

Uptake Data Fellows Natural Language Processing Workshop



Hello! I'm Michael Miller Yoder

Graduate student at the Language Technologies Institute,
Carnegie Mellon University, Pittsburgh, PA



1

What is natural language processing (NLP)?

Computational processing of human language.



“

Computational processing of human language.

Examples: machine translation, dialogue systems, question answering, speech recognition, search engines



“

Computational processing of human language.

Often involves applying machine learning
techniques to text or speech data



“



Use case: **free-text survey data**



Use case: **free-text survey data**

Many questions are categorical (yes/no/maybe) or on a numerical scale.



Use case: **free-text survey data**

Many questions are categorical (yes/no/maybe) or on a numerical scale. But others may be open text response ("please explain...")



Topic modeling



Topic modeling

Statistical models for finding "topics" that occur in a collection of documents.



Topic modeling

Statistical models for finding "topics" that occur in a collection of documents. Common approach is Latent Dirichlet Allocation (LDA) [Blei et al, 2003]



Exercise: **topic modeling**

tokenization

What is a word?

What words count?



Exercise: **topic modeling**

tokenization

What is a word?
What words count?

feature extraction

Words to numbers
(bag-of-words).



Exercise: **topic modeling**

tokenization

What is a word?
What words count?

feature extraction

Words to numbers
(bag-of-words).

LDA interpretation

Do some
unsupervised ML!
Play around, interpret
results.



Bag-of-words (unigram) features

- Represent documents as vector of counts (or presence) of words



Bag-of-words (unigram) features

- Represent documents as vector of counts (or presence) of words

- the cat by the door jumped off .*
- the dog jumped up .*





Bag-of-words (unigram) features

- Represent documents as vector of counts (or presence) of words



- the cat by the door jumped off.*
- the dog jumped up.*

- { by: 1, cat: 1, dog: 0, door: 1, jumped: 1, off: 1, the: 2, up: 0, .:1 }
- { by: 0, cat: 0, dog: 1, door: 0, jumped: 1, off: 0, the: 1, up: 1, .:1 }



Bag-of-words (unigram) features

- Represent documents as vector of counts (or presence) of words
- Oblivious to order information

- the cat by the door jumped off .*
- the dog jumped up .*



- { by: 1, cat: 1, dog: 0, door: 1, jumped: 1, off: 1, the: 2, up: 0, .:1 }
- { by: 0, cat: 0, dog: 1, door: 0, jumped: 1, off: 0, the: 1, up: 1, .:1 }



LDA (Latent Dirichlet Allocation)

- Unsupervised: no "true" topics



LDA (Latent Dirichlet Allocation)

topic 0
topic 1
topic 2
topic 3
topic 4



- Unsupervised: no "true" topics
- Each document mixture of topics



LDA (Latent Dirichlet Allocation)

topic 0
topic 1
topic 2
topic 3
topic 4



- Unsupervised: no "true" topics
- Each document mixture of topics
- Each topic mixture of words



LDA (Latent Dirichlet Allocation)

topic 0
topic 1
topic 2
topic 3
topic 4



- Unsupervised: no "true" topics
- Each document mixture of topics
- Each topic mixture of words
- Based on word co-occurrence



NLP tools and resources

- Python
 - NLTK
 - Gensim (what we'll be using today)
 - Scikit-learn
- R
 - tidytext

**BIKE
PGH!**



Data:
Autonomous vehicle survey





Autonomous vehicle survey from cyclists and pedestrians

Context

Pittsburgh is a testing ground for AVs from Uber, ArgoAI and other companies.

Bike Pittsburgh

Bike Pittsburgh, bike and pedestrian advocacy organization, made an online survey in [2017](#) and [2019](#).



Choose your environment

Python

Jupyter Notebook:

[https://github.com/michaelmilleryoder/av-survey-topic-modeling/
av-survey-topic-modeling_python.ipynb](https://github.com/michaelmilleryoder/av-survey-topic-modeling/av-survey-topic-modeling_python.ipynb)

R

Jupyter Notebook:

[https://github.com/michaelmilleryoder/av-survey-topic-modeling/
av-survey-topic-modeling_r.ipynb](https://github.com/michaelmilleryoder/av-survey-topic-modeling/av-survey-topic-modeling_r.ipynb)



Download:

bit.ly/2EJ3kLv





Choose a text field

- interaction_details
- positive_av_interaction
- negative_av_interaction
- other_av_regulations
- elaborate_bikepgh_position
- other_comments



Workflow

- Tokenize: split into words
- Extract features: words to word IDs (bag-of-words model)
- Run LDA with varying numbers of topics
- Interpret topics
 - Look at high-ranking words for each topic
 - Look at high-ranking documents for each topic



...If you get to it

- Correlate topics with categorical and numerical fields
- Predict non-text fields with a machine learning algorithm such as logistic regression from topic distributions or text features
- Look into the [Structural Topic Model](#) (R package)



Thanks!

Any **questions** ?

Email me at

• **yoder@cs.cmu.edu**