

Course Project

Accident topic modeling

Student: Tishun Peng

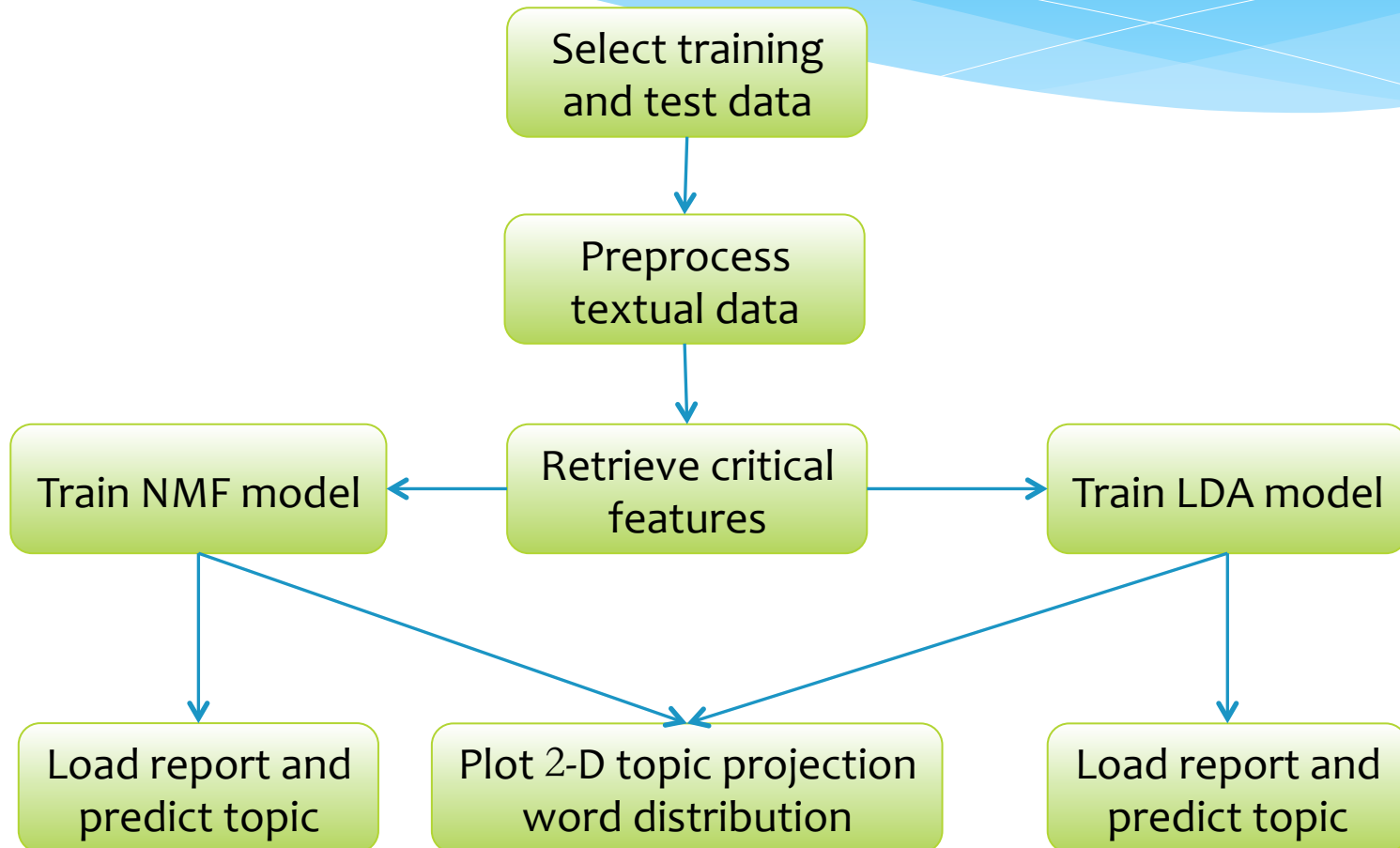
Objective

- ✓ Implement natural language processing and topic modeling algorithms to a practical problem
- ✓ Extract informative features (n-grams) from a collection of documents
- ✓ Identify the topics included in NTSB reports
- ✓ Predict the topic of another report in the test dataset

Why is it useful?

- ✓ NLP techniques can save the efforts of physically reading investigation documents, which may have hundreds of pages
- ✓ Find the reports that may be relevant to a specific type of accident using the trained topic model
- ✓ Programmatically extract critical information from a specific report

Development process

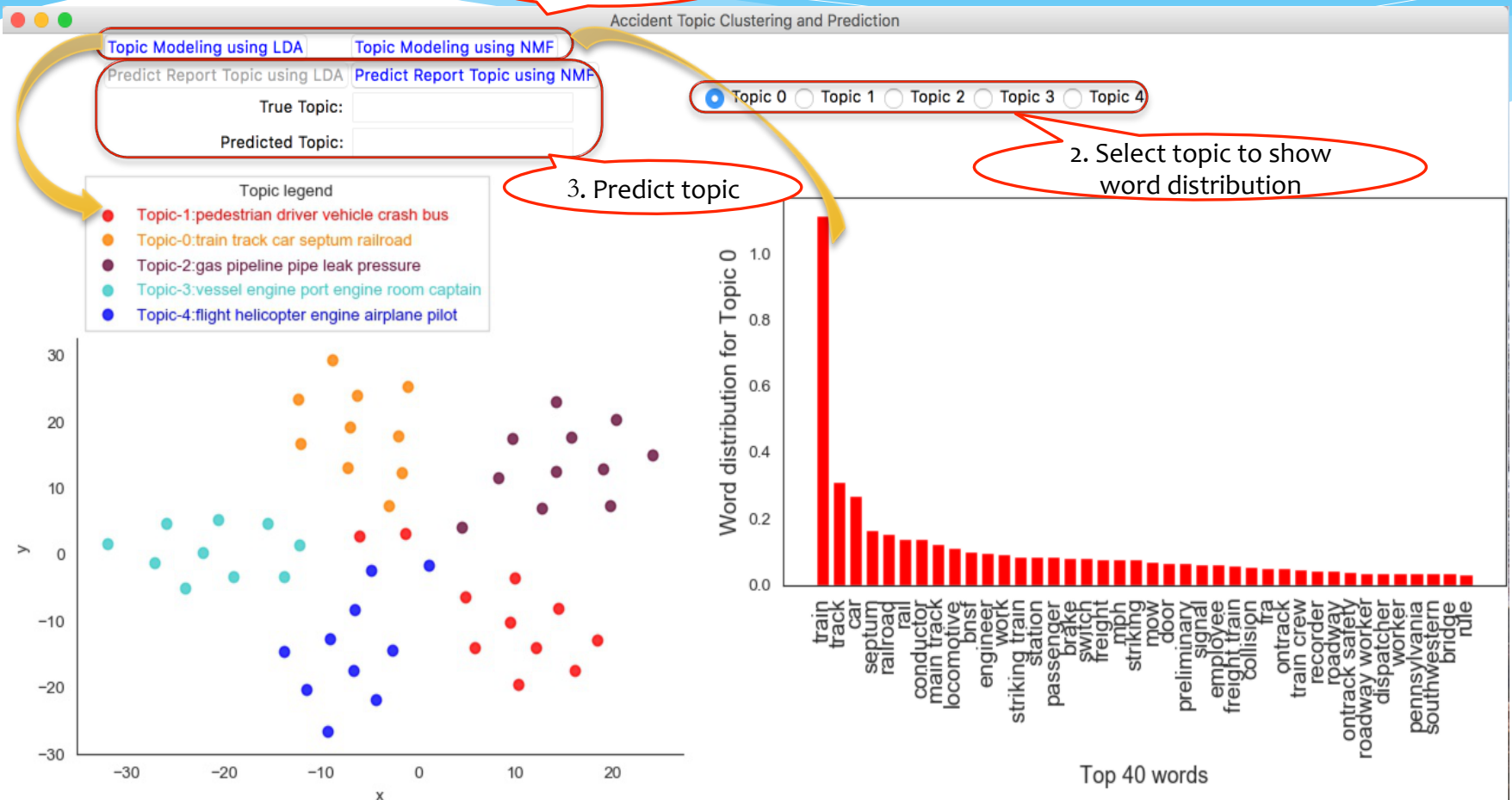


Technical details

- ✓ 52 accident training reports from NTSB website covering aviation, highway, marine, pipeline and railway industries
- ✓ 10 test reports for the above-mentioned topics
- ✓ Train NMF and LDA models for topic modeling
- ✓ Extract top 5 words to represent each topic and top 40 words to show word distribution
- ✓ A GUI is developed to encapsulate all functionalities

Graphical User Interface (GUI)

1. Train model

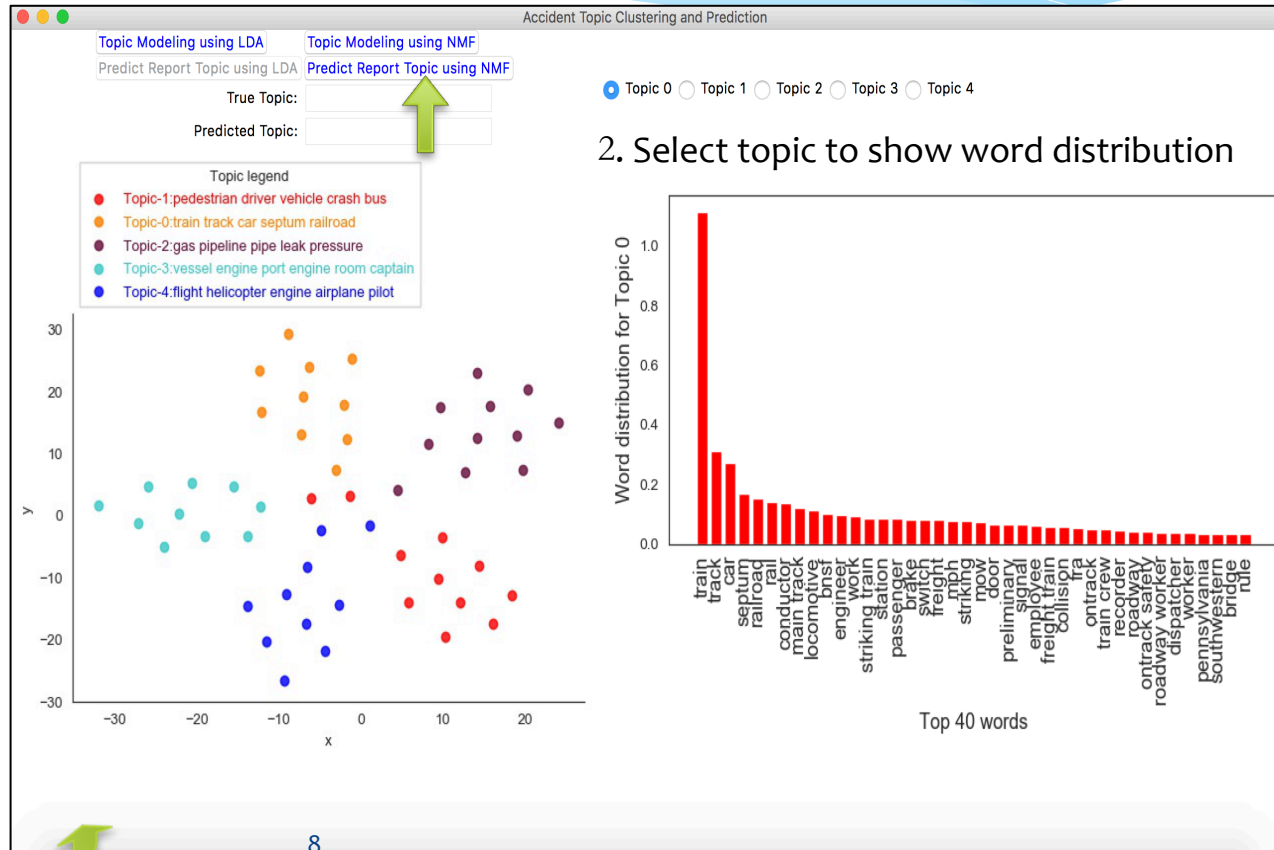
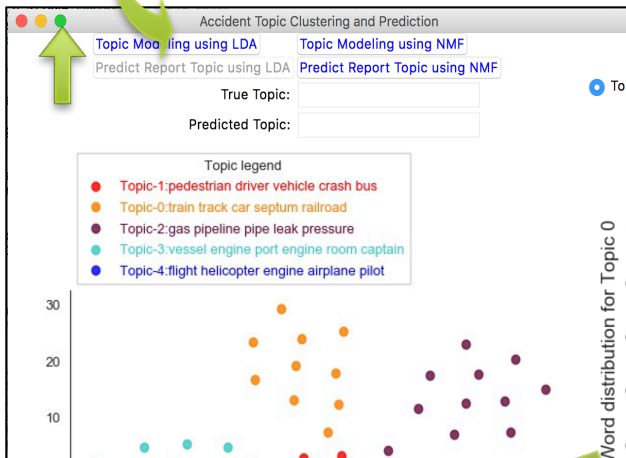
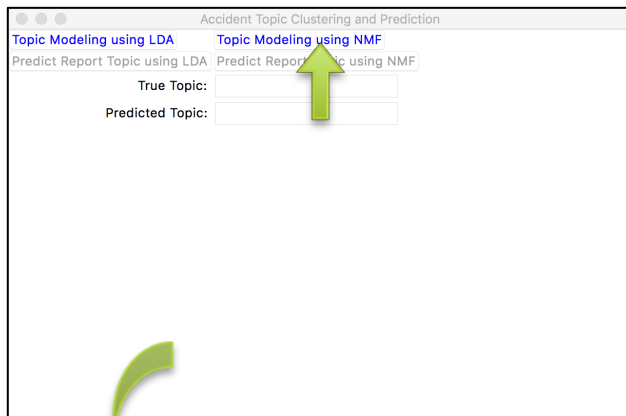


Setup and Run

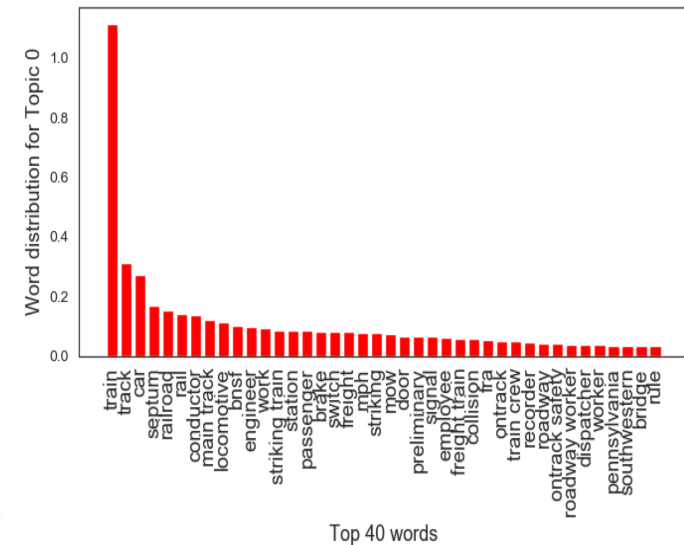
- Download and install Python 3.7
- Install required packages
 - matplotlib, scikit-learn, tkinter, Pillow
 - numpy, tika, glob, pandas, seaborn, string, regex
- Clone github repository
 - git clone https://github.com/tspeng/Course_Project.git
- Run application
 - cd Course_Project
 - python Application_code.py

Example

1. Train topic model



2. Select topic to show word distribution



Example

3. Select document to predict topic

Accident Topic Clustering and Prediction

Topic Modeling using LDA Topic Modeling using NMF

Predict Report Topic using LDA Predict Report Topic using NMF

True Topic:

Predicted Topic:

☒ Topic 0 ☐ Topic 1 ☐ Topic 2 ☐ Topic 3 ☐ Topic 4

test

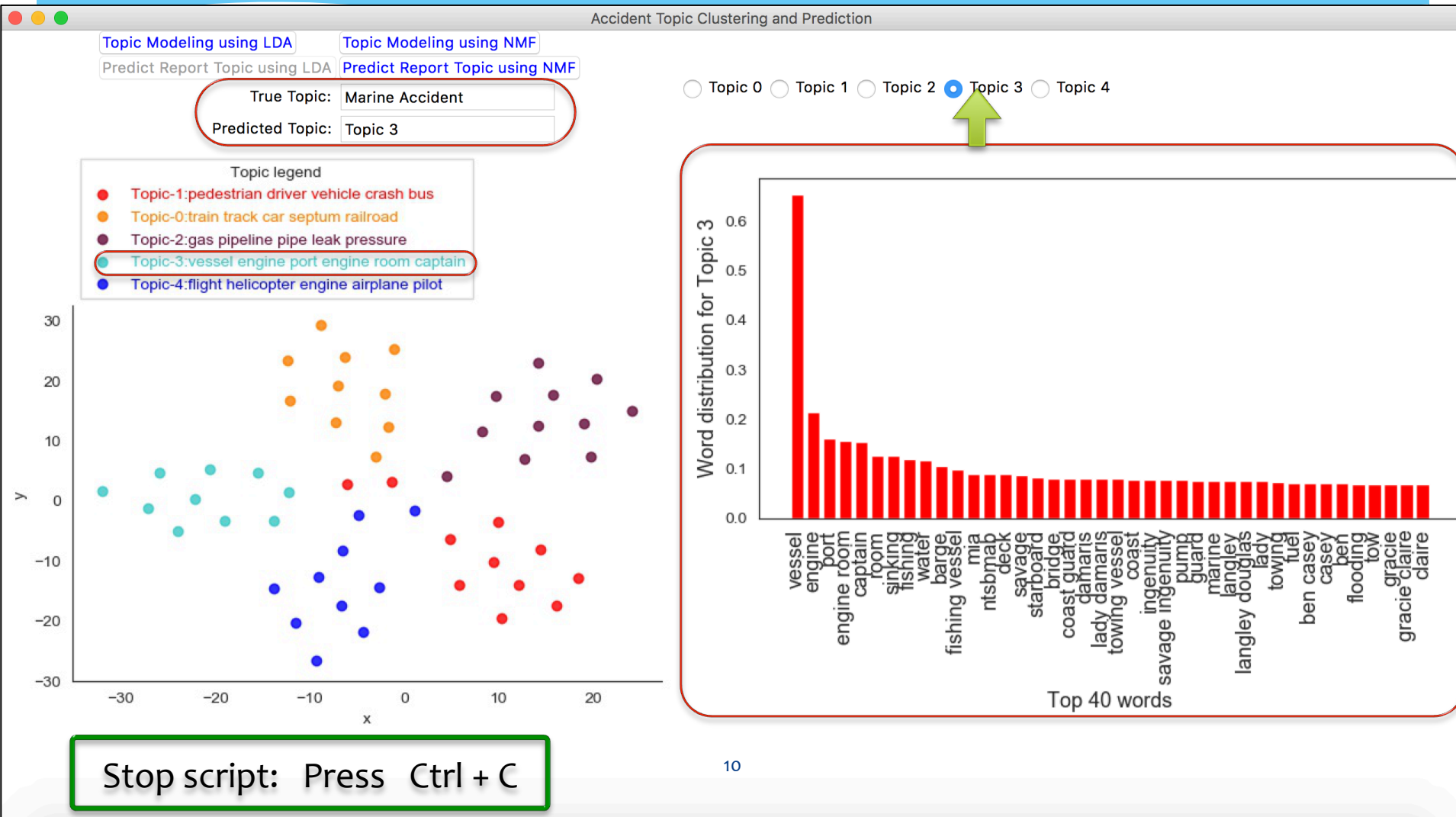
Name	Date Modified	Size
Railway Accident_RAB1713.pdf	Dec 11, 2018 at 4:10 PM	251 KB
Railway Accident_RAB1712.pdf	Dec 11, 2018 at 4:10 PM	342 KB
Pipeline Accident_PAB0801.pdf	Dec 11, 2018 at 4:09 PM	429 KB
Pipeline Accident_PAR0901.pdf	Dec 11, 2018 at 4:09 PM	1.3 MB
Marine Accident_MAB1812.pdf	Dec 11, 2018 at 4:09 PM	701 KB
Marine Accident_MAB1813.pdf	Dec 11, 2018 at 4:08 PM	716 KB
Highway Accident_MAB1803.pdf	Dec 11, 2018 at 4:08 PM	4.3 MB
Highway Accident_MAB1804.pdf	Dec 11, 2018 at 4:08 PM	505 KB
Aviation Accident_ASR1601.pdf	Dec 11, 2018 at 4:07 PM	2.2 MB
Aviation Accident_ASR1607.pdf	Dec 11, 2018 at 4:06 PM	307 KB

9

Cancel Open

preliminary
signal
employee
freight train
collision
ontrack
train crew
recorder
roadway
ontrack safety
roadway worker
dispatcher
pennsylvania
southwestern
bridge
rule

Example



Conclusions

- ✓ NMF can achieve relatively better topic modeling results than LDA
- ✓ Histograms of top ranking words can clearly represent the nature of each type of accident(topic)
- ✓ From the 2-D projection, intra-cluster similarity is much higher than inter-cluster similarity
- ✓ The trained model can provide pretty accurate prediction for report topic