Aviation Data Analysis Report

Yifei Wang

This is a data analysis report of Aviation Data provided by NTSB website. I build topic models on the event report text data using several approaches and find the possible topic(s) (cause) for each event.

Dataset

This dataset contains 2 different data, one is the structured event data of each aviation accident and the other one is the unstructured text data of event report. There are 76133 event, among which 75905 event has narrative report and 49789 event has report of probable cause. The latter report is more structured than the former one.

Preprocessing

I remove all the punctuation, digit and stop words from raw text, lower case them, then use stemming to tokenize each word and finally keep all tokens with frequency larger than 1.

Model - cause text

- I apply LSI to cause text to find the first 15 principal topics and their top 10 contributive words by treating each word as a feature. By using a relatively large numbers of topic, I could manually summarize possible topics with more details.
- Then I assume each document has a certain probability distribution of belonging to certain topics and each topic has a specific distribution of words. I train LDA on cause data with 8 (6/10/16) topics to find such distributions and make a word cloud for each topics to infer the clustering results.
- Since both LSI and LDA generate highly correlated topics, I use TFIDF to capture the distinctive details (tokens) for each topic. Then I fit the TFIDF vectors to KMeans with 8 topics and find it successfully capturing more detail information but using the same (or smaller) topic number than LDA or LSI.

Model - narrative text

- Compared to cause text, narrative text has more information but more unstructured with more noises.
- In order to accommodate to these differences, I try using doc2vec model to find the vector representation of each document and then fit KMeans model with 8 topics. However, unlike the cause data, the result here is not that appealing

Analysis

- I link the generated topics to weather information and find that more weather-related accident happed on worse weather.
- I link the generated topics to self-calculated fatal rates and find that high fatal rate accidents often result from pilot's negligence of environment and low fatal rate accidents often come from obvious and non-crucial damage to the airplane, so the pilots have enough time to react to the situations and avoid severe accident.