Topic Modeling for Exploratory Text Analysis

June 16, 2020

Agenda for today's workshop

- The concept (~ 20 min)
- Tutorial/ example (~ 20 min)
- Your turn

Learning objectives

- What topic modeling is
- Why it is useful
- How to build a topic model in Python

What is topic modeling?

- Natural language processing technique
- Unsupervised machine learning

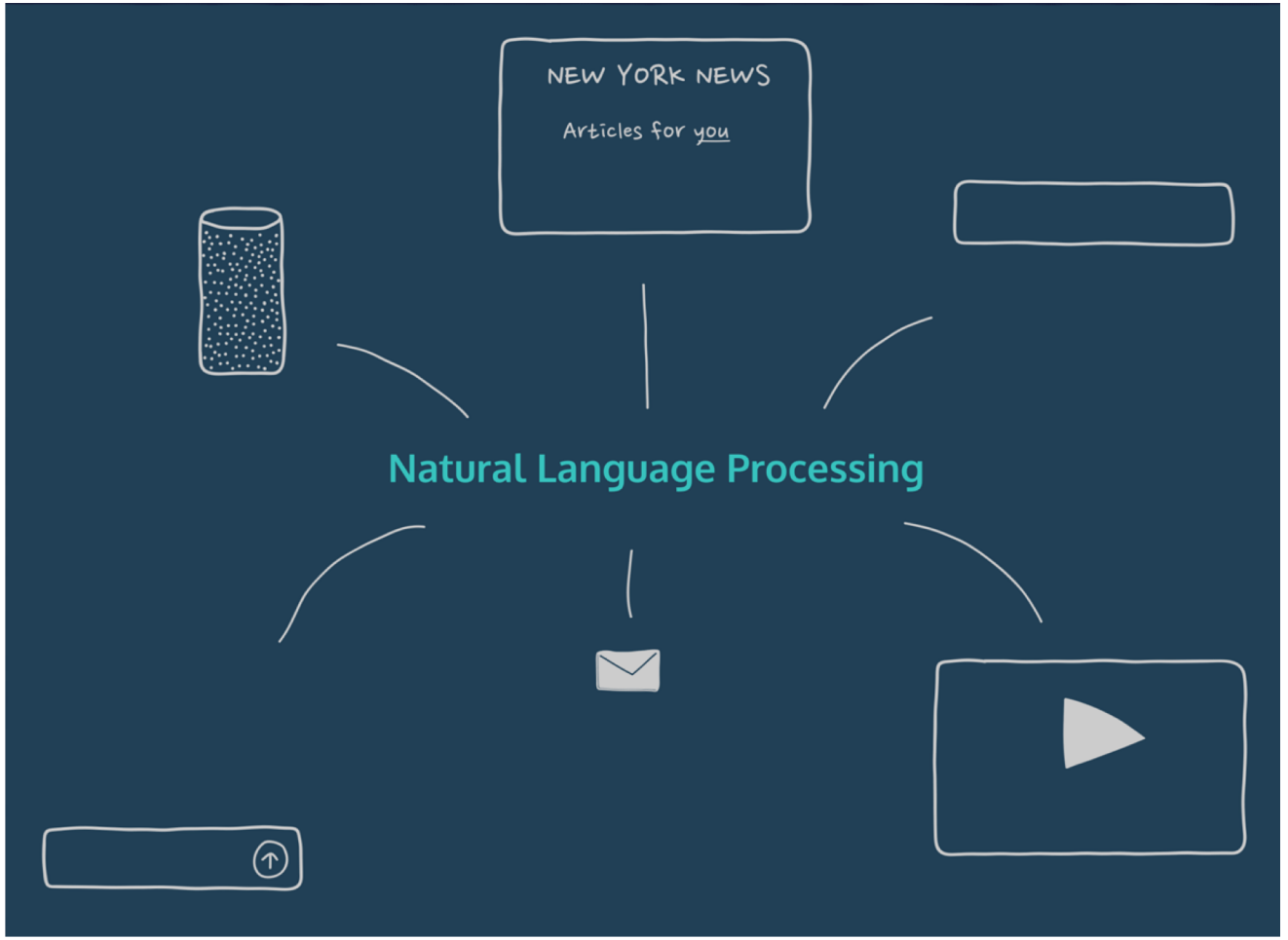


Image source: https://s3.amazonaws.com/codecademy-content/courses/NLP/Natural_Language_Processing_Overview.gif

What is topic modeling?

- The goal is to detect the latent topics among documents.
- Assumption: a document is a collection of topics and each topic is a collection of keywords.
- Popular algorithms: Non-negative Matrix Factorization (NMF), Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (pLSA), Latent Dirichlet Allocation (LDA), etc.

Latent Dirichlet Allocation

	text
0	Apples and oranges are delicious.
1	Michigan is a state in the United States.
2	Michigan apples are the best.

Topic 1

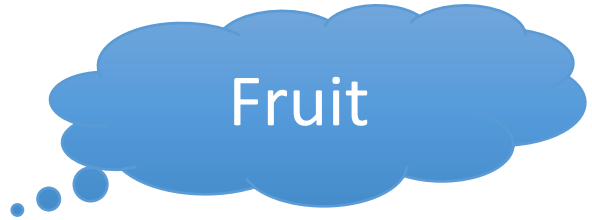
Topic 2

Topic 1, maybe also Topic 2

Latent Dirichlet Allocation

Topic 1

Apples, delicious, oranges, best …



Topic 2

State, Michigan …

Latent Dirichlet Allocation

	text
0	Apples and oranges are delicious.
1	Michigan is a state in the United States.
2	Michigan apples are the best.

Documents

Fruit (apples, oranges …)

Place (Michigan, state …)

Fruit (+ Place)

Topics

LDA auto-detects the topics among documents.

Why topic modeling?

- For big data
- For small data

Image source: https://medium.com/@souravboss.bose/comprehensive-topic-modelling-with-nmf-lsa-plsa-lda-lda2vec-part-1-20002a8e03ae

Steps for topic modeling

- Step 1: cleaning
- Step 2: tokenization
- Step 3: stemming / lemmatization
- Step 4: remove stop words
- Step 5: find topics
- Step 6: interpret the topics and improve the model