Aviation Accident Cause Codes

Text Analysis for aircraft crashes causes in R

Presentation Plan

- Introduction
- Getting and Cleaning Data
- Clustering
- Frequency Charts
- Conclusions

Introduction

- The goal is to try to extract the most common causes of planes crashes, by using text analysis on the context lines in the dataset. This website shows the actual most common causes since the 1960s.

- The data can be downloaded from https://opendata.socrata.com/Government/Airplane-Crashes-and-Fatalities-Since-1908/q2te-8cvq

- The code is developed using R.

- The libraries have been used:
  - tm
  - dplyr
  - stringr
  - tidyr
  - factoextra
  - ggplot2
  - wordcloud

Getting and Cleaning Data

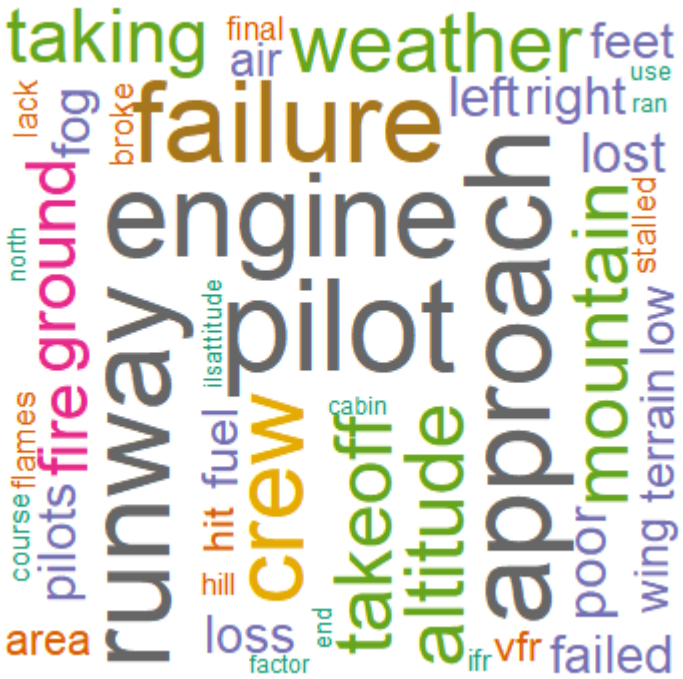
- Performed the following steps to clean data:
  - Remove Punctuation
  - Convert To Lower Case
  - Remove English Stopwords
  - Strip Whitespace
  - Remove removing generic terms related to airline industry
- Converted data into document-terms matrix.
- Remove sparse terms from document-terms matrix at 95% threshold.

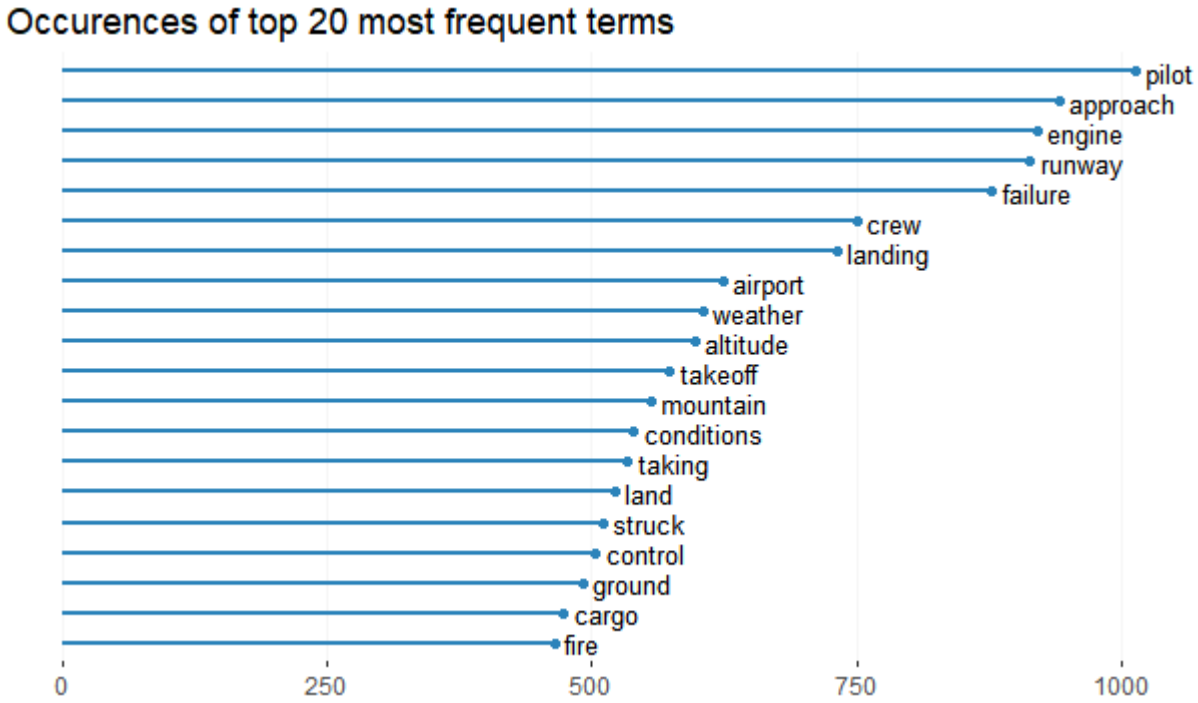
Clustering

- Used K-Means Cluster with N = 10 (total number of clusters– 10). The following clusters are formed:
  - Cluster01: engine
  - Cluster02: pilot
  - Cluster03: crew, landing
  - Cluster04: conditions, weather
  - Cluster05: altitude, cargo, control, ground, left, mountain, route, struck, taking
  - Cluster06: failure
  - Cluster07: approach, runway
  - Cluster08: accident, failed, feet, flying, fog, loss, lost, low, miles, poor, short, shortly
  - Cluster09: airport, fire, takeoff
  - Cluster10: attempting, land

Frequency Charts

- The word cloud chart shows the most frequent words in corpus.
- The second chart shows the top 20 most frequent terms.





Conclusions

- After performing the ‘Terms correlation’, the following can be concluded from the data:
  - **Pilot**: 'error' is one of the most correlated words, which is consistent with the fact that ~60% of crashes are due to pilot errors
  - **Approach**: the accidents in final approach phase seem to be often caused by confusion in reading instruments and low visibility ('instruments', 'visual', 'missed')
  - **Engine** seems related to shutdown of engine and/or loss of power
  - **Runway** is associated with 'short', 'end' and 'overran', that could be as well in takeoff or landing phases
  - **Failure**: we have more context here, suggesting that it can be pilot, maintenance, procedure or system failures
  - **Landing**: this shows that it is not necessarily about the standard landing phase, but rather about landing gears, or emergency landings
  - **Weather and Conditions** suggest that visibility is one of the most important crashes factors in bad weather

Thank You