Word Embeddings - Semantics: What is in my Documents?

Laura Dietz
University of New Hampshire
dietz@cs.unh.edu
The Problem
The Solution
Mr. Smith lives in the U.S.A. and reads crime novels.
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
Different techniques to inspect your documents.
- topic models
- word embeddings
- text classification
- entity linking
- entity aspects
- search index and retrieval (with entities)
- Same words are likely about the same topic.
- Words in the same document are likely about the same topic.
**Topic Models**

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

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

[Blei et al 09]
Topic Model Exercise - Apply rules to find topics!

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.

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
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....
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:
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.

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.
politicians read crime novels

people 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    novel    news    legal

\[
\begin{align*}
\text{crime} & : (3, 0.6, 2.6, -7) \\
\text{novel} & : (3, 0.5, 2.5, -7) \\
\text{news} & : (2, 0.5, 2.6, -6) \\
\text{legal} & : (1, 0.2, 1.5, -1)
\end{align*}
\]
"Read" can have the same distance to both words, without these words being similar to each other.

\[ \text{cosine similarity} = 0.7 \]
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)
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

people read crime novels

she likes crime

she wrote novels

polit. read crime novels

politics

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

polit. read reports

polit. read news

Laura Dietz dietz@cs.unh.edu - Summer School Series on Methods for Computational Social Science 2018
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]

many links between entities

similarity
names + context

popularity

Laura Dietz dietz@cs.unh.edu - Summer School Series on Methods for Computational Social Science 2018
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

Chocolate

Category: Food
Link to Theobromine

sweet brown dark

USA
Other resources define semi-structured entities

- **Chocolate**
  - has-name: Chocolate
  - of-category: Food
  - has-compound: Theobromine
  - description: sweet, dark brown food

- USA
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)
Harvested from sections of the entity's Wiki article.

Refine entity links with aspects that match context.

[Nanni et al 18]
Twitter classification into different aspects of Brexit.

- Brexit will kill UK's prosperity!
- I don't like foreigners - #Brexit!
- After brexit streets will be safe.

Economy: Brexit
Immigration: Brexit
Search Index

Create a search index with documents.

Create different descriptions of your topic.
Use description = query to retrieve top 10.

Brexit will kill UK's prosperity!
After brexit streets will be safe.
I don't like foreigners - #Brexit!
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)
Documents can have fields:

Query = "text:safe entity:brexit"
Relevant Entities

<table>
<thead>
<tr>
<th>Query</th>
<th>EU UK relations</th>
<th>dark chocolate health benefits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Query entities</td>
<td>EU, UK</td>
<td>chocolate, health</td>
</tr>
<tr>
<td>Latent entities</td>
<td>Brexit, Theresa May</td>
<td>Theobromine, circulatory system, heart</td>
</tr>
</tbody>
</table>

Named Entities

Concepts
Matching Entities in Documents

Q: dark chocolate health benefits

- chocolate
- health
- Theobromine
- circulatory system
- heart

Document relevant?
Q: dark chocolate health benefits

Document relevant?
Q: dark chocolate health benefits

Document relevant?
Matching Entities in Documents by Entity Links

Q: dark chocolate health benefits

Entity Link

Document relevant?
Matching Entities in Documents by Entity Links

Q: dark chocolate health benefits

Document relevant?
Q: dark chocolate
health benefits

Document relevant?
Q: dark chocolate health benefits

Entity Link

... dark chocolate ...

... Theobromine ...
circulatory system

Document relevant?
Using Entities as a Vocabulary of Concepts

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

use your favorite retrieval model here!
Using Entities as a Vocabulary of Concepts

How to find relevant entities?

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

use your favorite retrieval model here!
Query: dark chocolate health benefits

Category: Food

Theobromine

Chocolate

sweet brown dark
1. Retrieve documents with a query
2. Entity link documents
3. Derive distribution over \( X \) (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

Entities known -> to be relevant

Docs we want to rank

[Dalton et al 14]
Search Index with Entities

Query = "text:safe entity:brexit
entity: entity1 entity2 entity3 ...

Brexit will kill UK's prosperity!
I don't like foreigners - #Brexit!

After brexit streets will be safe.
I don't like foreigners - #Brexit!

After brexit streets will be safe.
Search Index (and Information Retrieval) Toolkits

IR:
- Lucene (Java) / PyLucene
- 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
tenentiy: entity1 entity2 entity3 ...
text: name1 name2 name3 ...
text: word1 word2 word3 ..."

Brexit will kill UK's prosperity!

I don't like foreigners - #Brexit!

After brexit streets will be safe.

I don't like foreigners - #Brexit!
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

Information Retrieval:
Entity Retrieval:
IR with Entities:
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!