Word Embeddings - Semantics: What is in my Documents?

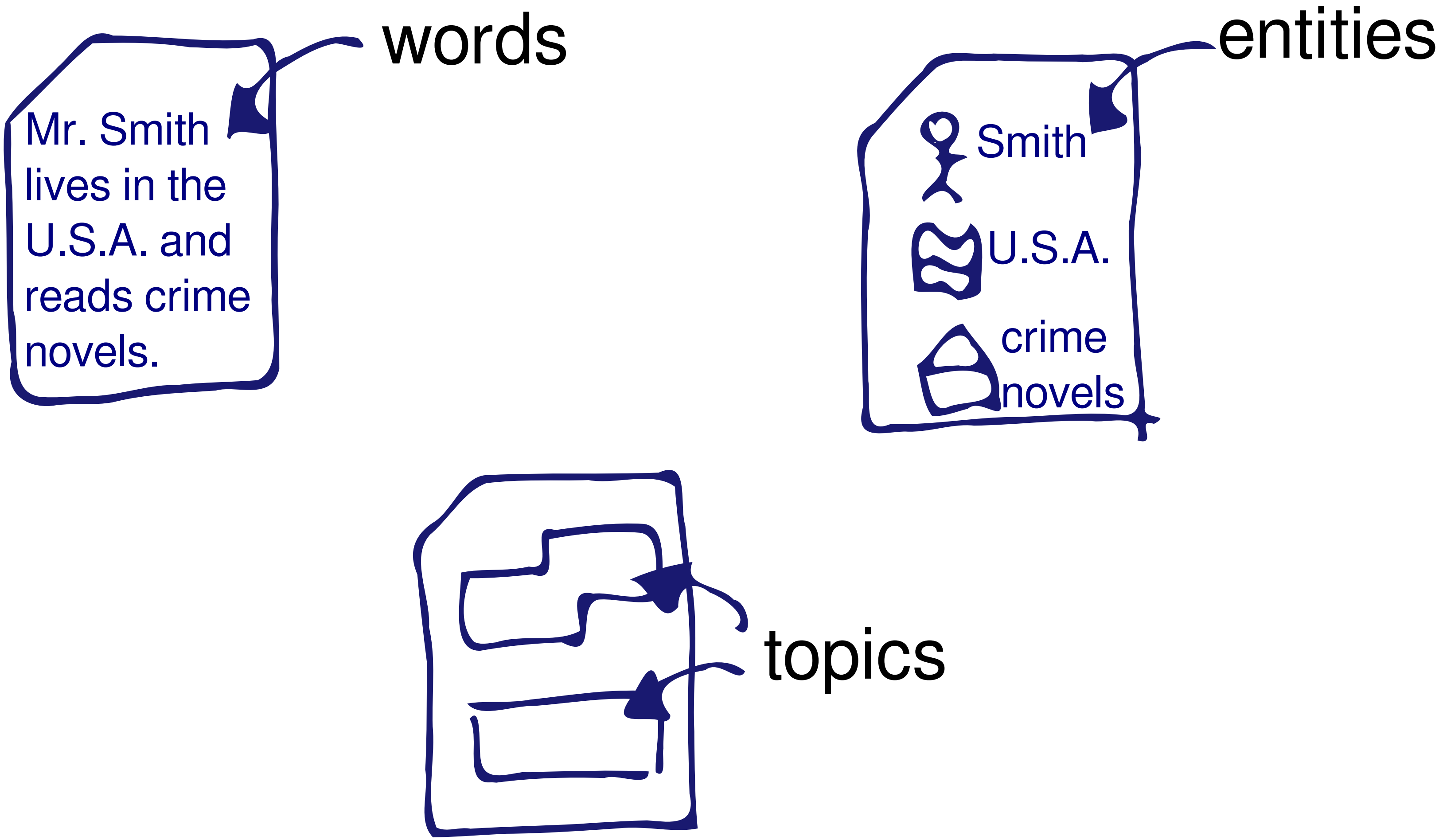
Laura Dietz
University of New Hampshire
dietz@cs.unh.edu

University of New Hampshire

The Problem

The Solution

Collections of Text



Outline

Different techniques to inspect your documents.
- topic models
- word embeddings
- text classification
- entity linking
- entity aspects
- search index and retrieval (with entities)

Why am I qualified to give this Talk?

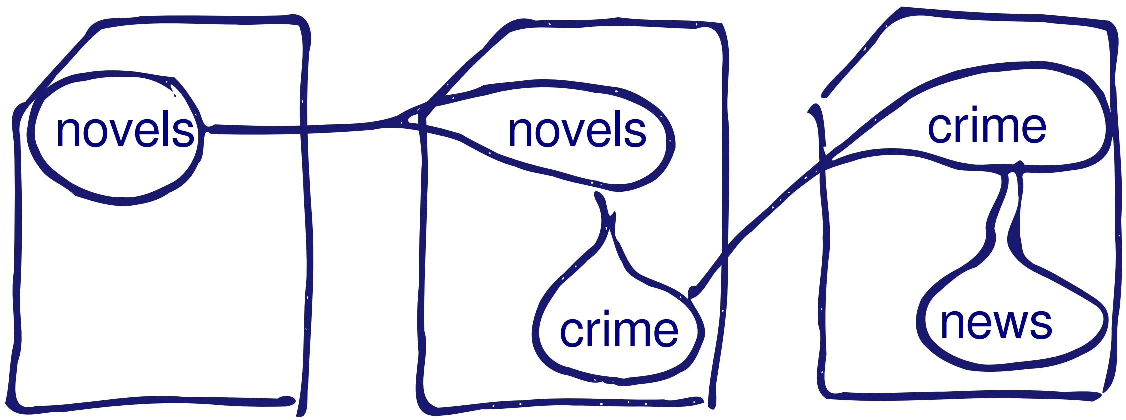
Laura Dietz - Computer Scientist

2000: Software Developer
2004: Semantic Web
2006: Machine Learning / Topic Models
2011: Natural Language Processing / Entity Linking
2013: Information Retrieval with KGs
2016: Assistant Professor

Outline: Topic Models

Different techniques to inspect your documents.
- topic models
- word embeddings
- text classification
- entity linking
- entity aspects
- search index and retrieval (with entities)

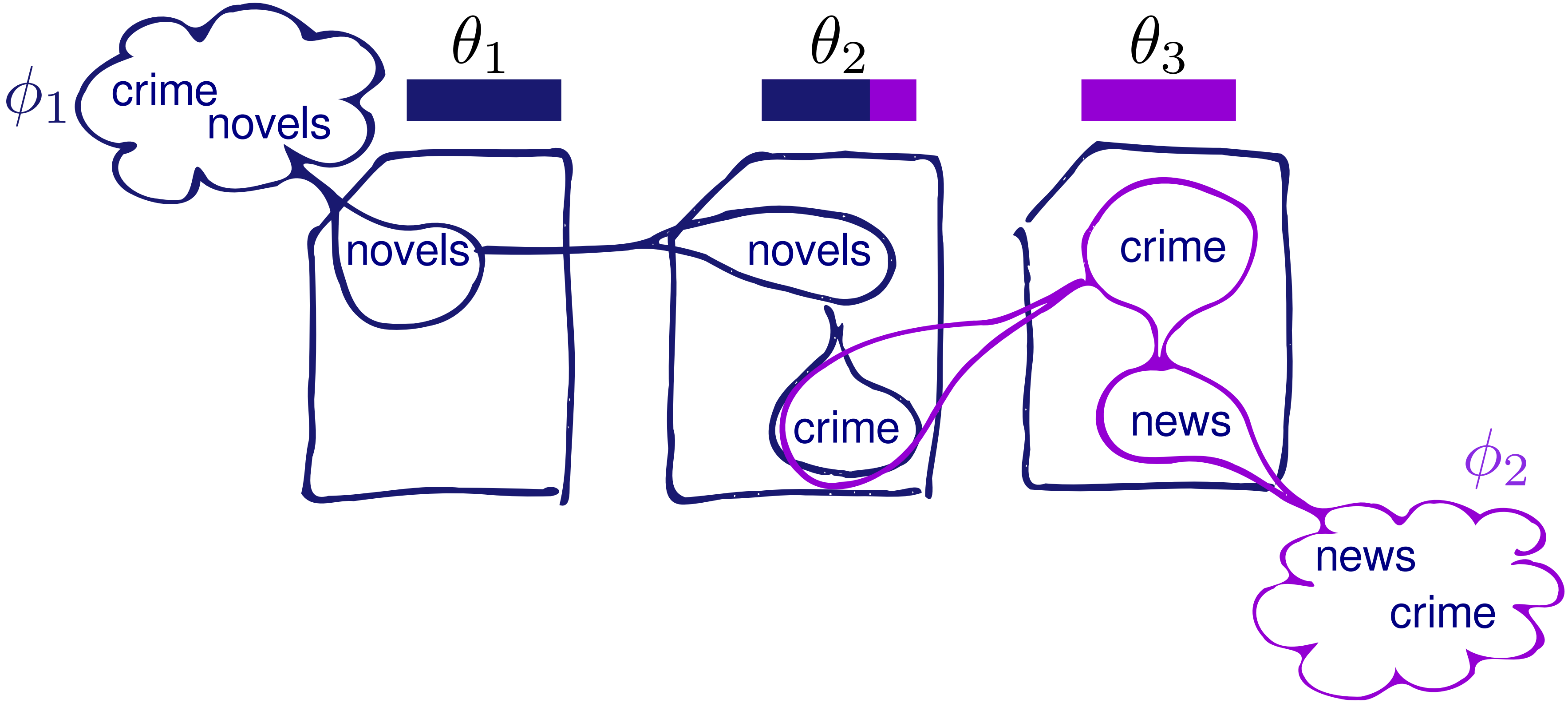
Topic Models



- Same words are likely about the same topic.
- Words in the same document are likely about the same topic.

Topic Models

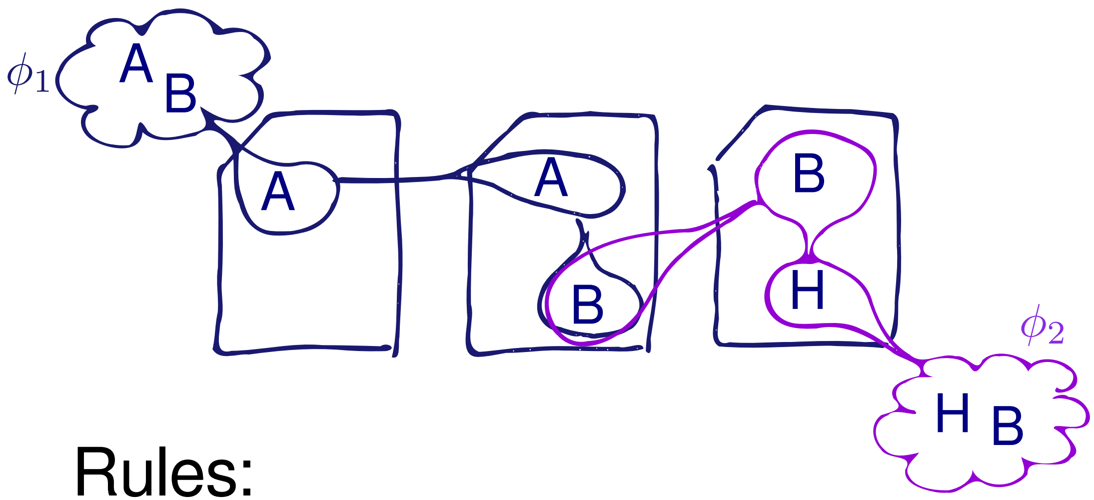
[Blei et al 09]



$\phi_{\text{topic}}(\text{word})$ high if word is important for topic.

$\theta_{\text{doc}}(\text{topic})$ high if topic important for doc.

Topic Model Exercise - Apply rules to find topics!



doc 1: B A G C

doc 2: B A D I

doc 3 : B A F H

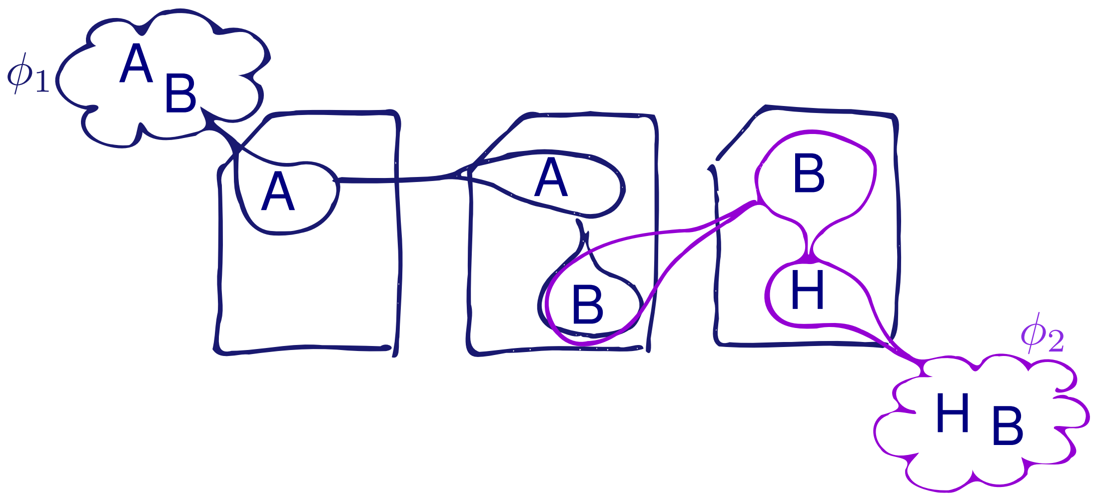
doc 4: E A F H

doc 5: E A G C

Rules:

0. Assign each word a random topic.
1. Assign two same words to the same topic.
2. Assign two words in the same document the same topic.

Topic Model Exercise - Solution



A: read
B: politicians
C: novels
D: legal
E: people
F: the
G: crime
H: news
I: texts

doc 1: B A G C
politicians read crime novels

doc 2: B A D I
politicians read legal texts

doc 3 : B A F H
politicians read the news

doc 4: E A F H
people read the news

doc 5: E A G C
people read crime novels

Topic Model Toolkits

- LDA-c
- Mallet
- Topic Model Toolbox
- Stanford Topic Modeling Toolbox
- Tomoto

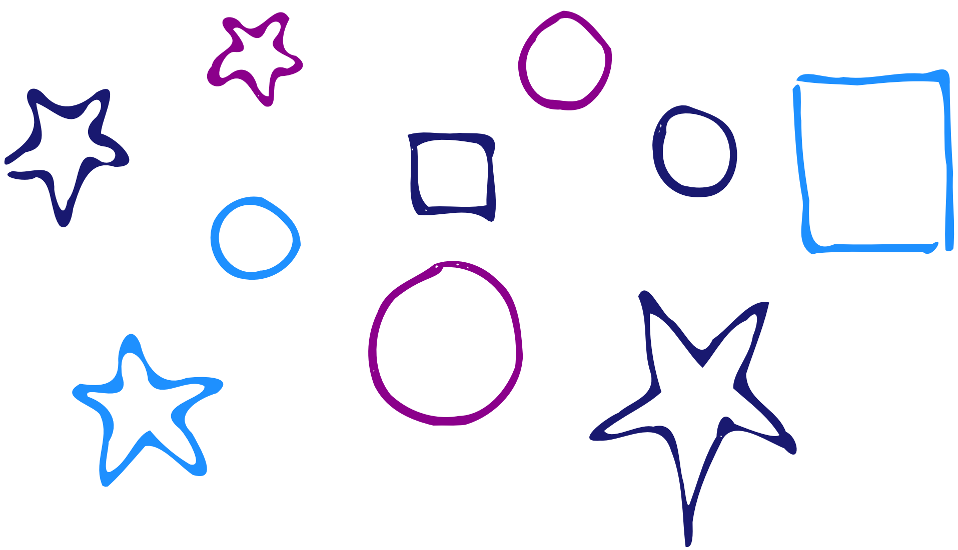
Extensions for:
Authors [Rosen-Zvi 04], Time [Wang 06],
Citation networks [Dietz 07], Ideal point [Gerrish 10]
Friend-networks [Dietz 12],
Taxonomies [Bakalov 12],
and so many more....

Topic Model Caveats

Some topics are great (aka "spot on")
others are merged/split or don't make sense.

It is impossible to know which topics are correct.

There are many correct solutions:



Please Evaluate Tools!

When your research relies on a tool
make sure it works in *your* domain
and for *your* task!

...otherwise you may draw wrong conclusions!

Issues of Topic Models

Topic models are based on assumptions that intuitively hold for topical words.
...but also for many "misleading" words.

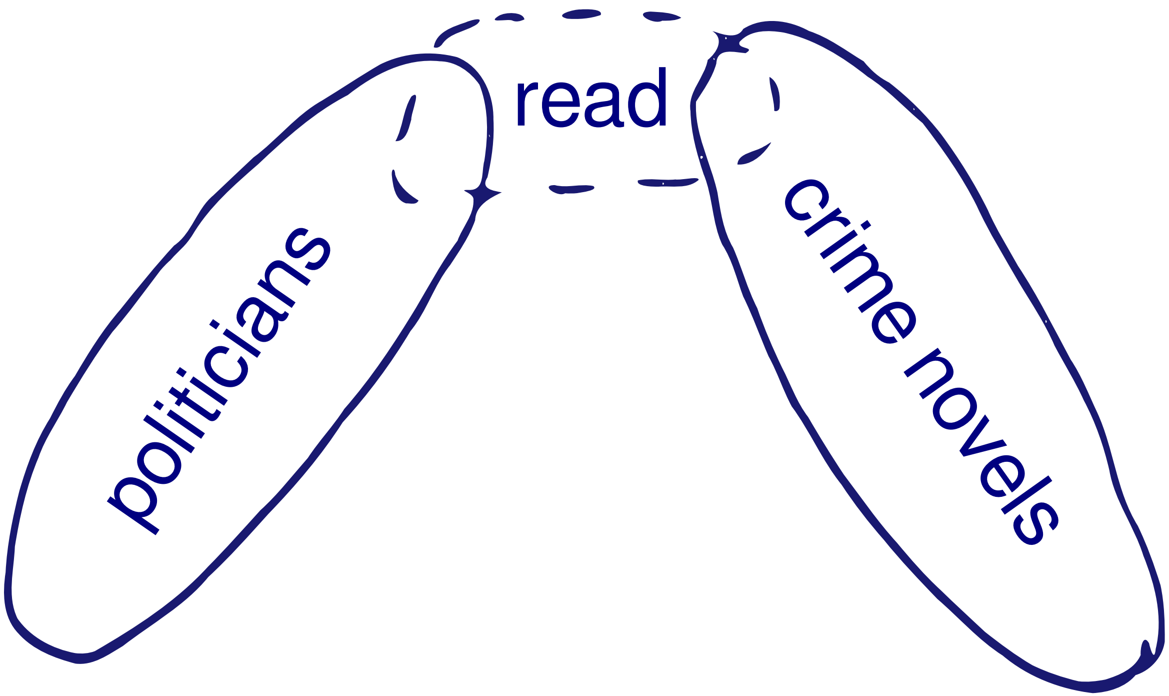
Which are the topical words?

Many **politicians** in the **U.S.A.** like to **read crime novels**.

We don't know which are the topical words we are looking for.

Topic Model Issues

politicians read crime novels



"read" bridges two topics

politicians read legal texts

politicians read the news

people read the news

people read crime novels

Outline: Word Embeddings

Different techniques to inspect your documents.

- topic models
- word embeddings
- text classification
- entity linking
- entity aspects
- search index and retrieval (with entities)

Word Embeddings

[Levy & Goldberg et al 14]

read crime novels

politicians read

All words that fit here are similar!

Word Embeddings

politicians read ____

crime

$$\begin{pmatrix} 3 \\ 0.6 \\ 2.6 \\ -7 \end{pmatrix}$$

novel

$$\begin{pmatrix} 3 \\ 0.5 \\ 2.5 \\ -7 \end{pmatrix}$$

news

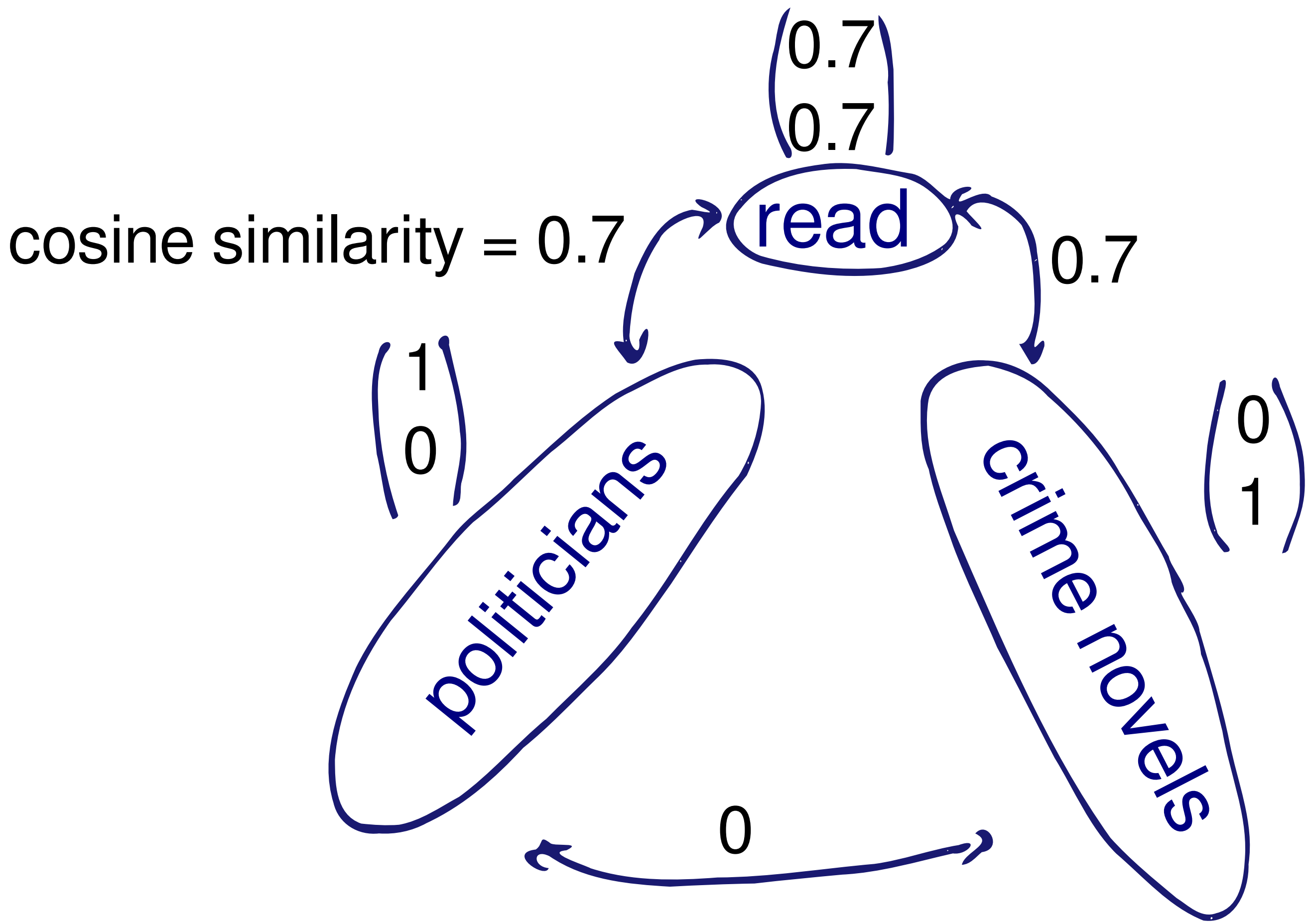
$$\begin{pmatrix} 2 \\ 0.5 \\ 2.6 \\ -6 \end{pmatrix}$$

legal

$$\begin{pmatrix} 1 \\ 0.2 \\ 1.5 \\ -1 \end{pmatrix}$$

Multiple Meanings with Word Embeddings

"Read" can have the same distance to both words, without these words being similar to each other.



Word Embedding Issues

Learns similarity according to **types**.
(e.g., crime novels, legal texts & news are similar).

But often does not learn **topical** similarity.
(e.g., a novel, its author, and its subject are different).

Word Embedding Toolkits

Word2vec
GloVE
Gensim

Download pre-trained embeddings
(avoid using embeddings from different domains)
or train embeddings yourself (all you need is text)

(you may also like SeqToSeq)

An Apology...

We computer scientists don't have a fool proof method for extracting topics from text.

(but next, a few things that work...)

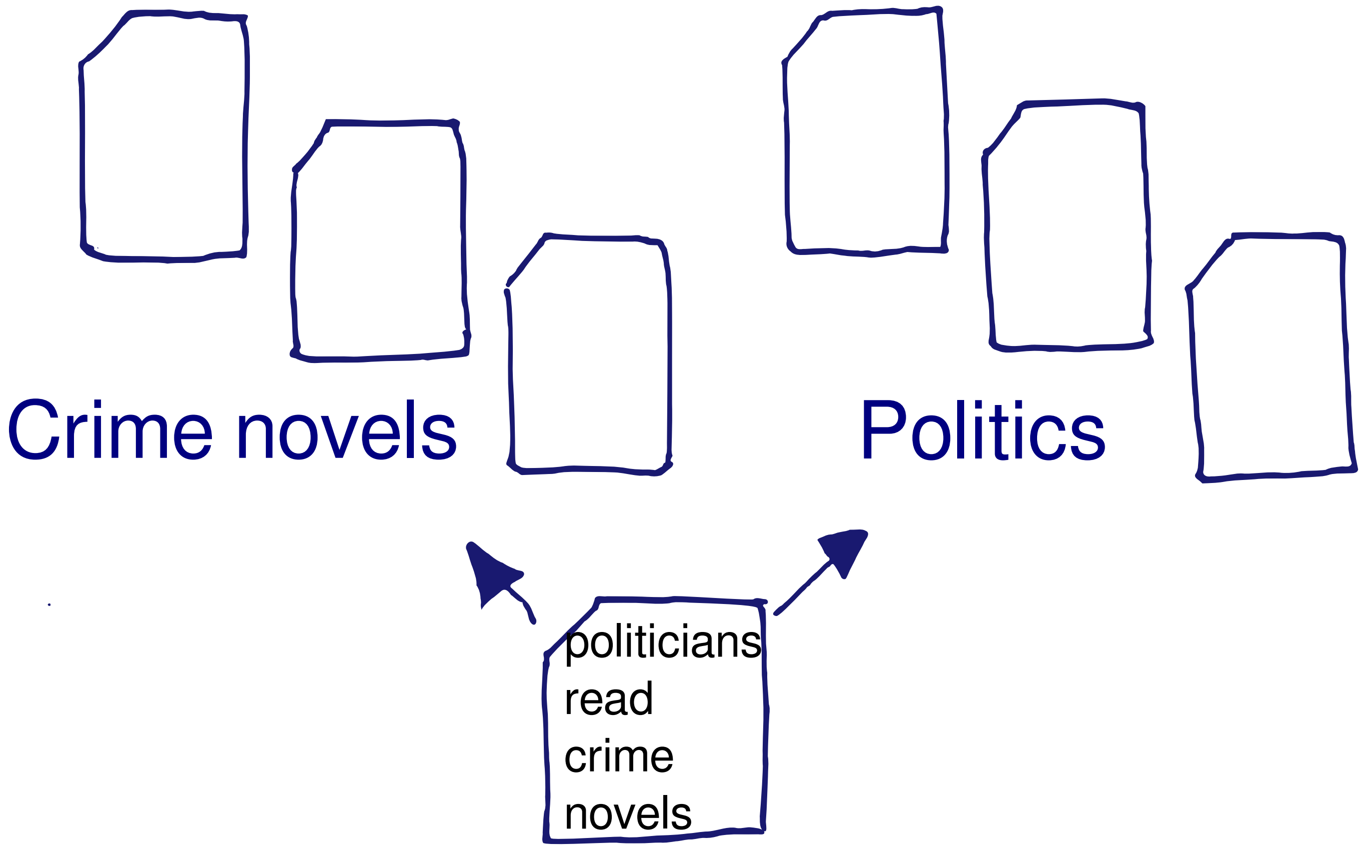
Outline: Text Classification

Different techniques to inspect your documents.

- topic models
- word embeddings
- text classification
- entity linking
- entity aspects
- search index and retrieval (with entities)

Text Classification

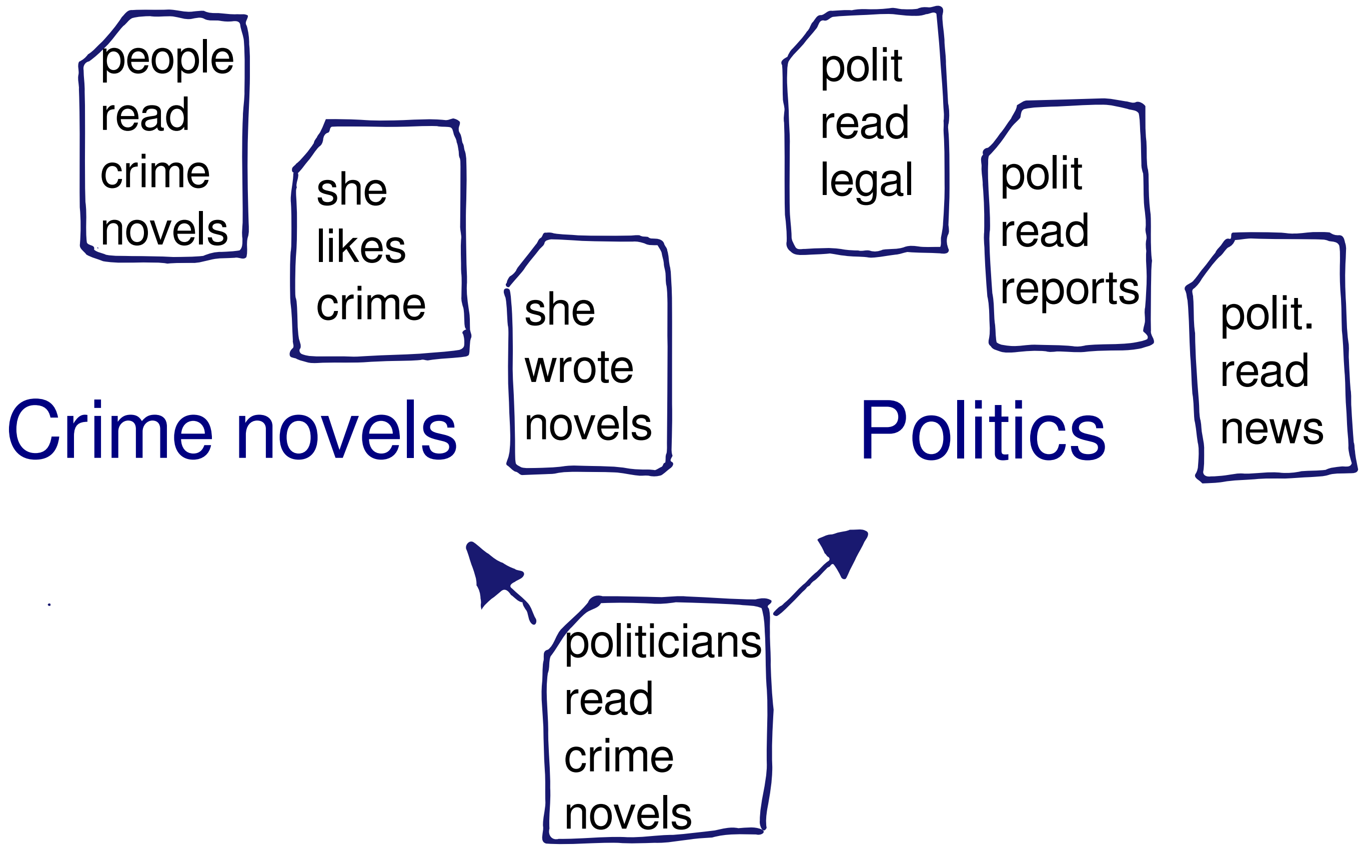
Caveat: Needs labeled training data
for *your* domain and *your* task!



Text Classification (Naive Bayes)

crime: 2
likes: 1
novels: 2
people: 1
read: 3

legal: 1
news: 1
read: 3
reports: 1
politicians: 3



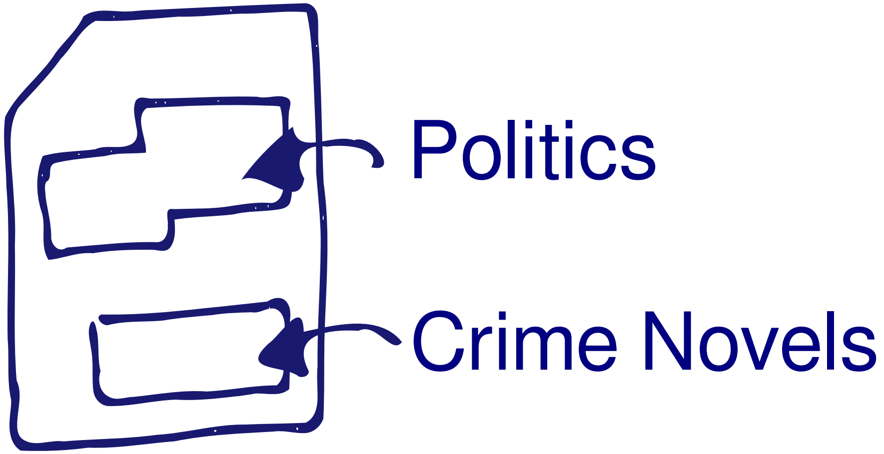
Text Classification Toolkits

Support Vector Machines (SVM)
Random Forests
Weka
Scikit.learn

Text Classification Issues

Requires a lot of manual training data.
(labor-intensive, not feasible for fine-grained topics).

Often only a portion of text is on topic.
(see Multi-label classification.)



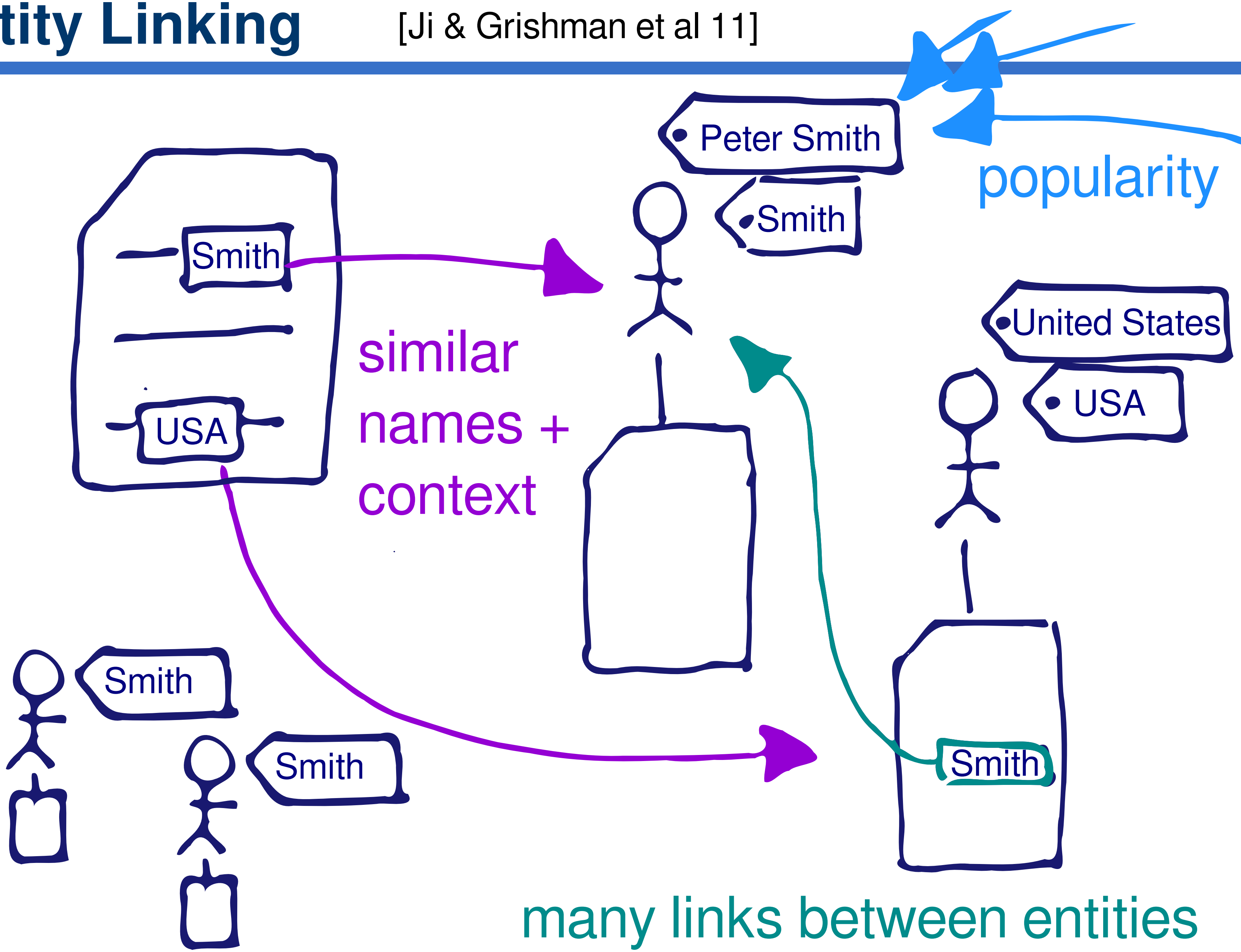
Outline: Entity Linking

Different techniques to inspect your documents.

- topic models
- word embeddings
- text classification
- entity linking
- entity aspects
- search index and retrieval (with entities)

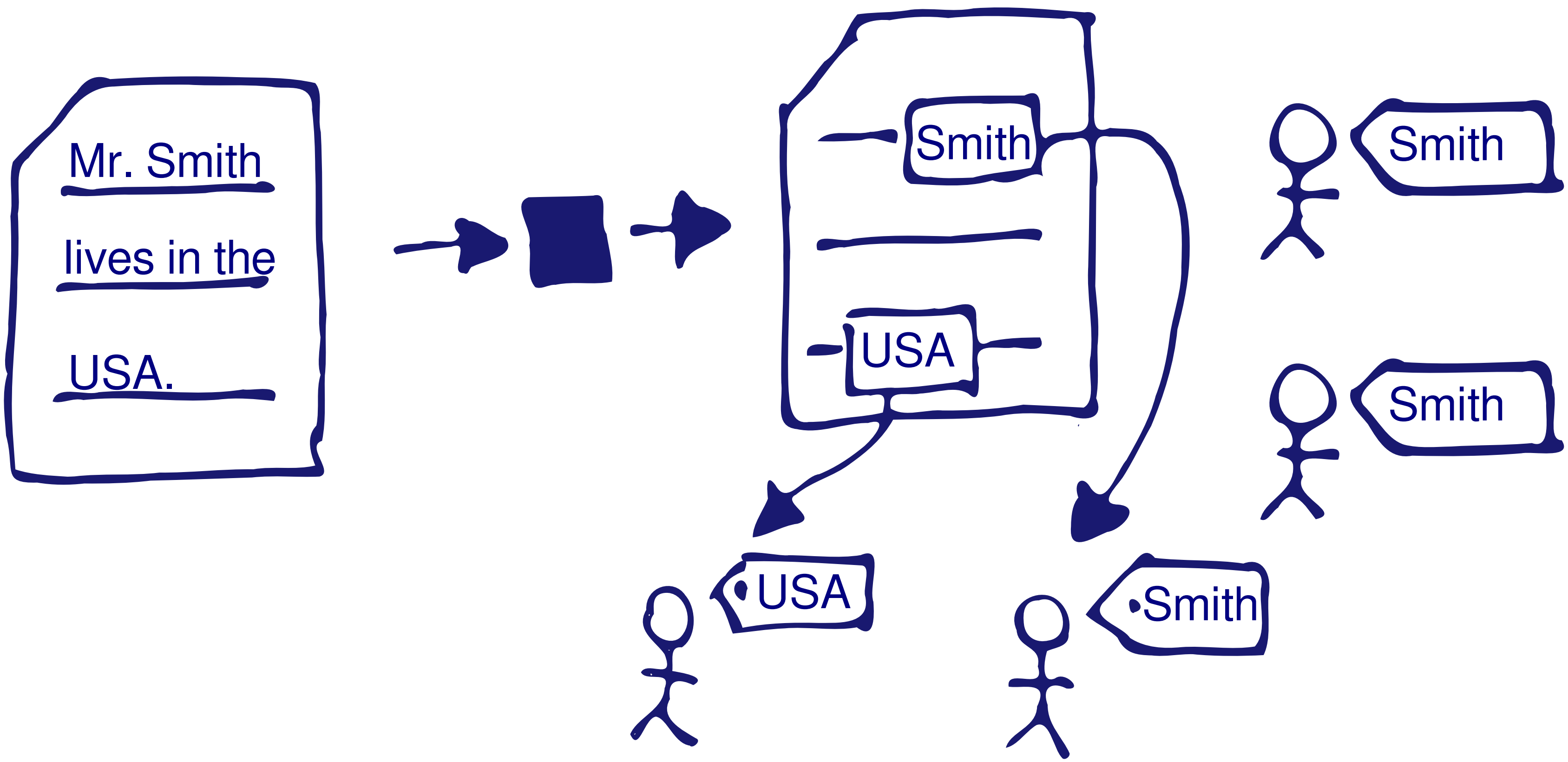
Entity Linking

[Ji & Grishman et al 11]



Entity Linking

A black box that takes text and...

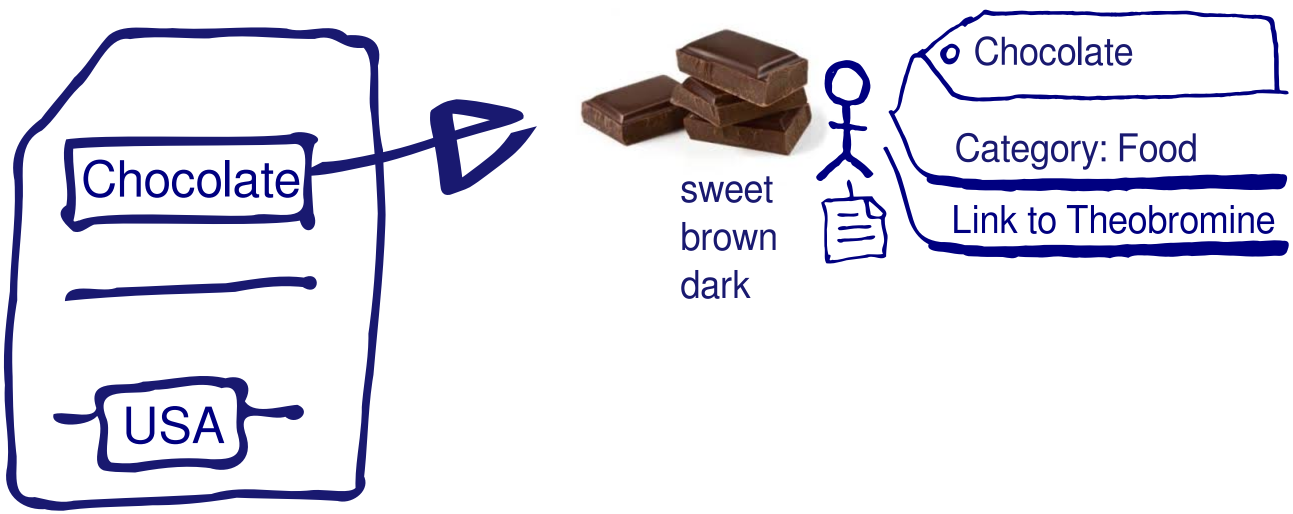


..spots mentions of (Wikipedia) entities in text
and disambiguates among similarly named entities.

Wikipedia Entities

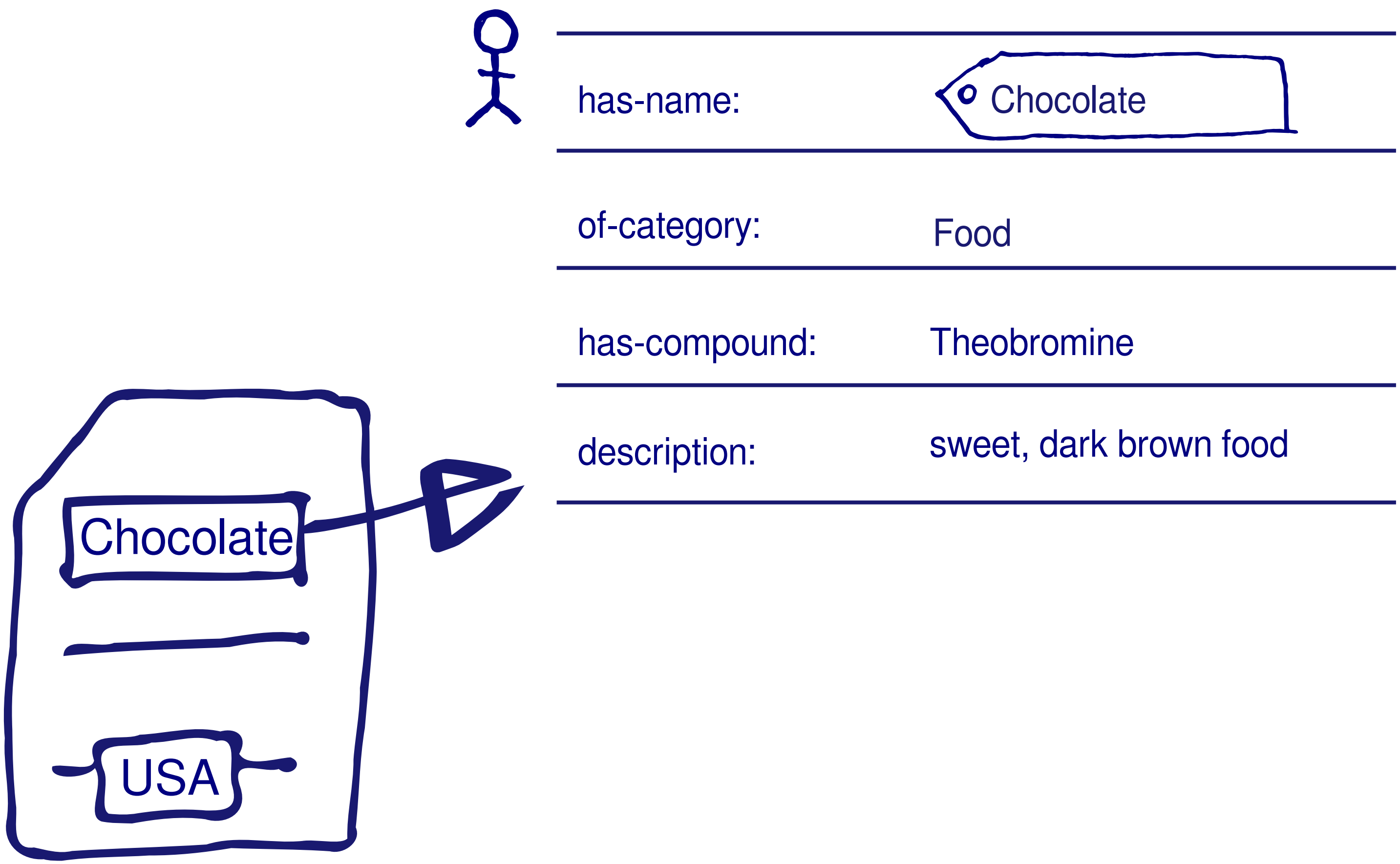
1 Wikipedia page = 1 entity

not just people, organization, and places
also: Brexit, Economy, Immigration, Chocolate

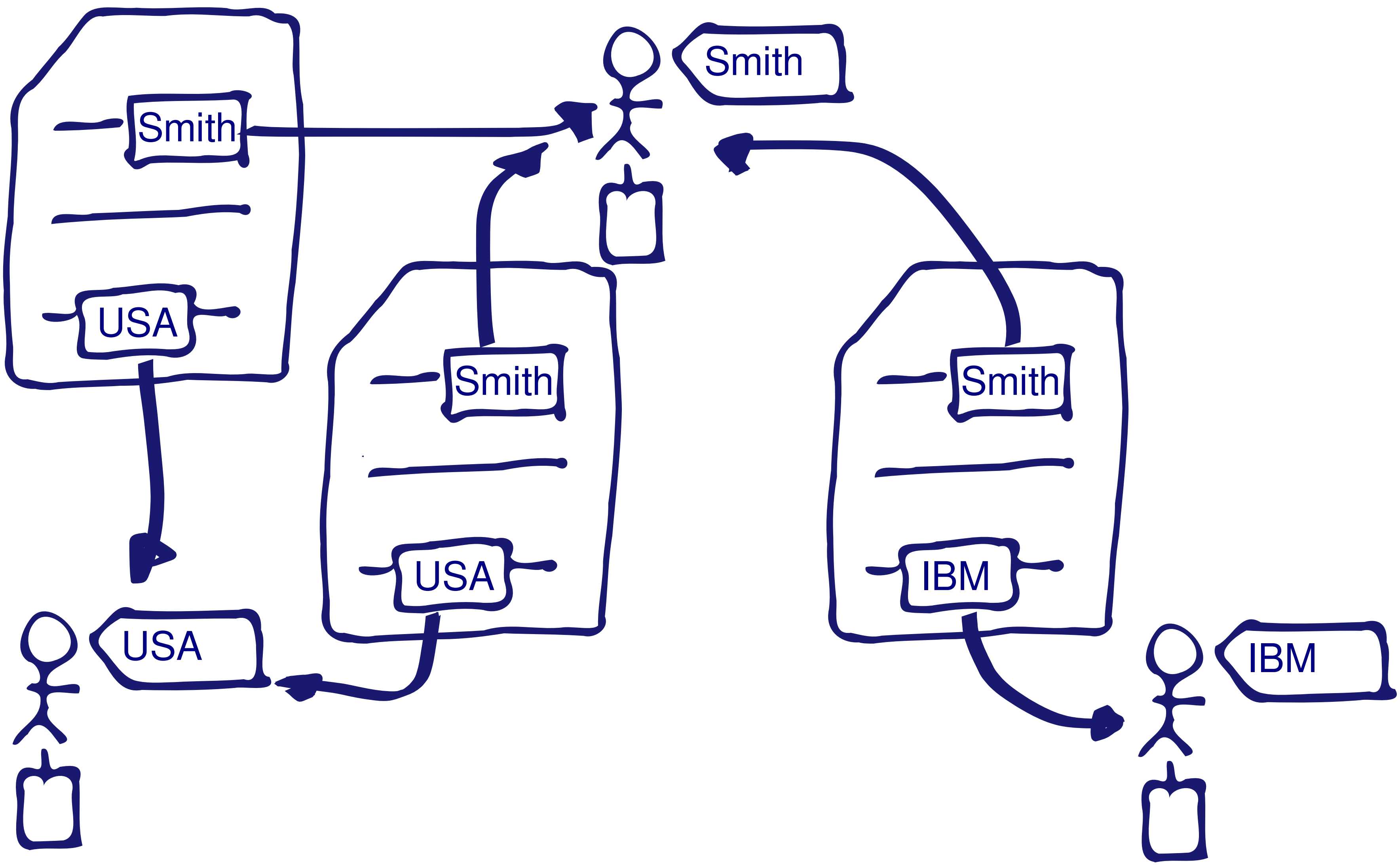


Entities from Ontologies / Knowledge Graphs

Other resources define semi-structured entities



Entity Linking to Inspect your Collection



Entity Linking Toolkits

- TagMe!
- Smaph
- DBpedia Spotlight
- AIDA
- ....

Idea: You can set up your own Wiki server define concepts important to your research and generate some training links.

Entity Linking Issues

Entity links are not saying much about topics.

One could use Wikipedia's categories.
But these are often very surprisingly incomplete, inconsistent, and too fine-grained.

Outline: Entity Aspects

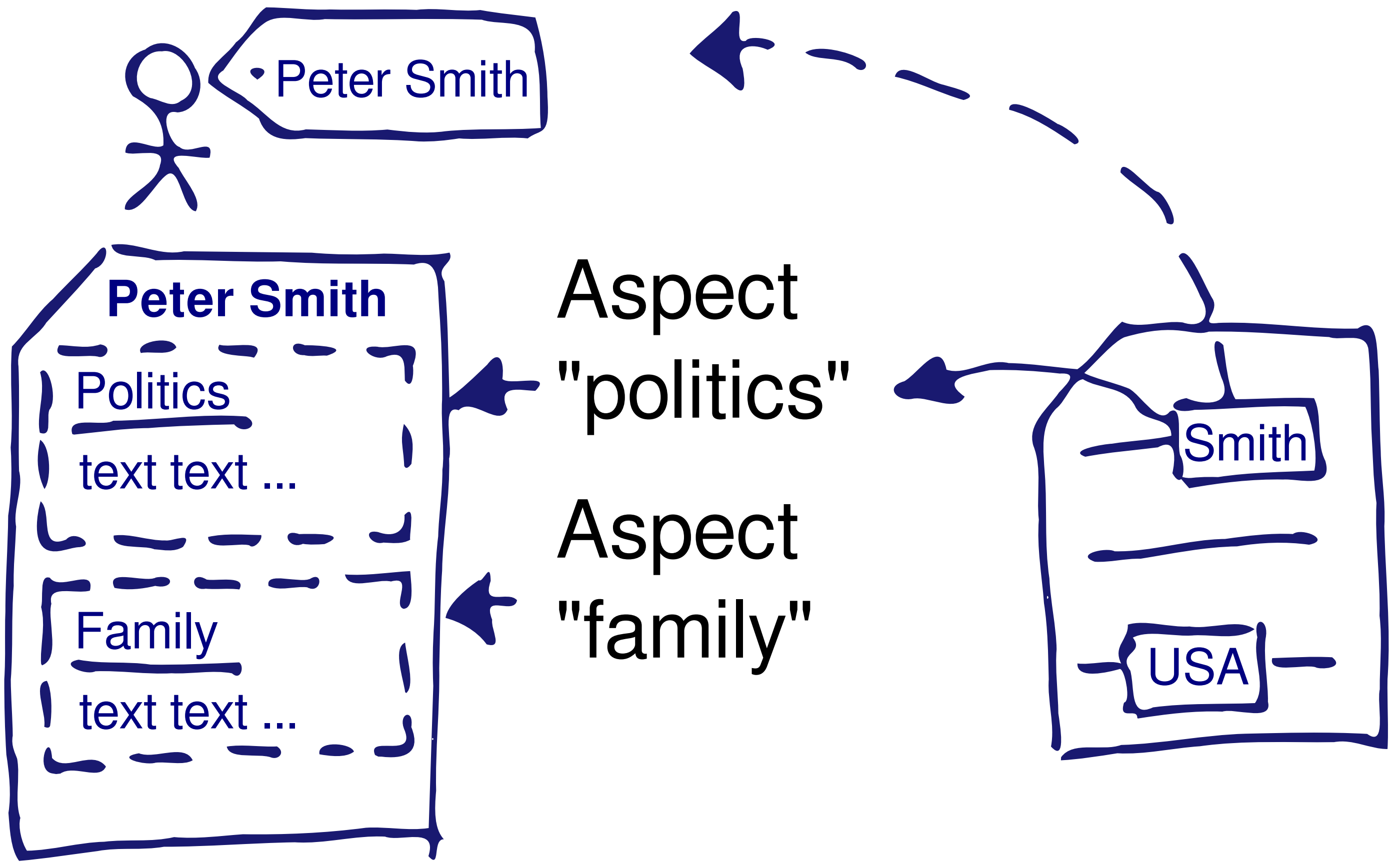
Different techniques to inspect your documents.

- topic models
- word embeddings
- text classification
- entity linking
- entity aspects
- search index and retrieval (with entities)

Entity Aspect Linking

[Nanni et al 18]

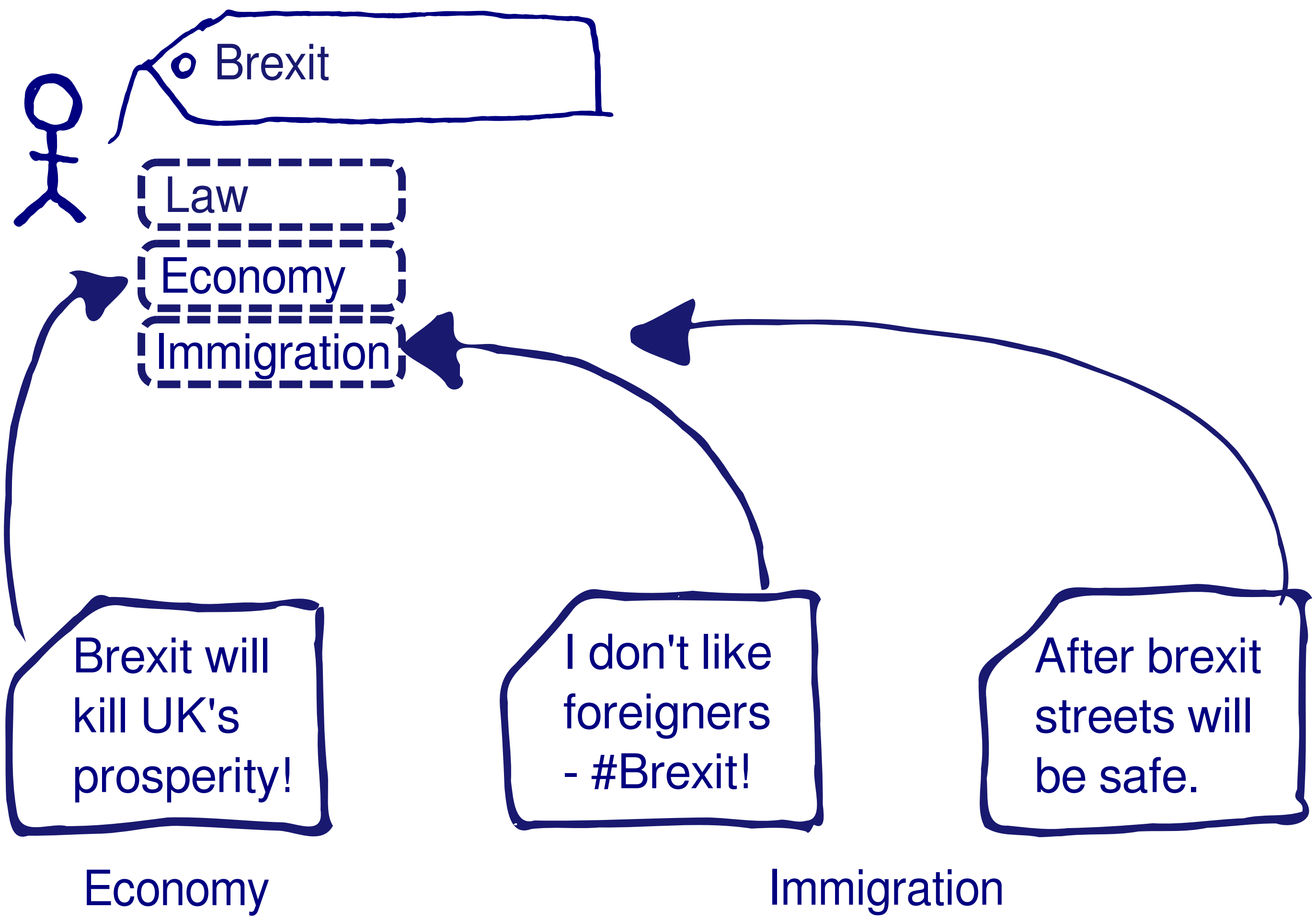
Harvested from sections of the entity's Wiki article.



Refine entity links with aspects that match context.

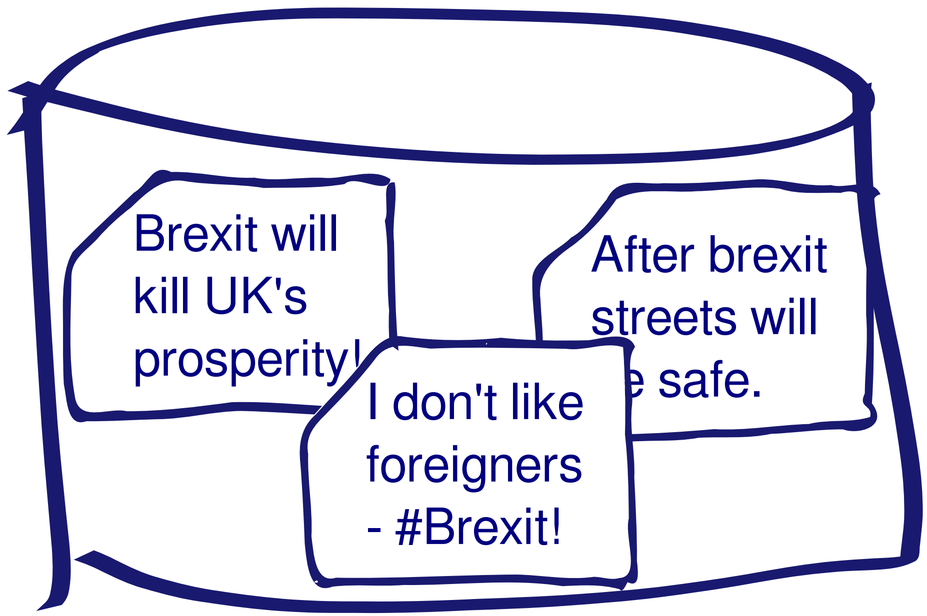
Entity Aspect Example Application

Twitter classification into different aspects of Brexit.



Search Index

Create a search index with documents.



Create different descriptions of your topic.

Use description = query to retrieve top 10.

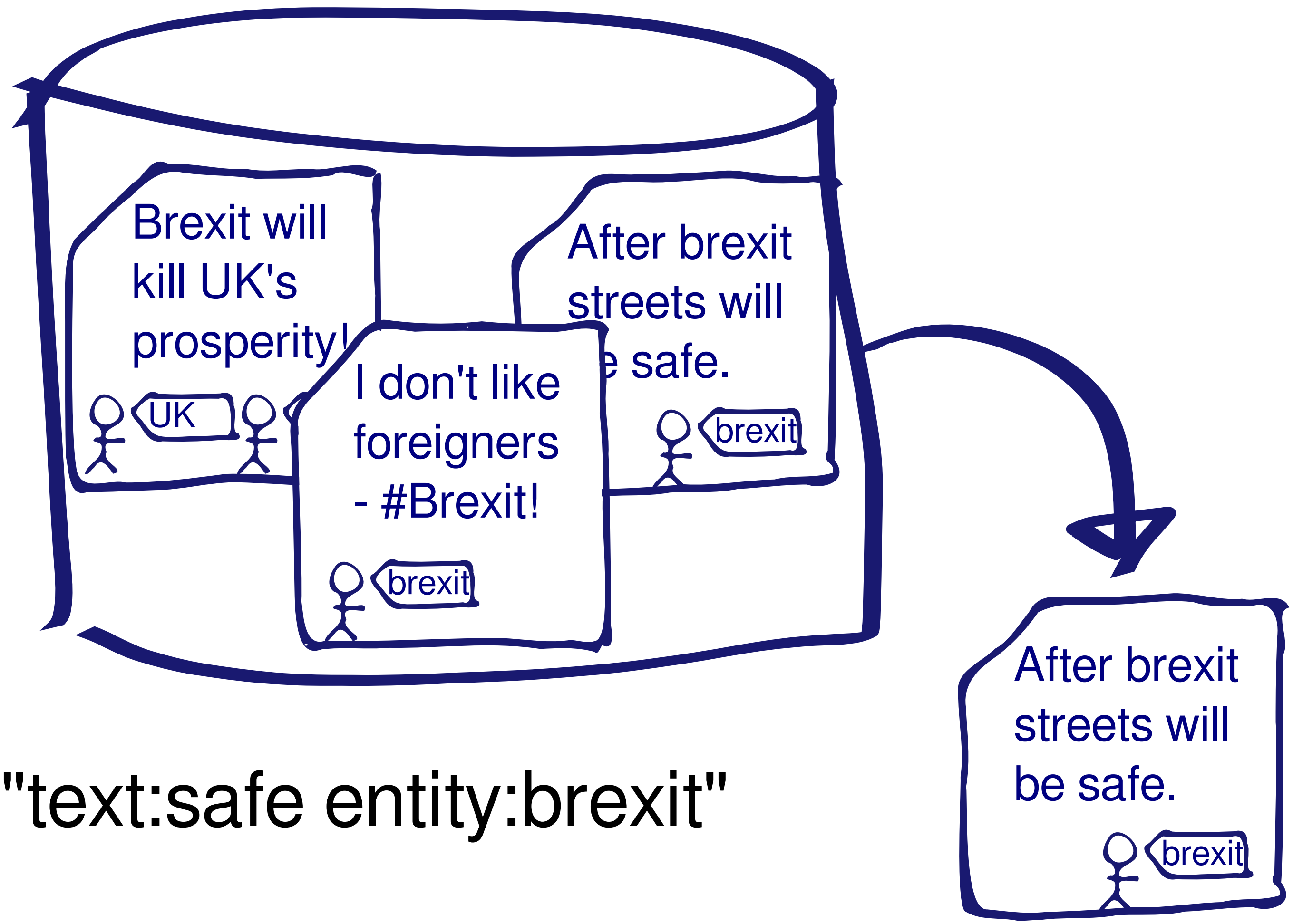
Outline: Search Index and Retrieval

Different techniques to inspect your documents.

- topic models
- word embeddings
- text classification
- entity linking
- entity aspects
- search index and retrieval (with entities)

Search Index with Entities

Documents can have fields:



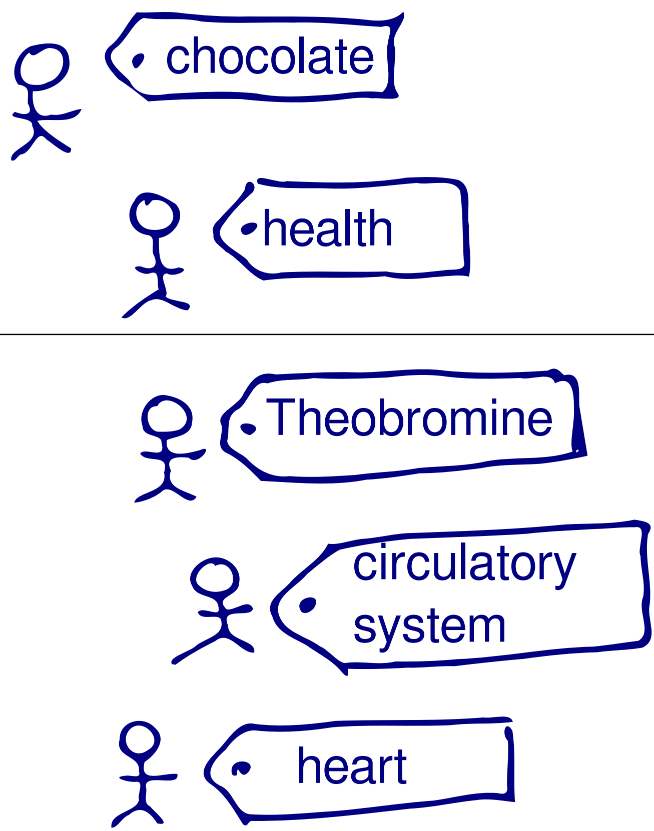
Query = "text:safe entity:brexit"

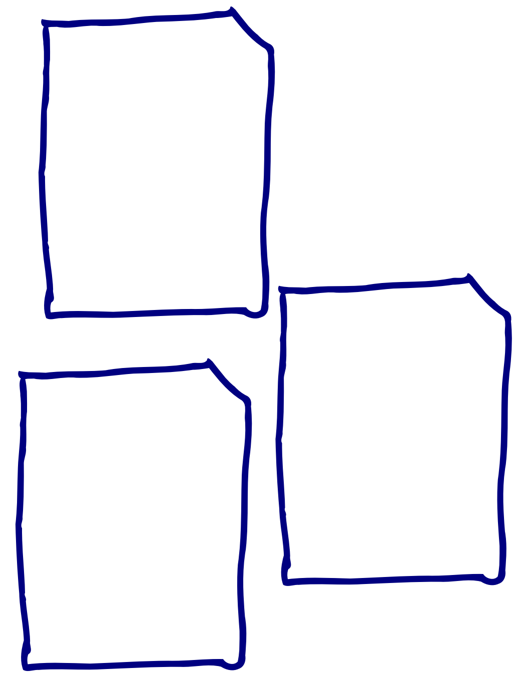
Relevant Entities

Query	EU UK relations	dark chocolate health benefits
Query entities	EU UK	chocolate health
Latent entities	Brexit Theresa May	Theobromine circulatory system heart
	Named Entities	**Concepts**

Matching Entities in Documents

Q: dark chocolate health benefits

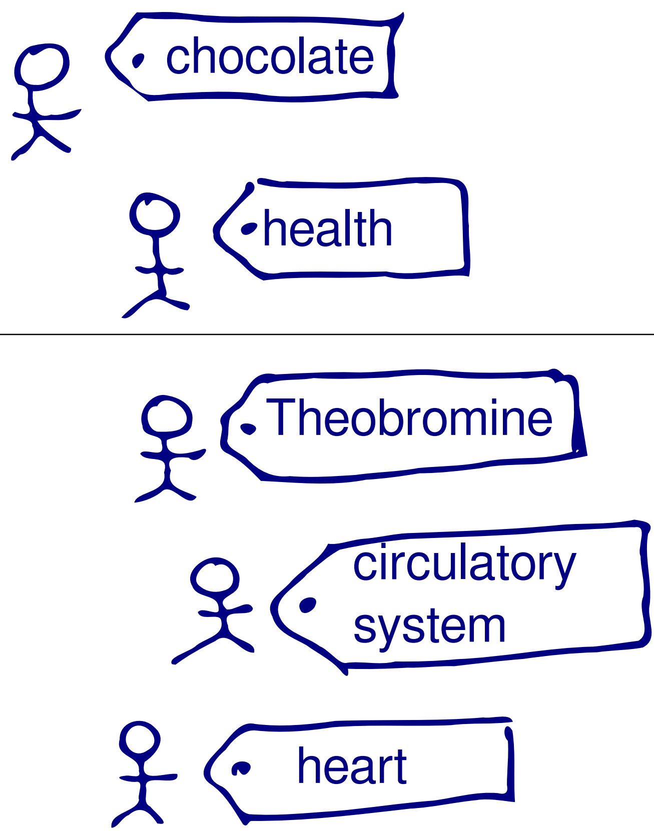


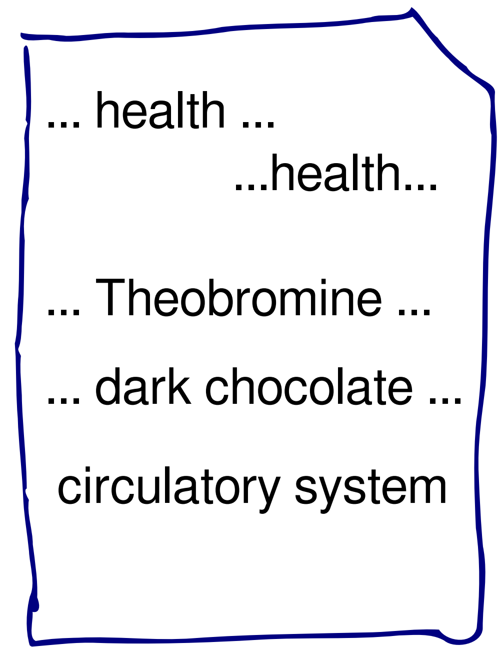


Document relevant?

Matching Entities in Documents by Name

Q: dark chocolate
health benefits

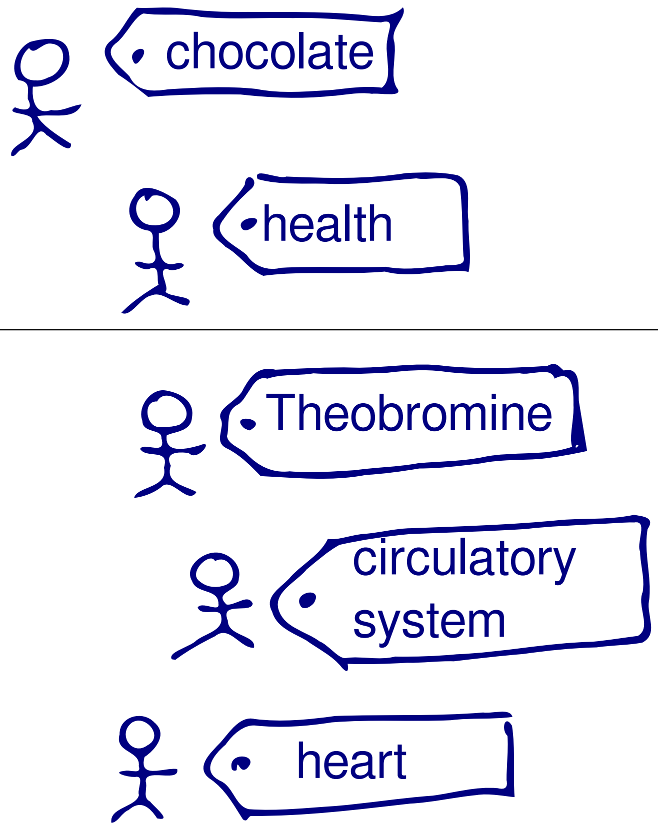


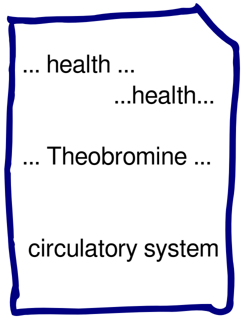


Document relevant?

Matching Entities in Documents by Name

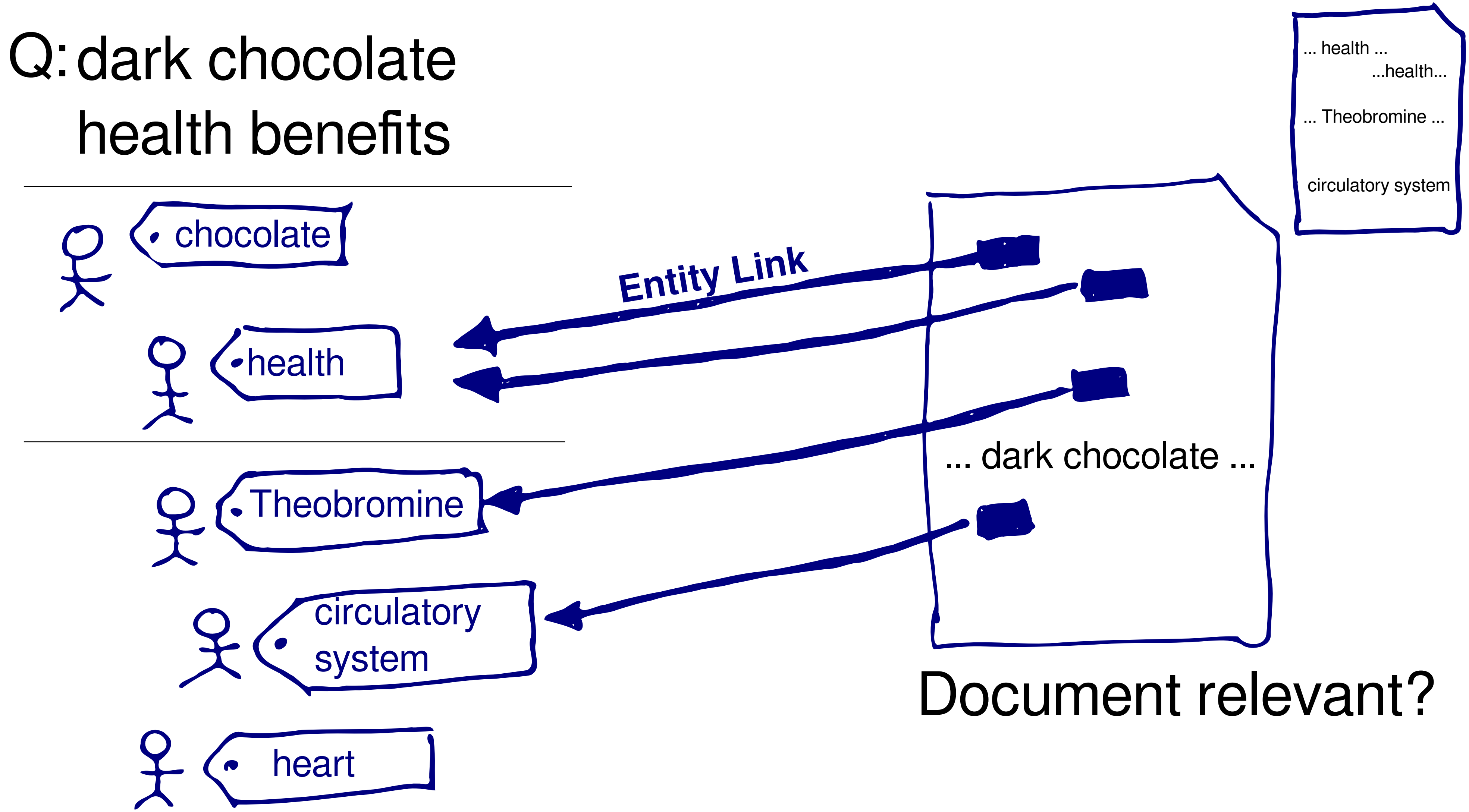
Q: dark chocolate health benefits





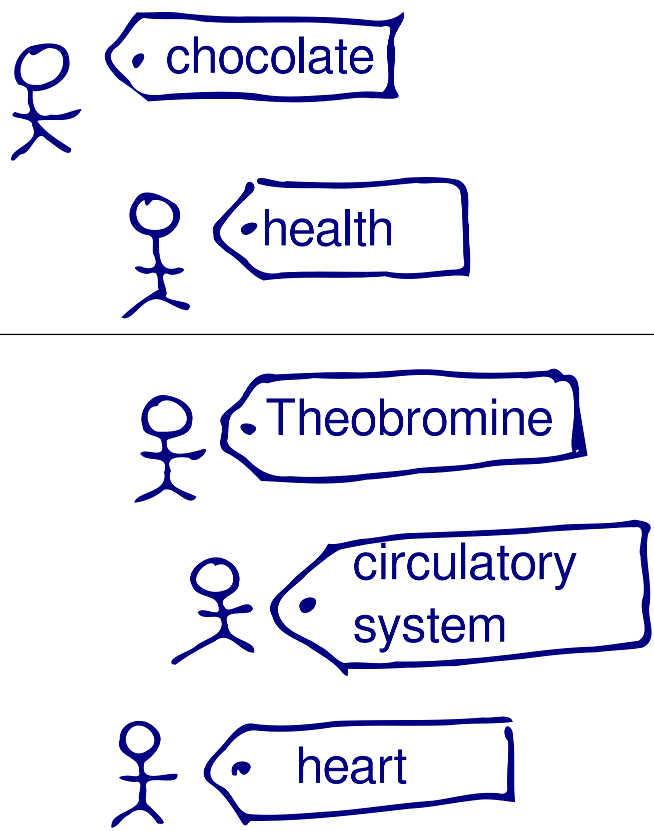
Document relevant?

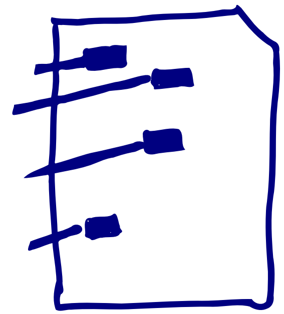
Matching Entities in Documents by Entity Links

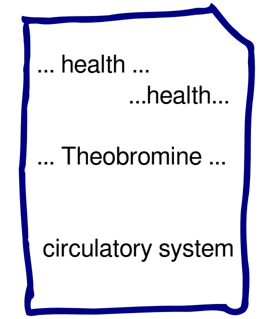


Matching Entities in Documents by Entity Links

Q: dark chocolate
health benefits



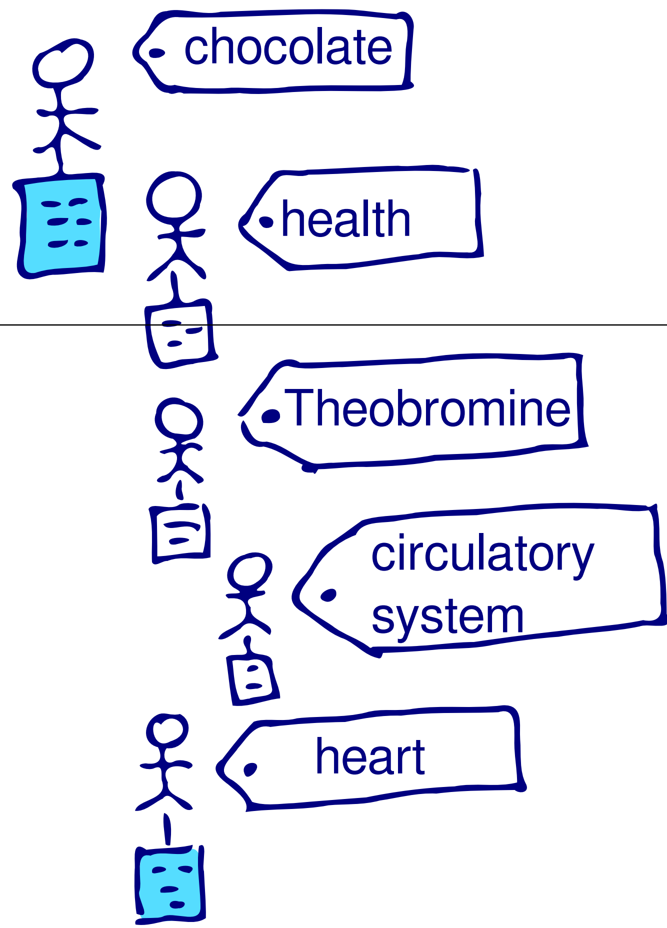


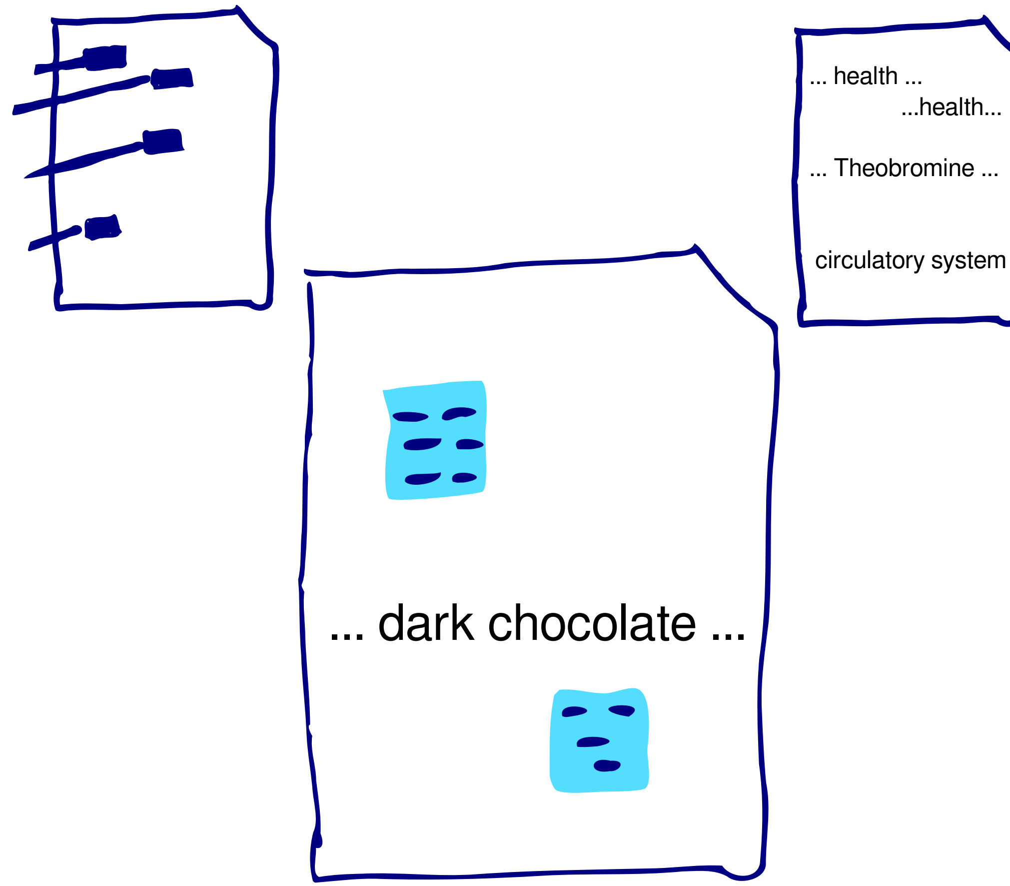


Document relevant?

Matching Entities in Documents by Article Terms

Q: dark chocolate
health benefits

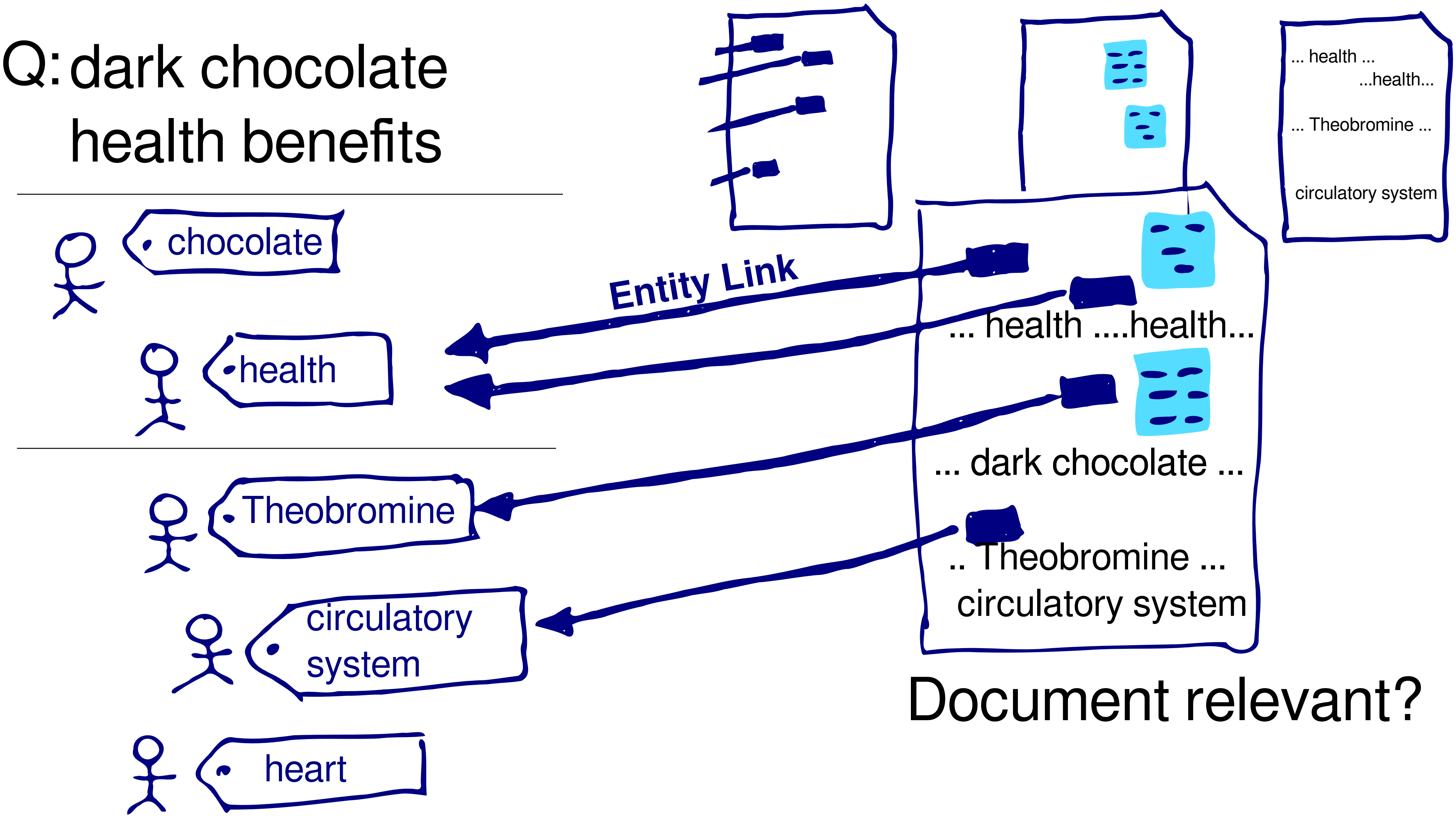




Document relevant?

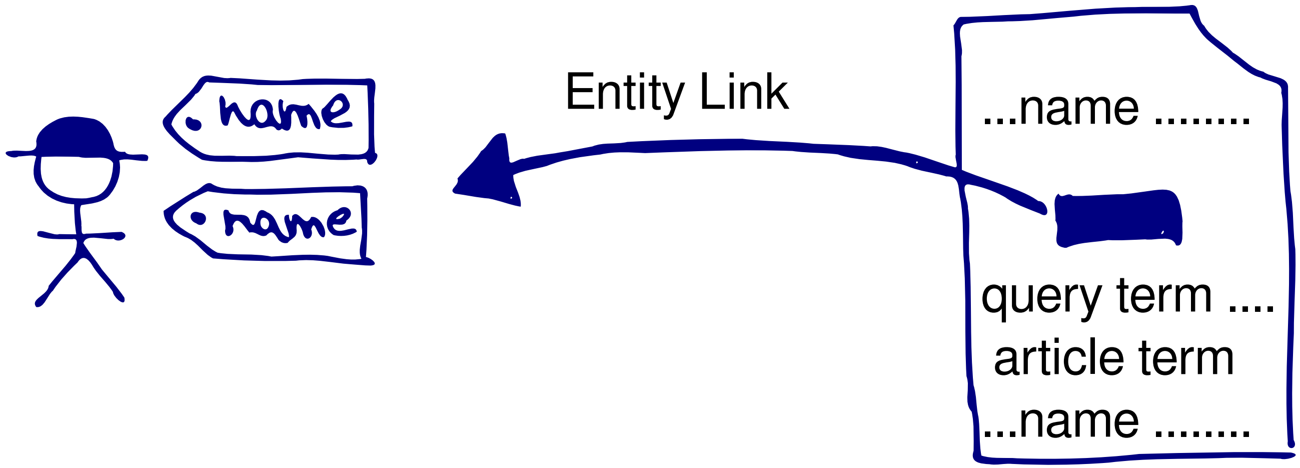
Combine All Names, Links, Terms

Q: dark chocolate health benefits



Document relevant?

Using Entities as a Vocabulary of Concepts

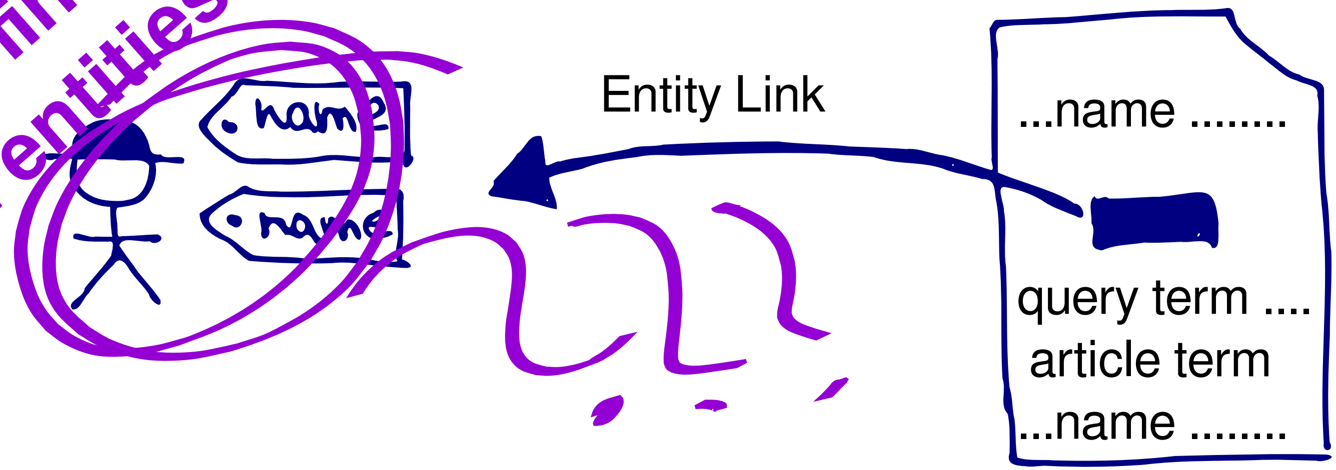


$$
\begin{aligned}
score(\square) = \quad & \lambda_1 \text{query terms} + \\
& \lambda_2 \text{names} + \\
& \lambda_3 \text{entity links} + \\
& \lambda_4 \text{article terms} + \ldots
\end{aligned}
$$

use your favorite retrieval model here!

Using Entities as a Vocabulary of Concepts

How to find relevant entities?

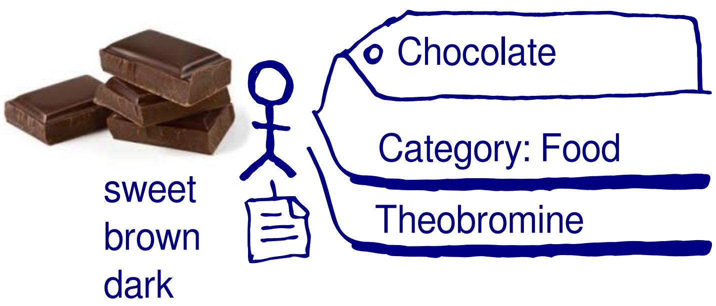


$$score(\square) = \lambda_1 \text{query terms} + \lambda_2 \text{names} + \lambda_3 \text{entity links} + \lambda_4 \text{article terms} + \ldots$$

use your favorite retrieval model here!

Query Entities through Entity Linking

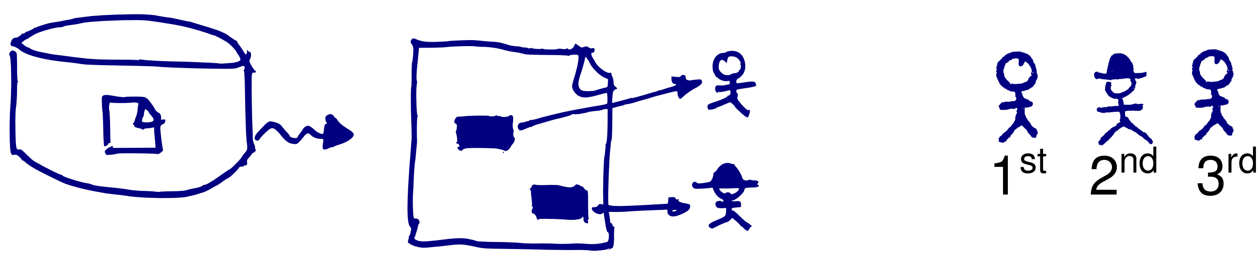
Query: dark chocolate health benefits



Latent Entities through Pseudo-Relev. Feedback

1. Retrieve documents with a query
2. Entity link documents
3. Derive distribution over  (bag of entities)
(see pseudo relevance feedback / RM3)

[Dalton et al 14, Liu & Fang 15]

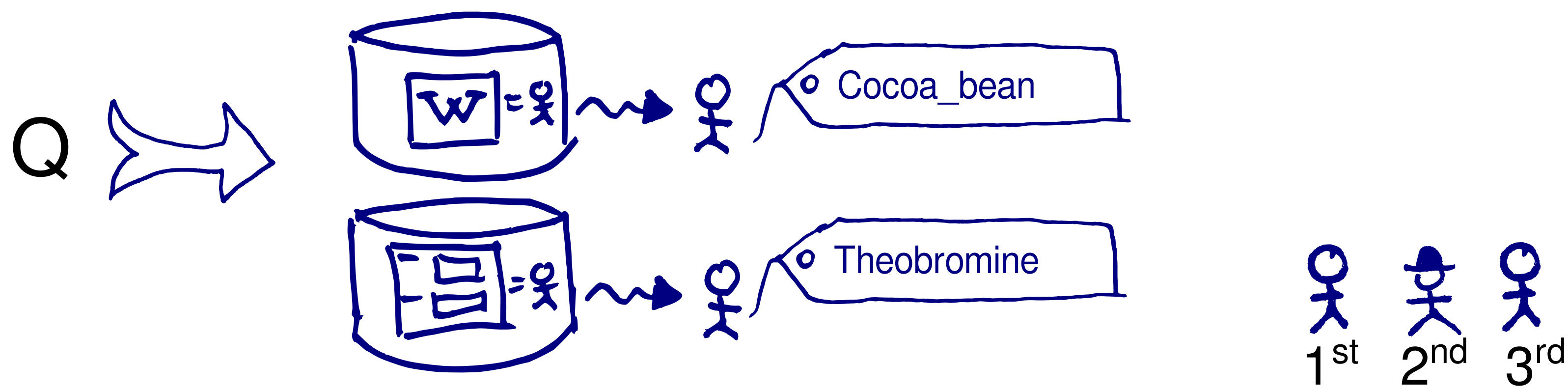


Latent Entities through Retrieval

Index Wikipedia pages or attribute sets of entities

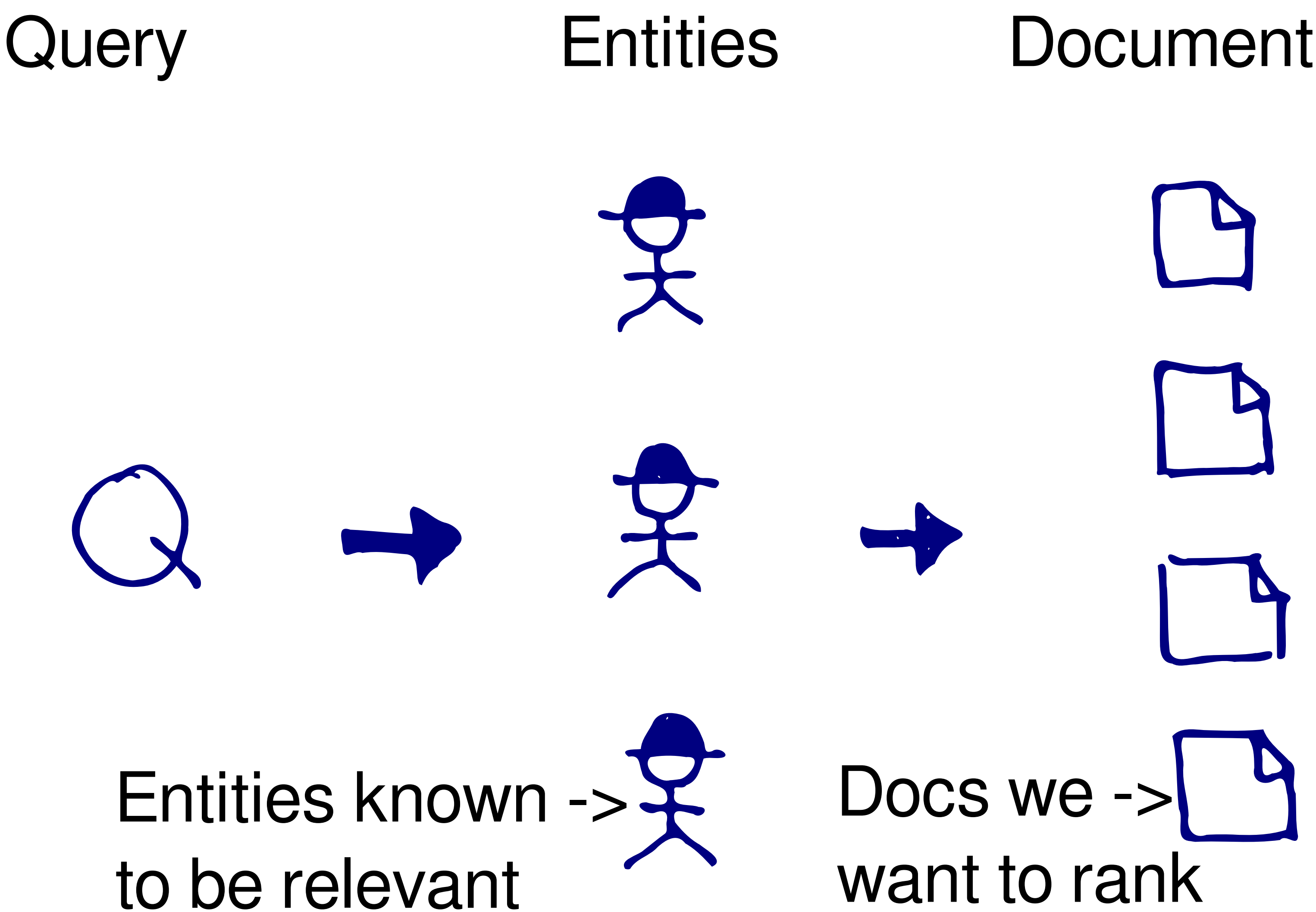
Retrieve entities from knowledge base
to obtain ranking of entities (with score)

[Pound et al 10, Niklaev et al 16, Balog 18]



Document Retrieval with Entities

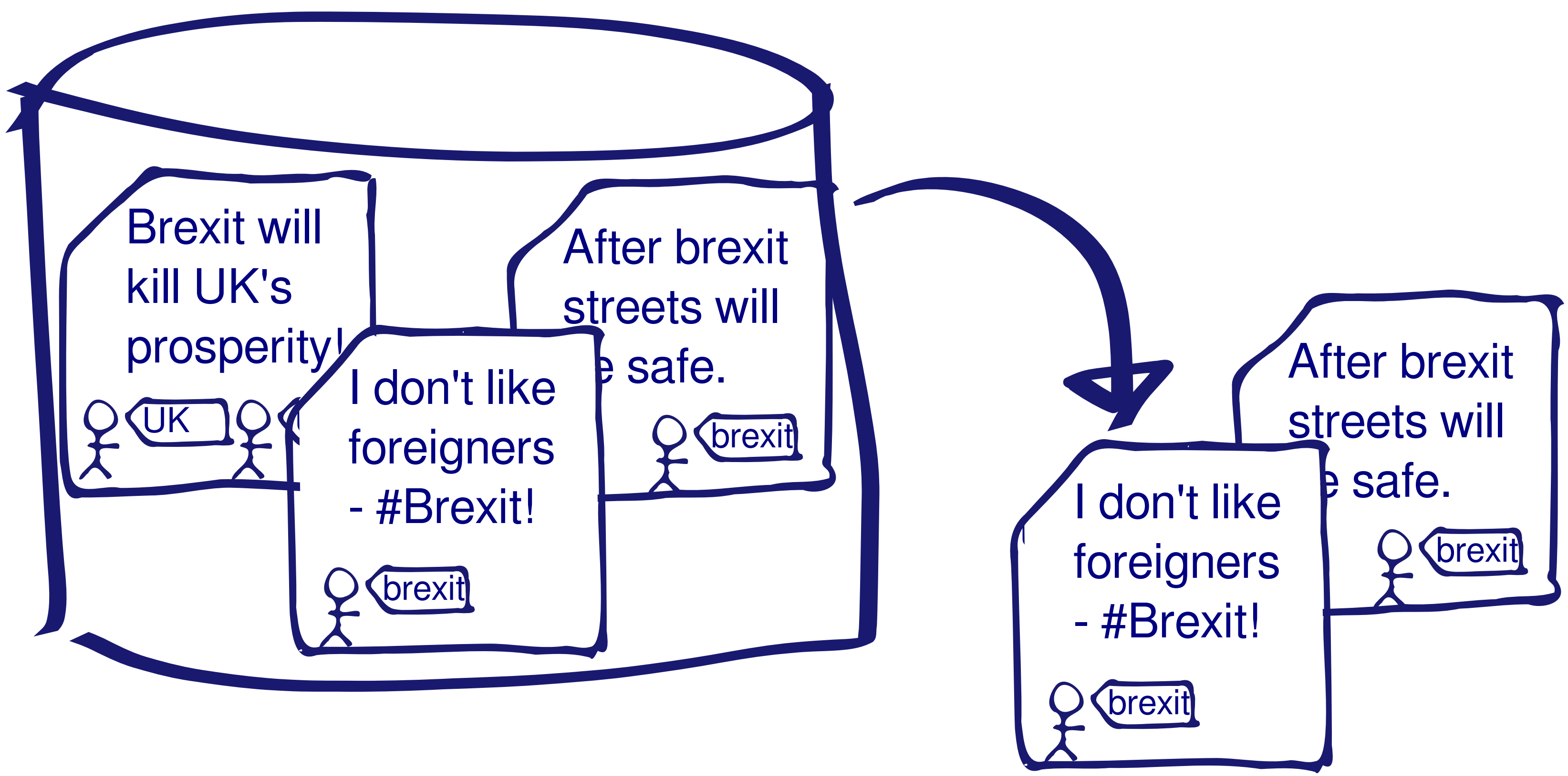
[Dalton et al 14]



Search Index with Entities

Query = "text:safe entity:brexit

entity: entity1 entity2 entity3 ...



Search Index (and Information Retrieval) Toolkits

IR:
- Lucene (Java) / PyLucene

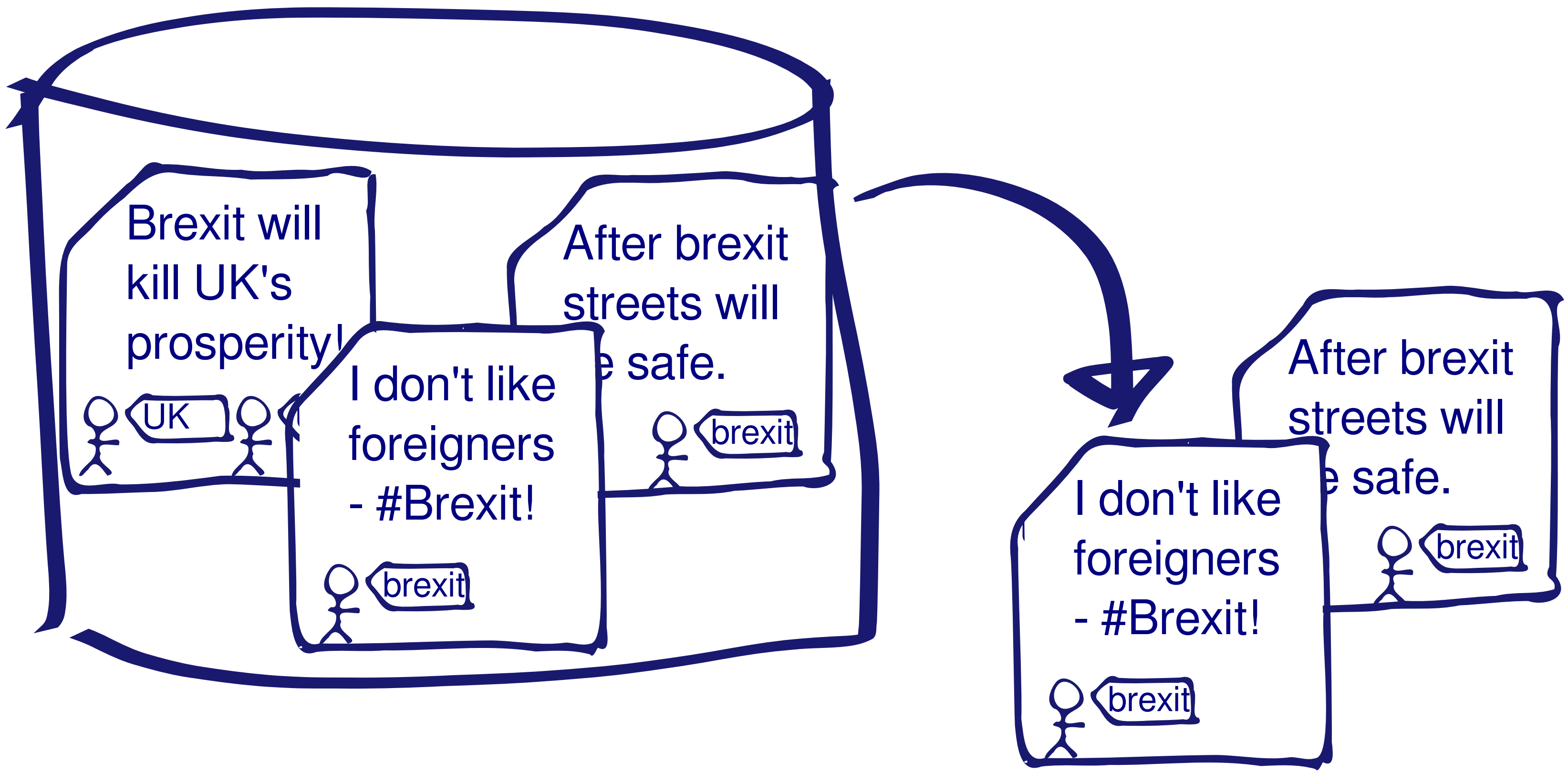
Combining different retrievals: Learning 2 Rank
- Ranking SVM, RankLib

Entity Retrieval: NordLys

Utilizing Knowledge Graphs for Information Retrieval:
- my tutorial: github.com/laura-dietz/tutorial-utilizing-kg
- "KG4IR" Workshop at SIGIR Conference
- Upcoming Special Issue

Search Index with Entities

Query = "text:safe entity:brexit
entity: entity1 entity2 entity3 ...
text: name1 name2 name3 ...
text: word1 word2 word3 ..."



Search Index (and Information Retrieval) Issues

Issue 1:

You need to guess a topic to look for.

Issue 2:

You still need to refine the results.

Citations

Topic Models: Blei & Lafferty. "Topic models." Text Mining. Chapman and Hall, 2009.
Word Embeddings: Levy & Goldberg. "Dependency-based word embeddings." ACL 2014.
Entity Linking: Ji & Grishman, "Knowledge base population: Successful approaches and challenges." NAACL-HLT, 2011.
Aspects: Nanni, Ponzetto, Dietz, "Entity-aspect linking: providing fine-grained semantics of entities i n context", JCDL 2018.
Information Retrieval:
Learning to Rank: Liu et al, "Learning to rank for information retrieval", FnTIR, 2009.
Entity Retrieval:
Pound, Mika, Zaragoza. "Ad-hoc object retrieval in the web of data", WWW, 2010.
Nikolaev, Kotov, Zhiltsov. "Parameterized fielded term dependence models for ad-hoc entity retrieval from knowledge graph", SIGIR 2016.
Balog. "Entity-Oriented Search (The Information Retrieval Series)", 2018, Springer.
IR with Entities:
Dietz, Kotov, Meij, "Utilizing Knowledge Graphs for Text-Centric Information Retrieval.", SIGIR 2018.
Dalton, Dietz, Allan, "Entity query feature expansion using knowledge base links", SIGIR 2014.
Liu & Fang, "Latent entity space: a novel retrieval approach for entity-bearing queries.", IRJ, 2015.
Xiong, Callan, Liu, "Word-Entity Duet Representations for Document Ranking", SIGIR, 2017.

Conclusions

Different techniques to inspect your documents.
- topic models
- word embeddings
- text classification
- entity linking
- entity aspects
- search index and retrieval (with entities)

There is no fool-proof method.
Make sure the tools are doing what you need!