Solving the Cold Start NLP Problem

Solving the *what*?

- **Natural Language Processing (NLP)** = Doing data science with text

- **Cold Start** = You want to do text classification…

Email message	Class
Buy these pills	Spam
Win cash prizes	Spam
Dear Mr. Atreides, please find attached…	Not Spam

https://jalammar.github.io/illustrated-bert/

- **Cold Start** = … but you don’t have any labeled data.

Email message	Class
Buy these pills	Spam
Win cash prizes	Spam
Dear Mr. Atreides, please find attached…	Not Spam

https://jalammar.github.io/illustrated-bert/

Case 1: You Know What You’re Looking For

- e.g. M & A in business news

Exclusive: UTC set to win EU approval for $23 billion Rockwell Collins deal

The deal, announced in September last year, would create a new player in the top echelon of suppliers to Boeing, Airbus, Bombardier, and other plane makers

Heintzelman & Johnson (2019), NLP from Scratch: Solving the cold start problem for natural language processing

- e.g. association between chemicals and diseases

NCBI Resources How To

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

PubMed

Advanced

Format: Abstract

Send to

Am J Dis Child. 1981 Oct;135(10):941-3.

Tricuspid valve regurgitation and lithium carbonate toxicity in a newborn infant.

Arnon RG, Marin-Garcia J, Peeden JN.

Abstract

A newborn with massive tricuspid regurgitation, atrial flutter, congestive heart failure, and a high serum lithium level is described. This is the first patient to initially manifest tricuspid regurgitation and atrial flutter, and the 11th described patient with cardiac disease among infants exposed to lithium compounds in the first trimester of pregnancy. Sixty-three percent of these infants had tricuspid valve involvement. Lithium carbonate may be a factor in the increasing incidence of congenital heart disease when taken during early pregnancy. It also causes neurologic depression, cyanosis, and cardiac arrhythmia when consumed prior to delivery.

PMID: 6794356 DOI: 10.1001/archpedi.1981.02130340047016

[Indexed for MEDLINE]

Ratner et al. (2017), Snorkel: Rapid Training Data Creation with Weak Supervision

- e.g. spam detection in YouTube comments

Alberto, Lochter & Almeida (2015), TubeSpam: Comment Spam Filtering on YouTube

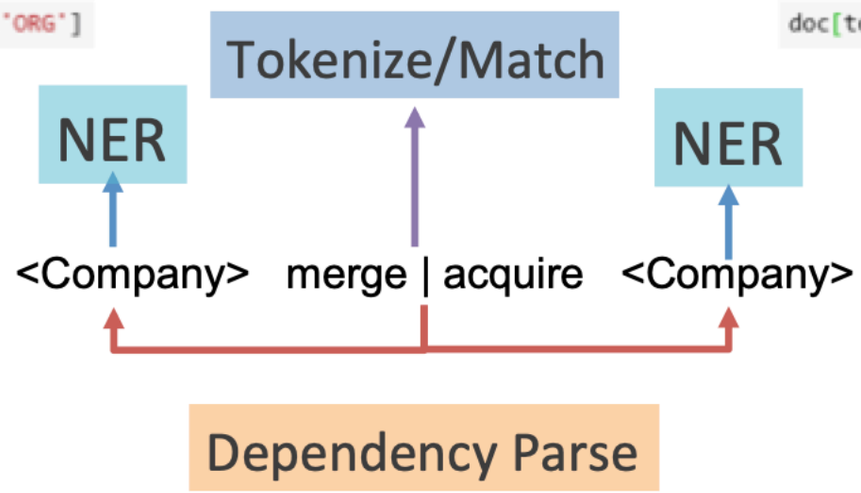
Idea: use rules/heuristics to create labels

- e.g. M & A in business news

```
matcher.add("MandA",None,
            [{'LEMMA': 'acquire'}],
            [{'LEMMA': 'merge'}],
            [{'LEMMA': 'acquisition'}])
```

```
doc[token].ent_type_ in ['PERSON','ORG']
```

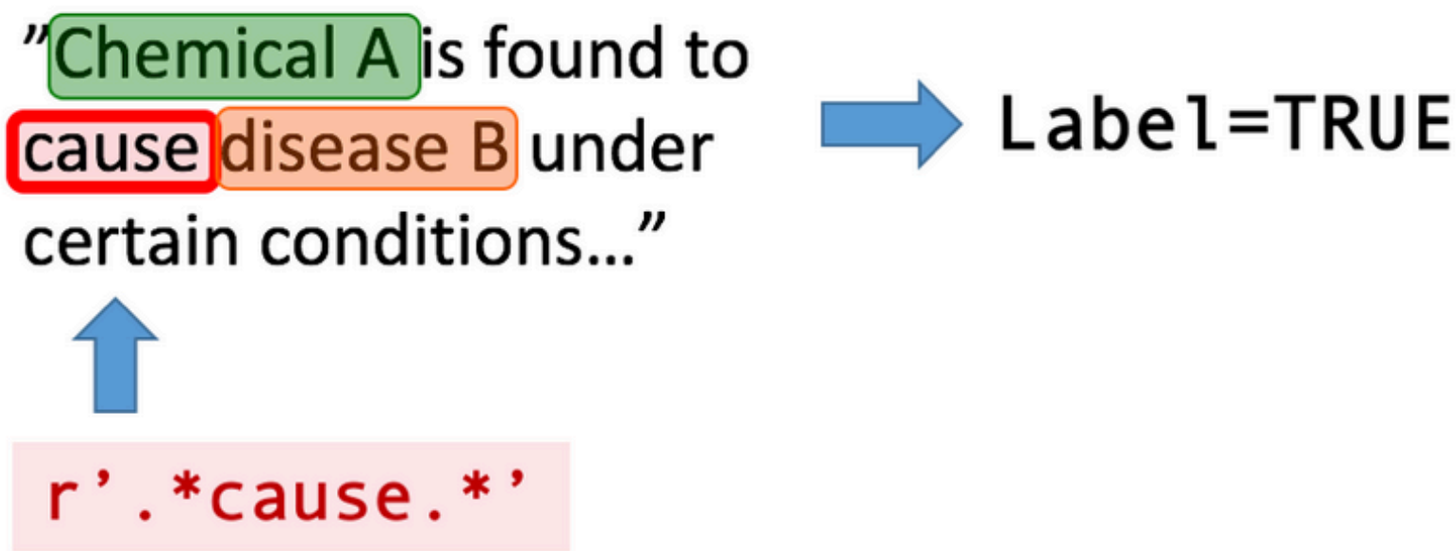
```
doc[token].ent_type_ in ['PERSON','ORG']
```




```
for word in MorA.rights:
    if(word.ent_type_) in['PERSON','ORG']:
        dependentCompany = Word
```

Heintzelman & Johnson (2019), NLP from Scratch: Solving the cold start problem for natural language processing

- e.g. association between chemicals and diseases



https://hazyresearch.github.io/snorkel/blog/weak_supervision.html

- snorkel by Hazy Research
- Write **labeling functions** that formalize rules/heuristics
- “weak supervision”
- Integrates seamlessly with pandas, spaCy, scikit-learn, keras, etc.

Ratner et al. (2017), Snorkel: Rapid Training Data Creation with Weak Supervision

- e.g. tag comments as spam if they tell you to ‘check’ or ‘check out’ something

AUTHOR	DATE	TEXT
zhichao wang	2013-11-29T02:13:56	i think about 100 millions of the views come f...
Santeri Saariokari	2014-09-03T16:32:59	Hey guys go to check my video name "growtopia ...
BeBe Burkey	2013-11-28T16:30:13	and u should.d check my channel and tell me wh...
Cony	2013-11-28T16:01:47	You should check my channel for Funny VIDEOS!!

https://www.snorkel.org/use-cases/01-spam-tutorial

- e.g. tag comments as spam if they tell you to ‘check’ or ‘check out’ something

```python
from snorkel.labeling import labeling_function



@labeling_function()
def check(x):
    return SPAM if "check" in x.text.lower() else ABSTAIN



@labeling_function()
def check_out(x):
    return SPAM if "check out" in x.text.lower() else ABSTAIN
```

https://www.snorkel.org/use-cases/01-spam-tutorial

- labeling functions return numpy arrays of {1, 0, -1}

```
array([[-1, -1],
       [-1, -1],
       [-1,  1],
       ...,
       [ 1,  1],
       [-1,  1],
       [ 1,  1]])
```

https://www.snorkel.org/use-cases/01-spam-tutorial

- e.g. tag comments as not spam if they mention people and are short

```
@labeling_function(pre=[spacy])
def has_person(x):
    """Ham comments mention specific people and are short."""
    if len(x.doc) < 20 and any([ent.label_ == "PERSON" for ent in x.doc.ents]):
        return HAM
    else:
        return ABSTAIN
```

https://www.snorkel.org/use-cases/01-spam-tutorial

- Write a bunch of labeling functions…

```
lfs = [
    keyword_my,
    keyword_subscribe,
    keyword_link,
    keyword_please,
    keyword_song,
    regex_check_out,
    short_comment,
    has_person_nlp,
    textblob_polarity,
    textblob_subjectivity,
]
```

https://www.snorkel.org/use-cases/01-spam-tutorial

- … and combine their outputs to produce class probabilities.

```
from snorkel.labeling import LabelModel

label_model = LabelModel(cardinality=2, verbose=True)
label_model.fit(L_train=L_train, n_epochs=1000, lr=0.001, log_freq=100, seed=123)
```

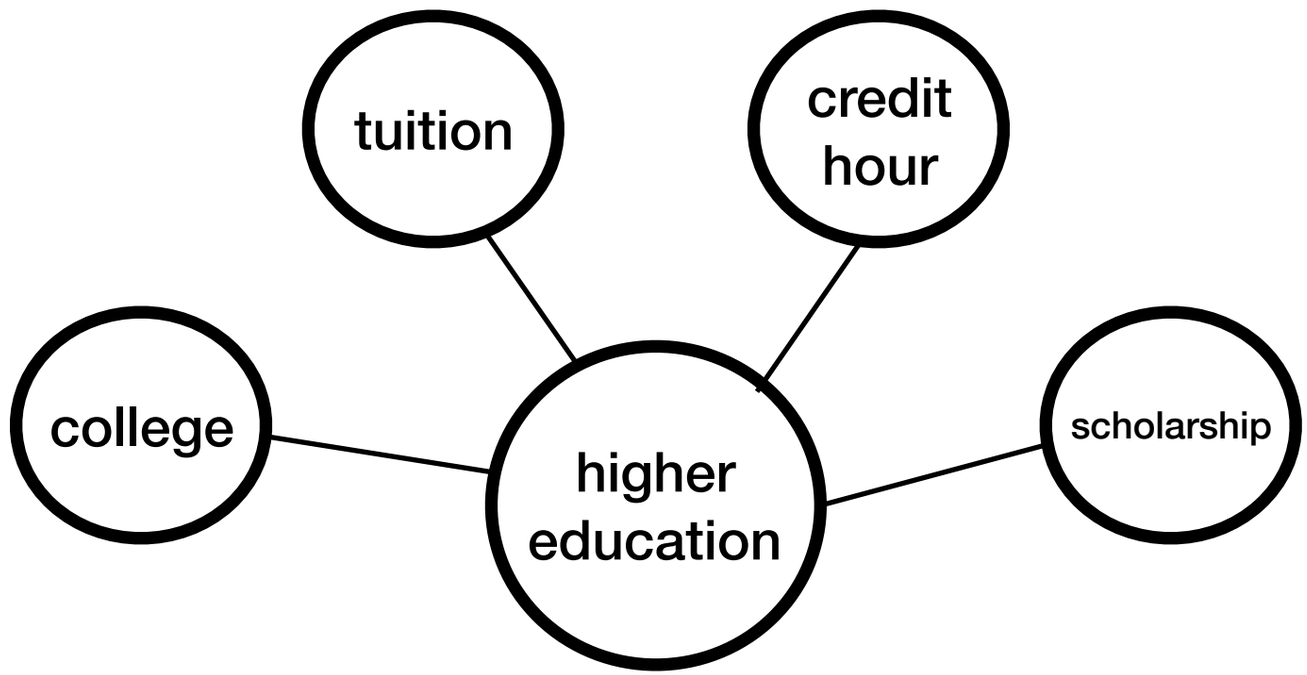
https://www.snorkel.org/use-cases/01-spam-tutorial

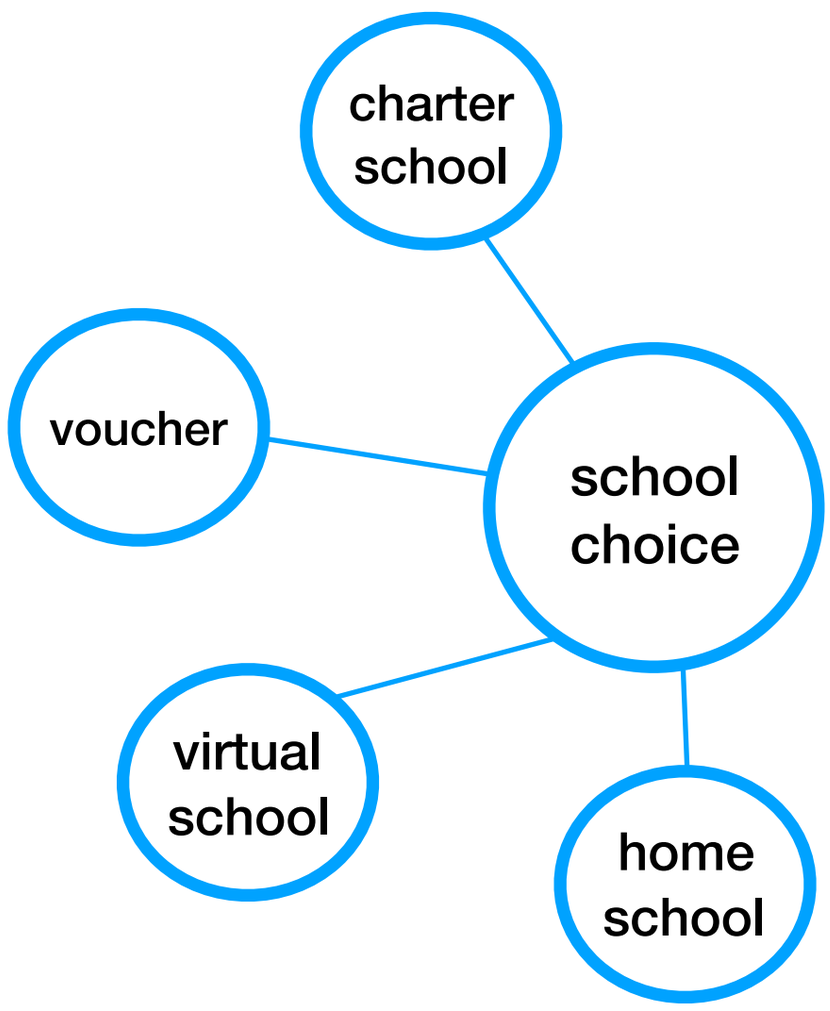
Case 2: You Don’t Know What You’re Looking For

- e.g. The Tennessee Educator Survey
- Yearly survey administered by the Department to teachers and administrators in TN public schools
- Open response to a question of the form “Is there anything you would like to communicate to the Department?”

- e.g. Search and Recommendation for Legislation
- Break down large policy areas (e.g. education) into smaller subtopics (higher education, school finance, school choice, etc.)

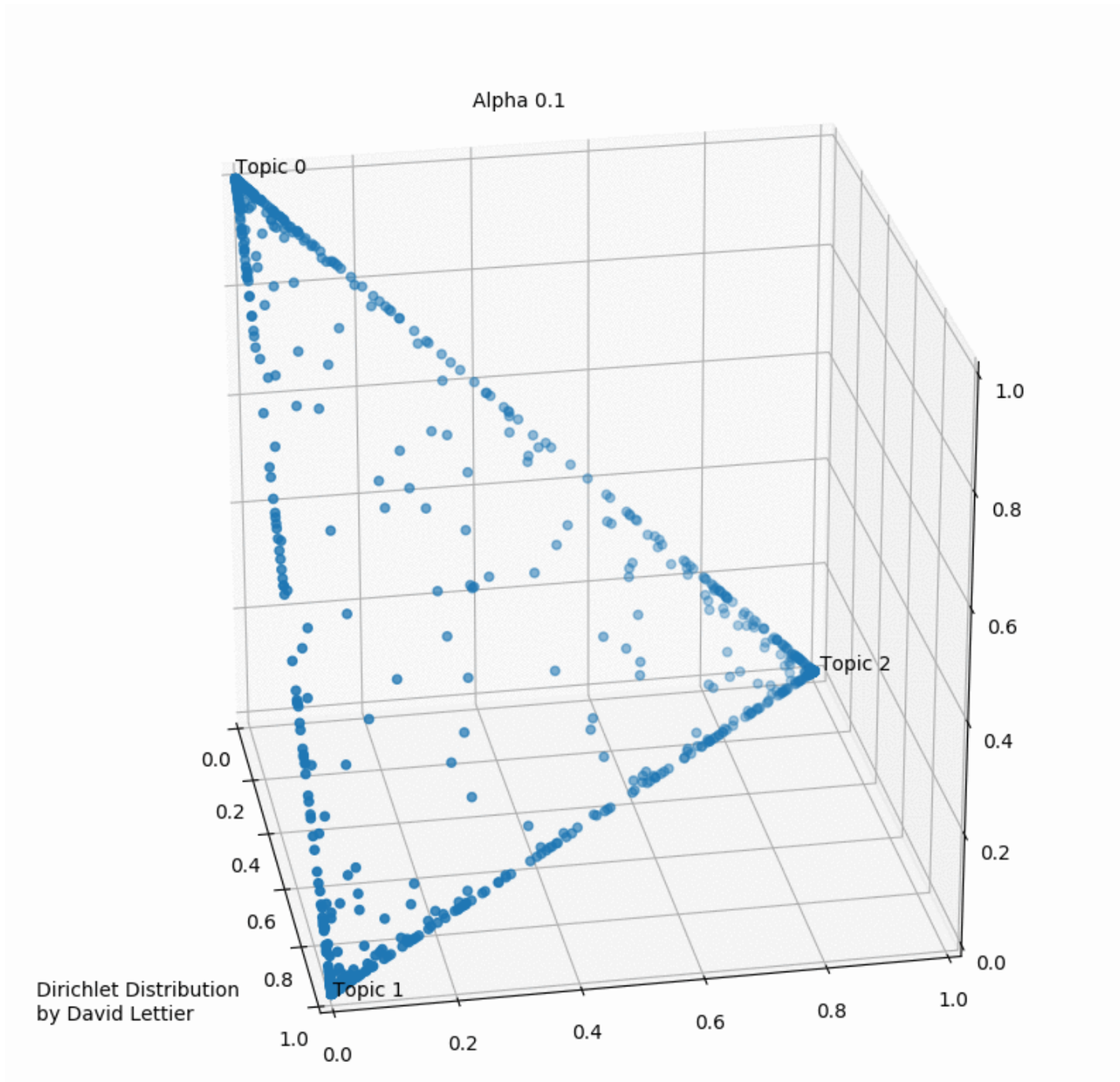
Idea: partition documents into groups based on presence of similar words


tuition
credit hour
college
higher education
scholarship


charter school
voucher
school choice
virtual school
home school

- Use Latent Dirichlet Allocation as a topic model

- Dirichlet = a probability distribution with one parameter α



https://medium.com/@lettier/how-does-lda-work-ill-explain-using-emoji-108abf40fa7d

- Latent = we assume that documents have an underlying distribution across topics

- Latent Dirichlet Allocation = assign documents across topics according to a Dirichlet distribution

- Start with a term frequency matrix:

	Word	Word	Word	…	Word
Doc	1	0	2		0
Doc					
Doc					
.					
.					
.					

- LDA represents each document as a combination of topics

	Topic	Topic	Topic	…	Topic
Doc	0.5	0	0	…	0.2
Doc					
Doc					
.					
.					
.					
.					
.					
.					

- LDA represents each document as a combination of topics and each topic as a combination of words

	Topic	Topic	Topic	…	Topic
Doc	0.5	0	0	…	0.2
Doc					
Doc					
.					
.					
.					
.					
.					
.					
.					

	Word	Word	.	.	.	Word
Topic						
Topic						
.						
Topic						

- LDA has three parameters of particular interest:
  - Number of topics
  - α: ~ how many topics a document can fall under
  - η: ~ how many words are associated with each document

- cf. clustering
- cf. dimensionality reduction

- Run a topic model via gensim

```
from gensim.corpora import Dictionary
from gensim.models import CoherenceModel
from gensim.models.ldamodel import LdaModel

n_topics = 20

dictionary = Dictionary(tokens)
corpus = [dictionary.doc2bow(b) for b in tokens]

lda_model = LdaModel(
    corpus=corpus,
    id2word=dictionary,
    num_topics=n_topics,
    passes=50,
    alpha='auto',
    eta='auto',
    random_state=79
)
```

- Evaluate a topic model with coherence

```
c = CoherenceModel(
    model=lda_model,
    texts=tokens,
    dictionary=dictionary,
    coherence='c_v'
)

print('Model Coherence:', c.get_coherence())
```

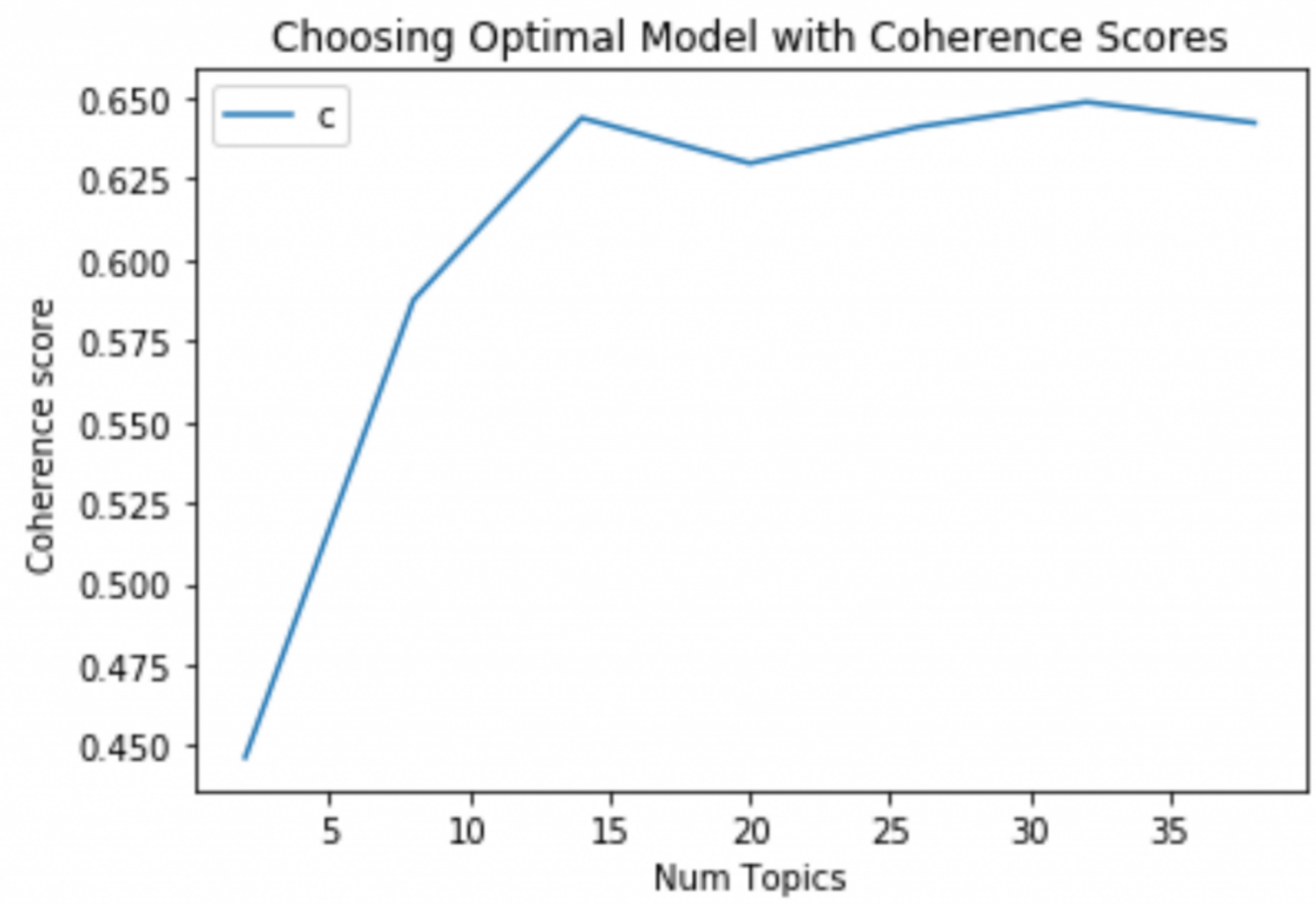
```
Model Coherence: 0.46699194637344477
```

- Evaluate a topic model with coherence

```
list(zip(range(20), c.get_coherence_per_topic()))
```

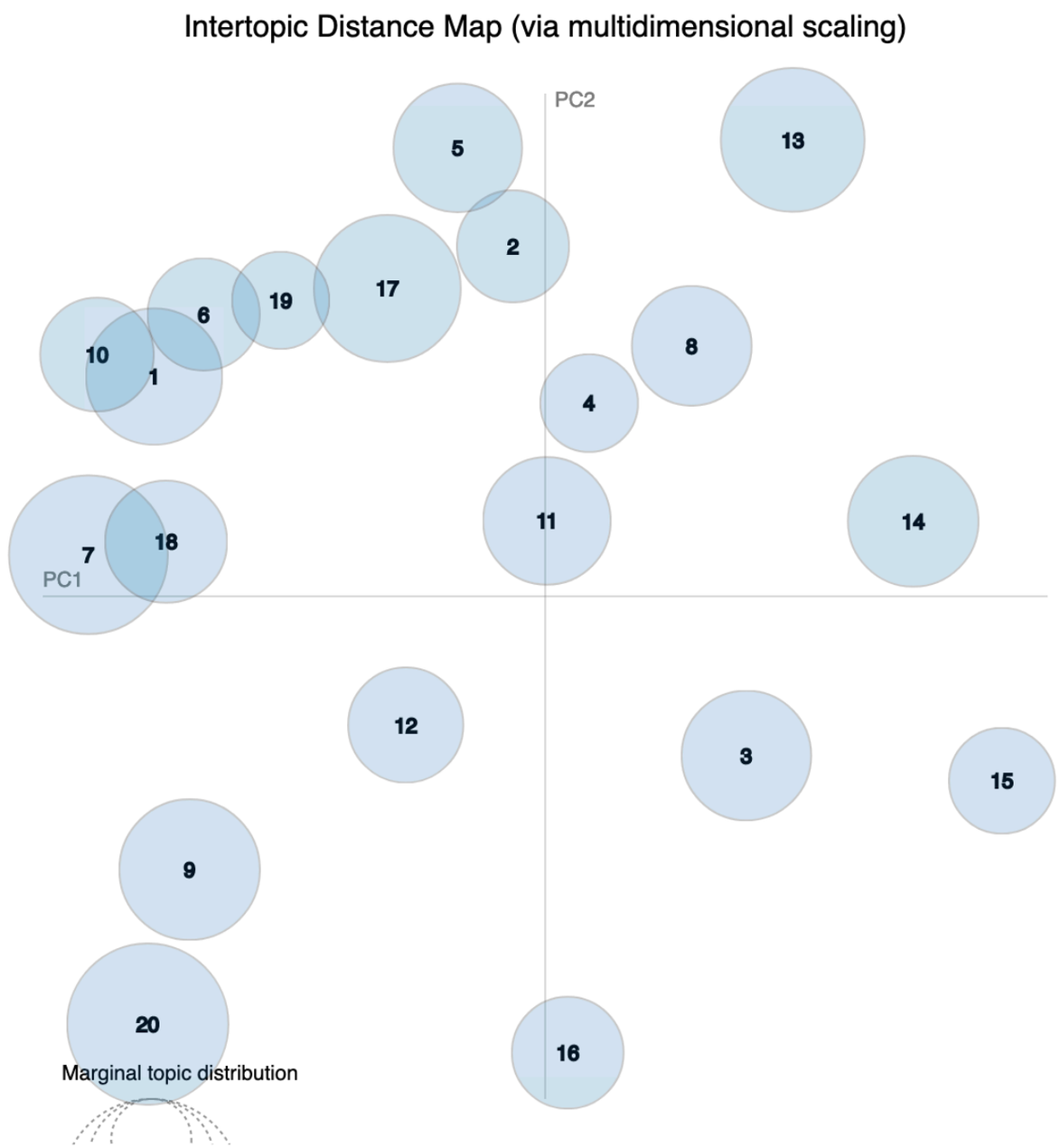
```
[(0, 0.40571702407814403),
 (1, 0.4698652721789071),
 (2, 0.41788896434340794),
 (3, 0.3548510763436543),
 (4, 0.38559117131862397),
 (5, 0.4244073421607143),
 (6, 0.4773018370456886),
 (7, 0.7630543726732141),
 (8, 0.41830949182171856),
 (9, 0.42243870761442504),
 (10, 0.5459883454752834),
 (11, 0.407525949914728),
 (12, 0.599162965392366),
 (13, 0.3640695171574774),
 (14, 0.5412834408395492),
 (15, 0.5022858720953146),
 (16, 0.4180497977979124),
 (17, 0.5250304088781431),
 (18, 0.3065546105589426),
 (19, 0.5904627597806802)]
```

- Evaluate number of topics with coherence



https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/

- Evaluate a topic model using LDAvis



- Interpret topics with keywords

```
lda_model.show_topic(7)
```

```
[('bond', 0.04953584),
 ('District', 0.041846886),
 ('note', 0.038327646),
 ('Board', 0.021514444),
 ('authorize', 0.021326357),
 ('Acts', 0.019330233),
 ('tax', 0.018274356),
 ('provide', 0.017869938),
 ('issue', 0.017210022),
 ('time', 0.01538261)]
```

- Interpret topics with keywords

	Coherence	Number of Bills	Keywords	Sample Title
7	0.763054	50	bond, District, note, Board, authorize, Acts, tax, provide, issue, time	School Districts, Special - As introduced, pursuant to the request of the Franklin special school district of Williamson County, permits the district to issue bonds or notes in an amount not to exceed $26.5 million and to issue bond anticipation notes.
12	0.599163	107	school, board, county, education, director, system, superintendent, elect, office, election	Education - As introduced, enacts the "Local School District Empowerment Act," which provides for reestablishment of elected office of school superintendent for county or city school systems upon two-thirds vote of county or city governing body and approval in an election on the question by the voters in 10 LEAs as a pilot program to allow the department to study the relevant procedures of reestablishing the office; provides for qualifications of candidates; adjusts duties of the local board of education in county or city school systems electing superintendents.
19	0.590463	275	student, scholarship, year, institution, semester, program, receive, HOPE, time, grant	Lottery, Scholarships and Programs - As introduced, sets awards from net lottery proceeds for certain postsecondary scholarships and grants at the amount the student initially received or at the amount awarded for initial recipients in the current semester of enrollment, whichever is greater.
10	0.545988	366	education, report, committee, representative, house, senate, department, year, study, commissioner	Education - As introduced, requires the director of the office of legislative budget analysis to provide the revised BEP funding formula to the speaker of the senate, the speaker of the house of representatives, and the education committees of the senate and the house of representatives, if the commissioner fails to provide the revised BEP funding formula for the ensuing fiscal year by January 1.
14	0.541283	85	member, board, term, serve, appoint, student, year, University, trustee, appointment	University of Tennessee - As introduced, reconstitutes the board of trustees of the University of Tennessee system.
17	0.525030	243	student, test, assessment, school, grade, score, year, education, state, examination	Education, Dept. of - As introduced, requires the department to release certain percentages of test questions and answers from the Tennessee comprehensive assessment program (TCAP) tests and end-of-course examinations to LEAs and public schools.

- Interpret topics with most representative documents

- Extract proportion that document falls into each topic

```
lda_model[corpus][0]
```

```
[(10, 0.9344231)]
```

```
lda_model[corpus][1]
```

```
[(10, 0.1524922), (13, 0.09571376), (14, 0.34877005), (18, 0.37520307)]
```

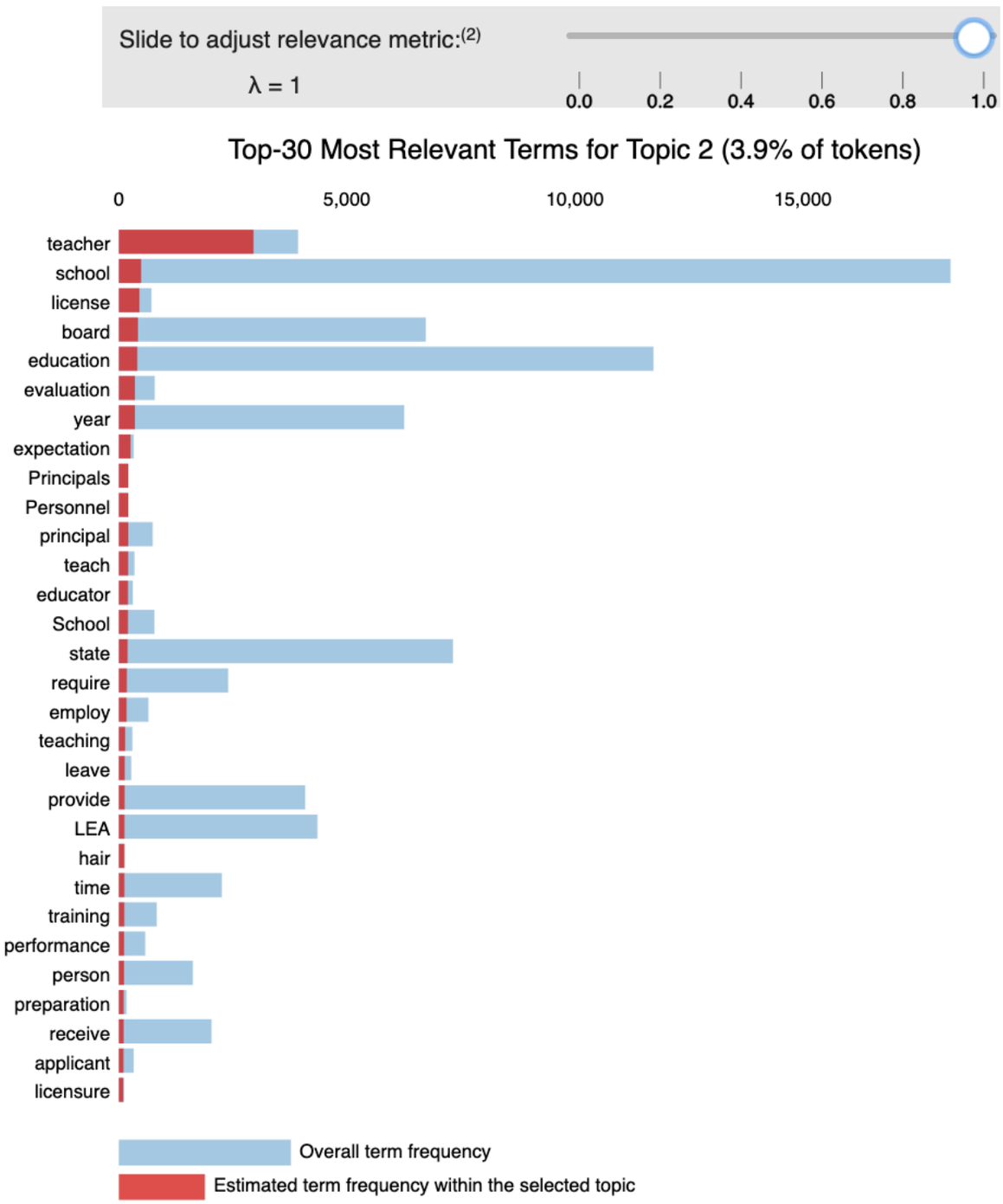
- Extract proportion that document falls into each topic

	session	bill_id	title	text	dominant_topic	max_perc	0	1	2	3	...	10	11	12	13	14	15	16
0	107	HB 1006	Education - As introduced, requires the commis...	An act to amend Tennessee Code Annotated, Titl...	10	0.934423	NaN	NaN	NaN	NaN	...	0.934423	NaN	NaN	NaN	NaN	NaN	NaN
1	107	HB 1027	Education, Higher - As introduced, requires th...	An act to amend Tennessee Code Annotated, Titl...	18	0.375270	NaN	NaN	NaN	NaN	...	0.152441	NaN	NaN	0.09553	0.348938	NaN	NaN

- Interpret topics with most representative documents

	title	text	max_perc
2578	Teachers, Principals and School Personnel - As enacted, revises compensation provisions and other provisions regarding substitute teachers.	An act to amend Tennessee Code Annotated, Section 49312, relative to retired teachers. Tennessee Code Annotated, Section 49312, is amended by deleting the language "certificate or permit" and substituting instead the word "license". Tennessee Code Annotated, Section 49312, is amended by deleting the subsection and substituting instead the following language: A substitute teacher who is a retired teacher is not required to continue to renew the teacher's license in order to work as a substitute teacher. The rate of compensation for a retired teacher without an active teaching license must not be less than the rate of compensation set by the LEA for a retired teacher with an active teaching license. This subsection only applies to retired teachers who retired after June 30, 2011.	0.984233
2222	Teachers, Principals and School Personnel - As enacted, revises compensation provisions and other provisions regarding substitute teachers.	An act to amend Tennessee Code Annotated, Section 49312, relative to retired teachers. Tennessee Code Annotated, Section 49312, is amended by deleting the language "certificate or permit" and substituting instead the word "license". Tennessee Code Annotated, Section 49312, is amended by deleting the subsection and substituting instead the following language: A substitute teacher who is a retired teacher is not required to continue to renew the teacher's license in order to work as a substitute teacher. The rate of compensation for a retired teacher without an active teaching license must not be less than the rate of compensation set by the LEA for a retired teacher with an active teaching license. This subsection only applies to retired teachers who retired after June 30, 2011.	0.984233
149	Teachers, Principals and School Personnel - As introduced, allows teachers evaluated as "meeting expectations" to be eligible for tenure on the same basis as those teachers evaluated as "above expectations" and "significantly above expectations."	An act to amend Tennessee Code Annotated, Title 49, Chapter 5, Part 5, relative to teacher eligibility for tenure. Tennessee Code Annotated, Section 49503, is amended by deleting subdivision in its entirety and substituting instead the following: Has received evaluations demonstrating an overall performance effectiveness level of "meets expectations," "above expectations" or "significantly above expectations" as provided in the evaluation guidelines adopted by the state board of education pursuant to § 49302, during the last two years of the probationary period; and Tennessee Code Annotated, Section 49504, is amended by inserting in subdivision the words and punctuation "meets expectations," between the words "effectiveness level of" and the words "above expectations".	0.984233

- Interpret topics with LDAvis



- Assign labels to documents where topics substantially fall under one topic

```
bill_topics['label'] = ''

bill_topics.loc[(bill_topics['dominant_topic'] == 1) &
                (bill_topics['max_perc'] > 0.4), 'label'] = 'Teachers; Employment'
```

Recap

- Use data programming approach to write labeling functions based on rules/heuristics
  - If you know what you are looking for
  - If there are relatively few classes
- Combine information across labeling functions to produce probabilistic class labels

- Use topic modeling to partition documents into groups based on presence of similar words/phrases
  - If you don’t know what you are looking for
  - Can handle many classes
- Evaluate topic model with coherence statistic, visualization
- Interpret topic model with keywords, representative documents
- Label topics

Questions, Comments, Ideas?

Alexander Poon
@ alexander.poon@pm.me
alexander-poon