

Wall Street Journal (1987-1992)
Documents: 174K
Tokens: 69M

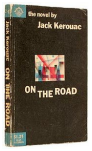
18

18

Disadvantages of Stopword Removal

What happens to these queries?

- To be or not to be
- Eye for an eye
- Let it be
- In the name of love
- On the road
- The Rite



Removing stopwords makes some queries difficult to satisfy

19

19

Query-Based Stopword Removal

An increasingly common solution

- **Store stopwords in the index**
  - Index becomes much larger, but maybe cost is less important
- **<u>Usually</u> discard stopwords from queries**
  - The Last Exorcism → Last Exorcism
- **<u>Occasionally</u> leave stopwords in the query**
  - E.g., if stopwords are more than x% of query terms
    - » The Rite → The Rite
  - E.g., if user indicates that they should be retained
    - » +the last → the last (+ indicates a required term)

20

20

Stopword Lists

Stopword lists are usually developed manually

- **Sort term dictionary based on frequency**
- **Examine the most frequent terms**
- **Examine a query log to see which frequent terms might be important**
  - E.g., "trading" and "prices" are very frequent in the Wall Street Journal
    …so they are potential stopwords
    …but they are important terms
    …so leave them in

21

21

First 60 Words From the Lemur Stopword List (418 Stopwords Total)

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

22

22

The Lucene Stopword List

a	in	the
an	into	their
and	is	then
are	it	there
as	no	these
at	not	they
be	of	this
but	on	to
by	or	was
for	such	will
if	that	with

23

23

Document Representation

How should this document be represented?

A Great Choice.
Review by topjimmy5150
April, 21 2003

I have been looking and looking for a new camera to replace our bulky, but simple and reliable (but only fair picture taker) Sony Mavica FD73. My other choice (Besides the more expensive Nikon Coolpix 3100) was the (also more expensive) Sony Cybershot P72. I recommend any of these cameras, and I was set to buy the Sony, but at the last minute I cheaped out and bought the 2100. No regrets. I bought the camera (along with 128mb memory card (the stock 16mb card will be kept in the bag as a spare) and carrying case) at the new Best Buy in Harrisburg, PA. I also bought a set of 4 Nickle-Metal Hydride rechargable batteries and charger at Walmart for less than $20. I keep 2 in the camera and two in the charger/in the camera bag along with the original Lithium battery pack as spares.

Hands down, the best feature of this camera is it's compact design. It is very small. My family likes to go camping during the summer, and last year we found the Mavica too

(topjimmy5150, Epinions.com)

24

24

Full-Text Indexing

Term	Tf	Term	Tf	Term	tf
the	78	up	8	pictures	6
to	35	for	7	red	6
i	31	have	7	digital	5
and	29	image	7	eye	5
a	19	like	7	not	5
camera	17	mode	7	on	5
is	17	much	7	or	5
in	12	software	7	shutter	5
with	11	very	7	sony	5
be	9	can	6	than	5
but	9	images	6	that	5
it	9	movies	6	after	4
of	9	my	6	also	4
this	9	no	6	: :	:

25

25

Full-Text Indexing: After Stopword Removal

Term	Tf	Term	Tf	Term	tf
camera	17	after	4	lcd	3
up	8	any	4	looking	3
image	7	auto	4	mavica	3
like	7	buy	4	problem	3
mode	7	flash	4	recorded	3
software	7	2100	3	reduction	3
images	6	bought	3	size	3
movies	6	button	3	zoom	3
pictures	6	down	3	15	2
red	6	feature	3	2mp	2
digital	5	focus	3	8x10	2
eye	5	included	3	98	2
shutter	5	lag	3	automatically	2
sony	5	last	3	batteries	2

26

26

Morphology

Concepts are often expressed by a family of words that are variations of a single root word

- **Morphology:** "a study and description of word formation (as inflection, derivation, and compounding) in language"

  -- *Merriam-Webster Dictionary*
- **Lemmatisation:** "the process of determining the lemma (canonical form) for a given word" -- *wikipedia*
  - Usually called **stemming** for English, because much of English morphology happens at the end of a word
- **Conflation:** Treating two entities as if they were the same entity
  - Example: conflate "computers" and "computer"

27

27

Conflating Morphological Variants

Inverted list for "image"

df: 109

docid=18, tf=3, locs=14, 39, 52

docid=92, tf=1, locs=79

: : :

Inverted list for "images"

df: 57

docid=18, tf=2, locs=27, 68

docid=58, tf=1, locs=19

: : :

Conflated inverted list for {"image", "images"}

df: 121

docid=18, tf=5, locs=14, 27, 39, 52, 68

docid=58, tf=1, locs=19

docid=92, tf=1, locs=79

: : :

Could also include "imaging", "imaged", "imager", …

28

28

Stemming Algorithms for English

Porter

- Many heuristics, not clear why they work well
- Often produces stems that aren't words
  - E.g., police → polic, executive → execut
- http://www.tartarus.org/martin/PorterStemmer/

KSTEM

- Rule-based, dictionary, heuristics, Porter
- Nearly always produces real words as stems
- http://lemurproject.org/ and http://lexicalresearch.com/

Very different behaviors, but about equally fast & effective

29

29

Stemming Examples

Original Text

Document will describe marketing strategies carried out by U.S. companies for their agricultural chemicals, report predictions for market share of such chemicals, or report market statistics for agrochemicals.

Porter Stemmer (stopwords removed)

market strateg carr compan agricultur chemic report predict market share chemic report market statist agrochem

KSTEM (stopwords removed)

marketing strategy carry company agriculture chemical report prediction market share chemical report market statistic agrochem

30

30

Is Stemming a Good Idea?

When might stemming be expected to improve results?

- **Enterprise search?**
  - Corpora are usually smaller, so Recall is usually important
  - Users are more likely to be tolerant of stemming mistakes because relevant documents are harder to find
- **Web search?**
  - Corpora are massive, so Recall is usually less important
  - Users are more likely to be intolerant of stemming mistakes because there are so many relevant documents
  - Originally Google didn't do stemming … now it seems to

31

31

More Advanced Morphology

Some languages make significant use of compound terms

- E.g., German, Dutch, Finnish, …
- E.g., computerviren ("computer viruses")

Treating the entire compound as a single term can reduce Recall

- "computer" won't match "computerviren"

The solution is <u>decompounding</u> to improve Recall

- E.g., conflate computerviren, computer, viren

This is a <u>different use of conflation</u>

- Instead of mapping the conflated terms to a common index term

  …pretend that the conflated terms occurred <u>at the same location</u>

32

32

More Advanced Morphology: German Decompounding

The parser can insert several index terms at each location

Text:	Ein	Computervirus	ist	ein	sich	selbst	verbreitendes ...	
Index terms:	*	computervirus computer virus	*	*	*	selbst	verbreitendes	
Location:	1	2	3	4	5	6	7	

Text:	... Computerprogramm,	welches	sich	in	andere …		
Index terms:	computerprogramm computer program	*	*	*	*		
Location:	8	9	10	11	12 ...		

*** is a stopword**

http://de.wikipedia.org/wiki/Computervirus

33

33

Effect of Decompounding on Accuracy

Experimental results indicate that decompounding greatly improves accuracy

- E.g., more than 25% in German
- E.g., from 10-28% in Dutch

34

34

Morphological Analysis: Summary

The good news:

- Conflating variations of a word
  - Provides a more accurate representation of the document
  - Enables a broader range of queries to (correctly) match

The bad news:

- Effects are inconsistent
- Terms can be grouped mistakenly (e.g., Apple, Apples)
- Sophisticated morphological analysis can be very slow

Final verdict: Done in most systems, but still a source of debate

35

35

Full-Text Indexing: Overview

Basic lexical processing

- Tokens
- Stopwords
- Morphological processing ("stemming")

Other representations

- Citations and inlink text, paths and urls

Multiple representations

36

36

Text Representation: Other Sources of Evidence

Full-text indexing is not restricted to text in the body of the document…

…useful clues about document content come from many sources

- Citations in "traditional" text
- Anchor text in hypertext (e.g., Web) documents ("inlink text")
- Word in a file name or path (e.g., URL)

Using multiple independent representations improves reliability

- If the title, body, url, and inlink representations all contain 'apple', it is very likely that the document is about apple

37

37

Text Representation: Citations

Citations are common in legal documents

When this Court held in *Artuz* v. *Bennett,* 531 U. S. 4, 8, 11, that time limits on postconviction petitions are "condition[s] to filing," such that an untimely petition would not be deemed "properly filed," it reserved the question …

-- U.S. Supreme Court case 03-9627

This citation provides clues about <u>what is significant</u> about Artuz v. Bennett

- Time limits on postconviction petitions
- An untimely petition would not be deemed properly filed
- Probably these are great index terms

38

38

Text Representation: Inlink Text

Citations are common on the web

Jamie Callan

This citation provides clues about <u>what is significant</u> about http://www.cs.cmu.edu/~callan>

– Jamie Callan

It is especially useful if the document doesn't contain text

– E.g., image, video, audio, software, …

39

39

Text Representation: File Paths and URLs

All computer files are described by file names and paths

- http://www.cs.cmu.edu/~callan/
- C:\Documents and Settings\callan\Desktop\Pictures\Birthday_0001.jpg

Principle: Word in a file name or path may describe the object

- A noisy representation, but important for some information needs
  - E.g., retrieving home pages
- **Issue:** "Stop tokens" such as "www" and "html"
- **Issue:** Are all tokens in a deep link equally useful?

No clear rules, but many effective heuristics

40

40

Full-Text Indexing: Overview

Basic lexical processing

- Tokens
- Stopwords
- Morphological processing ("stemming")

Other representations

- Citations and inlink text, paths and urls

Multiple representations

41

41

Multiple Representations on the Web

…. …. …. …. ….
...Little Jack Horner…
…. .. …. … … ….

…. …. …. …. ….
……the stealing of a deed …. .. …. … …

…. …. …. …. ….
…… 16th century …. nurseryrhyme ….

…. …. …. …. ….
… … .. Thomas …. Horner and the …. Abbot of …. …. …. Glastonbury …. …
… … …. …. …. ….

http://nurseryrhymes.org/jack_horner.html

<title> Little Jack Horner </title>
<body>
Little Jack Horner
Sat in the corner,
Eating of Christmas pie;
He put in his thumb
And pulled out a plumb,
And cried, What a good boy am I!
</body>

Representations

- Derived from the document
- Derived from anchor text
- Derived from URL

(Ogilvie, 2005)

42

42

Multiple Representations on the Web

Multiple representations are stored in document fields

	Document
Url terms	nurseryrhymes jack horner
Title terms	little jack horner
Body terms	little jack horner sat corner eat christmas pie put thumb pull out plumb cry good boy
Inlink terms	little jack horner steal deed 16th century nursery rhyme thomas horner abbot glastonbury

43

43

Full-Text Representation Summary

Search engines use a variety of heuristics to turn text into index terms (features)

- Derive index terms from the document
  - Tokenization, case conversion, stopword removal, stemming, …
- Derive index terms from citations
  - Traditional citations, inlink text
- Derive index terms from file names and paths
  - URLs
- …

44

44

Full-Text Representation Summary

The state of the art is to use multiple sources of evidence to determine what the document is about

- E.g., text from the title, body, metadata, url, inlink, …

Gather as many clues as possible about what the document means

Treat each type of evidence as a separate representation of the doc

- Store separately (later lecture)
- Enable the query to reference each type of evidence
  - E.g., #AND (cmu.url callan.title)
- Enable retrieval models to use many types of evidence

45

45

Document Representation Summary

Free-text or full-text index terms

- Basic lexical processing
  - Tokens
  - Stopwords
  - Morphological processing ("stemming")
- Other representations
  - Phrases, citations and inlink text, paths and urls
- Multiple representations

Controlled vocabulary index terms

46

46