Text Mining PubMed on Particulate Matter

Zhe Meng

University of Illinois at Chicago

Dec. 12, 2017

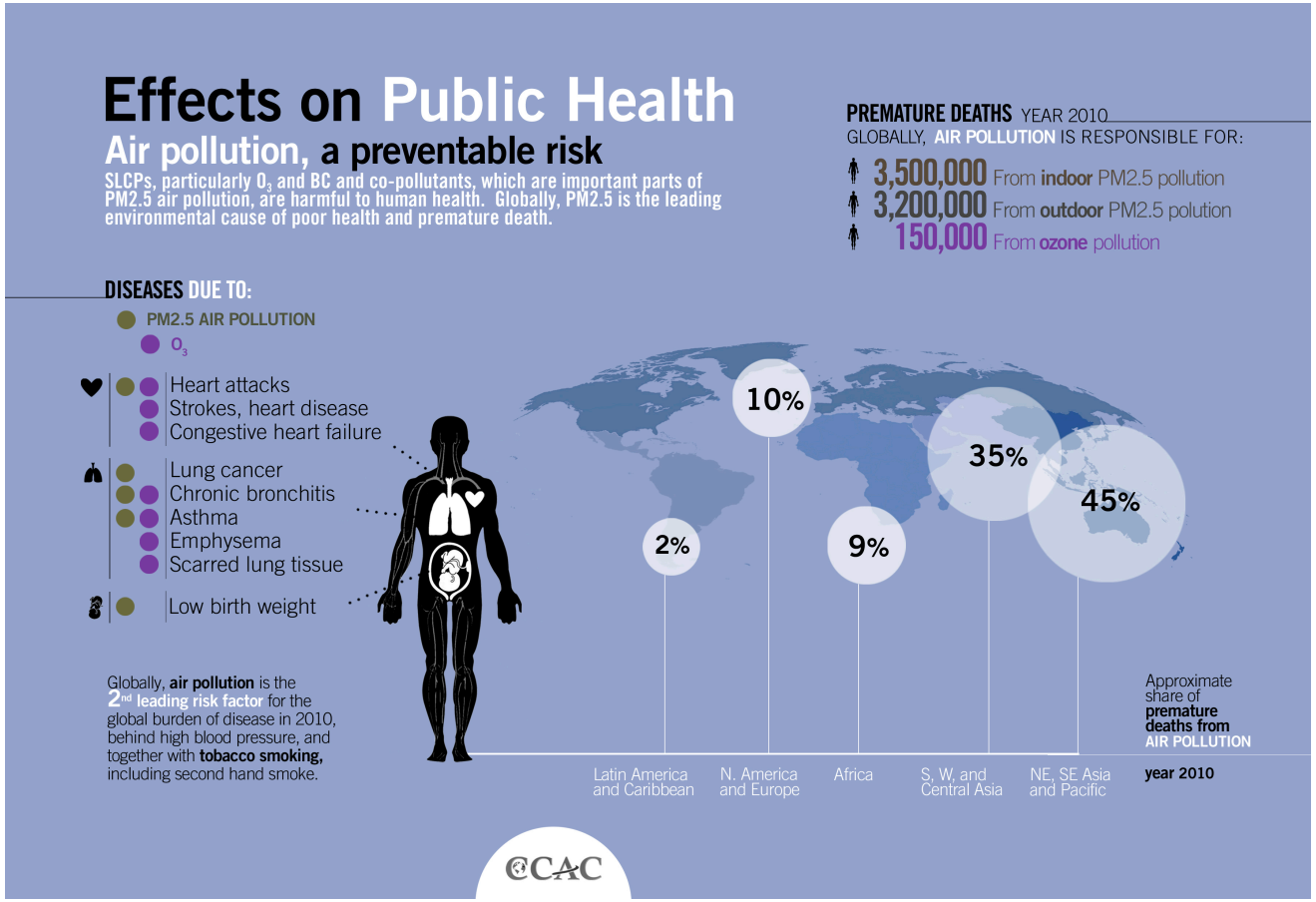
Outline

- 1. Introduction
- 2. Topic modeling
- 3. Clustering
- 4. Classification for Sentiment Analysis
- 5. Future Study

1. Introduction: Text Mining

- Text Mining refers to the process of extracting high quality of information from text.
- It widely covers a large set of related topics and algorithms for analyzing text, spanning various communities, including information retrieval, natural language processing, data mining, machine learning many application domains web and biomedical sciences.

1. Introduction: Problem Statement



- The World Health Organization estimates that PM air pollution contributes to approximately 800,000 premature deaths each year (Anderson JO, 2012). Thousands of articles on PM have been published.
- Using text mining one can attempt to predict the articles that are most relevant to what we are looking for and the key common concepts in those articles to help formulate hypothesis.

2. Topic Modeling: Concept Extraction

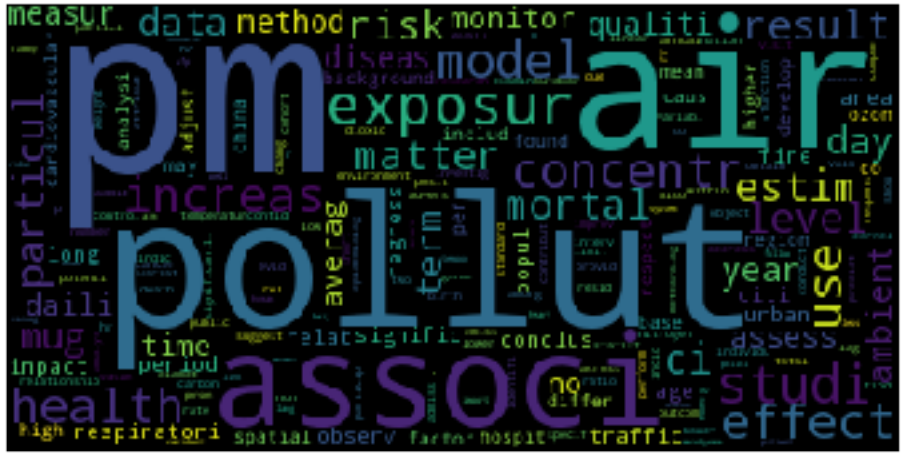
- 3721 abstracts published in 2016 were downloaded from PubMed.
- Use stop words such as ‘and’, ‘I’, ‘A’, ‘And’, ‘So’, ‘arnt’, ‘This’, ‘When’, ‘It’, ‘many’, ‘Many’, punctuation deletion, number 0-9 deletion, stemmer and lemmatization. 17226 unique tokens.
- Use Latent Dirichlet Allocation (LDA) to pre-process data, with num_topics=4.

2. Topic Modeling: Word Distribution

- **Topic 0**: 0.039∗"pm" + 0.021∗"pollut" + 0.020∗" air" + 0.017∗"associ”
- **Topic 1**: 0.017∗"cell" + 0.012∗"exposur" + 0.011 ∗"pm" + 0.010∗"induc”
- **Topic 2**: 0.012∗"concentr" + 0.009∗"particl" + 0.008∗"dust" + 0.008∗"sampl"
- **Topic 3**: 0.028∗"smoke" + 0.023∗"exposur" + 0.012∗"studi" + 0.010∗"use"

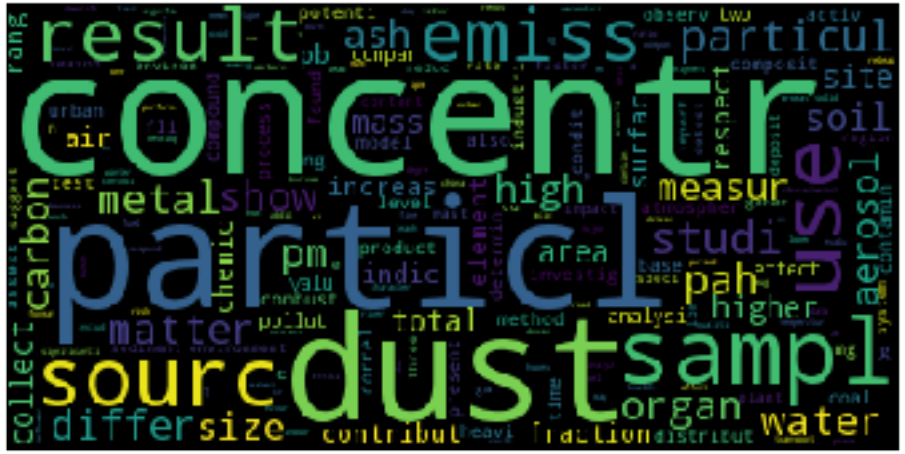
2. Topic Modeling: Word Cloud

Topic #0

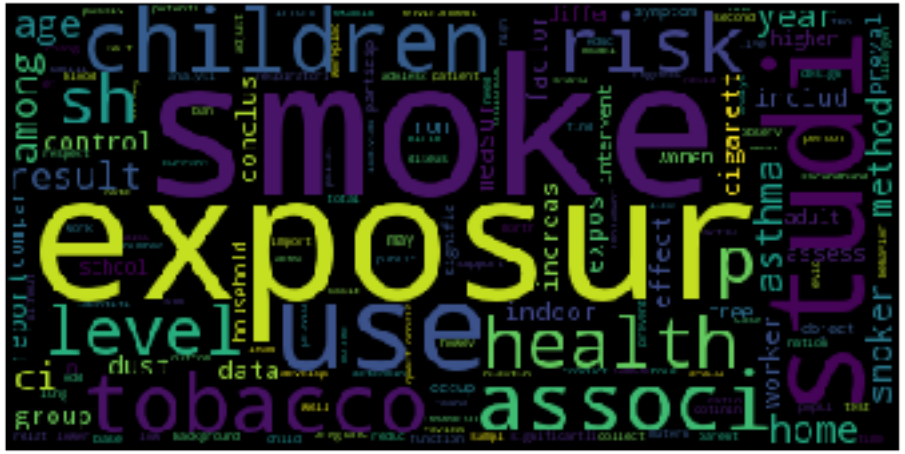


Topic #1

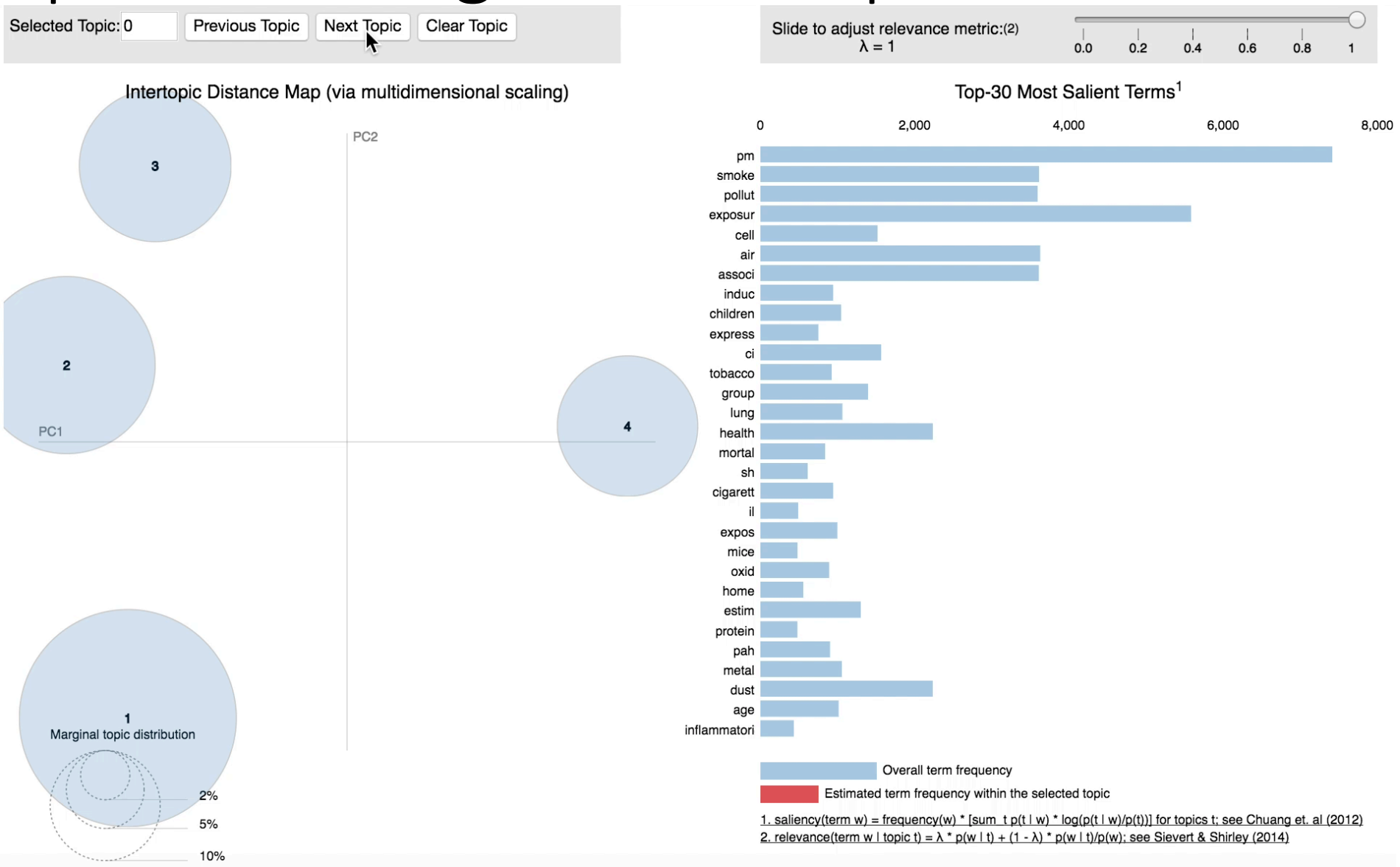
Topic #2



Topic #3



2. Topic Modeling: Plot of Topics and Words



3. Clustering: Kmeans with k = 4

Cluster 0	Cluster 1	Cluster 2	Cluster 3
exposure	pm2	exposure	concentration
air	concentration	smoke	particle
study	exposure	smoking	dust
95	air	cigarette	study
pollution	study	tobacco	particulate
ci	model	cell	result
association	pollution	study	sample
risk	level	group	air
associated	effect	shs	source
effect	particulate	child	matter

3. Clustering: Comparison of Cluster Labels

- Comparing cluster labels for the first 30 abstracts

Perform KMeans
3, 1, 3, 3, 1, 1, 0, 0, 3, 3, 3, 3, 0, 3, 3, 3, 1, 3, 1, 3, 1, 3, 1, 1, 3, 3, 3, 3, 1, 3
Perform PCA (Principal Component Analysis), then KMeans
3, 2, 3, 3, 2, 2, 1, 1, 3, 3, 3, 3, 3, 3, 3, 3, 2, 1, 2, 3, 2, 3, 2, 2, 3, 3, 3, 3, 2, 3
Perform LSA (Latent Semantic Analysis), then KMeans
1, 1, 1, 3, 1, 3, 1, 3, 1, 3, 0, 0, 0, 0, 3, 1, 3, 3, 1, 1, 1, 1, 3, 1, 1, 3, 1, 1, 1, 1

4. Predictive Modeling: Classification

- Classify the abstracts into two categories: those abstracts showing positive sentiment intensity (compound > 0) and those abstracts that are not positive (compound ≤ 0).
- Build three classifier: Naïve Bayes, Logistic Regression, and SVM (Support Vector Machines).
- Evaluate these three classifiers by using 10-fold cross-validation.

4. Predictive Modeling: Classifiers Comparison

scoring='accuracy'			
	Naïve Bayes	Logistic Regression	SVM
Mean	0.6594	0.7192	0.7524
Standard Deviation	0.0034	0.0237	0.0233

5. Future Study

- Try more stop words and other number of topics.
- For Kmeans, try other k values.
- Use other classification criterions.
- When performing cross-validation, try other scoring functions.