Course Project

Student: Tishun Peng

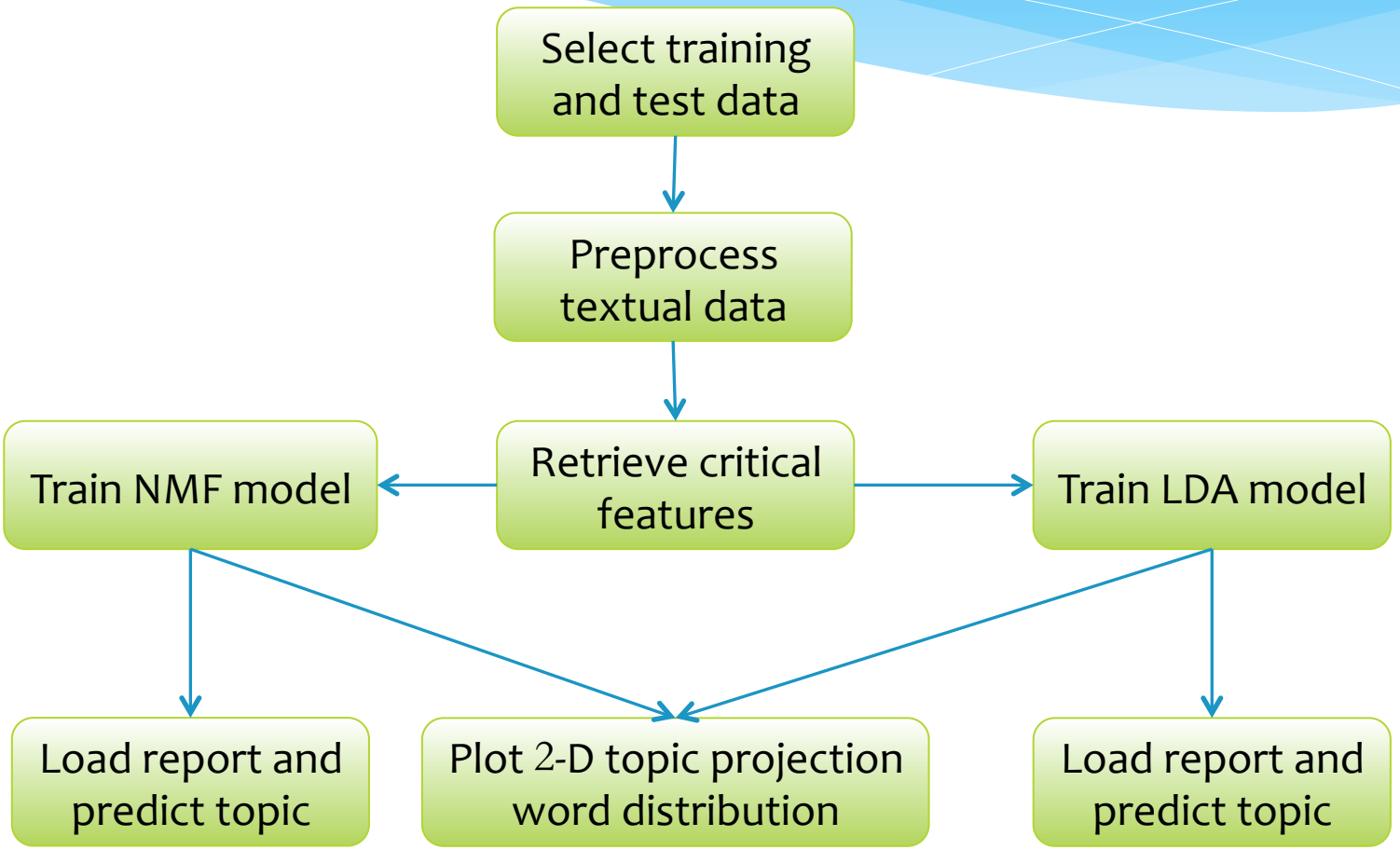
Objective

- ✓ 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 set

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

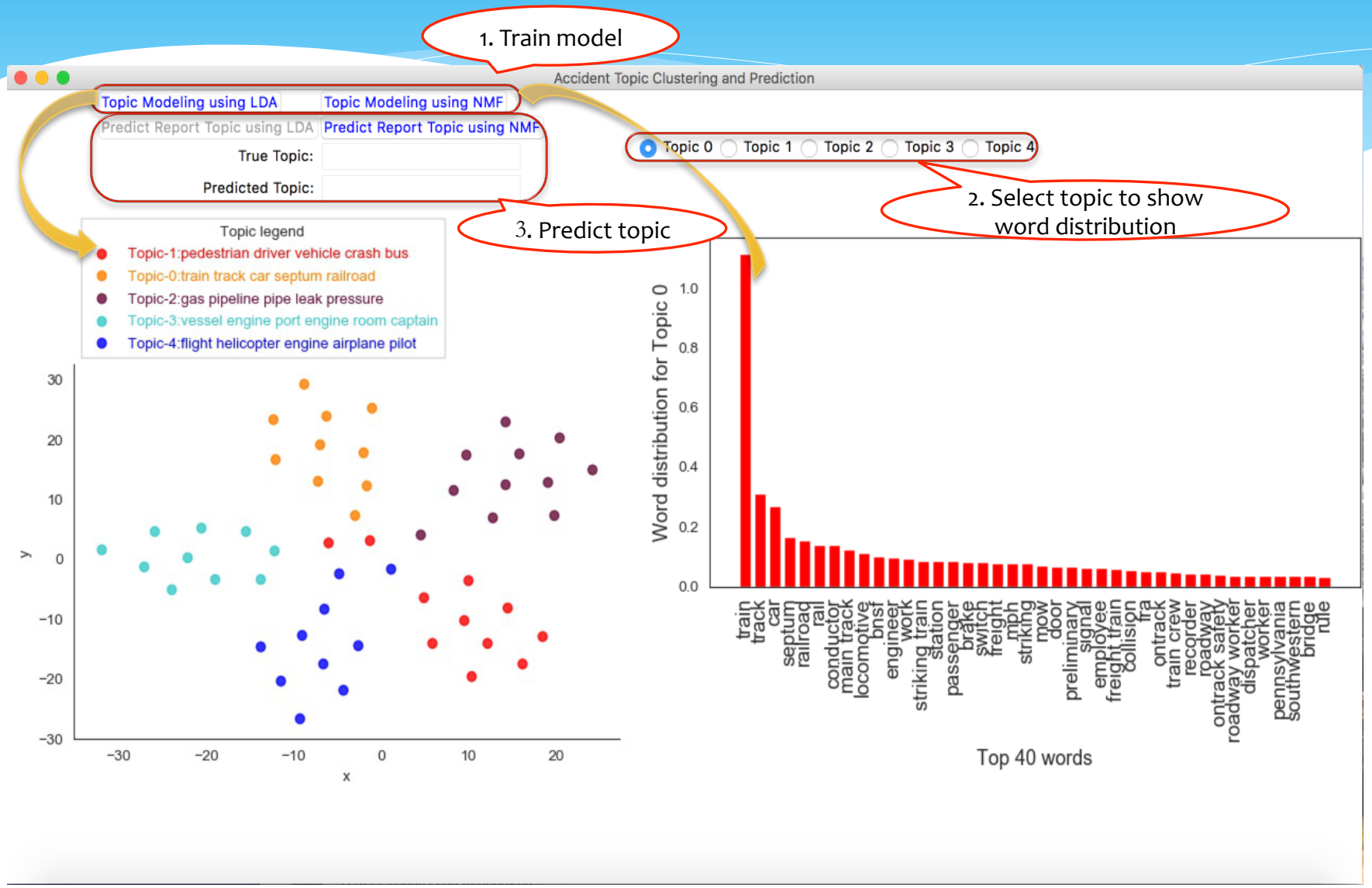
Development process



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)

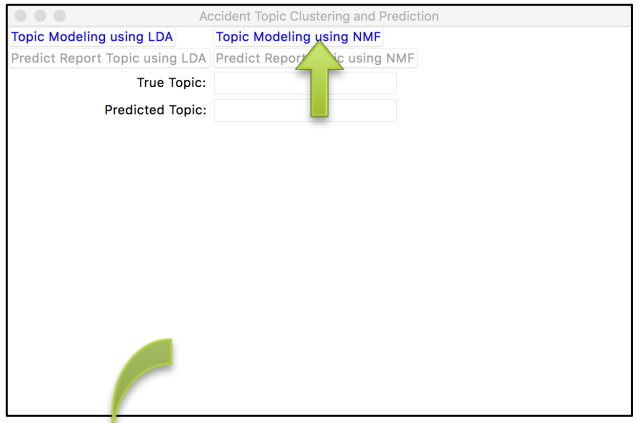


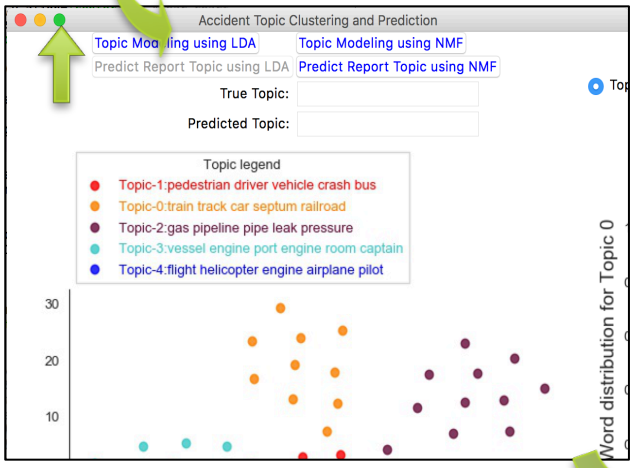
Setup and Run

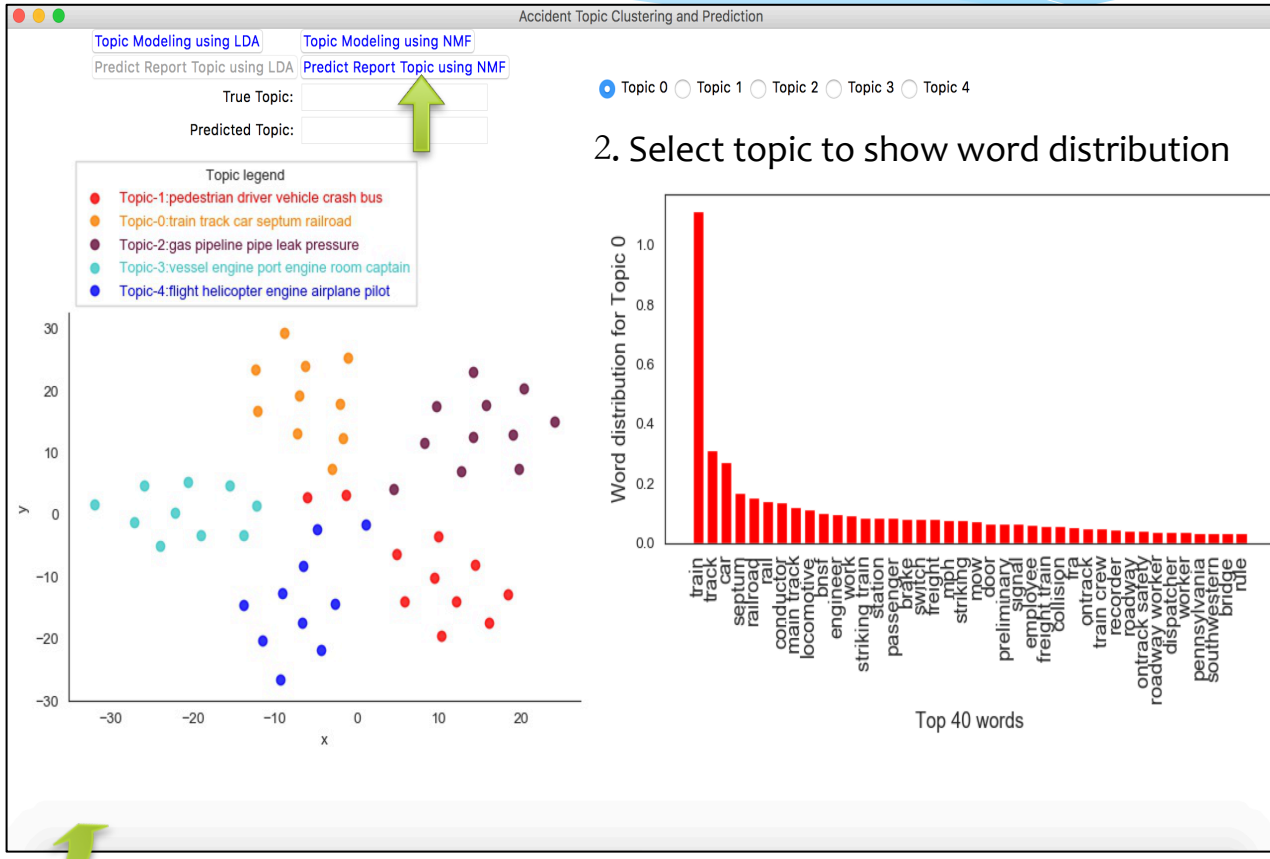
- Download and install Python 3.7
- Install required packages
  - matplotlib, scikit-learn, tkinter, Pillow
  - numpy, tika, glob, pandas, seaborn, string, regex
- Clone github repository
  - git clone https://github.com/tspeng/Course_Project.git
- Run application
  - cd Course_Project
  - python Application_code.py

Example

1. Train topic model



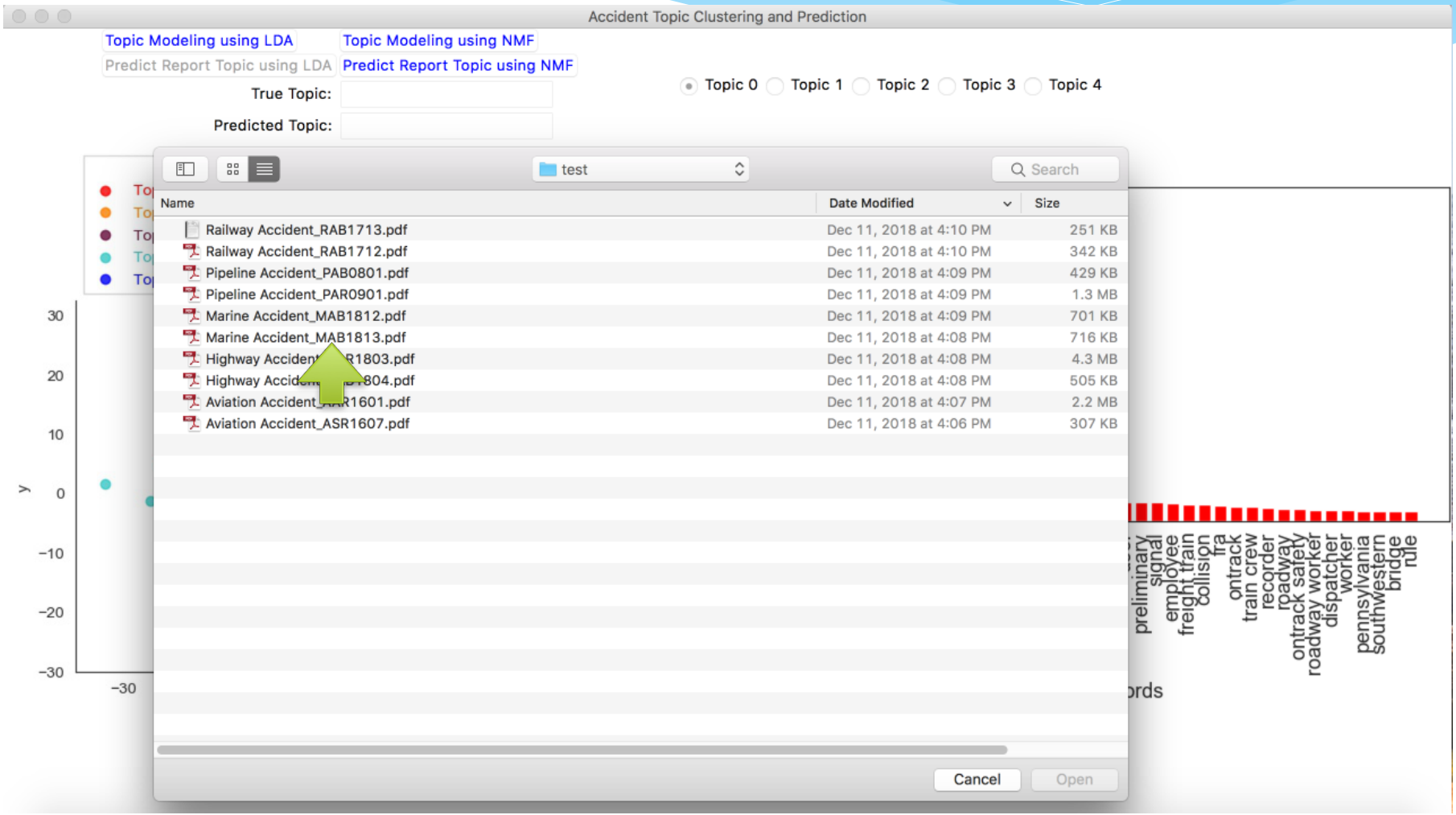




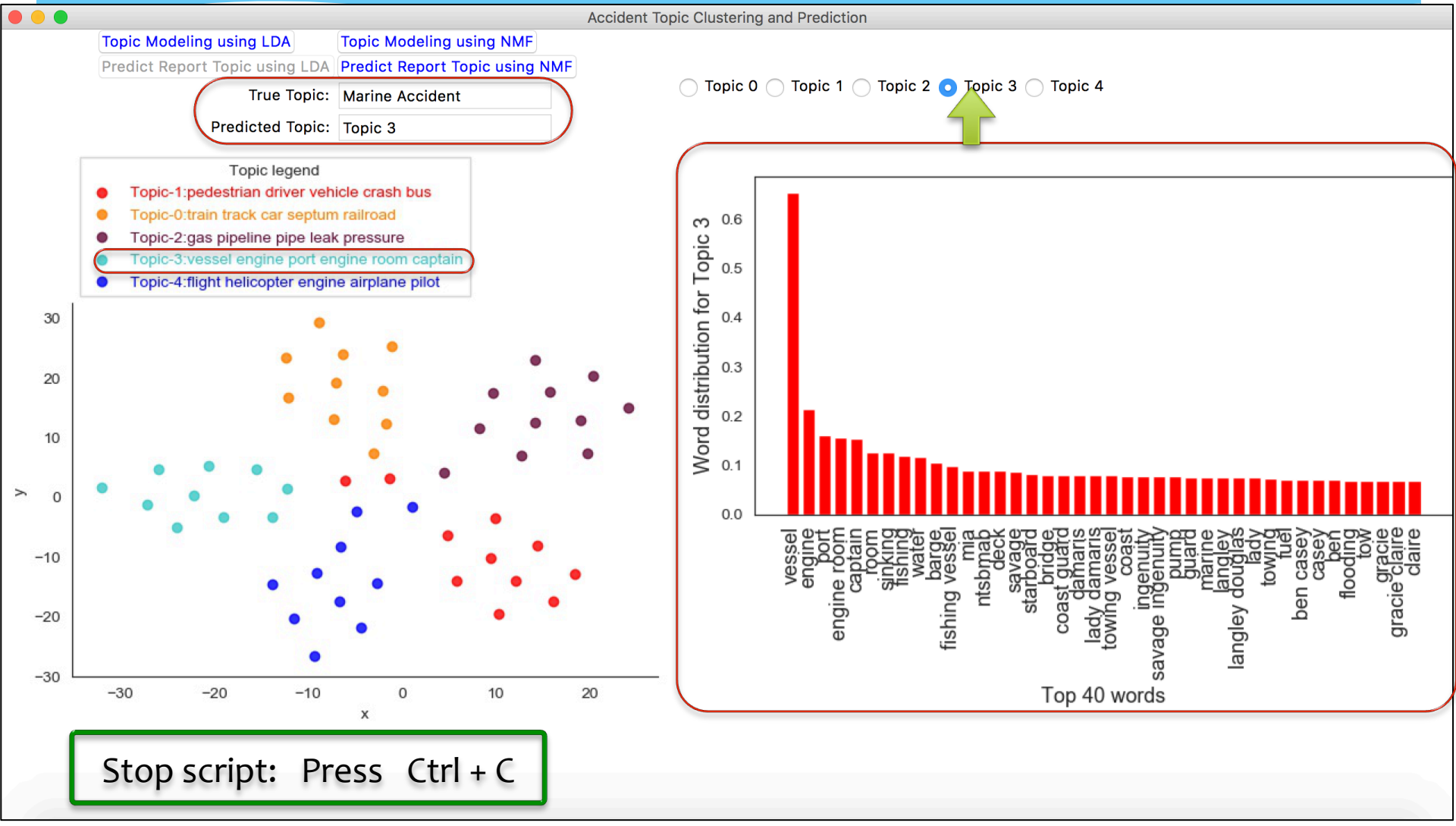
2. Select topic to show word distribution

Example

3. Select document to predict topic



Example


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: Marine Accident
Predicted Topic: Topic 3
Topic legend
Topic-1:pedestrian driver vehicle crash bus
Topic-0:train track car septum railroad
Topic-2:gas pipeline pipe leak pressure
Topic-3:vessel engine port engine room captain
Topic-4:flight helicopter engine airplane pilot
Topic 0 Topic 1 Topic 2 Topic 3 Topic 4
Word distribution for Topic 3
Top 40 words
vessel engine port engine room captain room sinking fishing water barge fishing vessel mia ntsbmab deck savage starboard bridge coast guard damaris lady damaris towing vessel coast ingenuity savage ingenuity pump guard marine langley langley douglas lady towing fuel ben casey casey ben flooding tow gracie gracie claire claire
Stop script: Press Ctrl + C

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 train model can provide pretty accurate prediction for report topic