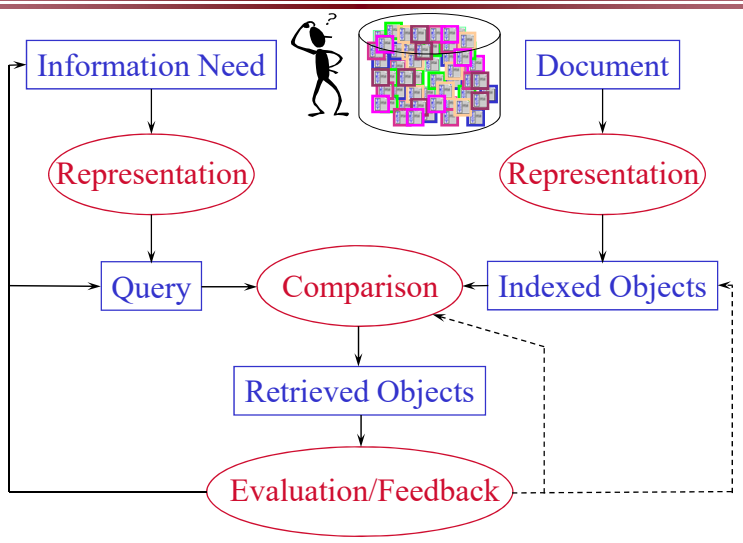
11-442 / 11-642:
Search Engines

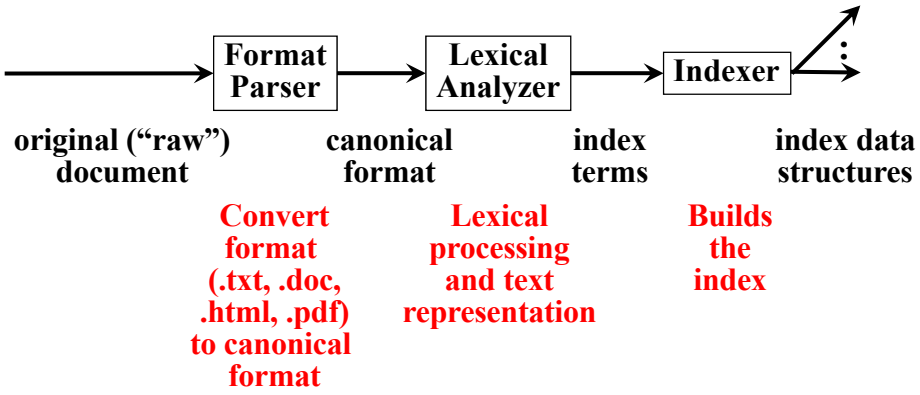
Document Representation

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

Overview of Information Retrieval Processes



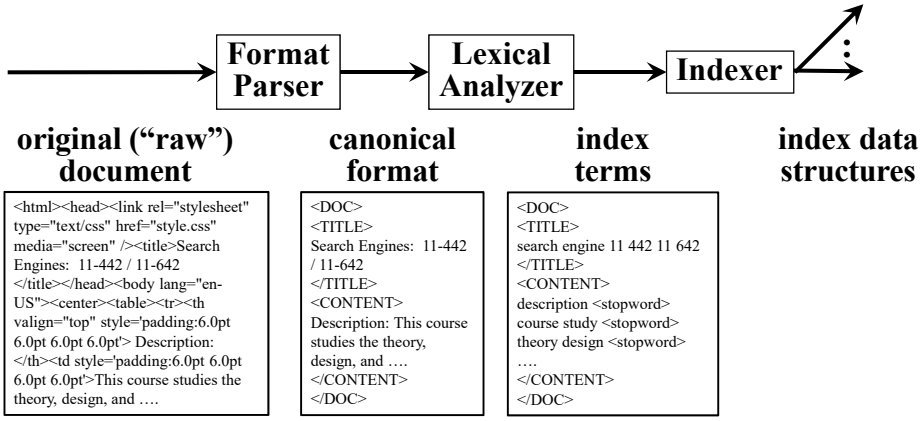
Lexical Processing and Text Representation: Overview



Task: (Quickly) convert document tokens into index terms

Issues: What makes a good index term?

Lexical Processing and Text Representation: Overview



Task: (Quickly) convert document tokens into index terms

Issues: What makes a good index term?

A Document is an Object That Contains Information

Metadata

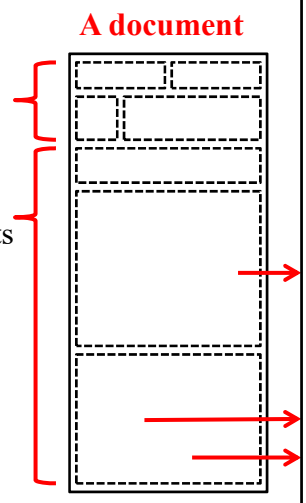
- Typically <attribute, value> data
- E.g., date, author, price, language, …

Content

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

Relations with other documents

- E.g., citations, hyperlinks (→)



5

Document Attributes

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

6

How is the Information Content in a Document Represented?

There are two approaches to representing information content

- **Controlled vocabulary index terms**
  - Terms selected from a well-defined classification scheme
- **Free-text or full-text index terms**
  - Terms selected from the text of the document
  - Terms selected from texts related to this document

7

Introduction to Controlled Vocabularies

Subject-based classification was the first approach to indexing

- The Library of Alexandria (3rd century B.C.E. to 30 B.C.E.)

Define a set of categories / labels / subject descriptors

- A controlled vocabulary of index terms
  - Only these terms can be used to represent document contents
- E.g., medicine, business, politics, entertainment, …

Assign 1-n controlled vocabulary term(s) to each document

Use controlled vocabulary term(s) to find desired information

- E.g., use controlled vocabulary terms to form a query
- E.g., browse the controlled vocabulary hierarchy to find documents

8

What is a Controlled Vocabulary?

Library Science defines a controlled vocabulary to have several components

- A set of rules for identifying the subject of a document
- Sometimes a thesaurus specifying different forms of a topic
- A group of indexing terms
- A set of instructions for assigning indexing terms

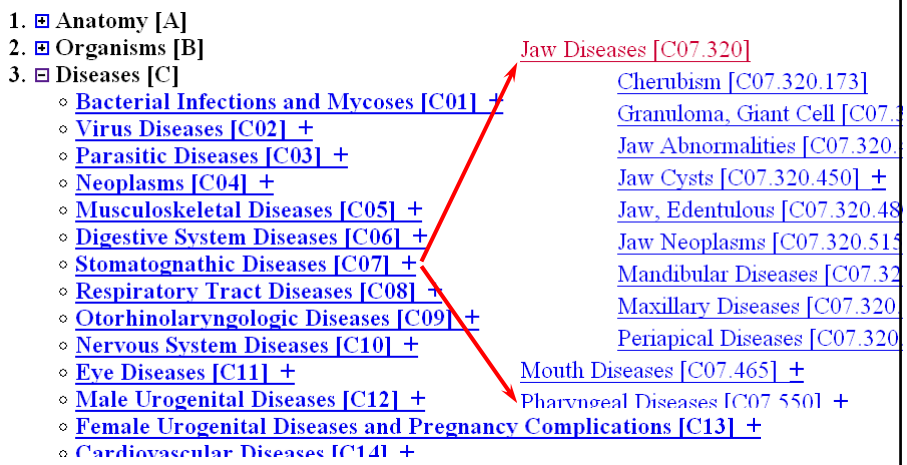
9

Controlled Vocabularies: Medical Subject Headings (MeSH)

1. ⊞ Anatomy [A]
2. ⊞ Organisms [B]
3. ⊞ Diseases [C]
4. ⊞ Chemicals and Drugs [D]
5. ⊞ Analytical, Diagnostic and Therapeutic Techniques and Equipment [E]
6. ⊞ Psychiatry and Psychology [F]
7. ⊞ Biological Sciences [G]
8. ⊞ Natural Sciences [H]
9. ⊞ Anthropology, Education, Sociology and Social Phenomena [I]
10. ⊞ Technology, Industry, Agriculture [J]
11. ⊞ Humanities [K]
12. ⊞ Information Science [L]
13. ⊞ Named Groups [M]
14. ⊞ Health Care [N]
15. ⊞ Publication Characteristics [V]
16. ⊞ Geographicals [Z]

10

Controlled Vocabularies: Medical Subject Headings (MeSH)



11

Document Text

How should this document be represented?

J Pak Med Assoc. 2015 Feb;65(2):225-7.

Artificial sweeteners: safe or unsafe?

Qurrat-ul-Ain, Khan SA.

Abstract

Artificial sweeteners or intense sweeteners are sugar substitutes that are used as an alternative to table sugar. They are many times sweeter than natural sugar and as they contain no calories, they may be used to control weight and obesity. Extensive scientific research has demonstrated the safety of the six low-calorie sweeteners currently approved for use in foods in the U.S. and Europe (stevia, acesulfame-K, aspartame, neotame, saccharin and sucralose), if taken in acceptable quantities daily. There is some ongoing debate over whether artificial sweetener usage poses a health threat .This review article aims to cover thehealth benefits, and risks, of consuming artificial sweeteners, and discusses natural sweeteners which can be used as alternatives.

12

Controlled Vocabulary Indexing: How PubMed Indexes the Document

Metadata

AU- Qurrat-ul-Ain
LA- eng
PT- Journal Article
PT – Review
: : : : :

Chemical Abstracts Service (CAS) terms

RN - 0 (Dipeptides)
RN - 0 (Sweetening Agents)
RN - 0 (Thiazines)
RN - 56038-13-2 (trichlorosucrose)
RN - 57-50-1 (Sucrose)
RN - FST467XS7D (Saccharin)
RN - MA3UYZ6K1H (acetosulfame)

Medical Subject Heading (MeSH) terms

MH - Aspartame/adverse effects
MH - Diabetes Mellitus, Type 2 …
MH - Dipeptides/adverse effects
MH - Humans
MH - Neoplasms/*chemically induced
MH - Obesity/*chemically induced
MH - Saccharin/adverse effects
MH - Sucrose/adverse effects/analogs …
MH - Sweetening Agents/*adverse effects
MH - Weight Gain

13

Introduction to Controlled Vocabularies

There are many controlled vocabularies

- Broad vocabularies describe many topics at a general level
- Detailed vocabularies describe a fewer topics in great detail
- There is a coverage vs. detail tradeoff (you can't have both)

Many types of representations have controlled vocabularies

- Taxonomies, ontologies, semantic web, knowledge bases, …
- Key characteristics: Predefined index terms, defined semantics

The next few slides show examples of controlled vocabularies

- Some are formal and well-defined
- Some are informal and less well-defined

14

Introduction to Controlled Vocabularies: Library of Congress Subject Headings

- A: General Works
- B: Philosophy. Psychology. Religion
- C: Auxiliary Sciences Of History
- D: World History And History Of Europe, Asia, Africa, Australia, New Zealand, Etc.
- E: History Of The Americas
- F: History Of The Americas
- G: Geography. Anthropology. Recreation
- H: Social Sciences
- J: Political Science
- K: Law
- L: Education
- M: Music And Books On Music
- N: Fine Arts
- P: Language And Literature
- Q: Science
- R: Medicine
- S: Agriculture
- T: Technology
- U: Military Science
- V: Naval Science

(U.S. Library of Congress, 2012)

15

Introduction to Controlled Vocabularies: Library of Congress Subject Headings

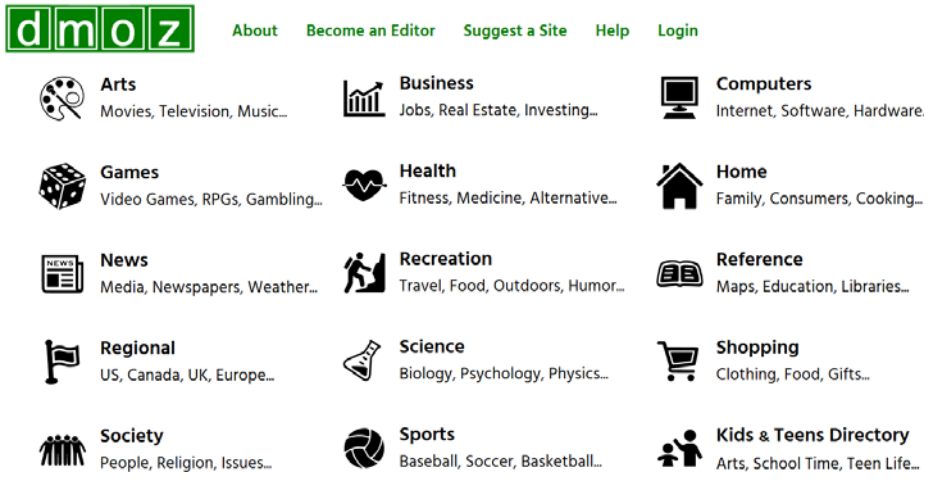
Subclass M

M1-5000	Music
M1-1.A15	Music printed or copied in manuscript in the United States or the colonies before 1860
M1.A5-3.3	Collections
M1.A5-Z	Miscellaneous
M2-2.3	Musical sources
M3-3.3	Collected works of individual composers
M5-1480	Instrumental music
M5	Collections
M6-175.5	One solo instrument
M176	Motion picture music
M176.5	Radio and television music
M177-990	Two or more solo instruments
M1000-1075	Orchestra
M1100-1160	String orchestra
M1200-1270	Band
M1350-1366	Other ensembles
M1375-1420	Instrumental music for children
M1470-1480	Aleatory music. Electronic music. Mixed media

(U.S. Library of Congress, 2012)

16

Controlled Vocabularies: DMOZ (The Open Directory Project)



http://dmoz.org/

17

Controlled Vocabulary Indexing: Freebase

American football
Amusement Parks
Architecture
Astronomy
Atom Feeds
Automotive
Aviation
Awards
Baseball
Basketball
Bicycles
Biology
Boats

Books
Boxing
Broadcast
Business
Celebrities
Chemistry
Comics
Common
Community
Computers
Conferences and Conventions
Cricket
: : :

/business/advertising_slogan
/business/asset
/business/asset_owner
/business/board_member_title
/business/brand
/business/business_operation
/business/competitive_space
/business/consumer_company
/business/consumer_product
/business/customer
/business/employer
/business/endorsed_product
/business/industry
: : : :

18

(http://www.freebase.com, 2012)

Controlled Vocabularies: Summary

Advantages

- Index terms have clear semantics, consistent usage
  - Concepts rather than words enables higher Recall
- Supports both browsing and search

Disadvantages:

- Coverage vs. detail tradeoff
- Expensive to create and maintain
- Difficult for people to assign to documents consistently
- Not easy for most people to use for search

Popular in some fields (e.g., medicine, law, patent)

19

Free-Text Indexing

Main Idea: Select a few index term from the document

- **Free-text indexing uses 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”

20

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, …

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

Document Representation

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)

23

Full-Text Indexing: Overview

Basic lexical processing

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

Other representations

- Phrases, citations and inlink text, paths and urls

Multiple representations

24

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

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 really fast, 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)

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

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)

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 is the part of the system that most affects accuracy

- Often poor performance is due to a poor text representation

29

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

30

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

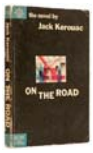
31

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

32

Query-Based Stopword Removal

An increasingly common solution…

- **Store stopwords in the index**
  - Index becomes much larger, but disks are inexpensive (maybe!)
- **Usually discard stopwords from queries**
  - The Last Exorcism → Last Exorcism
- **Occasionally leave stopwords in the query**
  - E.g., if stopwords are more than half the query terms
    - » The Rite
  - E.g., if user indicates that they should be retained
    - » +the last (the + indicates a required term)

Stopword Lists

Stopword lists are usually developed manually

- **Sort 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 would be potential stopwords
    - …but they are important terms
    - …so leave them in

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

35

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

36

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	:	:

37

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

38

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"

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", …

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/

Very different behaviors, but about equally fast & effective

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

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

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 decompounding

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

This is a different use of conflation

- Instead of mapping the conflated terms to a common index term …pretend that the conflated terms occurred at the same location

More Advanced Morphology: German Decompounding

Text:

Ein Computervirus ist ein sich selbst verbreitendes Computerprogramm, welches sich in andere …

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

Index terms produced by the parser (<term, location>):

<computervirus, 2> <computer, 2> <virus, 2>

<selbst, 6>

<verbreitendes, 7>

<computerprogramm, 8> <computer, 8> <program, 8>

<sich, 10>

…

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

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

Full-Text Indexing: Overview

Basic lexical processing

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

Other representations

- Phrases, citations and inlink text, paths and urls

Multiple representations

Phrases

Should these phrases be stored in the index?

- "white house"
- "interest rate",
- "roe v. wade",
- "american idol"
- "purple house"
- "table rate"
- "row the boat"
- "american idiot"

49

How Are Phrases Used?

Precoordinate

Store phrases in the index

- e.g., interest_rate
- Possibly insert constituents into index
  - e.g., "interest", "rate"

Replace query phrase with phrase index term

- e.g., "interest rate" → interest_rate

Look up phrase matches

Postcoordinate

Use a query operator to construct phrases

- e.g., #NEAR/1 (interest rate)

Retrieve inverted lists of terms

Combine them to find phrase matches

50

Phrase Recognition Methods: Part of Speech Tagging

Annotate each word using a part of speech tagger

- **Example**: "recent/**JJ** interest/**NN** rate/**NN** hikes/**NNS** have/**VBP**"
- Part of speech taggers are usually very fast

JJ: Adjective
NN: Noun
NNS: Plural noun
NNP: Singular proper noun
VBP: Verb

Match phrases by POS patterns

- **Example:** (NN | NNS | NNP | NNPS){2,8}

Matching phrases

- interest rate hikes, official interest rates, student loans

Maybe generate index terms for sub-phrases, to improve Recall

- interest rate hikes → "interest rate" "rate hikes"

51

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

52

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 what is significant about Artuz v. Bennett

– Time limits on postconviction petitions

– An untimely petition would not be deemed properly filed

53

Text Representation: Inlink Text

Citations are common on the web

Jamie Callan

This citation provides clues about what is significant 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, …

54

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

Full-Text Indexing: Overview

Basic lexical processing

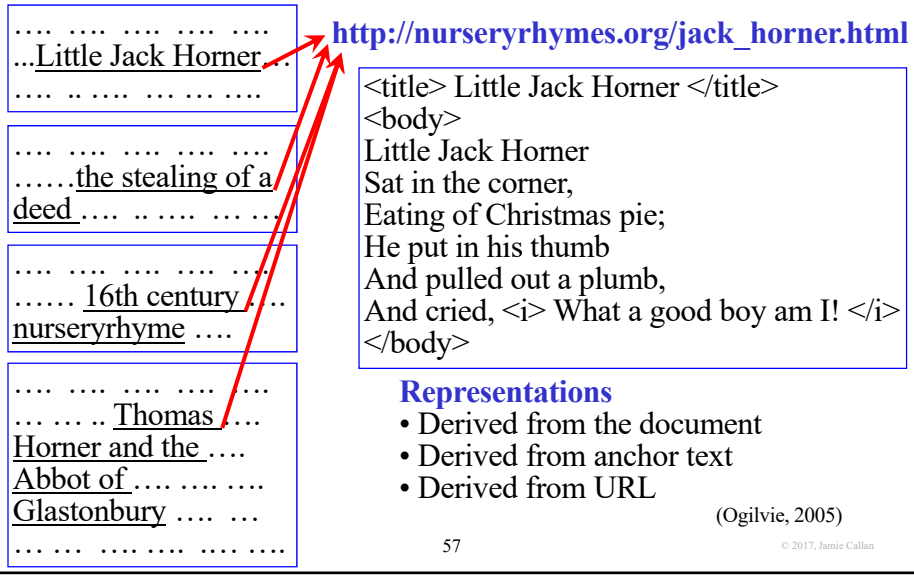
- Tokens
- Stopwords
- Morphological processing ("stemming")

Other representations

- Phrases, citations and inlink text, paths and urls

Multiple representations

Multiple Representations on the Web



Multiple Representations on the Web

Multiple representations are stored in document fields

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

Full-Text Representation Summary

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

- Derive index terms from the document
  - Stopword removal, stemming, phrases, …
  - Named entity and part-of-speech annotations (covered later)
- Derive index terms from citations
  - Traditional citations, inlink text
- Derive index terms from file names and paths
  - URLs
- …

Document Representation Summary

Controlled vocabulary index terms

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

Document Representation Summary

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

- E.g., controlled vocabulary terms
- 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 meaning

- 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 (later lecture)