Lexical Knowledge Bases

Natural Language Processing

Lexical Knowledge Bases

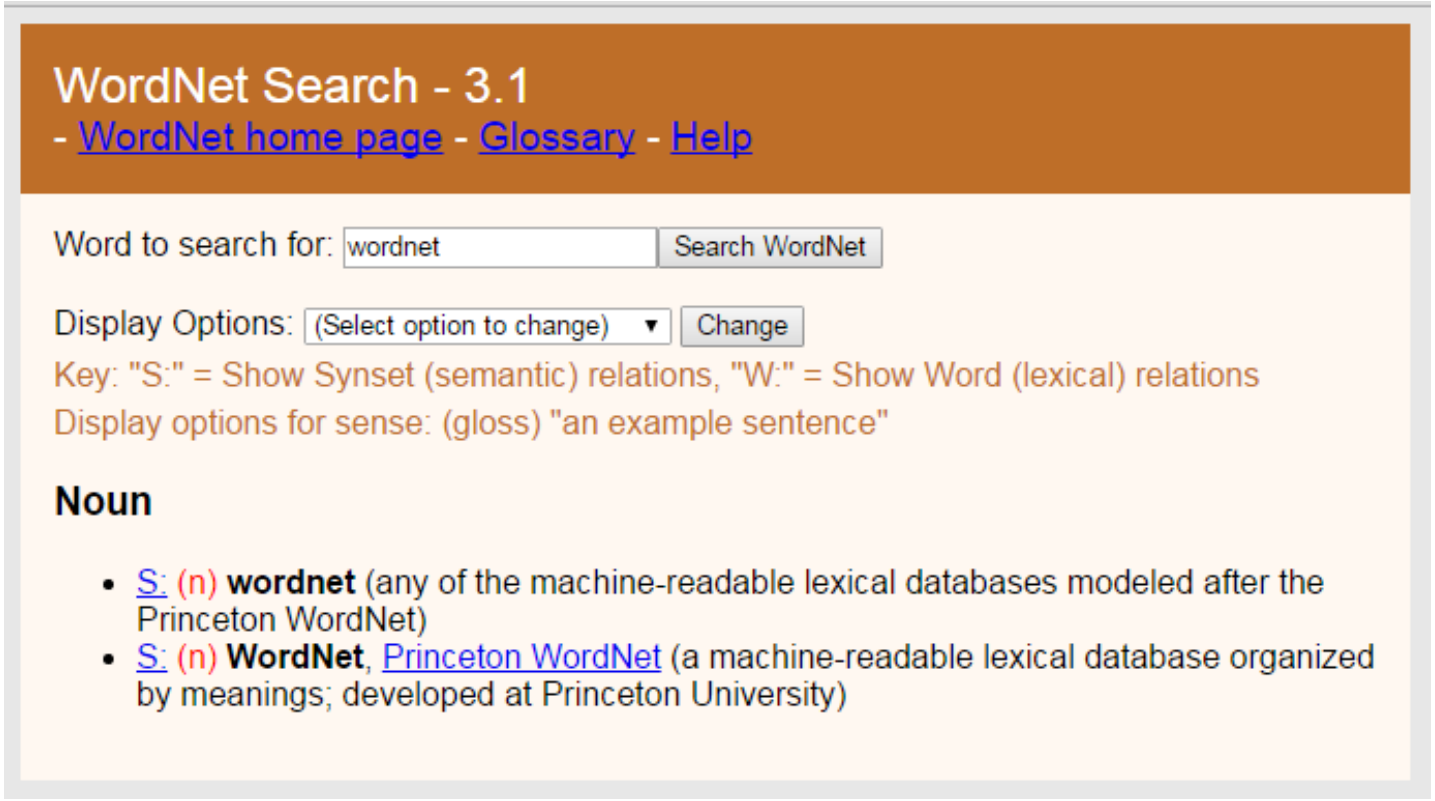
- A *lexicon* goes beyond a “dictionary” in that it is machine-readable and carries the information needed to perform major NLP functions, such as:
  - Parts of speech
  - Inflections
  - Transitive vs. intransitive verbs

Lexical Knowledge Bases

- A *lexical knowledge base ("lexical KB")* goes even further, breaking words into senses, and linking senses to senses via relations such as:
  - Synonym/antonym
  - Hyponym/hypernym
  - Holonom/meronym

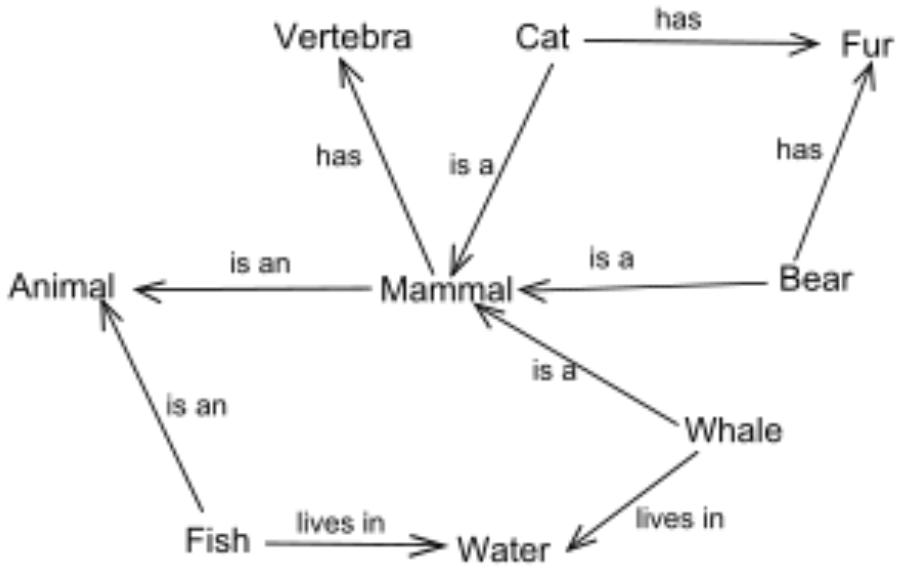
WordNet

Our “go-to” lexical knowledge base (KB)



WordNet

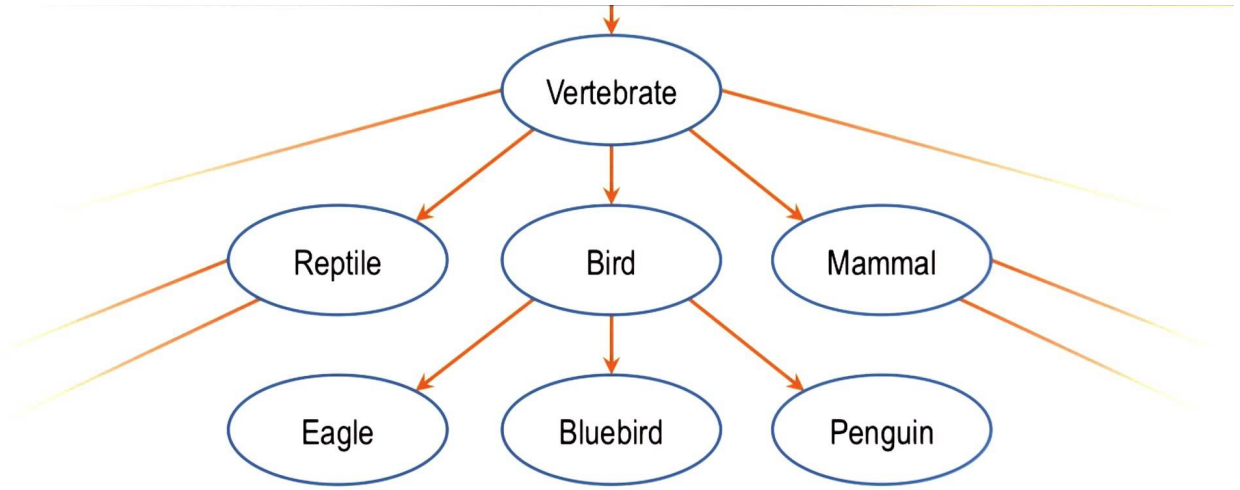
WordNet is an example (in limited form) of a *semantic network,* which means it can tell you relations similar to these:



But WordNet is limited largely to just the “is a” and “has part / is part” relations.

Concept Hierarchies

It's possible to put a semantic network into a strict hierarchy like this, using just the "is-a" relation.



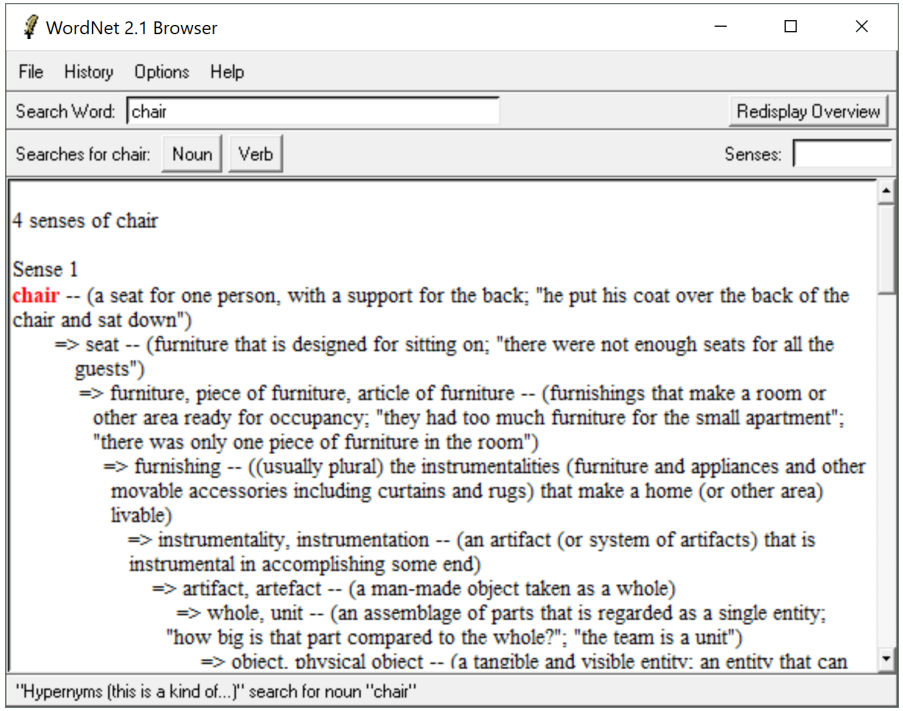
Improving Feature Extraction

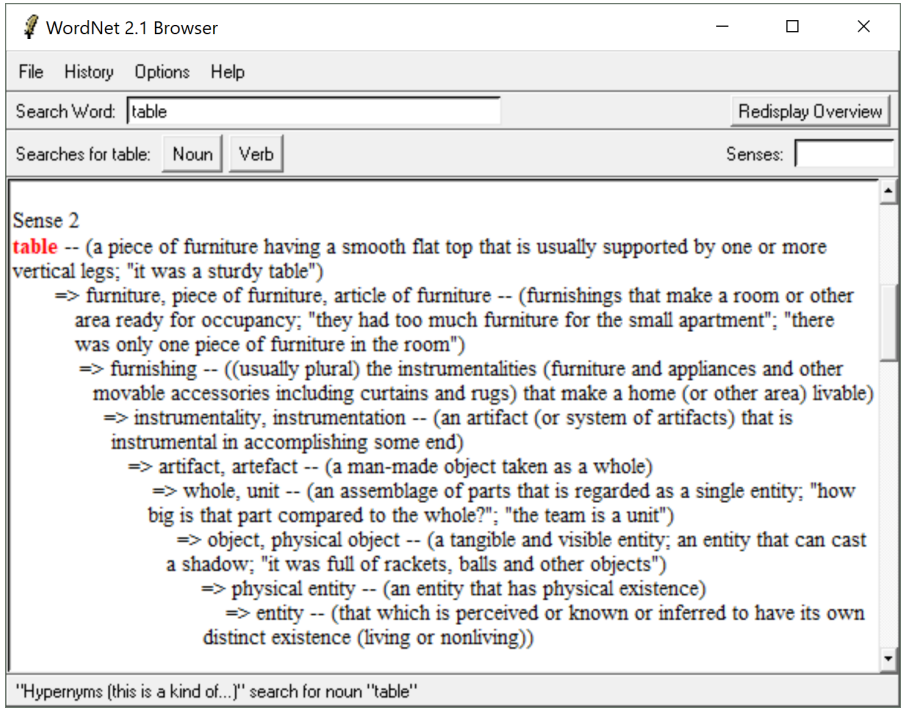
- We can use a lexical KB's concept hierarchy; we can improve our low-level feature extraction.
- Default ontological distance (number of nodes "travelled" through the hierarchy to get from node A to node B):
  - Compute default ontological distance of any two headwords in WordNet.
  - Without knowing which word sense, just check all senses and find the closest/longest distance via hyponym trees.
  - This gives us a rough sense of distance in "lexical space" between terms.

Can you guess the ontological distance in WordNet of “chair” and “table”?

Example of Lexical Distance

Can you guess the ontological distance in WordNet of "chair" and "table"?
Comparing noun sense 1 of "chair" with noun sense 2 of "table"

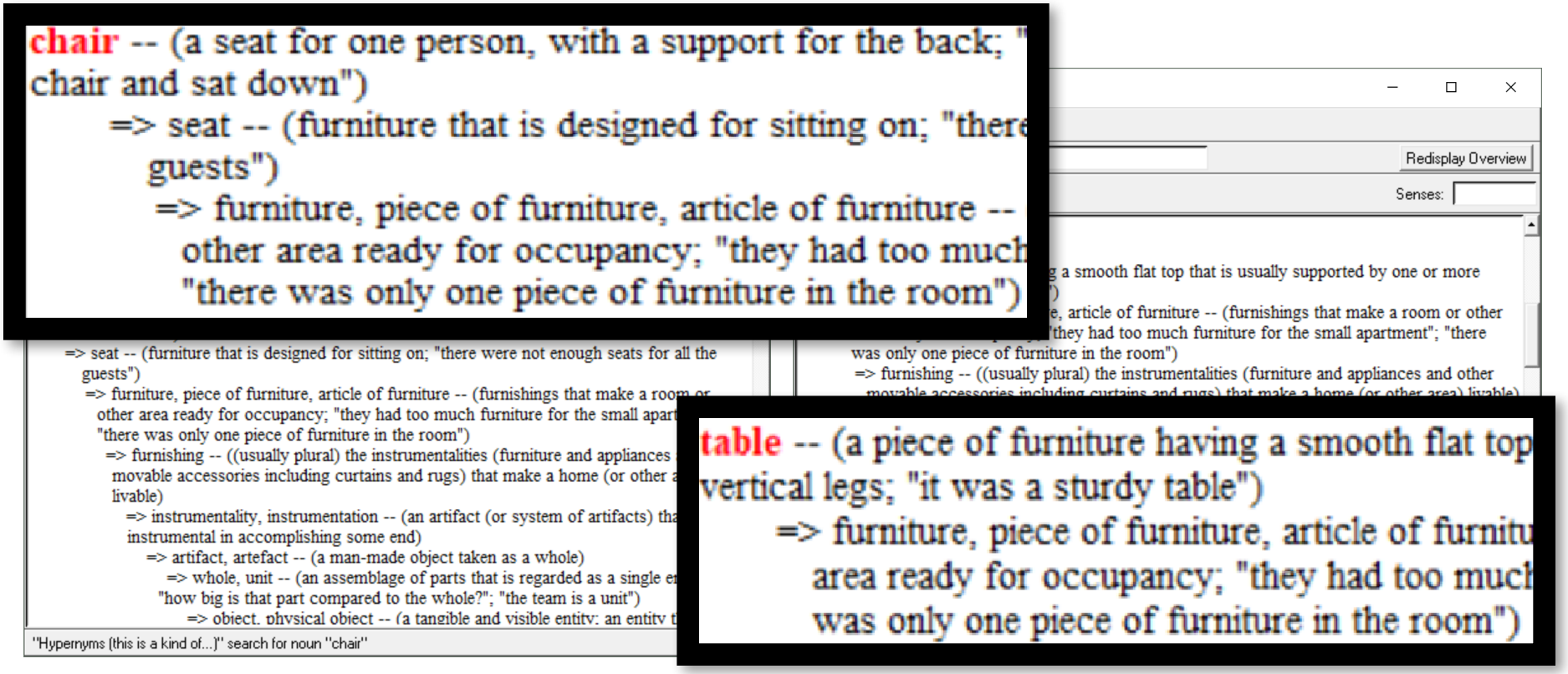




Distance = 3

Example of Lexical Distance

Can you guess the ontological distance in WordNet of “chair” and “table”?
Comparing noun sense 1 of “chair” with noun sense 2 of “table”



Distance = 3

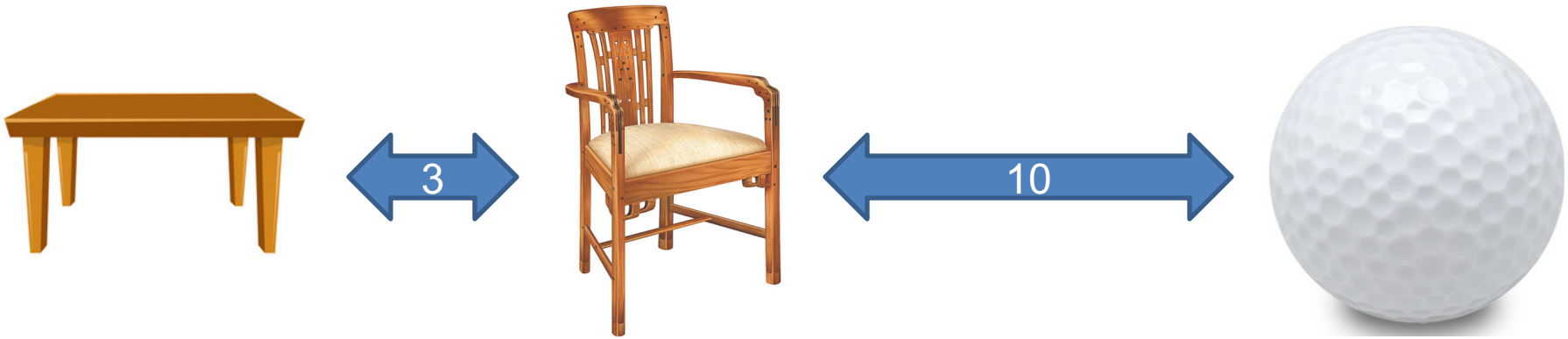
DataScience@SMU

Lexical Knowledge Bases

Natural Language Processing

Comparative Term Distance

- So, in line with our intuitions, “chair” is closer to “table” than it is to “golf ball.”



If you were to check this distance between the most frequent terms across two different documents, you could add up the cumulative distance, and *voila!* You’d have an automated measure of the general *document distance*.

Limits of This Approach

- We don't know which sense of "chair" was intended in each occurrence.
- We would have to disambiguate word senses, which we'll talk about later.

Furniture to sit on

Leader of an academic department

Instrument of capital punishment

Polysemy vs. Monosemy

- Monosemy = having only one sense
- Polysemy = having more than one sense

Unfortunately, many words in English are polysemous.

Polysemy in WordNet 3.0

POS*	Monosemous Words and Senses	Polysemous Words	Polysemous Senses
Noun	101863	15935	44449
Verb	6277	5252	18770
Adjective	16503	4976	14399
Adverb	3748	733	1832
Totals	128391	26896	79450

**part of speech*

Polysemy in WordNet 3.0

POS*	Average Polysemy Including Monosemous Words	Average Polysemy Excluding Monosemous Words
Noun	1.24	2.79
Verb	2.17	3.57
Adjective	1.40	2.71
Adverb	1.25	2.50

**part of speech*

DataScience@SMU

Building or Extending a Lexical Knowledge Base

Natural Language Processing

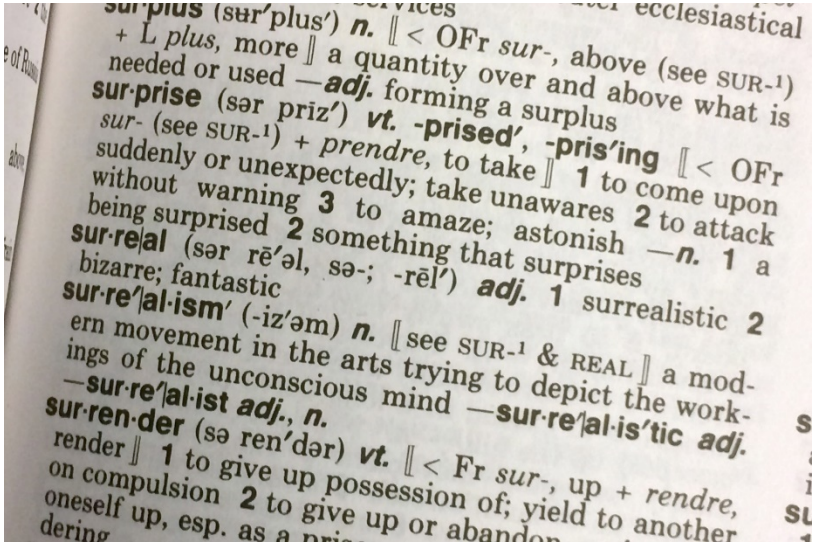
Resources for Creating or Extending Lexical Knowledge Bases

WordNet's a good start, but it's not enough.

How do we expand our lexical KB?

Three types of resources:

- Dictionaries
- Encyclopedias
- Taxonomies



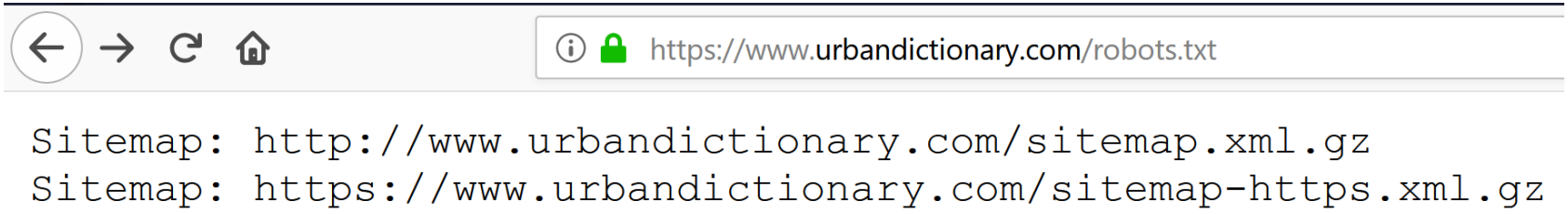
Resources for Creating or Extending Lexical Knowledge Bases

- Main vocabulary resources
  - Baseline: Princeton WordNet
  - Supplementary: Getty Vocabularies, Wiktionary, Urban Dictionary

Can We Utilize Urban Dictionary?

First we want to check their “robots.txt” file to see if it permits crawling the site.

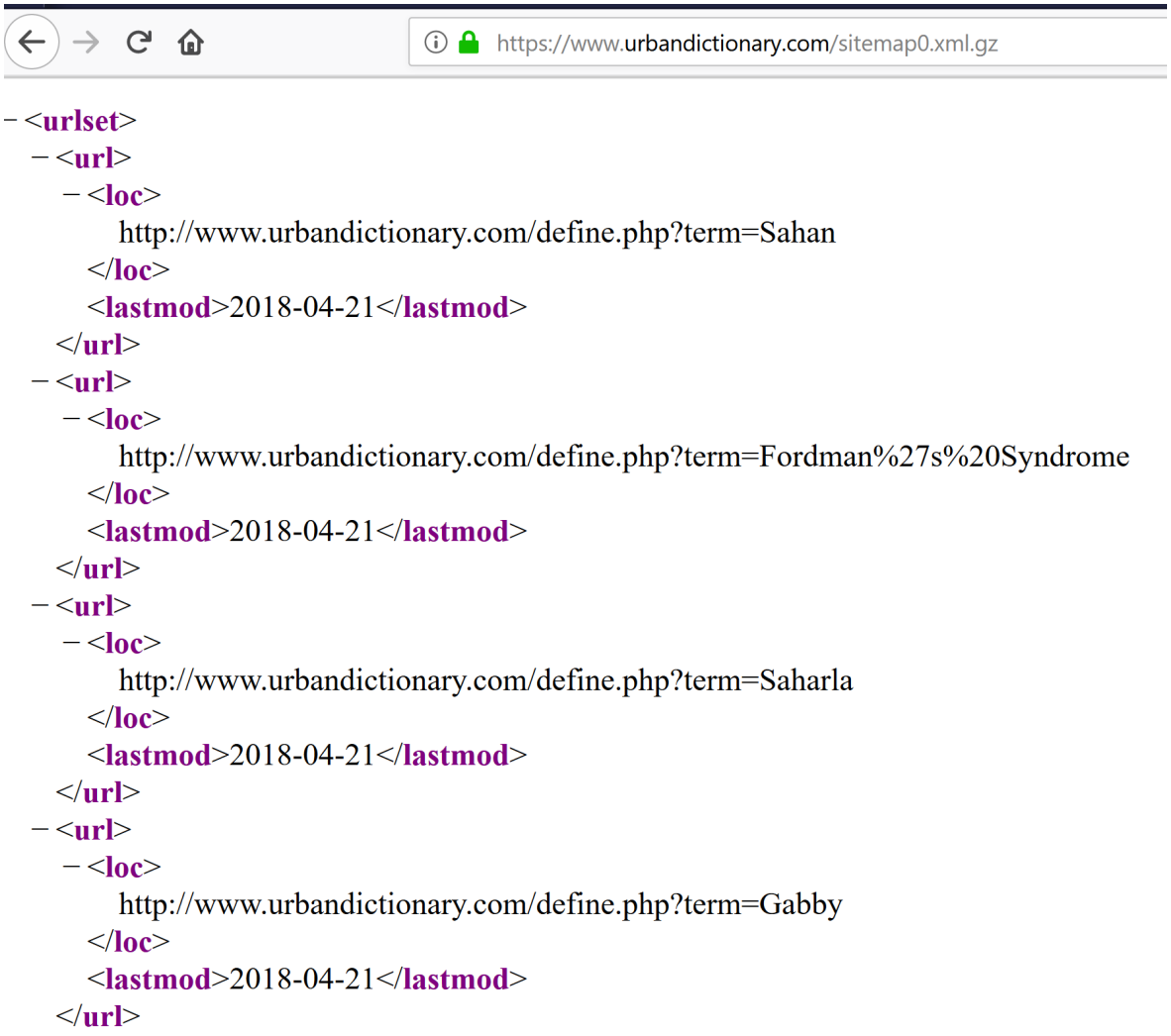
Fortunately, it does!



It sends us to an xml sitemap for further crawling.

Can We Utilize Urban Dictionary?

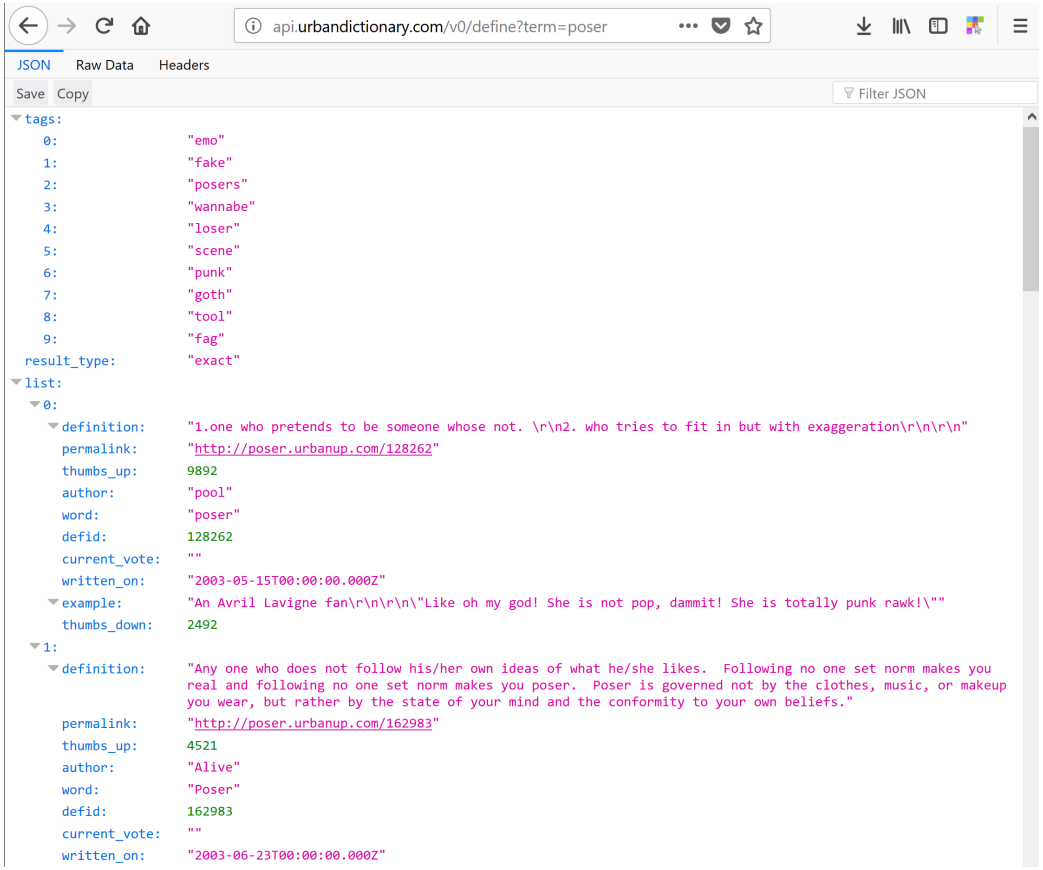
Looks like it's in no particular order, but it gives us a link to every single word in the urban dictionary.



How Else Can We Access These?

You must be prepared to tackle a variety of access methods.

- Getty is available as JSON, JSONLD, or RDF.
- Wiktionary can be downloaded as XML.
- Urban Dictionary has an undocumented API.



How Do We Access These?

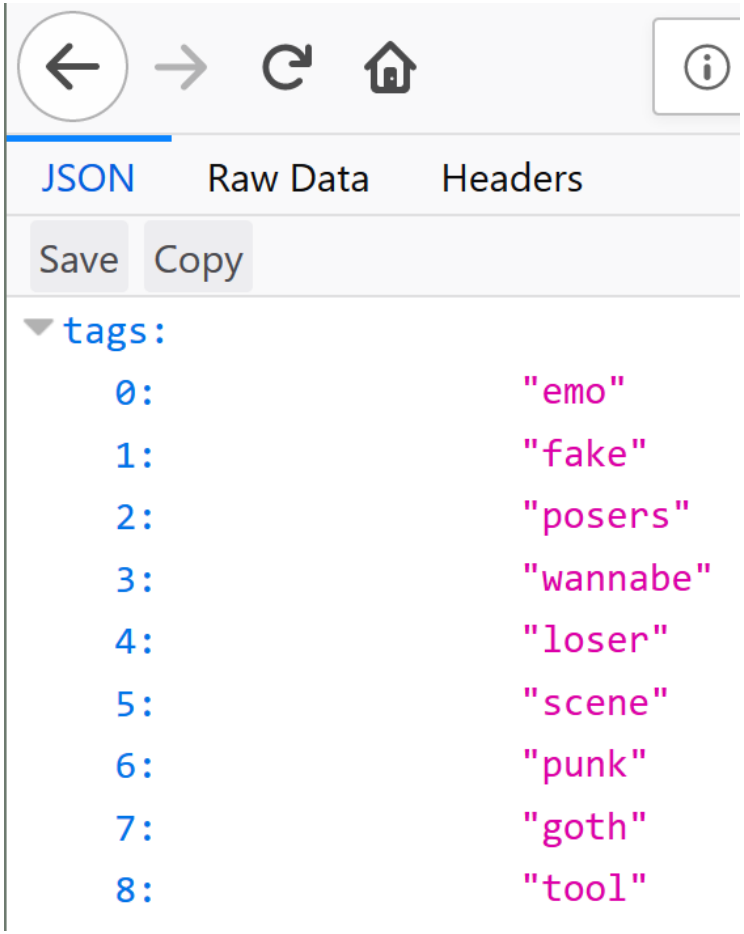
Here is a call to look up the word “poser”:

- Note that this will give you tags, and then you can look up all headwords for any given tag, for example:

https://www.urbandictionary.com/tags.php?tag=emo

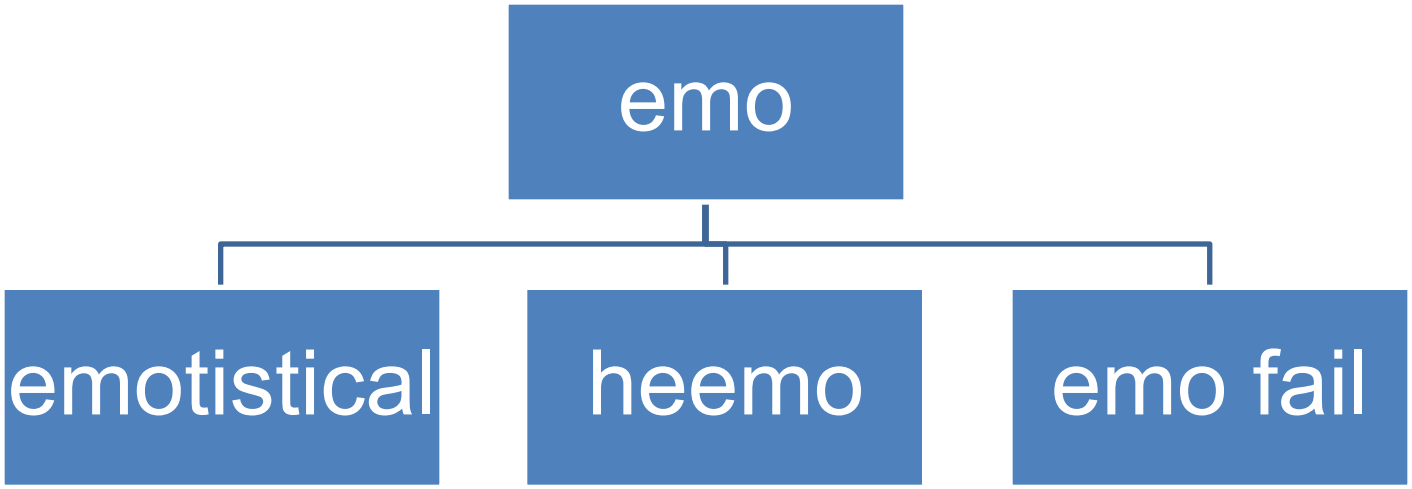
You could then use a parser like BeautifulSoup to find all the words sharing the “emo” tag where

```
class = “word”
```



How Could We Use This?

In this manner, you could build up families of slang terms with a common “parent” tag:

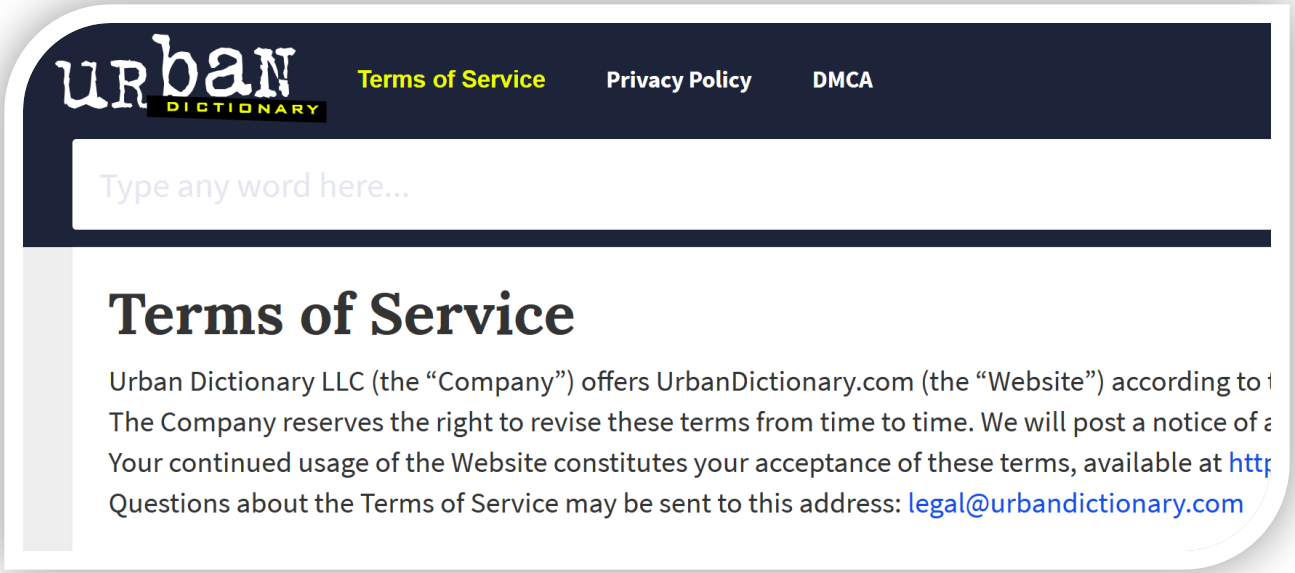


And *that’s* starting to look like WordNet!

Word of Warning

I’m not a lawyer!

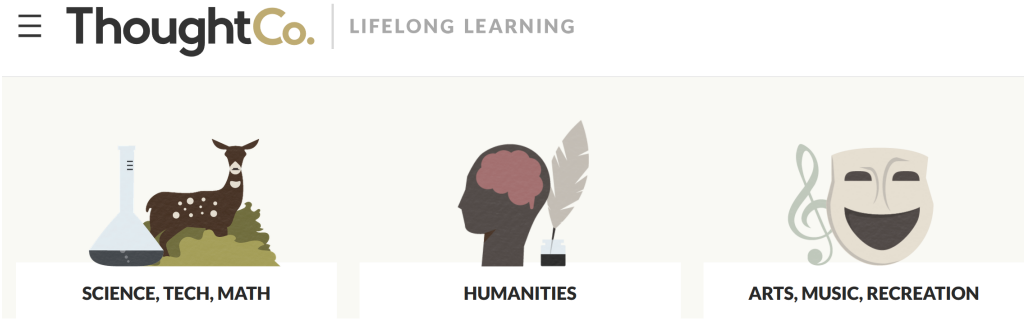
Be sure to consult a site’s ToS (“Terms of Service”) to see how you are allowed to use their site data—especially if you intend to commercially exploit it.



Resources for Creating or Extending Lexical Knowledge Bases

Encyclopedic resources

- Baseline: Wikipedia
- Supplementary: IMDB, DotDash (formerly About.com), more!



Resources for Creating or Extending Lexical Knowledge Bases

Encyclopedic resources

Resources for Creating or Extending Lexical Knowledge Bases

Taxonomical organizations of web pages

- Baseline: Curie (formerly DMOZ)
- Supplementary: Sitemaps (or RSS taxonomies) for sites like CNN Money, LA Times, Vogue, SFGate, etc.

CNNMoney VOGUE

Los Angeles Times SFGATE

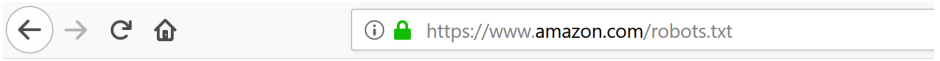
And Don’t Forget Amazon…

Books are published on everything, and Amazon has all the books…

…organized into a lot of categories.

And Don’t Forget Amazon…

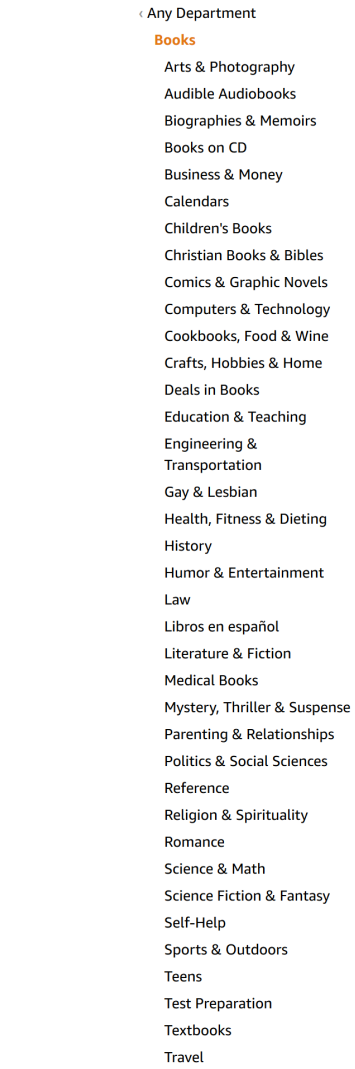
Robots.txt has a lot of Disallow statements, but “bestsellers” is not among them, so…




```
User-agent: *
Disallow: /exec/obidos/account-access-login
Disallow: /exec/obidos/change-style
Disallow: /exec/obidos/flex-sign-in
Disallow: /exec/obidos/handle-buy-box
Disallow: /exec/obidos/tg/cm/member/
Disallow: /gp/aw/help/id=sss
Disallow: /gp/cart
Disallow: /gp/flex
Disallow: /gp/product/e-mail-friend
Disallow: /gp/product/product-availability
Disallow: /gp/product/rate-this-item
Disallow: /gp/sign-in
Disallow: /gp/reader
Disallow: /gp/sitbv3/reader
Disallow: /gp/richpub/syltguides/create
Disallow: /gp/gfix
Disallow: /gp/associations/wizard.html
Disallow: /gp/dmusic/order
Disallow: /gp/legacy-handle-buy-box.html
Disallow: /gp/aws/ssop
Disallow: /gp/yourstore
Disallow: /gp/gift-central/organizer/add-wishlist
Disallow: /gp/vote
Disallow: /gp/voting/
Disallow: /gp/music/wma-pop-up
```

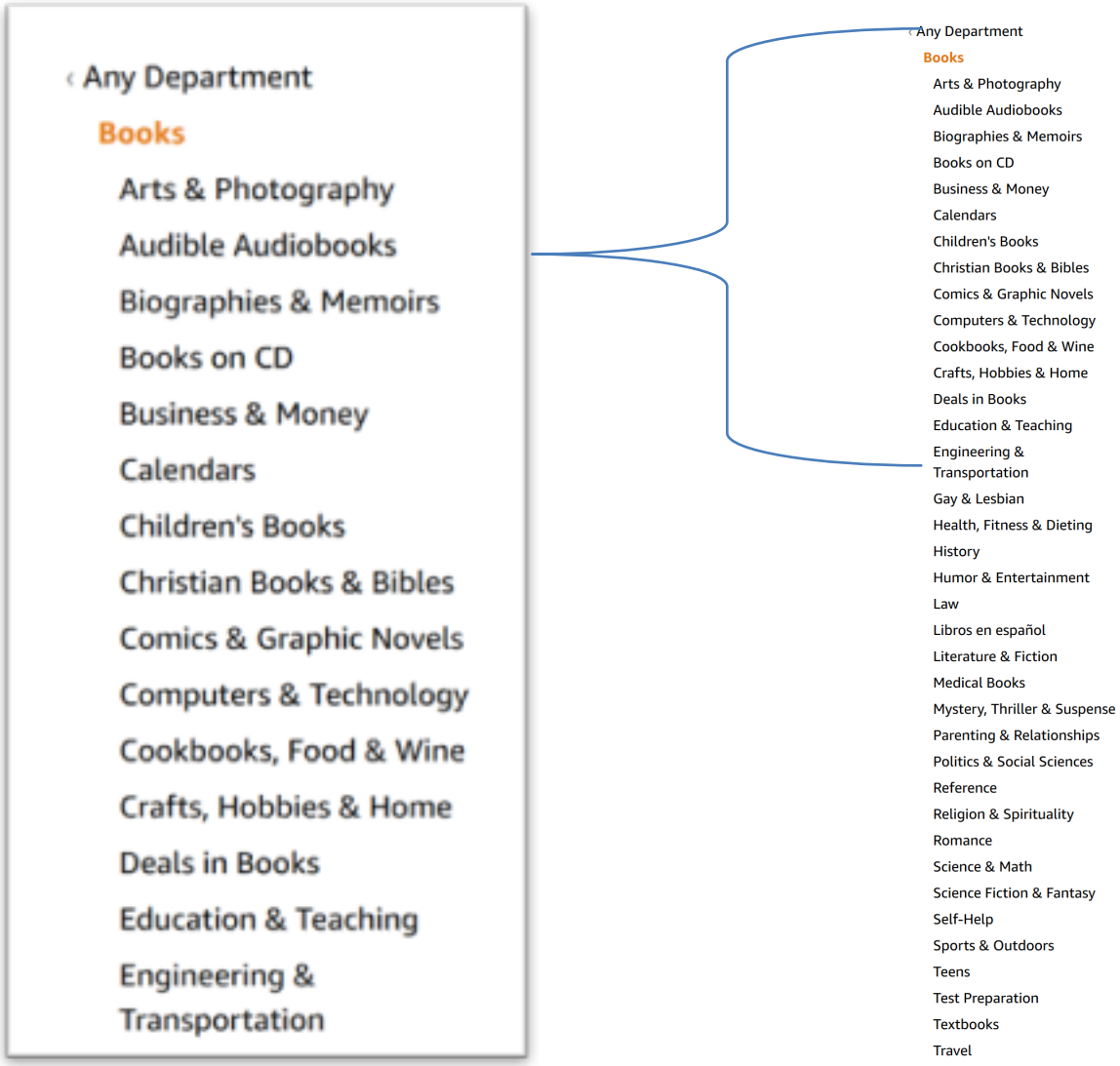
And Don’t Forget Amazon…

There are really a lot of bestseller lists…



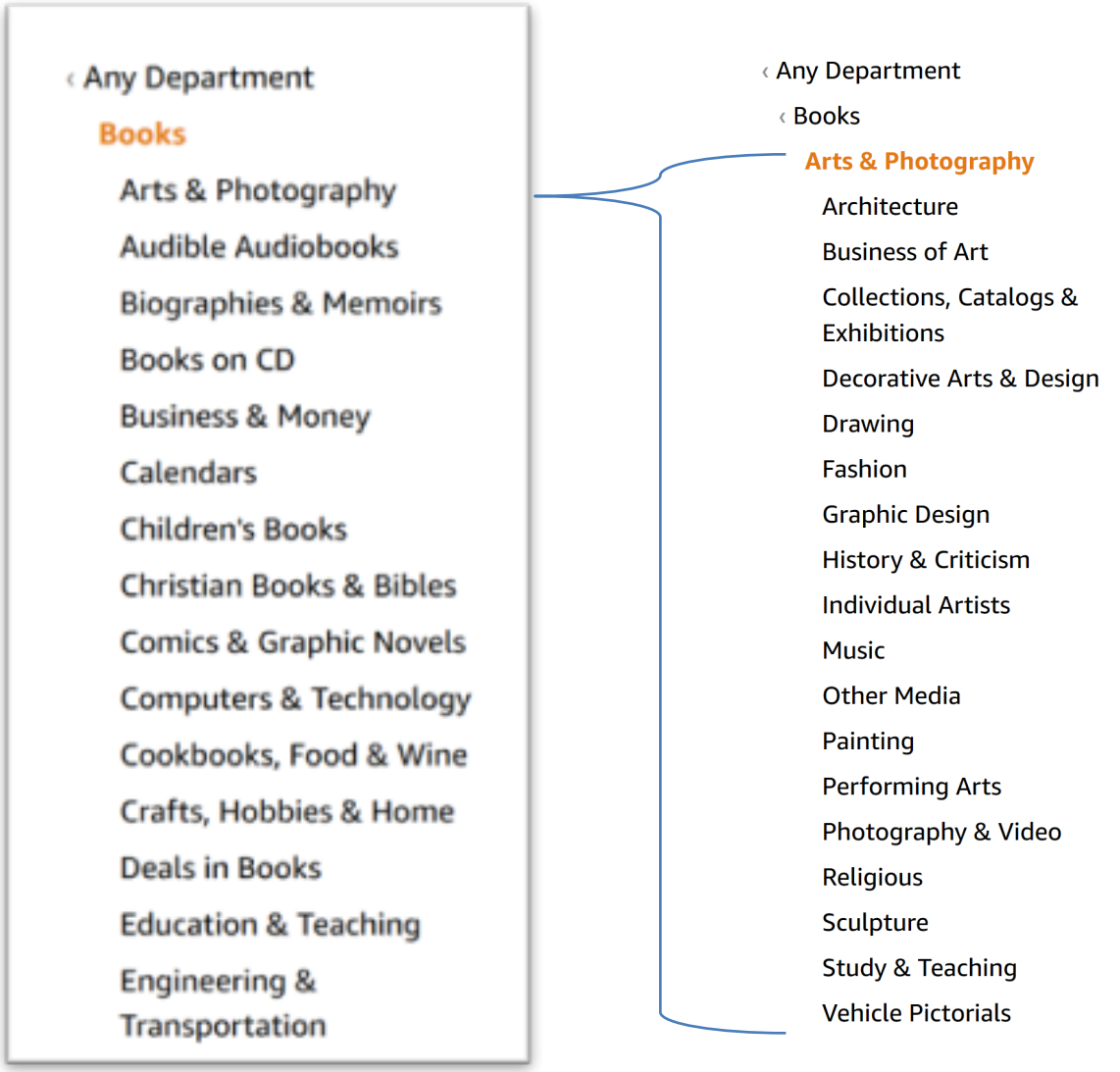
And Don’t Forget Amazon…

There are really a lot of bestseller lists…



And Don’t Forget Amazon…

There are really a lot of bestseller lists…



Benefits

Amazon Best Sellers

Our most popular products based on sales. Updated hourly.

- Any Department
- Books
- Arts & Photography
- Architecture
- Buildings
- Criticism
- Decoration & Ornament
- Drafting & Presentation
- Historic Preservation
- History
- Individual Architects & Firms
- Interior Design
- Landscape
- Project Planning & Management
- Regional
- Security Design
- Sustainability & Green Design
- Urban & Land Use Planning
- Vernacular

Best Sellers in Architecture

#1

The Dot
Peter H. Reynolds
★★★★★ 427
Hardcover
$10.19 prime

#2

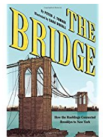
The Last Castle: The Epic Story of Love, Loss, and...
› Denise Kiernan
★★★★☆ 126
Kindle Edition
$12.99

#3

2018 - 2019 Weekly & Monthly Planner: 2018...
Nicole Planner
★★★★½ 3
Paperback
$8.98 prime

#4

The Finer Things: Timeless Furniture, Textiles,...
› Christiane Lemieux
★★★★½ 34
Hardcover
$34.53 prime

#5

The Bridge: How the Roeblings Connected...
Peter J. Tomasi
Hardcover
$18.81 prime

#6

Styled: Secrets for Arranging Rooms, from...
› Emily Henderson
★★★★½ 232
Hardcover
$22.70 prime

#7

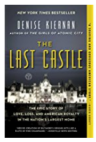
Essential Perennials: The Complete Reference to...
› Ruth Rogers Clausen
★★★★½ 30
Kindle Edition
$1.20

#8

Structures: Or Why Things Don't Fall Down
› J. E. Gordon
★★★★½ 96
Paperback
$16.44 prime

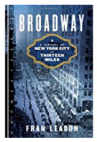
#9

The Last Castle: The Epic Story of Love, Loss, and...
› Denise Kiernan
★★★★☆ 126
Hardcover
$17.17 prime

#10

The Last Castle: The Epic Story of Love, Loss, and...
› Denise Kiernan
★★★★☆ 126
Paperback
$17.00 prime
Release Date: May 1, 2018

#11

India Hicks: A Slice of England
India Hicks
★★★★½ 3
Hardcover
$34.25 prime

#12

Broadway: A History of New York City in Thirteen...
Fran Leadon
Hardcover
$24.20 prime

#13

A Guide to the Good Life: The Ancient Art of...
› William B. Irvine
★★★★½ 420
Hardcover
$17.75 prime

#14

A Pattern Language: Towns, Buildings,...
› Christopher Alexander
★★★★½ 211
Hardcover
$44.02 prime

#15

Ray Booth: Evocative Interiors
Ray Booth
★★★★½ 5
Hardcover
$37.00 prime

#16

The Well-Tended Perennial Garden: The Essential...
Tracy DiSabato-Aust
★★★★½ 23
Kindle Edition
$1.99


New Releases
in ARCHITECTURE


Most Wished For
in ARCHITECTURE


Gift Ideas
in ARCHITECTURE

#17

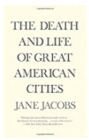
Rachel Ashwell: My Floral Affair: Whimsical...
› Rachel Ashwell
★★★★★ 17
Hardcover
$25.50 prime

#18

Veranda Decorating
Mario López-Cordero
★★★★☆ 4
Hardcover
$28.12 prime

#19

Rosa's Castle
› Deanna Edens
★★★★½ 27
Kindle Edition
$0.99

#20

The Death and Life of Great American Cities
› Jane Jacobs
★★★★½ 196
Paperback
$11.52 prime

#21

A Field Guide to American Houses (Revised): The...
Virginia Savage McAlester
★★★★½ 278
Paperback
$20.36 prime

#22

Henri Samuel: Master of the French Interior
› Emily Evans Eerdmans
★★★★½ 6
Hardcover
$46.39 prime

#23

Planting in a Post-Wild World: Designing Plant...
Thomas Rainer
★★★★½ 70
Kindle Edition
$1.13

#24

Garden Revolution: How Our Landscapes Can Be...
Larry Weaner
★★★★½ 32
Kindle Edition
$1.20

#25 #26 #27 #28 #29 #30 #31 #32

Benefits of This

Now you know (after a bit of HTML parsing), for just about every subject matter category imaginable:

- Some major subcategories
- Popular book titles
- Popular author names
- Content words from those titles, and descriptions thereof

But Amazon is just an example. You can use any content network that allows you to crawl (and as always, check terms of service!).

DataScience@SMU

Applications of a Lexical Knowledge Base

Natural Language Processing

Applications of Lexical Knowledge Bases

Right away and without any higher-level syntax or semantic analysis, we can already use lexical KBs to improve a variety of applications.

Examples are the following:

- Enhance usability of search engines
- Writing evaluation and advice
- Smarter tag clouds

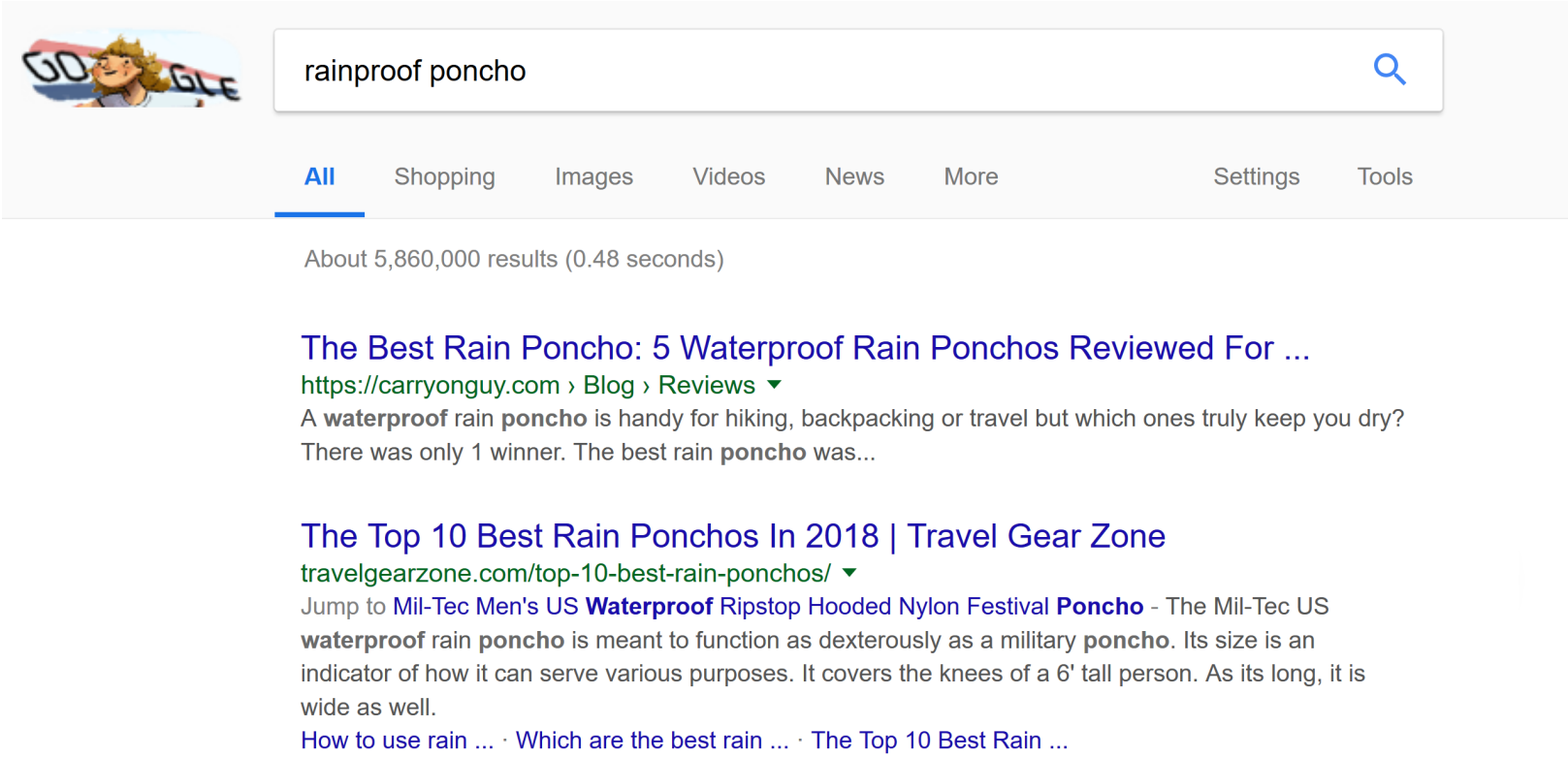
Enhancements to Search Engines

We can use lexical KBs to improve the following:

- Query expansion
- Related searches
- More like this

Query Expansion

- Offer to add synonyms and hyponyms *only* that appear in search result capsules—(pseudorelevance feedback).
- If you dare—automatically show results using expanded terms (like Google does).



Pseudorelevance Feedback

Simple procedure:

1. Retrieve initial search results and assume the top *k* results are relevant.
2. Utilize the top *k* results as though they were user input (marked as relevant).
3. Expand or modify initial query and/or initial results accordingly.

In our case, we usually have *too many* synonyms and hyponyms and many of them are *irrelevant*—but we can use the above procedure to narrow the list down to the probably-more-relevant ones.

Results for "berry pie recipes"

Blah blah **pie** blah **strawberry** blah

Blah blah blah **raspberry tart** blah

Blah blah **cobbler** blah **blueberry**

Marionberry pie blah blah blah

Note: "patty" is a hyponym of "pie," but we won't use it; "currant" is a hyponym of "berry," but we won't use it—because there's no support here.

Related Searches

- If initially you have *sparse* results, replace one term at a time with a high-frequency member of a *synset* (validated with pseudorelevance feedback)
- Eliminate any zero-result queries
- Suggest in descending weighted order
  - Weightings can include number of results found, median score of top results found, etc.

On this one, we swapped the order of the query terms. Can you guess why?

Related Searches

- If initially you have *too many* results, replace one term at a time with a *hyponym* (validated with pseudorelevance feedback).
- Again, eliminate any zero-result queries, and suggest in descending weighted order.

Showing results for "dog behavior."

Narrow your search:

puppy behavior

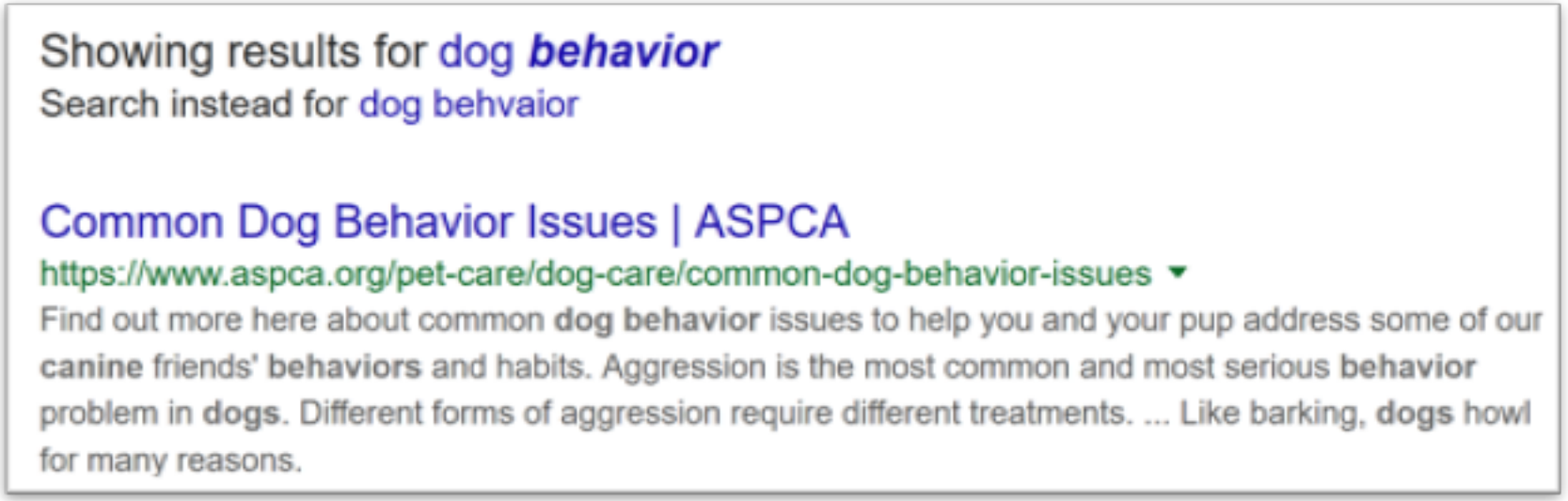
Dalmatian behavior

pug behavior

dog aggression

More Like This

- Which words from the capsule are relevant?
  - Answer: Those related, in our lexical knowledge base, to the query terms
  - Can use those terms to fashion a Boolean search string




```
(dog | canine) & (behavior | aggression | habits)
```

Writing Advice

Remember the application to writing advice? Now we can give suggested word alternatives when we notice redundancy.

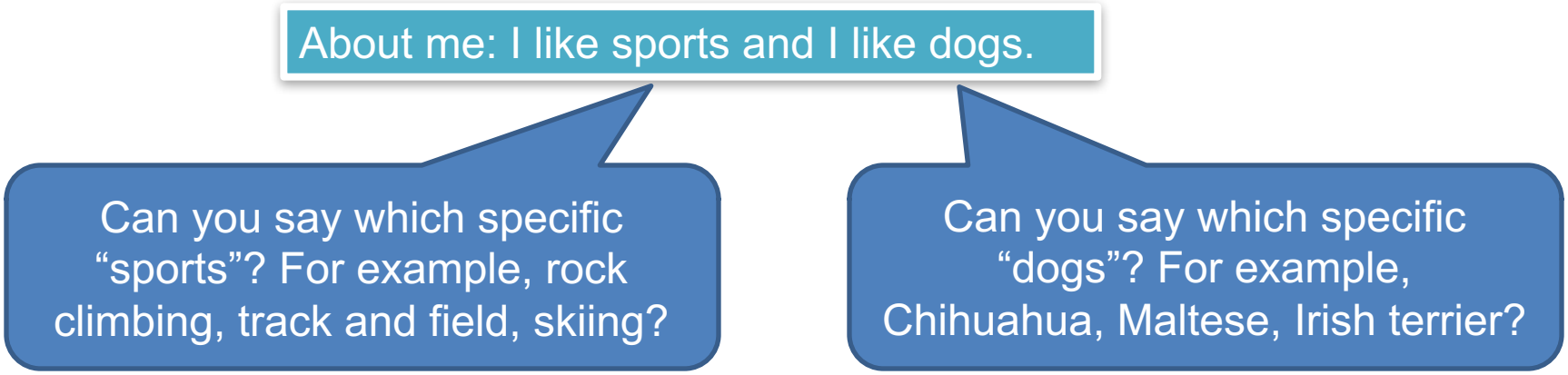
I like Alex because he is fun. When I want to have fun then I have Alex come over and then we have fun together.

You said “fun” three times…. Could you rewrite using words like “play” or “sport” or “laughs”?

Of the many synonyms and hyponyms of “fun,” these were selected because they were the only ones with high frequency in the relevant corpus (same-grade-level essays).

Writing Advice

- We can also notice a lack of specificity and give suggestions.



We decide *which words* should be more specific by noticing which nouns are *content words* that have a lot of hyponyms. The words "dogs" and "sports" both have a lot of hyponyms at multiple levels.

Filtering those hyponyms by frequency within a reference corpus yields the suggestions shown here.

Smarter Tag Clouds

We can avoid having semiredundant words in a tag cloud by checking synonymy, hypernymy, etc.

Smarter Tag Clouds

What else could you do?

- Color-code the tags according to hypernym tree:

  player receiver quarterback halfback wide-out
  coach trainer
  agreement compliance contract

- Size the tags by hyponym-tree-depth

  player receiver tight-end wide-out

Applications, and More Applications

We've just explored a few application areas.

You can brainstorm around these:

- Dating service
  - Match free-form essays prompted by "describe yourself and your interests."
- Job-seeker service
  - Match resumes to job opening descriptions.
- Taxonomy mapping
  - Match node names from two taxonomies that are not exact word matches.

What would be the overall procedure for each of the above?

What other applications can you think of?

DataScience@SMU