# RTI Technical Exercise: Classifying Aviation Accident Text

Chris Coxen

## Introduction

The National Transportation Safety Board (NTSB) aviation accident database contains information about civil aviation accidents and incidents within the United States, its territories and possessions, and international waters. NTSB provides reports for these cases that contain descriptive metadata and full narrative descriptions based on its investigations.

The goal of this exercise is to evaluate whether various analytical techniques can be used to effectively classify the main topics of the NTSB report narratives.

## Methods

Text generally always needs to be preprocessed before it can be analyzed. Many words, and sometimes entire sentences, are removed from the body of the text to ensure you only analyze the words that best represent the topic or sentiment of that text. Our first step was to review a random selection of the NTSB narratives to get a feel for the narrative structure and the type of information conveyed in the text. This led us to remove the first sentence from every narrative because nearly all of them began with a scripted boilerplate introductory sentence.

Our next step was to remove uninformative words through a process called stop word removal. We then standardized words by reducing them to their base "dictionary" form (or lemma) through lemmatization. We tagged all of the words by part of speech to ensure the lemmatization algorithm properly reduced each word based on its relative part of speech.

We then grouped the narratives into decades and analyzed the word frequency over time. While we could not detect any obvious trends through an "eyeball" analysis, we did learn that our topic clustering analysis may benefit from a stricter term frequency threshold. We then generated a full document-word matrix, turning each document into a numeric vector so that we could perform topic modeling with Latent Dirichlet Allocation (LDA).

LDA is a popular topic modeling method because it is simple to implement and produces effective results. LDA works to find the topic (or topics) a document belongs to based on the words that make up that document. Through tuning, we learned that our LDA model would perform best by classifying the narratives between five different topics.

## Results

Our model generated five distinct topic clusters. The five major topics were:

1. Weather or miscellaneous crash
2. Fuel issue
3. Engine or mechanical failure
4. Helicopter or student pilot accident
5. Steering or landing issue

Examining the top 15 words for each topic was helpful for assigning the topic themes and to begin to develop a narrative around how the model grouped words into each topic. To provide additional context, we implemented a pyLDAvis visual to explore the relationship between word frequency and the five clusters in our LDA model. Explore the visual here. A screenshot is pictured in Figure 1 below.

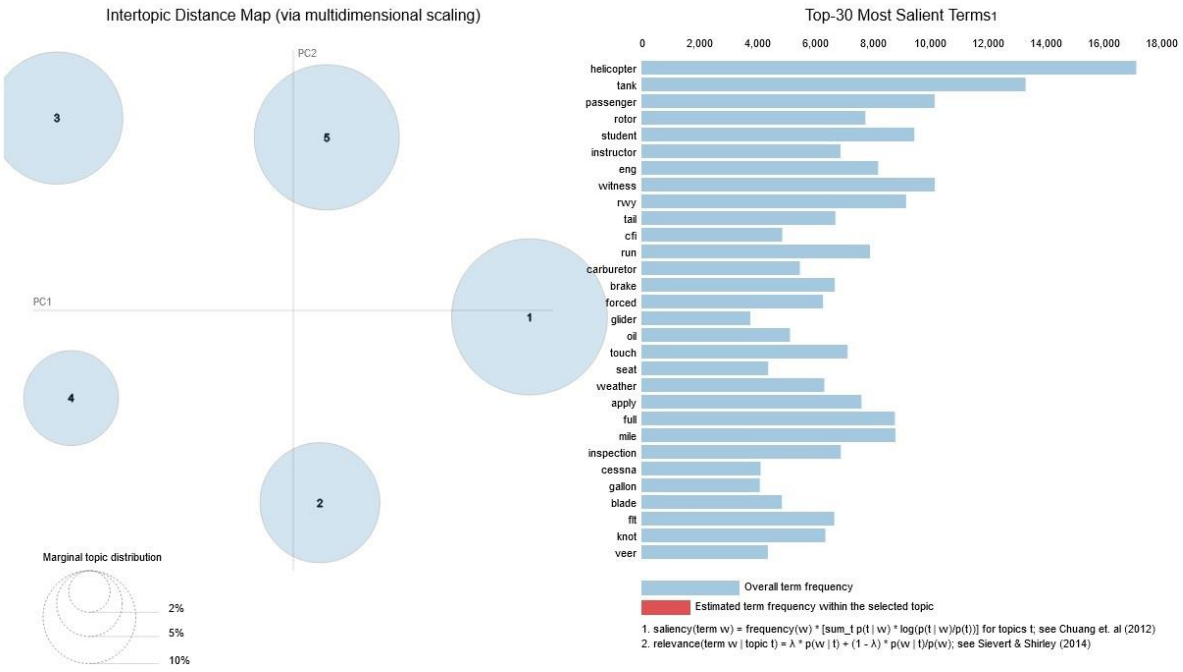


**Figure 1.** Still image of the interactive pyLDAvis visual for our LDA model

The distance between each cluster (shown in the Intertopic Distance Map on the left of the visual) suggests the model performed well and generated five distinct topics. This is further supported by the fact that the dominant terms in each cluster are almost entirely contained in that cluster and do not spill over into other clusters (shown in the Top-30 Most Salient Terms frequency chart on the right of the visual).