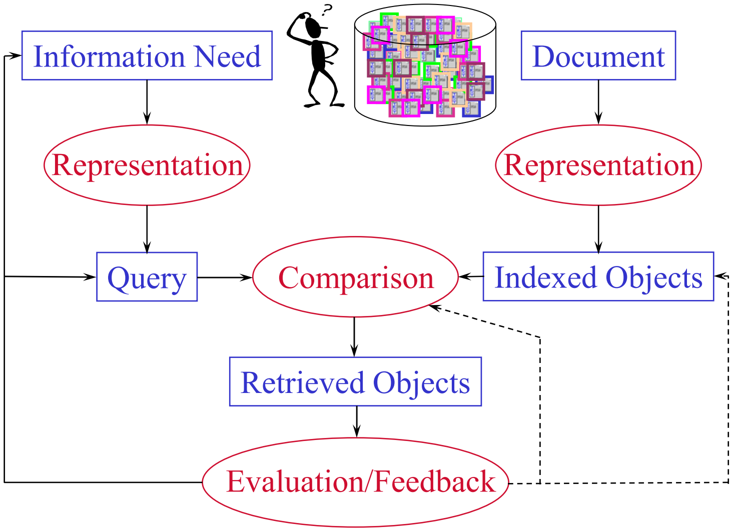
11-442 / 11-642 / 11-742: Search Engines

Document Representation (And Related Topics)

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

Overview of Information Retrieval Processes

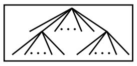


Indexing: Outline

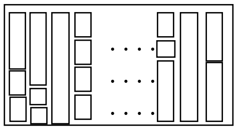
- Building inverted lists on a single processor
- Inverted lists and inverted files
  - – Inverted list compression
  - – Inverted list optimizations
- Forward indexes
- **Index updates**

Inverted File Management: Static File (No Updates)

Access Information (Small File)



Inverted Lists (Large File)



- Create files when inverted list fragments are merged
- There is no empty space between inverted lists
- Lists are stored in canonical order (e.g., alphabetic)
- Easy to create, very space efficient
- Very difficult to update; easier to rebuild
  - – Update by merging fragments with file to create new file

Updating Indexes

Indexes are expensive to update

- Suppose a new document contains 100 unique terms
- Adding that document means updating 100 inverted lists
  - – Acquire lock, read list, write list, release lock
  - – A lot of complexity, a lot of I/O
- Adding one document is tolerable, adding several is expensive

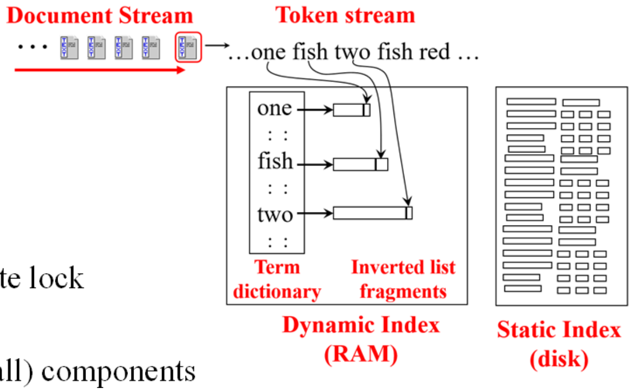
Updates are often done in batches

- Update every day, or after N documents arrive, or …
- Parse documents to generate index modifications
- Update each inverted list for all documents in the batch

Updating Indexes

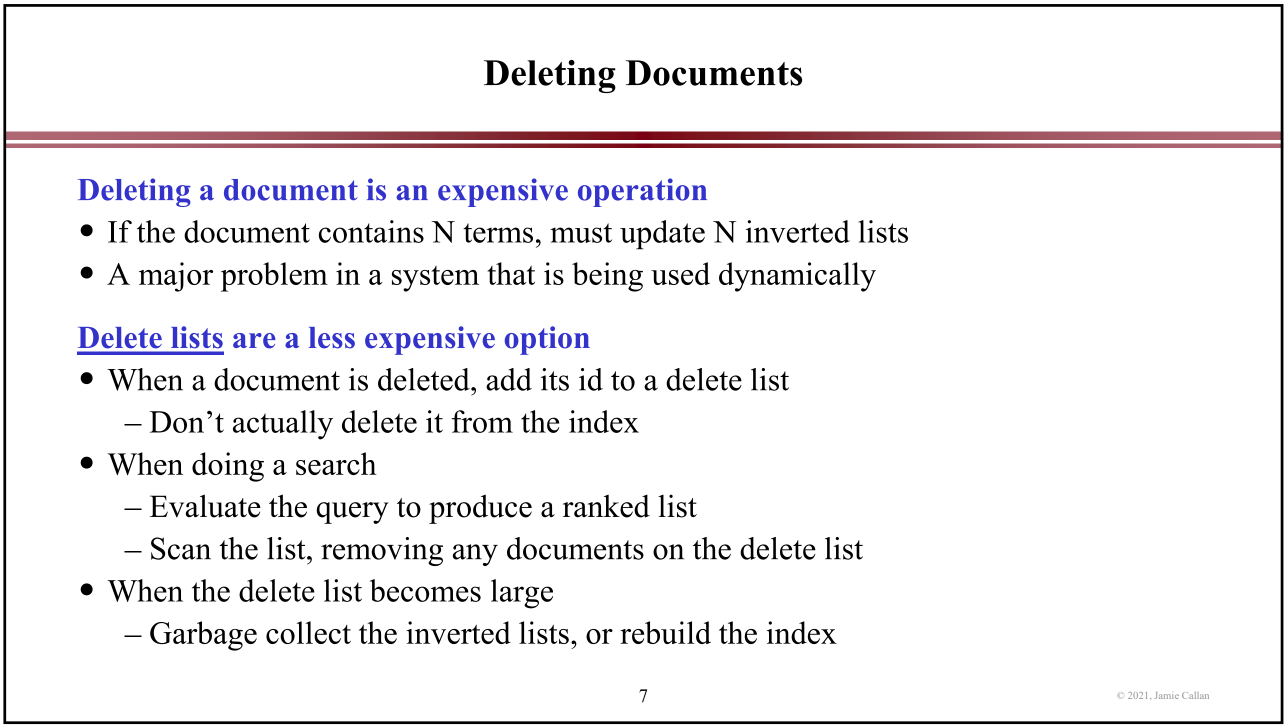
Sometimes dynamic updates are unavoidable

- E.g., news, Twitter, …



Split index into dynamic and static parts

- The dynamic index is small
- The static index is big
- Make updates to the dynamic index
  - – Acquire lock, read list, update list, write lock
  - – Faster because lists are small
- Search both static (big) and dynamic (small) components
- Periodically merge dynamic into static



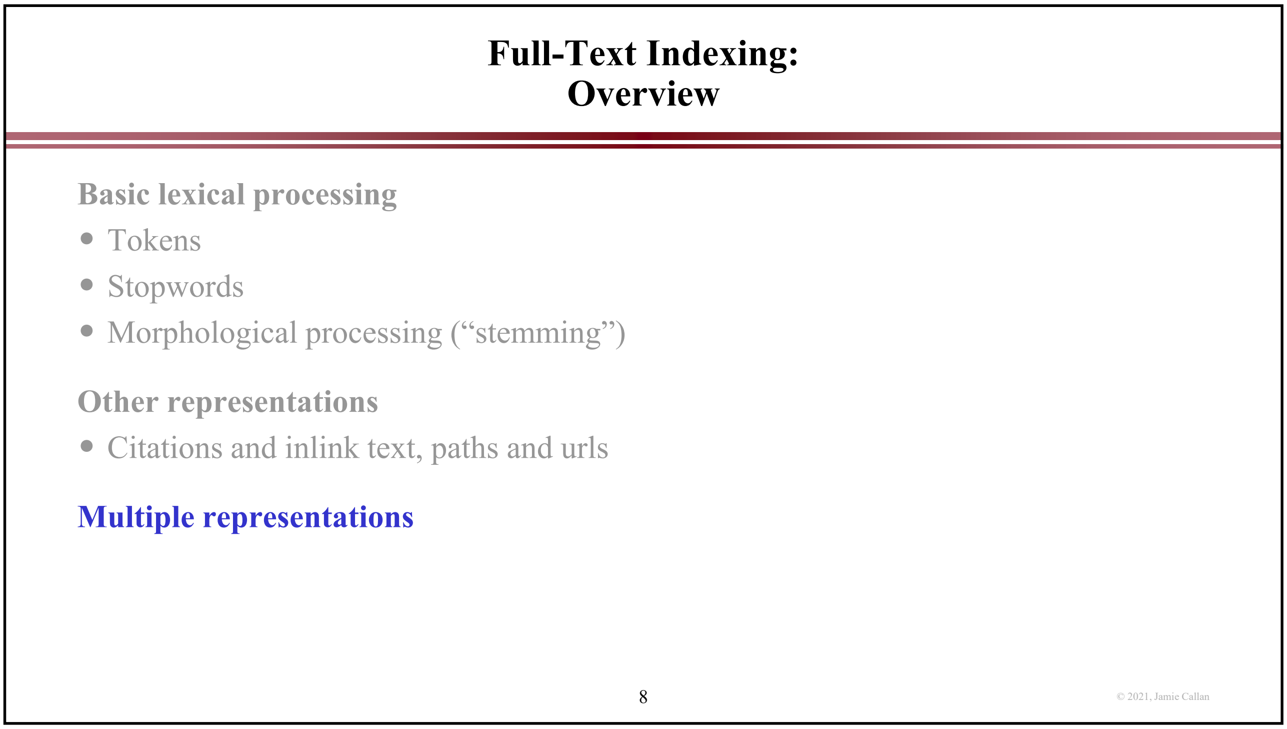
Deleting Documents

Deleting a document is an expensive operation

- If the document contains N terms, must update N inverted lists
- A major problem in a system that is being used dynamically

Delete lists are a less expensive option

- When a document is deleted, add its id to a delete list
  - Don't actually delete it from the index
- When doing a search
  - Evaluate the query to produce a ranked list
  - Scan the list, removing any documents on the delete list
- When the delete list becomes large
  - Garbage collect the inverted lists, or rebuild the index



Full-Text Indexing: Overview

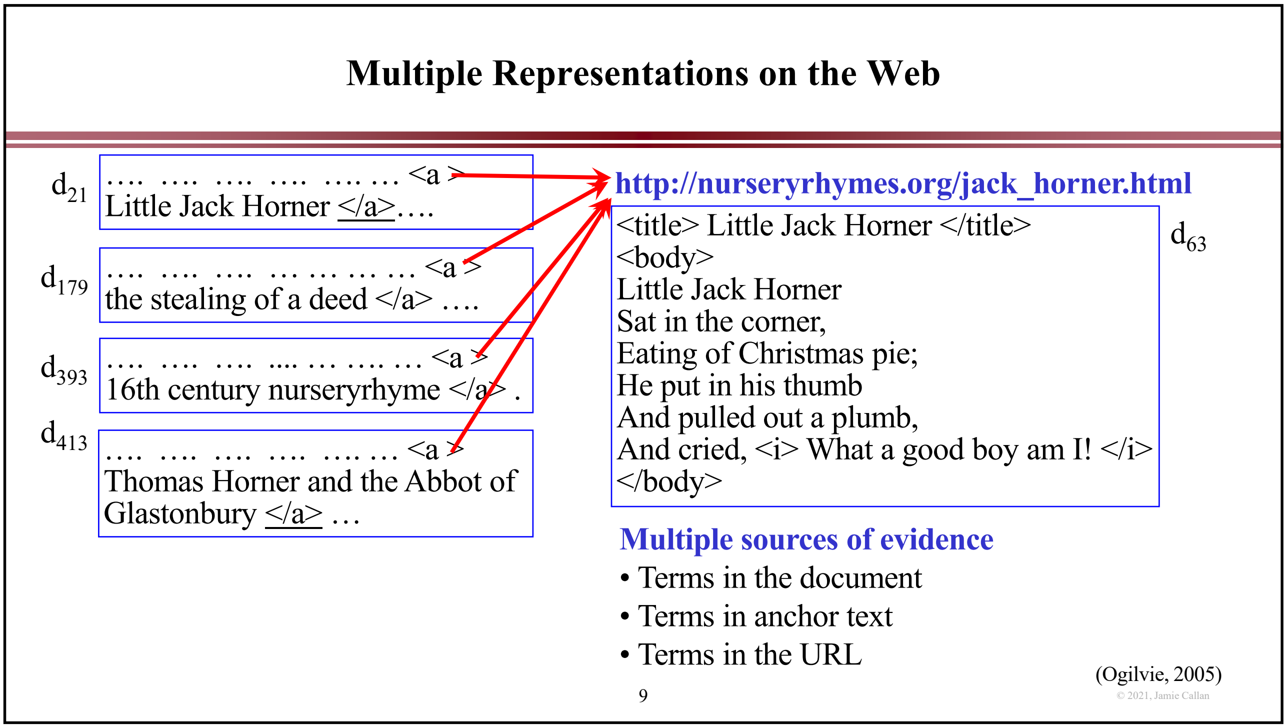
Basic lexical processing

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

Other representations

- Citations and inlink text, paths and urls

Multiple representations



Multiple Representations on the Web

d21 …. …. …. …. …. … <a >
Little Jack Horner </a>….

d179 …. …. …. … … … … <a >
the stealing of a deed </a> ….

d393 …. …. …. .... … …. … <a >
16th century nurseryrhyme </a> .

d413 …. …. …. …. …. … <a >
Thomas Horner and the Abbot of
Glastonbury </a> …

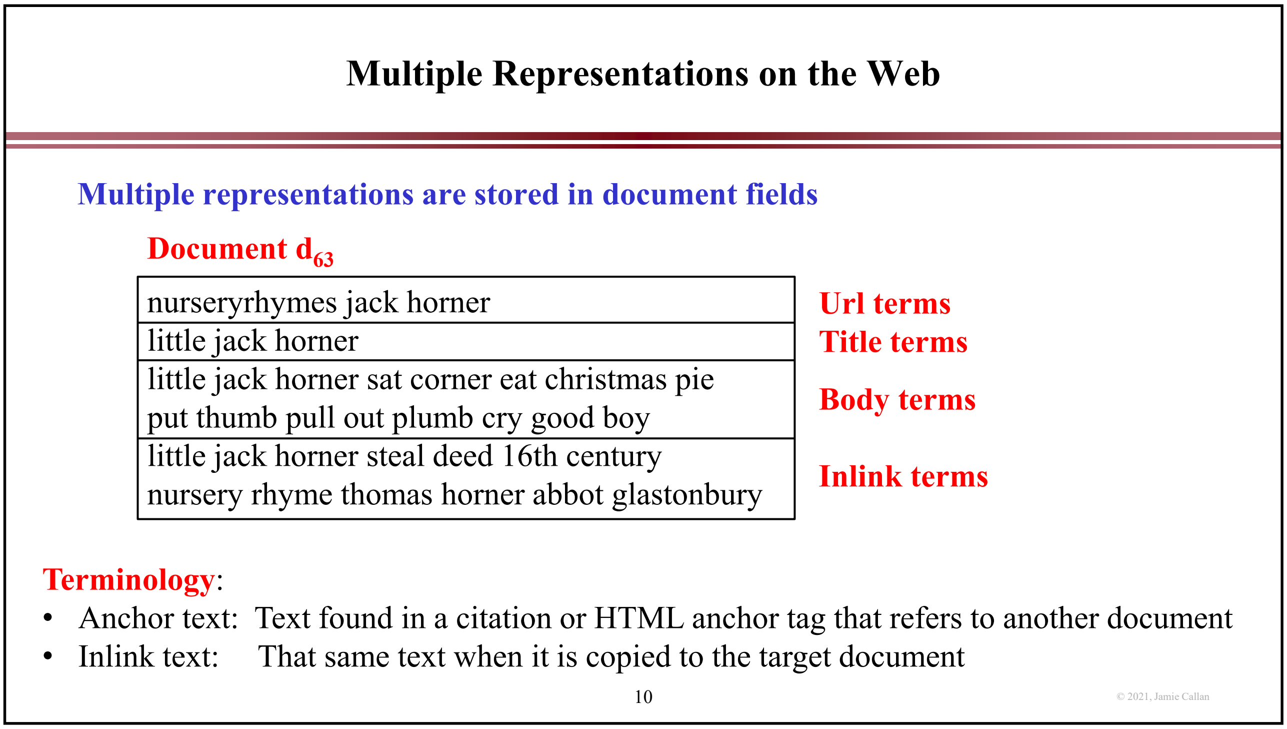
http://nurseryrhymes.org/jack_horner.html

d63
<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, <i> What a good boy am I! </i>
</body>

Multiple sources of evidence

- Terms in the document
- Terms in anchor text
- Terms in the URL

(Ogilvie, 2005)



Multiple Representations on the Web

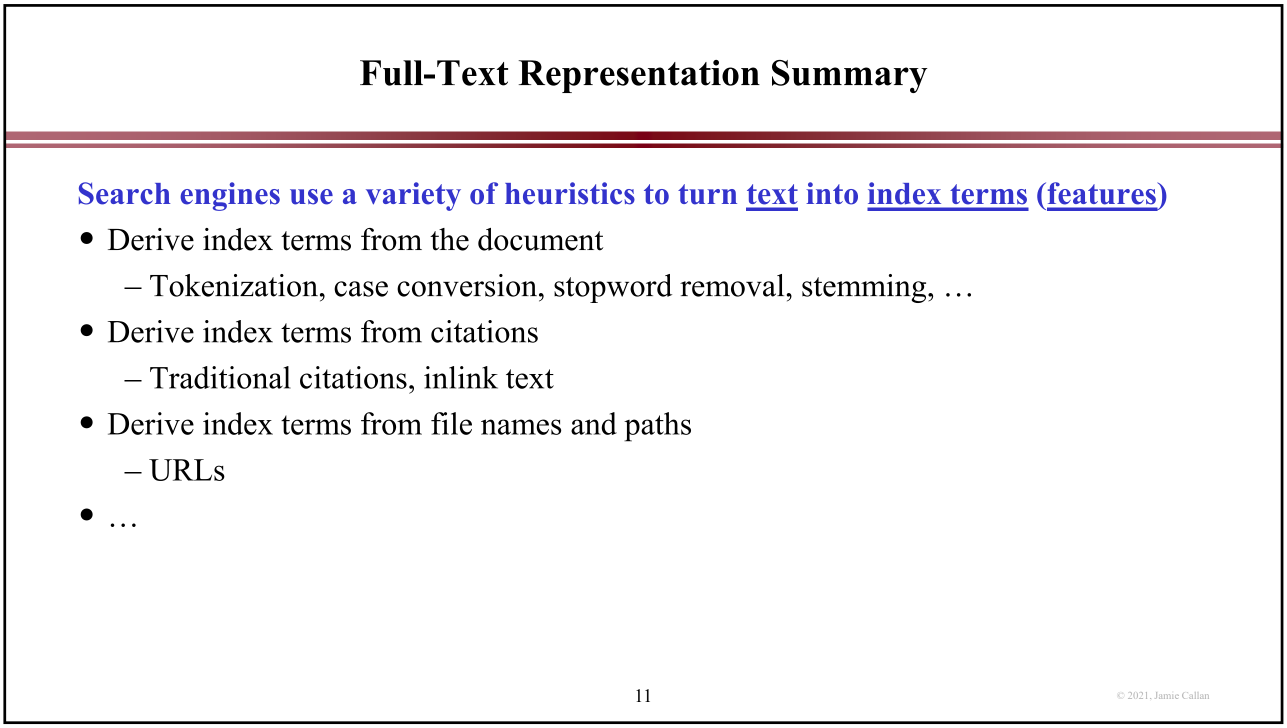
Multiple representations are stored in document fields

Document d_{63}

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

Terminology:

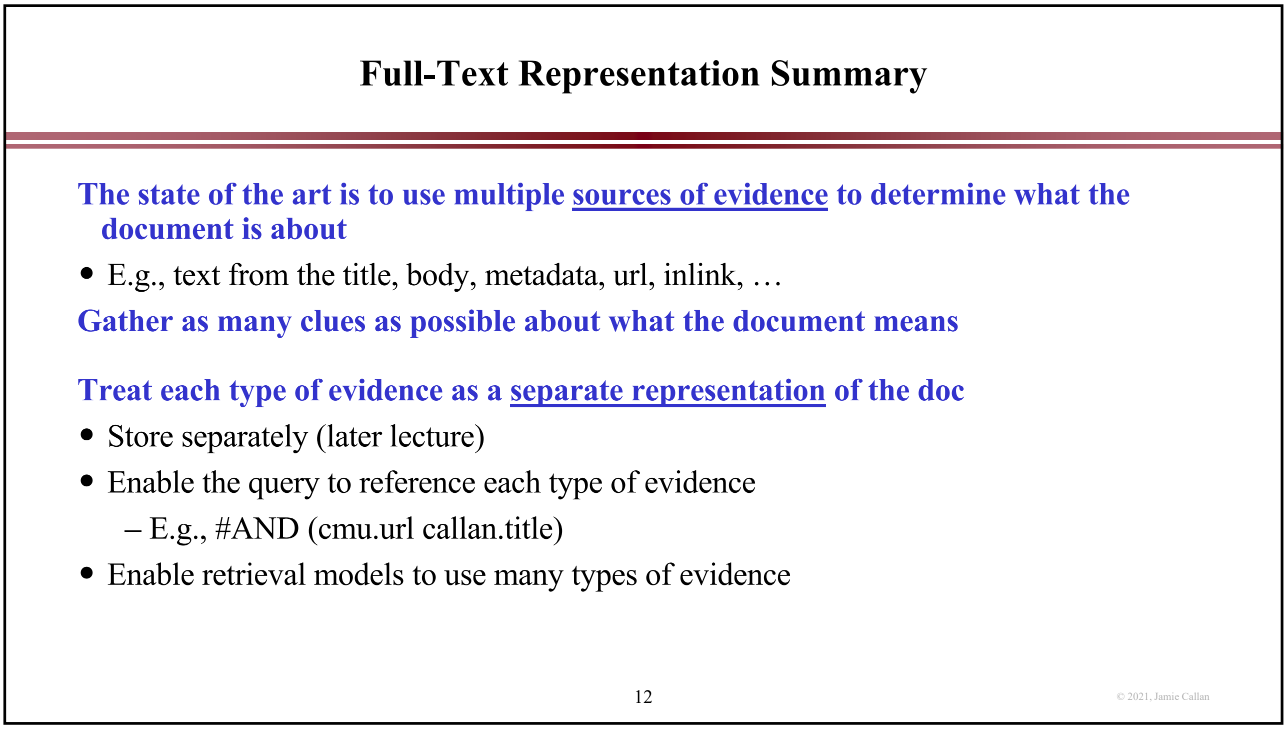
- Anchor text: Text found in a citation or HTML anchor tag that refers to another document
- Inlink text: That same text when it is copied to the target document



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
- …



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

Document Representation Overview

Free-text or <u>full-text</u> index terms

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

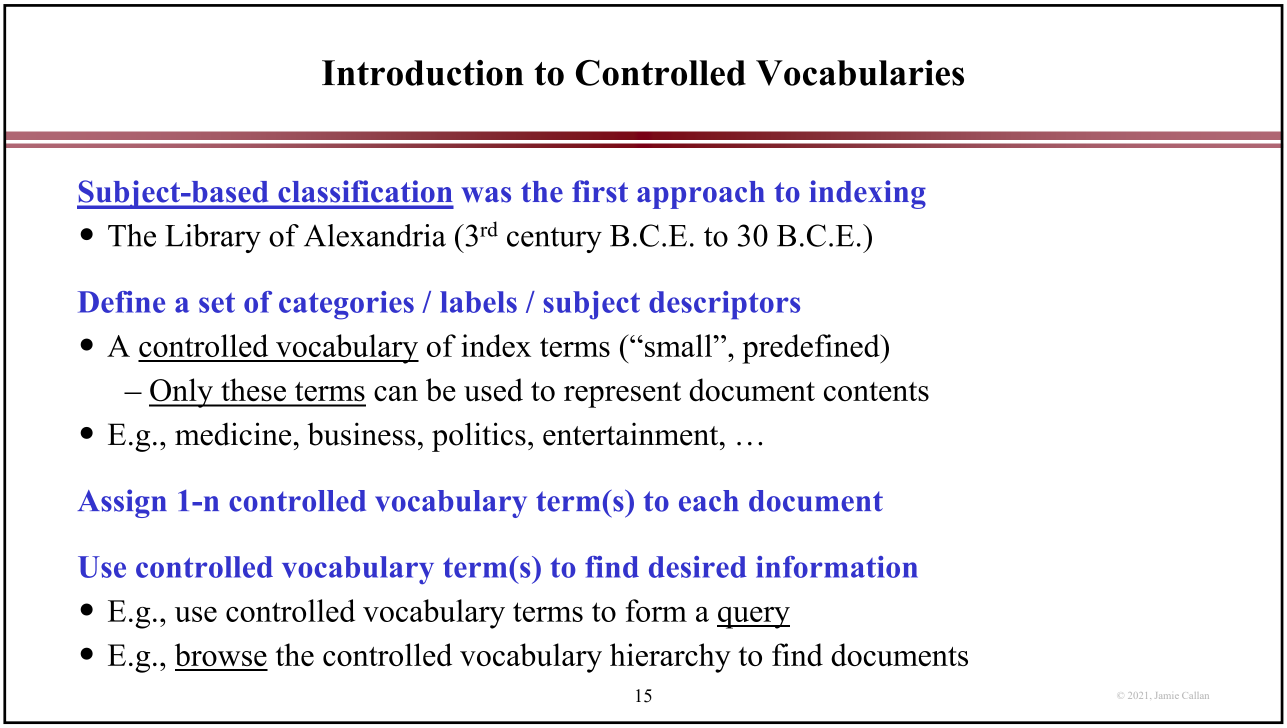
<u>Controlled vocabulary</u> index terms

Introduction to Controlled Vocabularies

<u>Subject-based classification</u> was the first approach to indexing

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

(Thanks to Scott Fahlman)



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.)

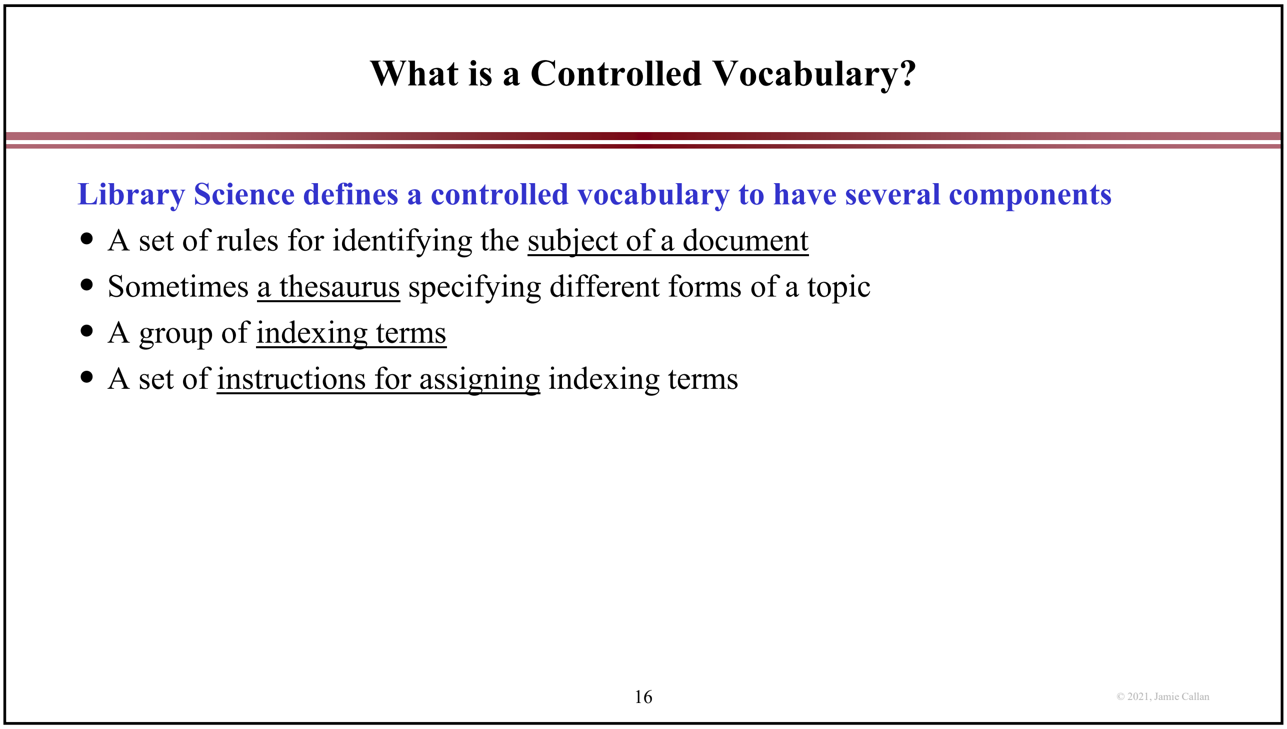
Define a set of categories / labels / subject descriptors

- A controlled vocabulary of index terms ("small", predefined)
  - 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

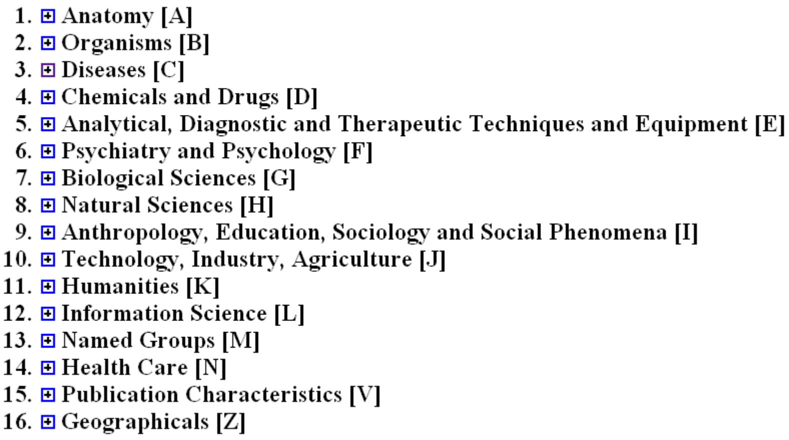


What is a Controlled Vocabulary?

Library Science defines a controlled vocabulary to have several components

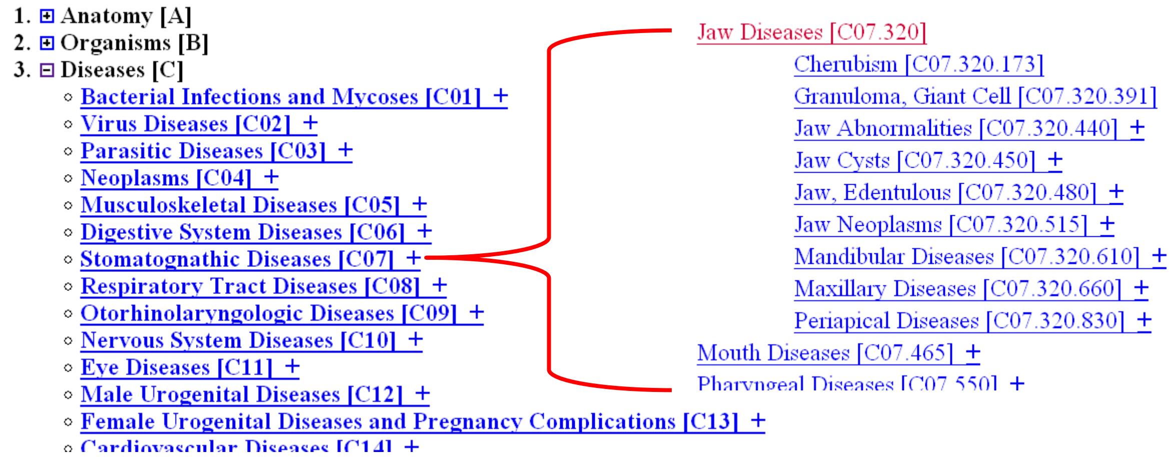
- A set of rules for identifying the subject of a document
- Sometimes a thesaurus specifying different forms of a topic
- A group of indexing terms
- A set of instructions for assigning indexing terms

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]

Controlled Vocabularies: Medical Subject Headings (MeSH)



1. ⊞ Anatomy [A]
2. ⊞ Organisms [B]
3. ⊟ Diseases [C]
   - Bacterial Infections and Mycoses [C01] +
   - Virus Diseases [C02] +
   - Parasitic Diseases [C03] +
   - Neoplasms [C04] +
   - Musculoskeletal Diseases [C05] +
   - Digestive System Diseases [C06] +
   - Stomatognathic Diseases [C07] +
   - Respiratory Tract Diseases [C08] +
   - Otorhinolaryngologic Diseases [C09] +
   - Nervous System Diseases [C10] +
   - Eye Diseases [C11] +
   - Male Urogenital Diseases [C12] +
   - Female Urogenital Diseases and Pregnancy Complications [C13] +
   - Cardiovascular Diseases [C14] +

Jaw Diseases [C07.320]

- Cherubism [C07.320.173]
- Granuloma, Giant Cell [C07.320.391]
- Jaw Abnormalities [C07.320.440] +
- Jaw Cysts [C07.320.450] +
- Jaw, Edentulous [C07.320.480] +
- Jaw Neoplasms [C07.320.515] +
- Mandibular Diseases [C07.320.610] +
- Maxillary Diseases [C07.320.660] +
- Periapical Diseases [C07.320.830] +

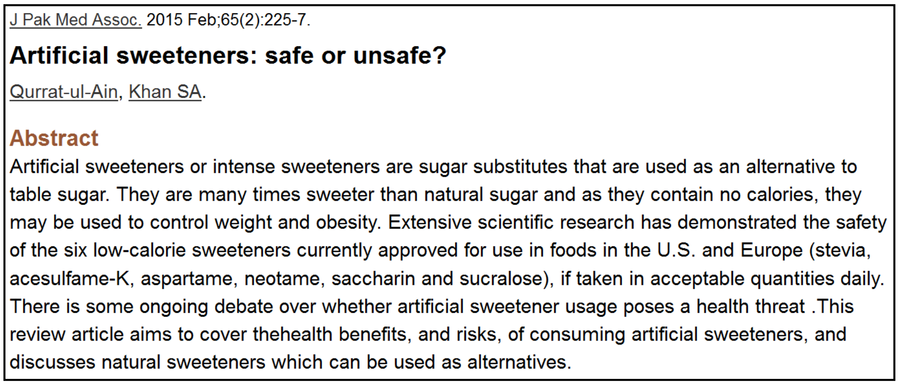
Mouth Diseases [C07.465] +

Pharyngeal Diseases [C07.550] +

The MeSH controlled vocabulary contains about 27,000 index terms

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.

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

Controlled Vocabulary Indexing:
How PubMed Indexes the Document

Treat the document as having controlled and free-text vocabulary fields

artificial sweeteners safe or unsafe **Title Field**

Dipeptides
Sweetening_Agents
Thiazines
trichlorosucrose
Sucrose
Saccharin
Acetosulfame

CAS Terms

Aspartame_adverse_effects
Diabetes_Mellitus_Type_2
Dipeptides_adverse_effects
Humans
Neoplasms_chemically_induced
Obesity_chemically_induced
Saccharin_adverse_effects
Sucrose_adverse_effects_analogs
Sweetening_Agents_adverse_effects
Weight_Gain

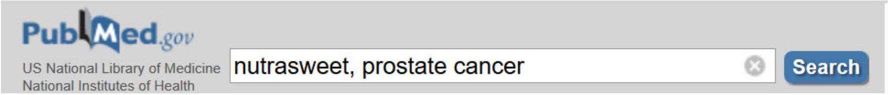
MeSH Terms

Abstract Field

artificial sweeteners intense sweeteners sugar substitute alternative table sugar many times sweeter natural sugar contain calories control weight obesity extensive scientific research demonstrate safety six low calorie sweetener current approve foods u.s. europe …

Controlled Vocabulary Indexing:
Query Reformulation

PubMed converts the user query to a structured query

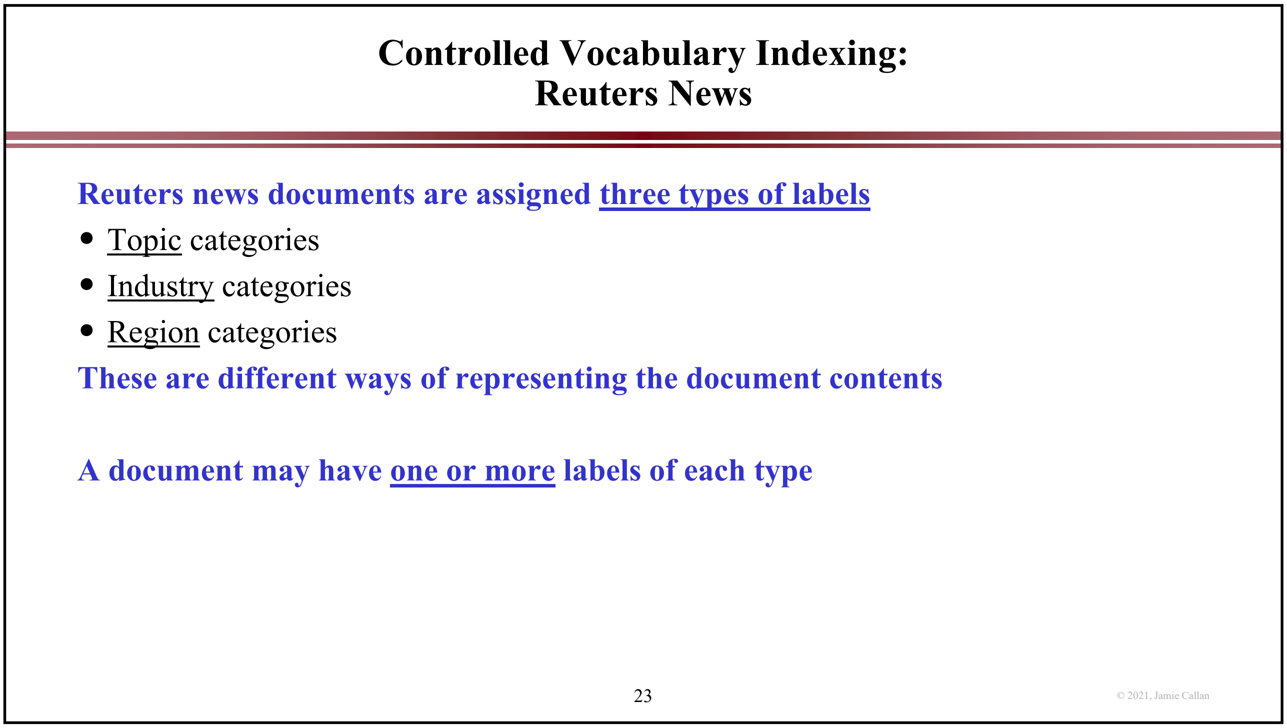

PubMed.gov
US National Library of Medicine
National Institutes of Health
Search

nutrasweet, prostate cancer

Search details

```
(   "nutrasweet"[All Fields]
    "aspartame"[MeSH Terms] OR
    "aspartame"[All Fields] OR]    )               AND
(   "prostate cancer"[All Fields] OR
    (    "prostate"[All Fields] AND "cancer"[All Fields]    ) OR
    "prostatic neoplasms"[MeSH Terms] OR
    (    "prostatic"[All Fields] AND "neoplasms"[All Fields]    ) OR
    "prostatic neoplasms"[All Fields])
```

The syntax is: "Query term"[Document field]



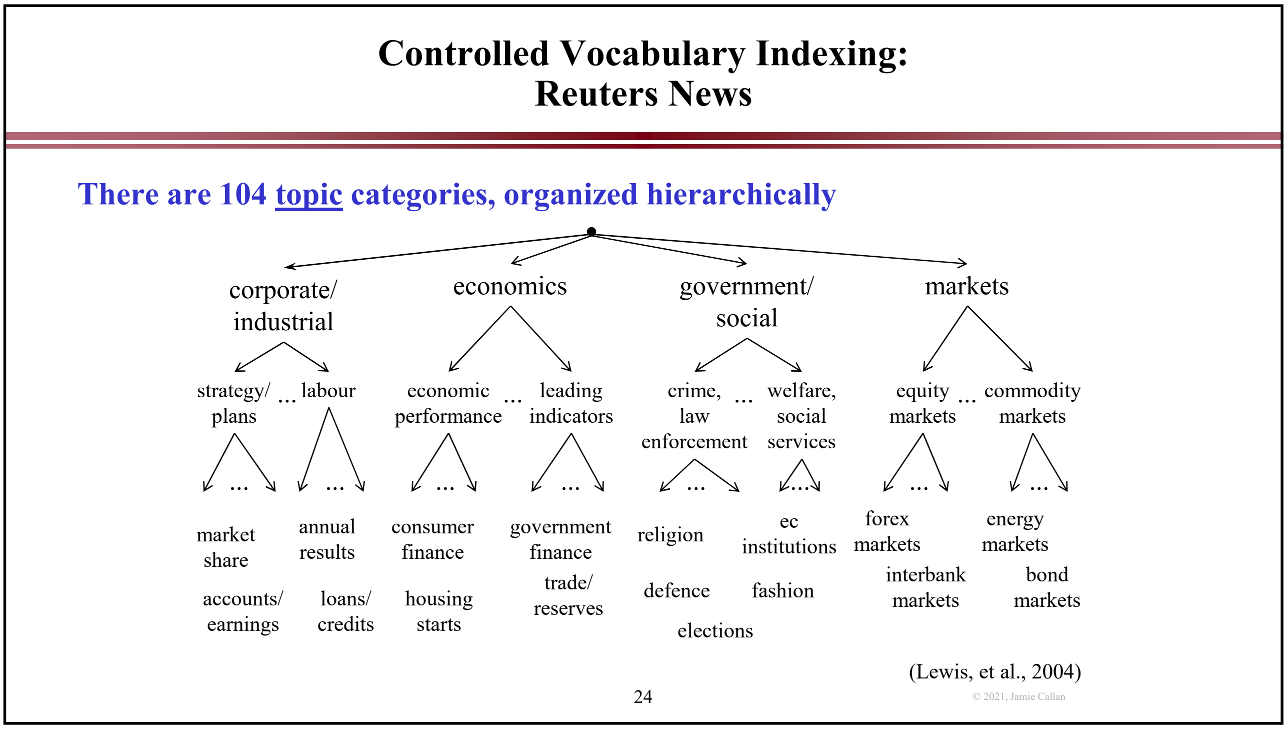
Controlled Vocabulary Indexing: Reuters News

Reuters news documents are assigned three types of labels

- Topic categories
- Industry categories
- Region categories

These are different ways of representing the document contents

A document may have one or more labels of each type



Controlled Vocabulary Indexing: Reuters News

There are 104 topic categories, organized hierarchically

corporate/ industrial

strategy/ plans … labour

market share

accounts/ earnings

annual results

loans/ credits

economics

economic performance … leading indicators

consumer finance

housing starts

government finance

trade/ reserves

government/ social

crime, law enforcement … welfare, social services

religion

defence

elections

ec institutions

fashion

markets

equity markets … commodity markets

forex markets

interbank markets

energy markets

bond markets

(Lewis, et al., 2004)

Controlled Vocabulary Indexing:
How Reuters Indexes the Document

FOCUS - Fed stays put, but rate hike seen inevitable

While the Federal Reserve's Federal Open Market Committee left interest rates unchanged at its meeting Tuesday, the odds of monetary tightening are likely to mount as 1996 comes …

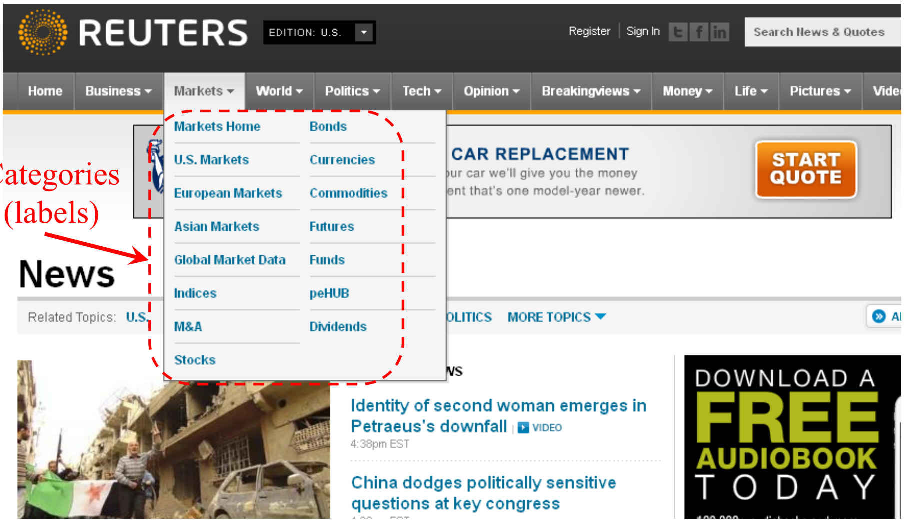
Fed policy-makers astutely resisted pressures earlier this summer to raise short-term rates, but economic data ….

“My guess is (Fed tightening) is going to be right back on tap after the August data start coming out,” said Salomon Brothers ….

Topic categories assigned to this document:

- Economics (ECAT), Monetary/Economic (E12)
- Markets (MCAT), Money Markets (M13), Interbank Markets (M131)

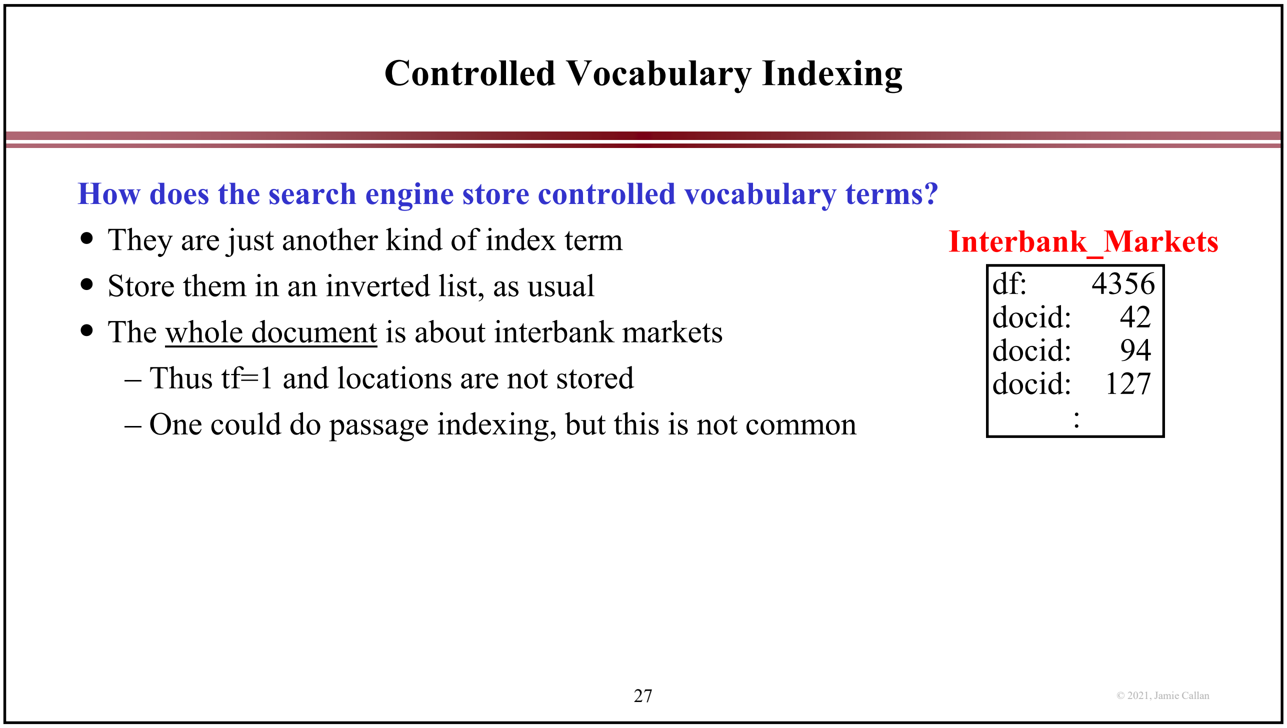
Controlled Vocabulary Indexing:
Reuters News



Identity of second woman emerges in Petraeus's downfall VIDEO
4:38pm EST

China dodges politically sensitive questions at key congress

Categories (labels)



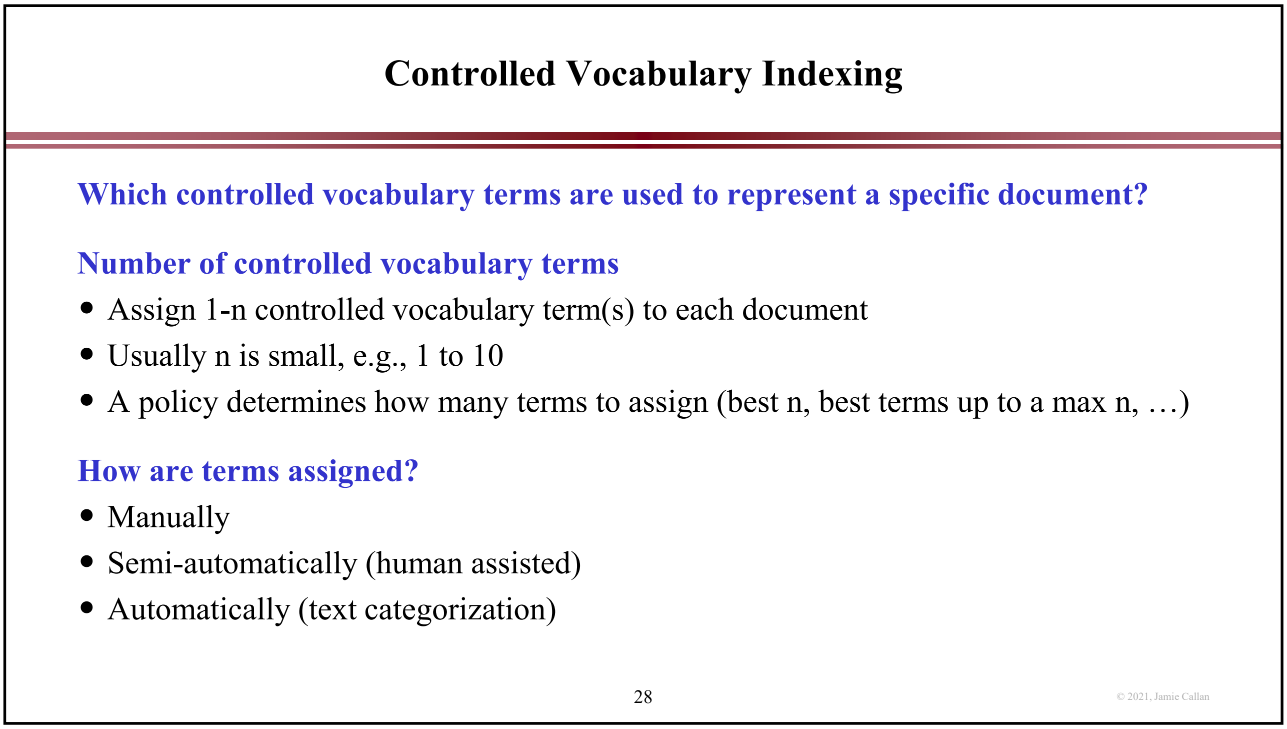
Controlled Vocabulary Indexing

How does the search engine store controlled vocabulary terms?

- They are just another kind of index term
- Store them in an inverted list, as usual
- The whole document is about interbank markets
  - Thus tf=1 and locations are not stored
  - One could do passage indexing, but this is not common

Interbank_Markets

df:	4356
docid:	42
docid:	94
docid:	127
:	



Controlled Vocabulary Indexing

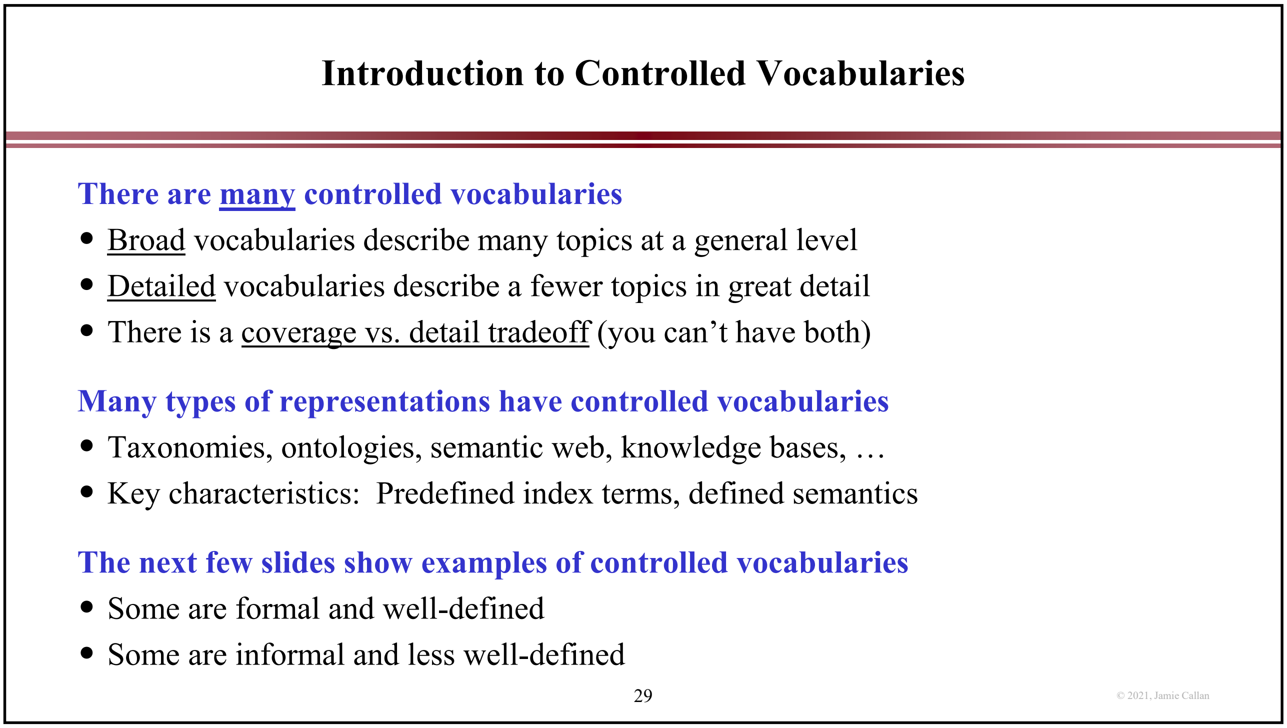
Which controlled vocabulary terms are used to represent a specific document?

Number of controlled vocabulary terms

- Assign 1-n controlled vocabulary term(s) to each document
- Usually n is small, e.g., 1 to 10
- A policy determines how many terms to assign (best n, best terms up to a max n, …)

How are terms assigned?

- Manually
- Semi-automatically (human assisted)
- Automatically (text categorization)



Introduction to Controlled Vocabularies

There are many controlled vocabularies

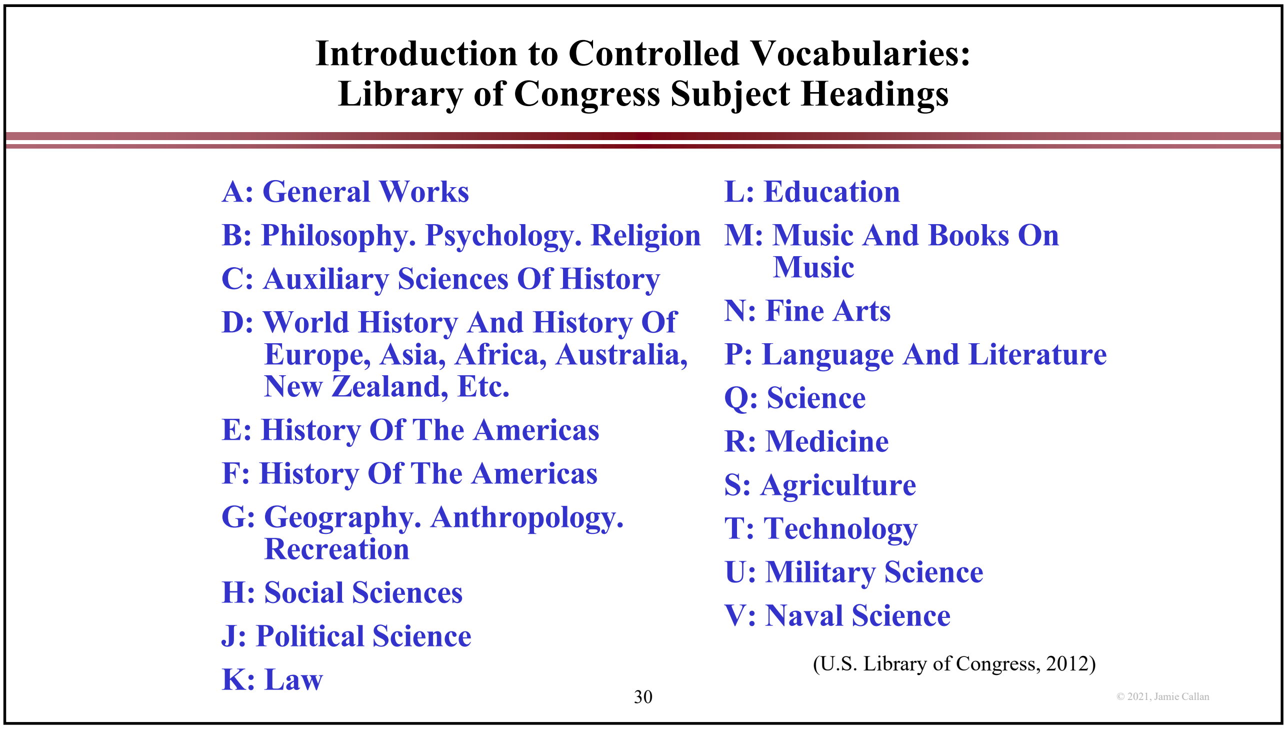
- Broad vocabularies describe many topics at a general level
- Detailed vocabularies describe a fewer topics in great detail
- There is a coverage vs. detail tradeoff (you can't have both)

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



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

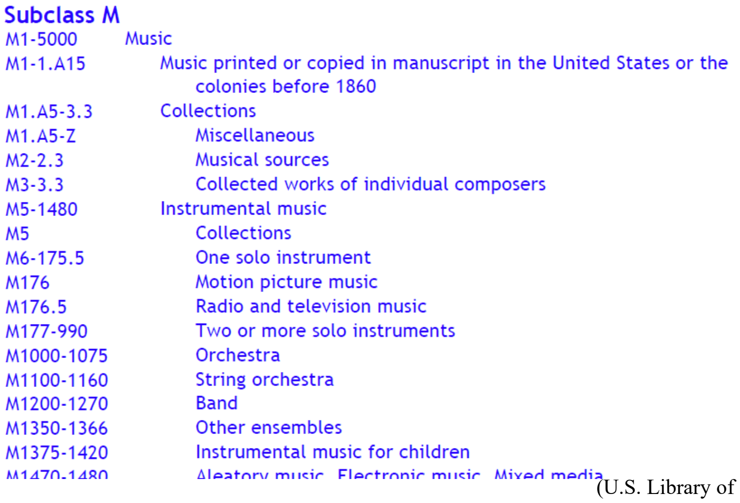
T: Technology

U: Military Science

V: Naval Science

(U.S. Library of Congress, 2012)

Introduction to Controlled Vocabularies: Library of Congress Subject Headings

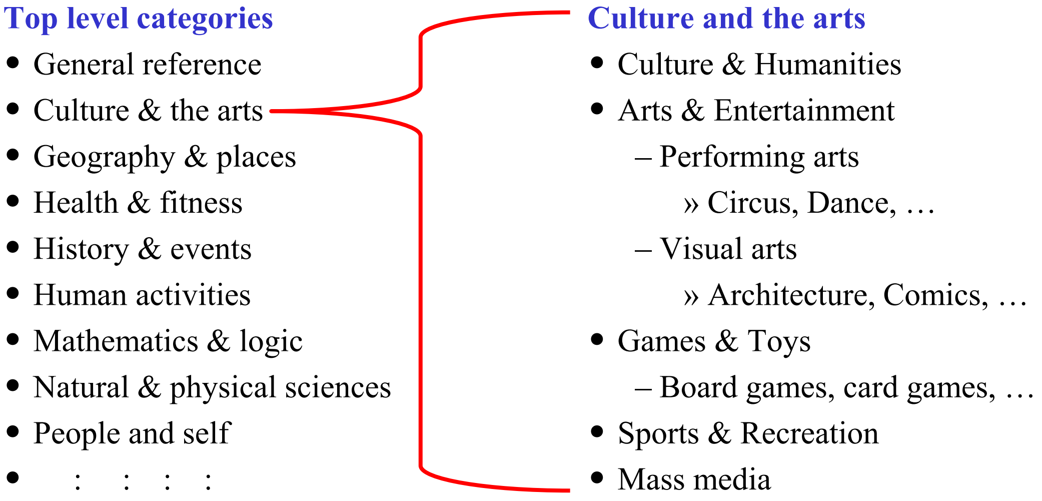


Subclass M

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)

Controlled Vocabularies: Wikipedia

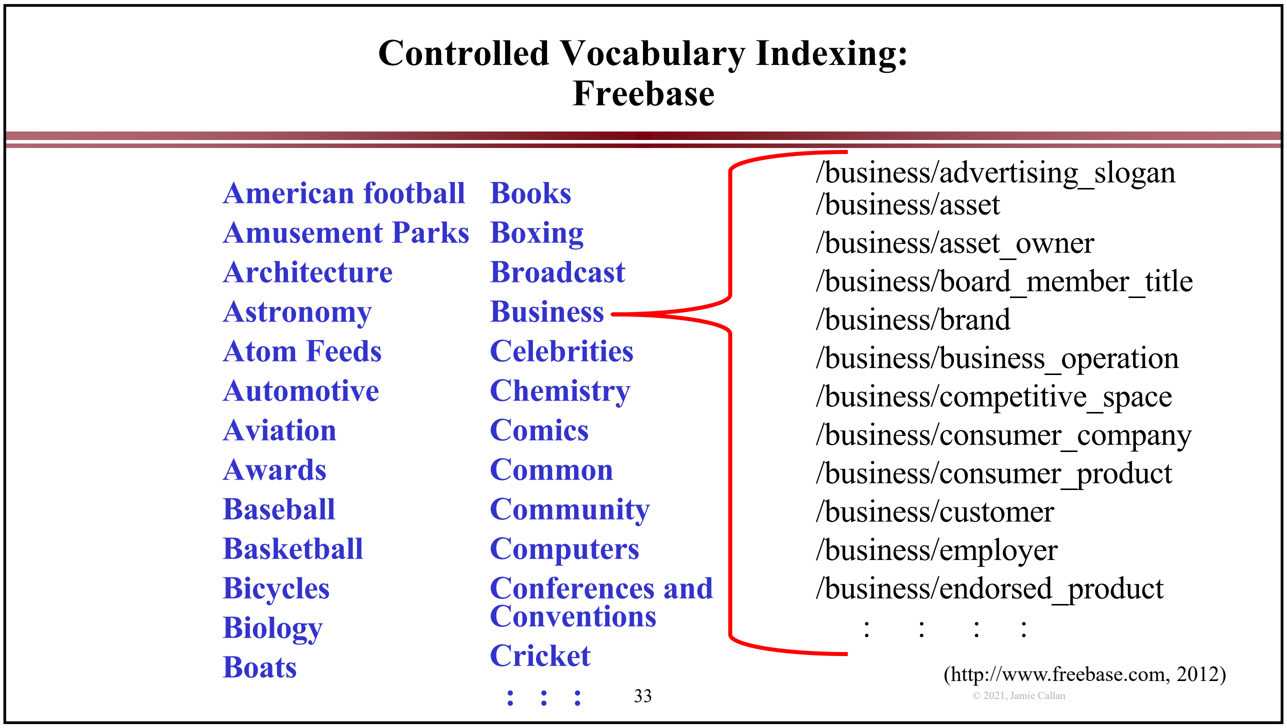


Top level categories

- General reference
- Culture & the arts
- Geography & places
- Health & fitness
- History & events
- Human activities
- Mathematics & logic
- Natural & physical sciences
- People and self
- : : : :

Culture and the arts

- Culture & Humanities
- Arts & Entertainment
  - – Performing arts
    - » Circus, Dance, …
  - – Visual arts
    - » Architecture, Comics, …
- Games & Toys
  - – Board games, card games, …
- Sports & Recreation
- Mass media



Controlled Vocabulary Indexing: Freebase

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

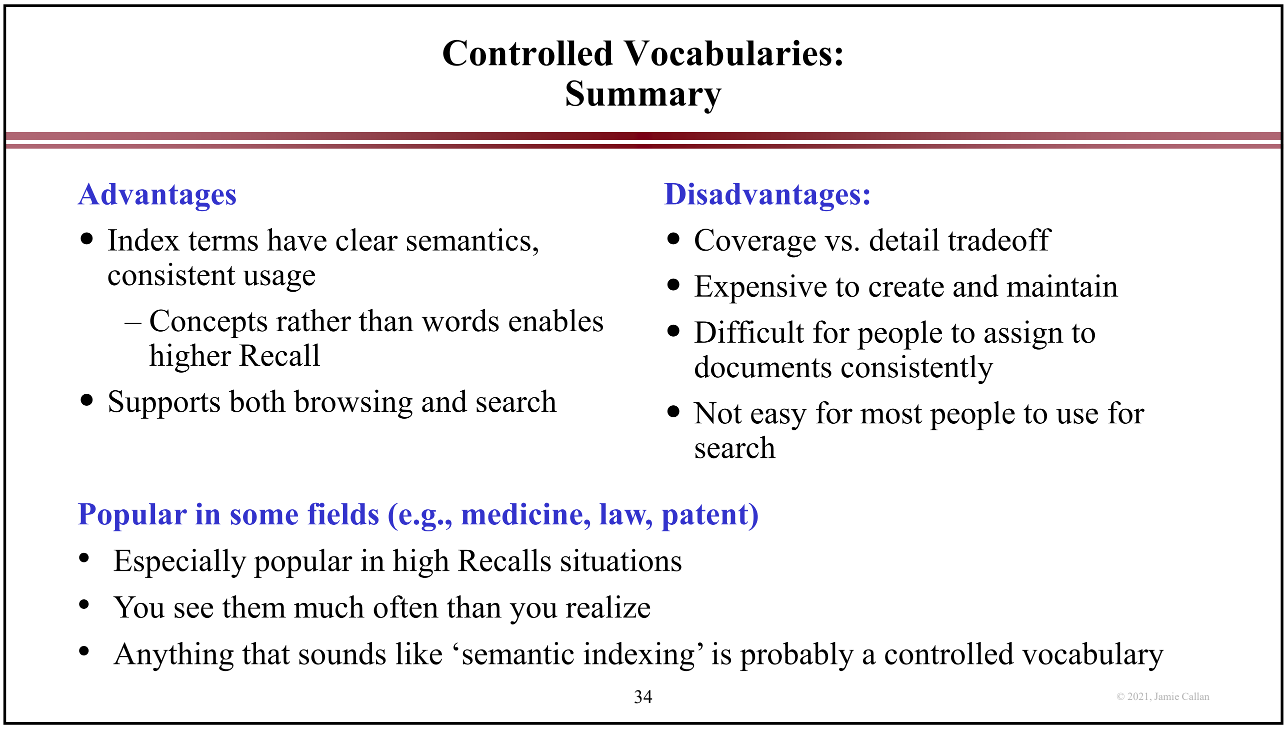
Books
Boxing
Broadcast
Business
Celebrities
Chemistry
Comics
Common
Community
Computers
Conferences and Conventions
Cricket
: : :

/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
: : : :

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

33

33



Controlled Vocabularies: Summary

Advantages

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

Disadvantages:

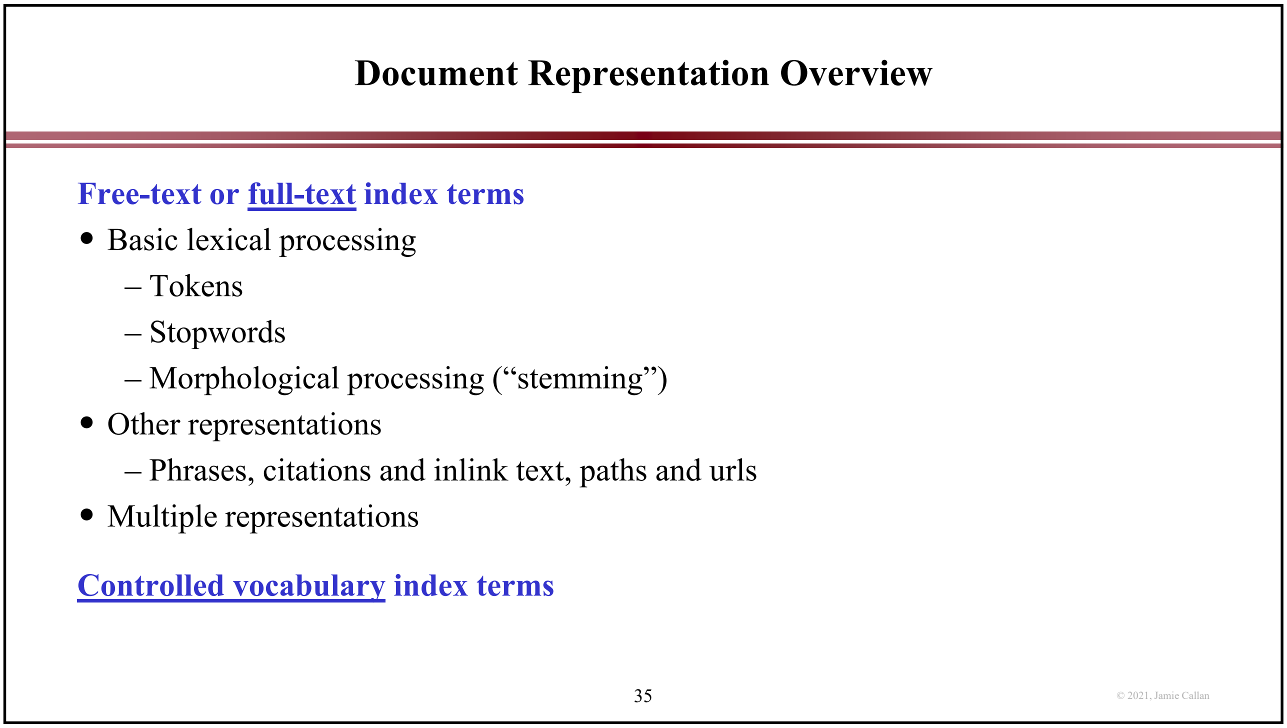
- Coverage vs. detail tradeoff
- Expensive to create and maintain
- Difficult for people to assign to documents consistently
- Not easy for most people to use for search

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

- Especially popular in high Recalls situations
- You see them much often than you realize
- Anything that sounds like ‘semantic indexing’ is probably a controlled vocabulary

34

34

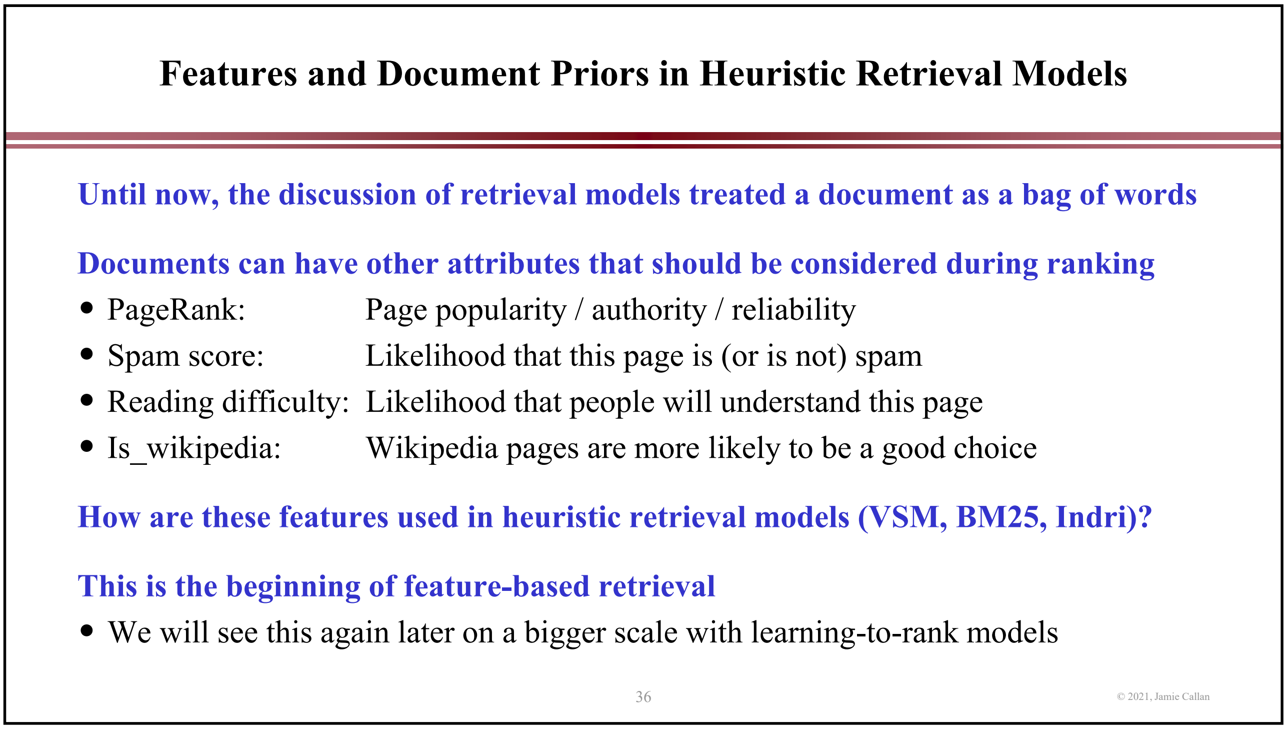


Document Representation Overview

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



Features and Document Priors in Heuristic Retrieval Models

Until now, the discussion of retrieval models treated a document as a bag of words

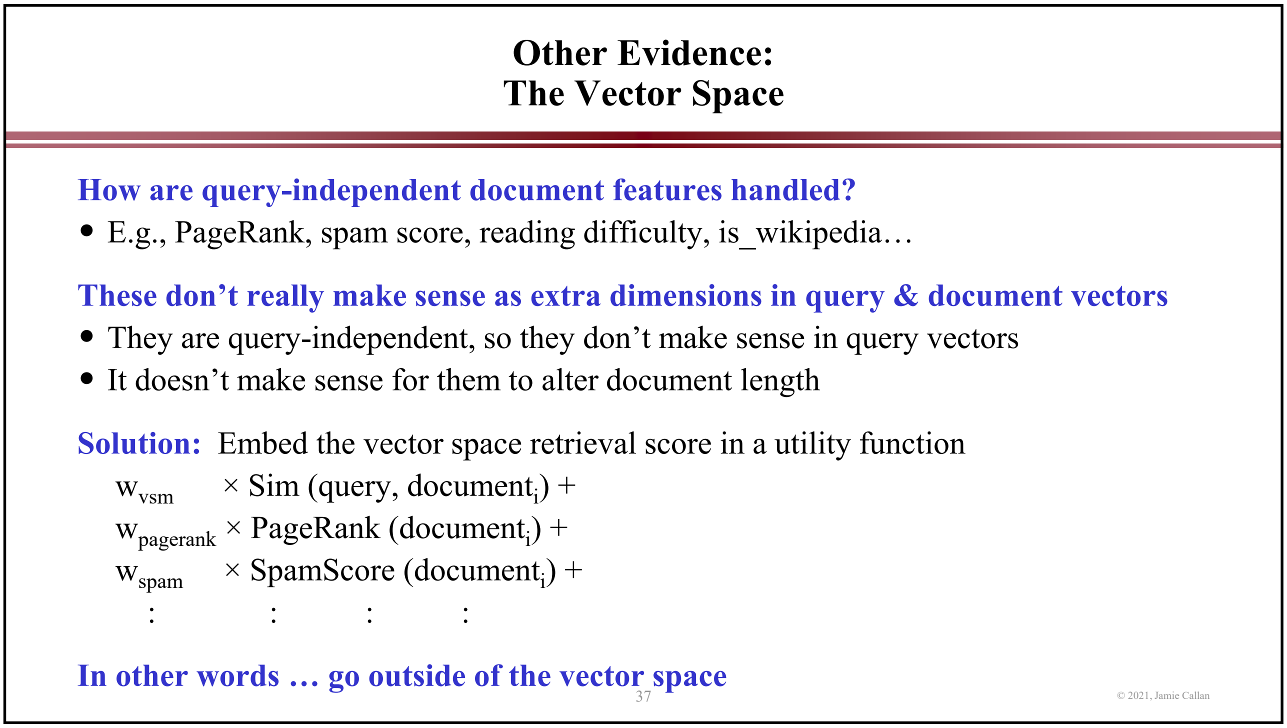
Documents can have other attributes that should be considered during ranking

• PageRank:	Page popularity / authority / reliability
• Spam score:	Likelihood that this page is (or is not) spam
• Reading difficulty:	Likelihood that people will understand this page
• Is_wikipedia:	Wikipedia pages are more likely to be a good choice

How are these features used in heuristic retrieval models (VSM, BM25, Indri)?

This is the beginning of feature-based retrieval

- We will see this again later on a bigger scale with learning-to-rank models



Other Evidence: The Vector Space

How are query-independent document features handled?

- E.g., PageRank, spam score, reading difficulty, is_wikipedia…

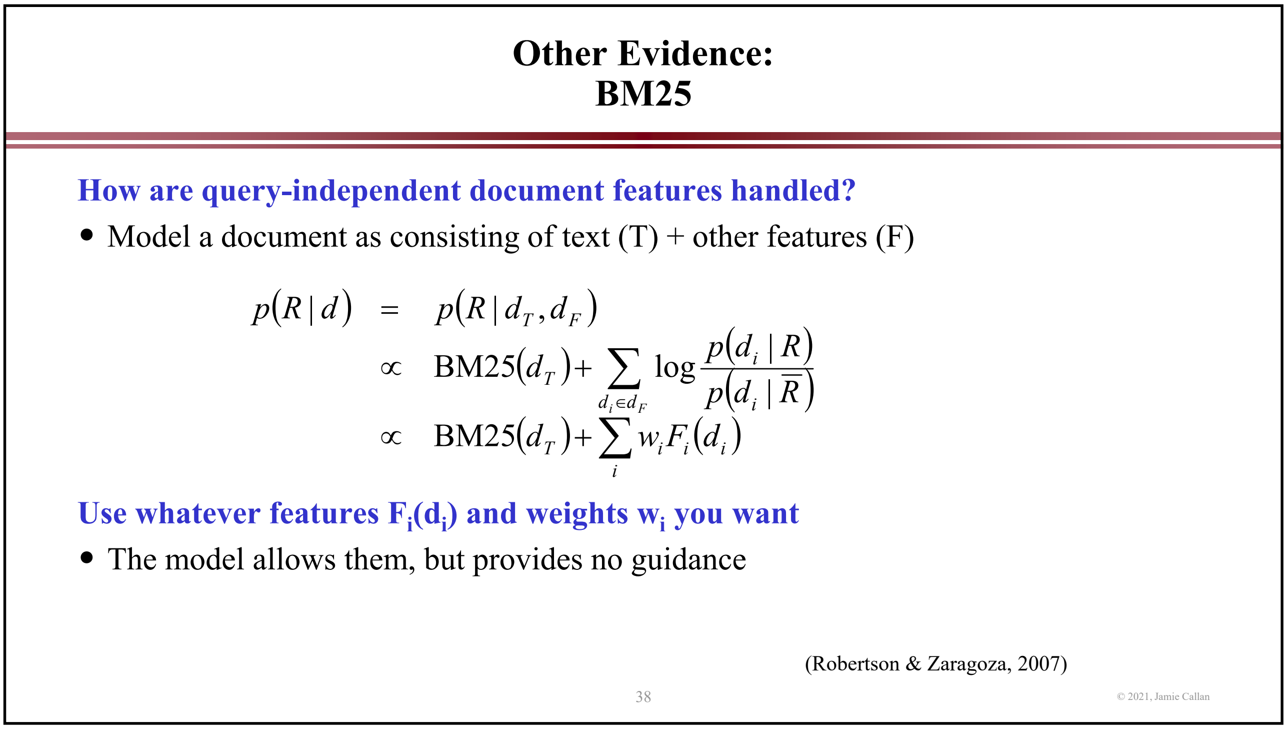
These don't really make sense as extra dimensions in query & document vectors

- They are query-independent, so they don't make sense in query vectors
- It doesn't make sense for them to alter document length

Solution: Embed the vector space retrieval score in a utility function

w_{vsm} × Sim (query, document$_i$) +
$w_{pagerank}$ × PageRank (document$_i$) +
w_{spam} × SpamScore (document$_i$) +
: : : :

In other words … go outside of the vector space



Other Evidence: BM25

How are query-independent document features handled?

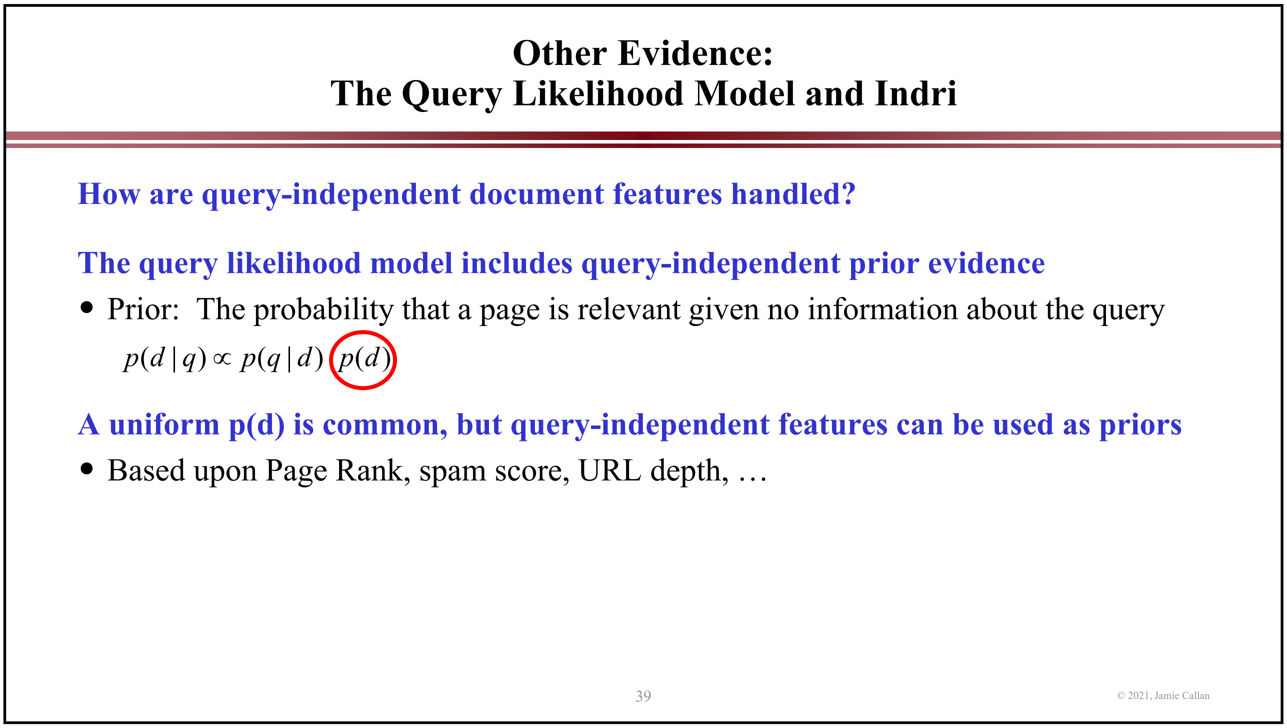
- Model a document as consisting of text (T) + other features (F)

$$\begin{aligned} p(R|d) &= p(R|d_T, d_F) \\ &\propto \text{BM25}(d_T) + \sum_{d_i \in d_F} \log \frac{p(d_i|R)}{p(d_i|\bar{R})} \\ &\propto \text{BM25}(d_T) + \sum_i w_i F_i(d_i) \end{aligned}$$

Use whatever features $F_i(d_i)$ and weights w_i you want

- The model allows them, but provides no guidance

(Robertson & Zaragoza, 2007)



Other Evidence: The Query Likelihood Model and Indri

How are query-independent document features handled?

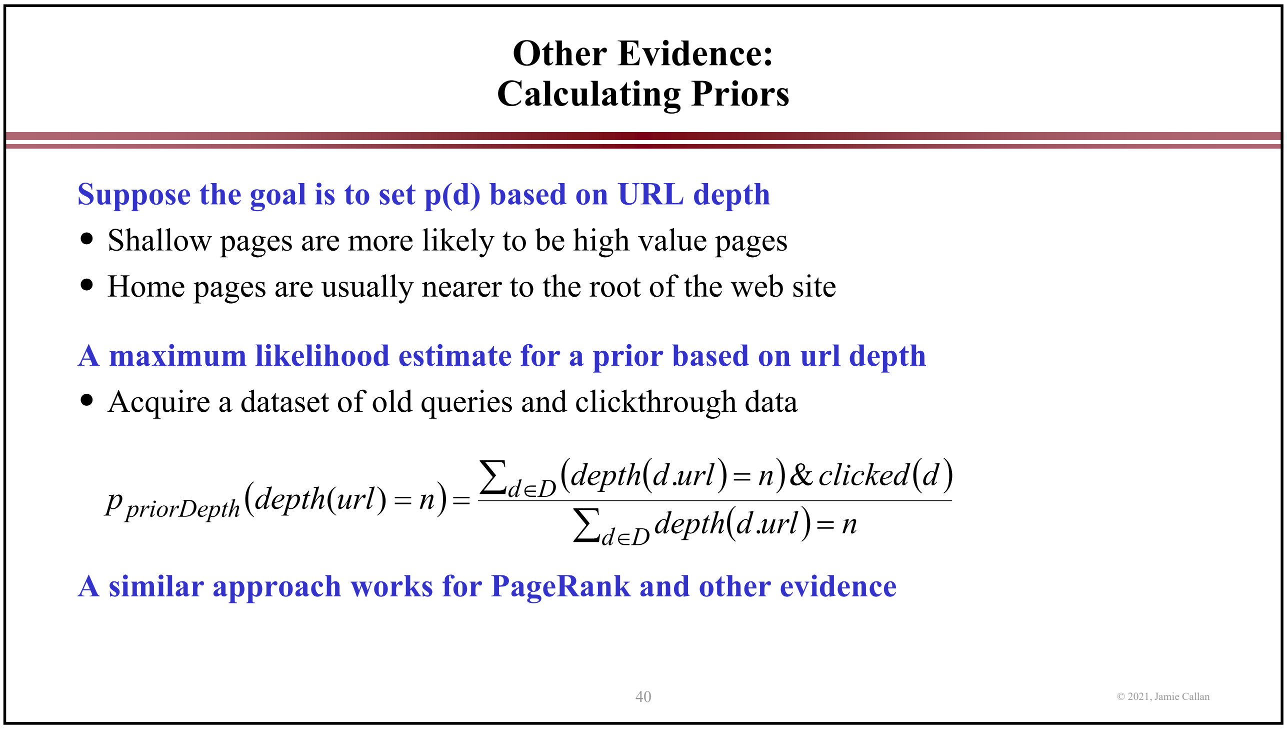
The query likelihood model includes query-independent prior evidence

- Prior: The probability that a page is relevant given no information about the query

$$p(d \mid q) \propto p(q \mid d)\, p(d)$$

A uniform p(d) is common, but query-independent features can be used as priors

- Based upon Page Rank, spam score, URL depth, …



Other Evidence: Calculating Priors

Suppose the goal is to set p(d) based on URL depth

- Shallow pages are more likely to be high value pages
- Home pages are usually nearer to the root of the web site

A maximum likelihood estimate for a prior based on url depth

- Acquire a dataset of old queries and clickthrough data

$$p_{priorDepth}\left(depth(url)=n\right)=\frac{\sum_{d\in D}\left(depth\left(d.url\right)=n\right)\&\,clicked\left(d\right)}{\sum_{d\in D} depth\left(d.url\right)=n}$$

A similar approach works for PageRank and other evidence

Other Evidence: Different Approaches to Priors

Query Likelihood and KL Divergence are similar

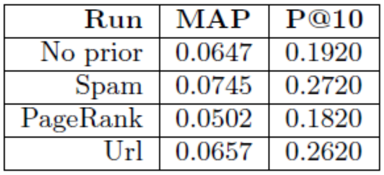
…until priors are introduced

- **Query likelihood** $p(d|q) \propto \log p(d) + \sum_{q_i \in Q} \log p(q_i|d)$
  - – Expressed in Indri as #and (#prior (url) a b c)
- **KL Divergence** $p(d|q) \propto \log p(d) + \frac{1}{|Q|} \sum_{q_i \in Q} \log p(q_i|d)$
  - – Expressed in Indri as #and (#prior (url) #and (a b c))
- **On long queries, priors have a much larger effect on the KL divergence model than on the query likelihood model**

Other Evidence: Are Document Priors Important?

Document priors are a convenient way of introducing query-independent evidence

- E.g., spam score, PageRank, url depth, …

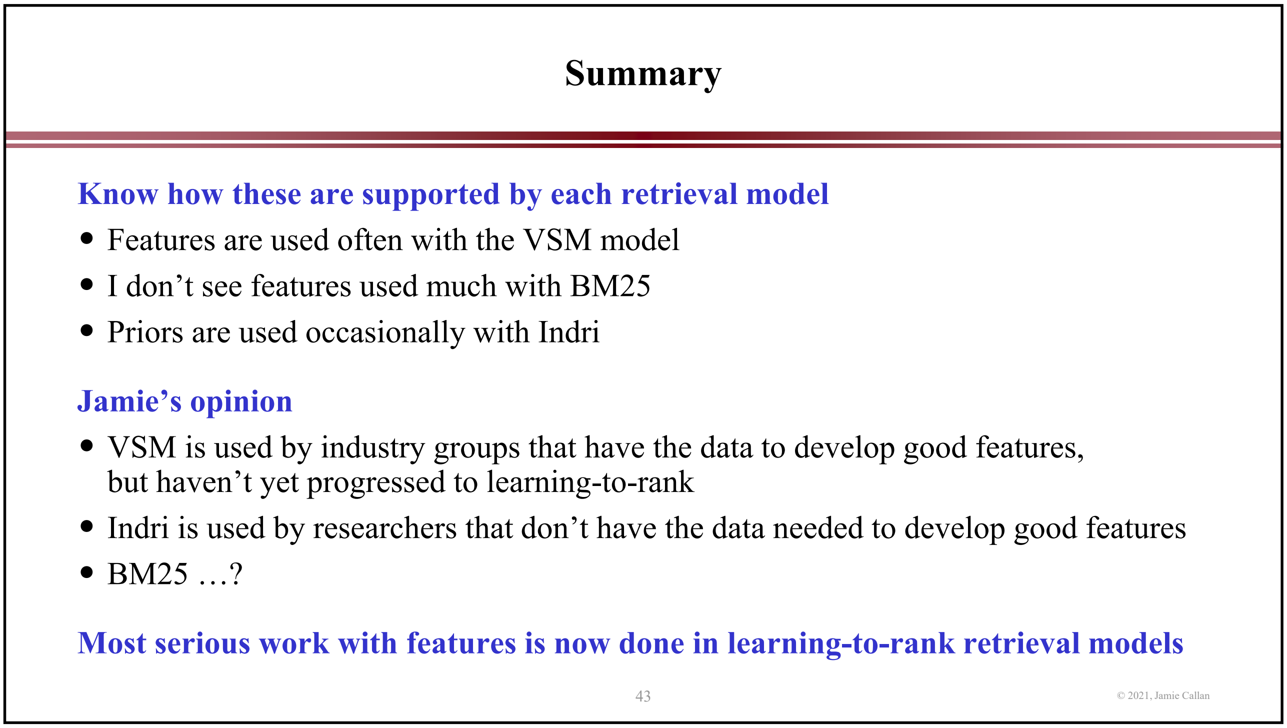


Run	MAP	P@10
No prior	0.0647	0.1920
Spam	0.0745	0.2720
PageRank	0.0502	0.1820
Url	0.0657	0.2620

Perhaps better theory than in the vector space and Okapi

- But … similar effects can be achieved with those models

(Nguyen and Callan, 2011)



Summary

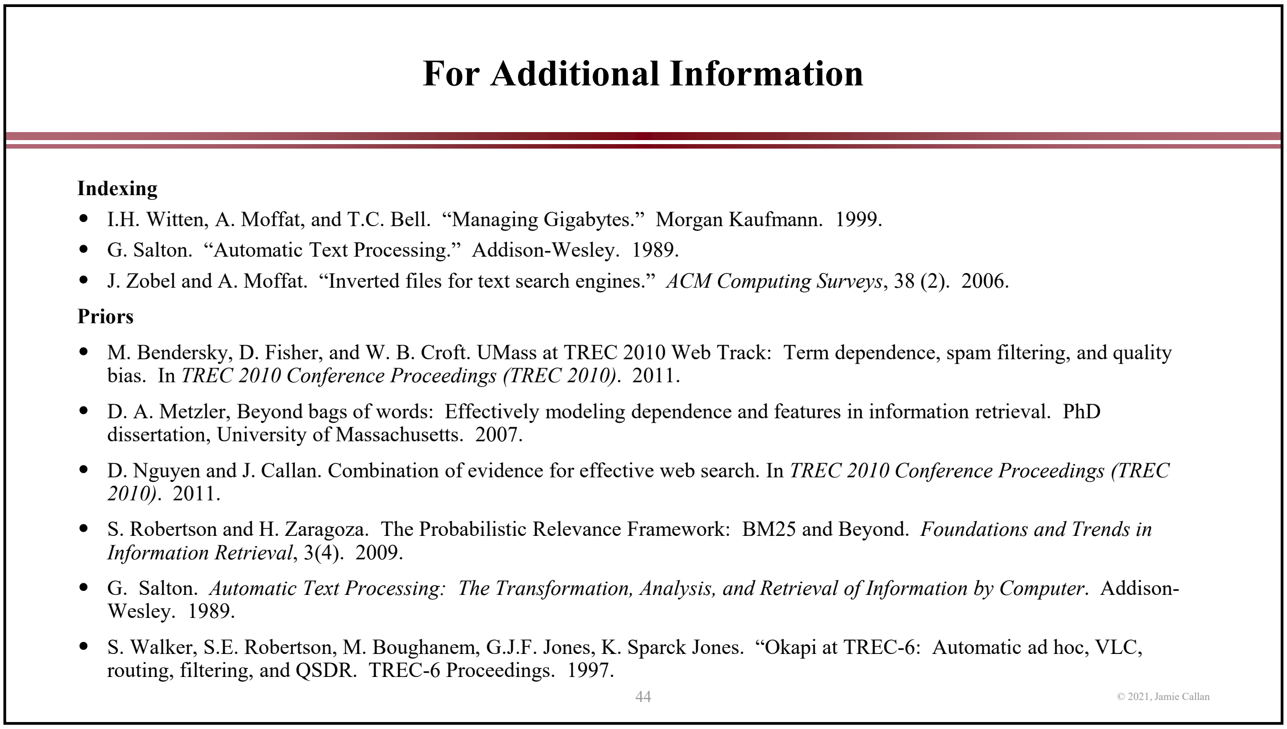
Know how these are supported by each retrieval model

- Features are used often with the VSM model
- I don’t see features used much with BM25
- Priors are used occasionally with Indri

Jamie’s opinion

- VSM is used by industry groups that have the data to develop good features, but haven’t yet progressed to learning-to-rank
- Indri is used by researchers that don’t have the data needed to develop good features
- BM25 …?

Most serious work with features is now done in learning-to-rank retrieval models



For Additional Information

Indexing

- I.H. Witten, A. Moffat, and T.C. Bell. “Managing Gigabytes.” Morgan Kaufmann. 1999.
- G. Salton. “Automatic Text Processing.” Addison-Wesley. 1989.
- J. Zobel and A. Moffat. “Inverted files for text search engines.” *ACM Computing Surveys*, 38 (2). 2006.

Priors

- M. Bendersky, D. Fisher, and W. B. Croft. UMass at TREC 2010 Web Track: Term dependence, spam filtering, and quality bias. In *TREC 2010 Conference Proceedings (TREC 2010)*. 2011.
- D. A. Metzler, Beyond bags of words: Effectively modeling dependence and features in information retrieval. PhD dissertation, University of Massachusetts. 2007.
- D. Nguyen and J. Callan. Combination of evidence for effective web search. In *TREC 2010 Conference Proceedings (TREC 2010)*. 2011.
- S. Robertson and H. Zaragoza. The Probabilistic Relevance Framework: BM25 and Beyond. *Foundations and Trends in Information Retrieval*, 3(4). 2009.
- G. Salton. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley. 1989.
- S. Walker, S.E. Robertson, M. Boughanem, G.J.F. Jones, K. Sparck Jones. “Okapi at TREC-6: Automatic ad hoc, VLC, routing, filtering, and QSDR. TREC-6 Proceedings. 1997.