Analysis Report of Aviation Accidents

In this report, I analyzed the aviation data from NTSB and tried to answer one of the greatest concerns of stakeholders - the fatalness of an accident and the corresponding factors.

The NTSB data have records of 76133 events from 1982 to 2014. There are 77257 accidents, since multiple accidents may be filed for the same event. The injury severity is categorized into three types – fatal, incident, and non-fatal – in the raw data, and I combine the incident and non-fatal as non-fatal.

[Task 1] Analysis of the relationship between the story and the fatalness of the events.

Problem and Approach:
Textual content contains information of how an event happened. A stakeholder may want to know the relationship between the cause and the fatalness so that further actions (e.g., train pilot better or maintain a component more often) can be taken to prevent an accident in the future.

Therefore, I formalized the problem as a binary classification problem. I use topic modeling (Latent Dirichlet Allocation) to vectorize the textual description of each event. Then I use Logistic Regression to test if an accident in the vector space can be correlated to a fatalness label.

Result:
I split the dataset into a training set (68031 events) and a testing set (7559 events). The performance of logistic regression is as below:

	precision	*recall*	*f1-score*	*support*
nonfatal	0.91	0.96	0.93	6159
fatal	0.75	0.56	0.64	1400
weighted avg	0.88	0.88	0.88	7559

Since the dataset is imbalanced, I use another metrics Area Under Receiver Operating Characteristic Curve (AUC-ROC) which achieves a score of 0.906. It indicates that using topic modeling is effective to assess the fatalness of an accident.

[Task 2] Visualization of the probable causes of fatal and non-fatal events.

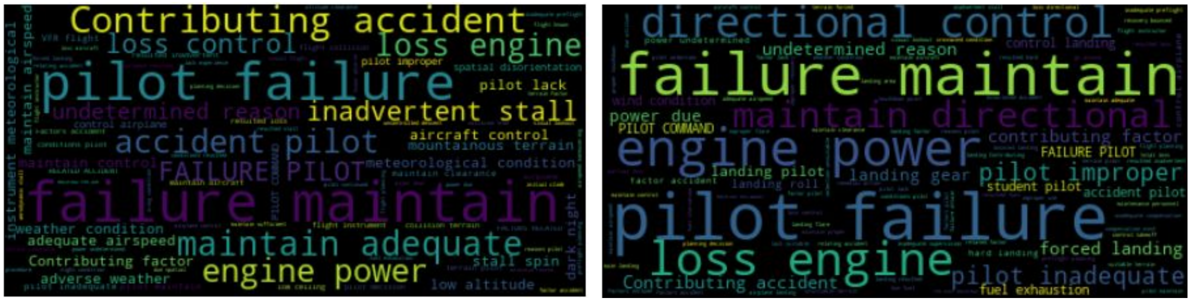
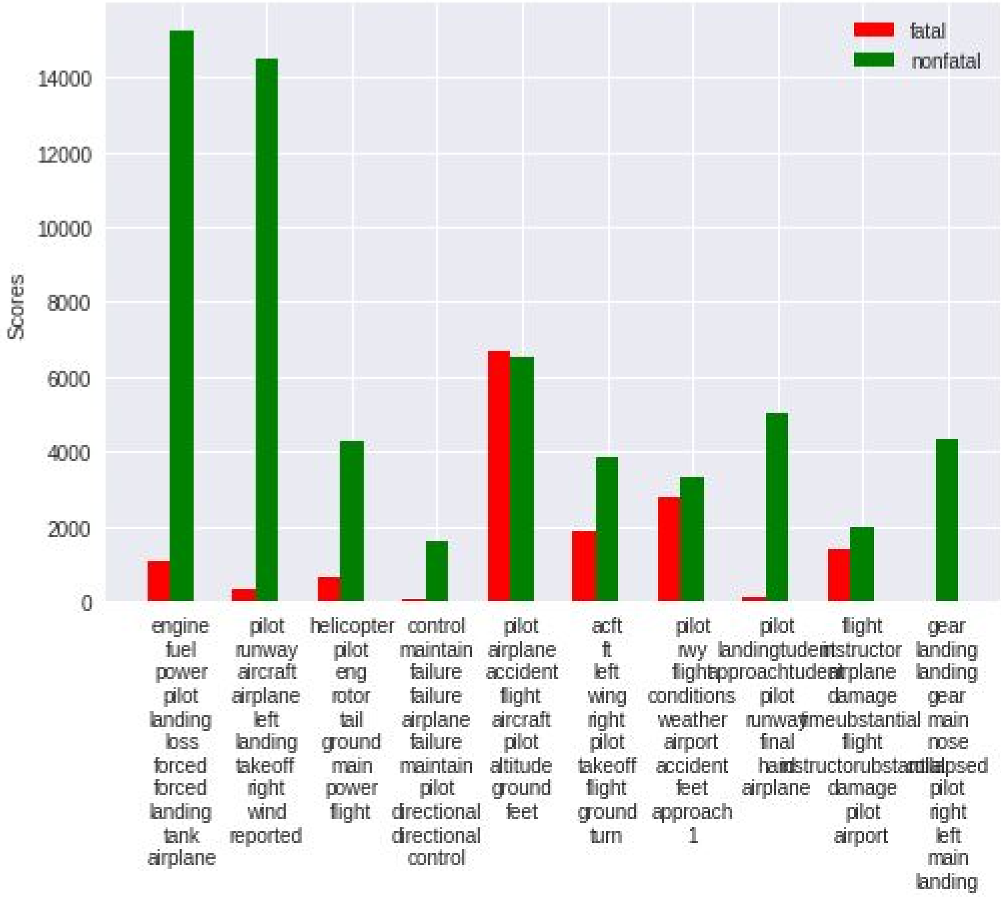


Figure 1 (a) Keywords of causes of fatal accidents; (b) Keywords of causes of nonfatal accidents.

Figure 1 shows the word cloud representation of fatal events and nonfatal events. The font size of the word cloud indicates the frequency of the occurrence of the keywords. It indicates that pilot failure is the dominant probable cause in both fatal and nonfatal events.

The following bar chart shows the counts of events in each topic and the corresponding keywords. It may need domain knowledge to interpret each topic. However, concepts such as “pilot”, “altitude”, and “weather” play an important role in fatal accidents.


fatal
nonfatal
Scores
engine fuel power pilot landing loss forced forced landing tank airplane
pilot runway aircraft airplane left landing takeoff right wind reported
helicopter pilot eng rotor tail ground main power flight
control maintain failure failure airplane failure maintain pilot directional directional control
pilot airplane accident flight aircraft pilot altitude ground feet
acft ft left wing right pilot takeoff flight ground turn
pilot rwy flight conditions weather airport accident feet approach 1
pilot landingtudent approachtudent pilot runway final hand airplane
flight instructor airplane damage timeubstantial flight instructorubstantial damage pilot airport
gear landing landing gear main nose collapsed pilot right left main landing