This exercise details an examination of aviation accident data from the National Transportation Safety Board, including exploration of both structured data and text (in the form of narrative text and probable cause statements). In this report, I will describe the structured dataset before diving into the textual analysis to contrast and compare findings.

The initial structured dataset is comprised of 77,257 records of aviation “accidents” (non-fatal, incidents, or fatal), with 31 recorded variables, all with varying levels of *missingness.* The overwhelming majority of the records are listed as “accidents” (as opposed to incidents). Geographically, approximately 95% of the cases occurred in the United States (understandably, since the data was collected by a U.S. organization), although figure 1 shows that a number incidents occurred at sea (based on lat and long figures). Moreover, a small percentage of accidents occurred regionally (Canada & Mexico, Bahamas) as well as in Europe. Figure 2 shows that within the United States, The greatest percentage of accidents occurred in California, followed by Florida, Texas, and Arkansas. Although a plurality of the records do not list an airport name (31,136 missing), the top 10 common airport listings indicate a “private” airstrip.

Regarding fatalities, figure 3 shows that nearly 80% of cases were listed as non-fatal accidents, whereas nearly 10% were listed as single-fatality (presumably the pilot), followed two fatalities at 6% of cases (with the remainder of cases being higher fatality or missing). However, when examining aircraft damage, we note that 72% of records indicate substantial damage to the aircraft, with a further 22% indicating the aircraft was destroyed and only 3% listing minor damage. The make of the aircraft is of particular importance, since the top three (accounting for 55% of accidents, and 53% of fatal accidents) are all small airplane manufactures: Cessna, Piper, and Beech. These companies typically produce small aircraft for 4-6 people, and have a max-capacity of between 11 and 14 depending on the company. The high proportion of fatal cases being single- and double-fatalities also indicates that small planes are the riskiest. Although visible weather conditions are a factor in many accidents (89% of all accidents vs. 70% in fatal accidents), instrumental weather conditions appear to lead to fatalities, as the weather was listed as obstructing instruments in 21% of fatal cases (as opposed to 7% for all accidents). Finally, for all accidents, landing and takeoff comprise a plurality of incident-times (43%). However, for fatal accidents, the most common phase of flight is when the pilot is maneuvering, cruising, or taking off.

Examination of the textual data reveals a few salient themes[[1]](#footnote-1). The most common words in the narrative data (seen in figure 4) were “pilot”, “airplane/aircraft”, “accident”, “landing”, and either “engine” or “fuel.” Figure 5 shows that the most common words in the probable cause records were for “pilot’s”, “failure”, and “landing”, indicating that pilot error during landing may be a prominent feature in accidents. In fact, the top word for narrative text by year varies by little, with pilot being the top word in the 90’s while airplane is the most popular word in the 2000’s and beyond. This trend is seen in the probable cause, where “pilot’s” (e.g. pilot’s failure to.., or pilot’s loss of…) is the most commonly occurring over the years.

For topic model analysis, I chose to model 10 topics to gain a broad sense of themes in the data. In the narrative text, theme 1 focuses on pilot accidents due to weather (evidenced by “weather” and “conditions”) as seen in Figure 6. Theme 2 touches on mechanical issues, with “fuel”, “engine”, “tank”, “power”, and “tanks” weighted heavily. Theme 4 depicts landing gear issues, as the common words are “gear” and “landing”. Theme 6 may account for student/instructor accidents, with accidents common during landing. Lastly, theme 10 appears to portray helicopter accidents, as “helicopter”, “rotor”, and “ground” are commonly mentioned.[[2]](#footnote-2)

The probable cause topic models reveals similar themes but are more poignant as to the conditions present during the accident. Topic 1 appears to indicate a failure in landing figure (seen in Figure 7). Topic 2 addresses weather conditions. Topic 4 describes fuel loss or failure of engine power (potentially due to pre-flight oversight). Theme 7 appears to depict a failure to maintain airspeed resulting in a stall. In both cases, the topic models add richness not present in the traditional structural data.

Thus, although this dataset would be ill-fitted for predictive modeling (given the lack of variation – i.e. all cases are accidents and the universe/sample contains no non-accident cases for comparison), it can provide a good understanding of risk factors present during flights. The “riskiest” flights appear to occur during inclement weather (specifically weather that affects instruments); a finding born out in the structural data as well as in the narrative text and the probable cause. Structural data tells us that small planes carrying few passengers may be risky as well, as these low-fatality cases account for the overwhelming majority of fatal accidents. Additional data-collection efforts to include non-accidental flights, however, would be a valuable proposition in order to predict risky flights.

Figures

Figure 1 Figure 2

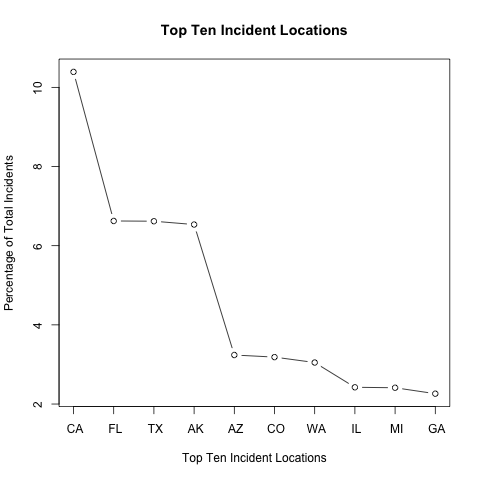
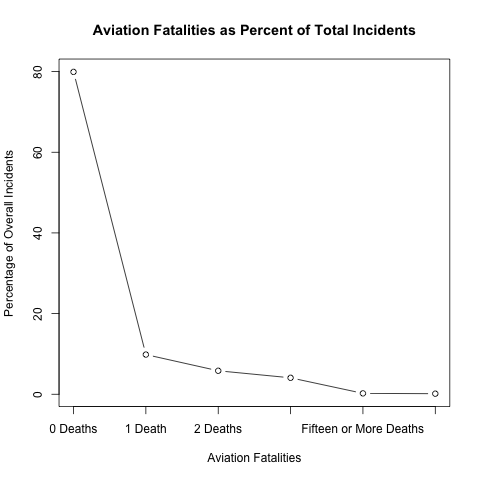


Figure 3 Figure 4



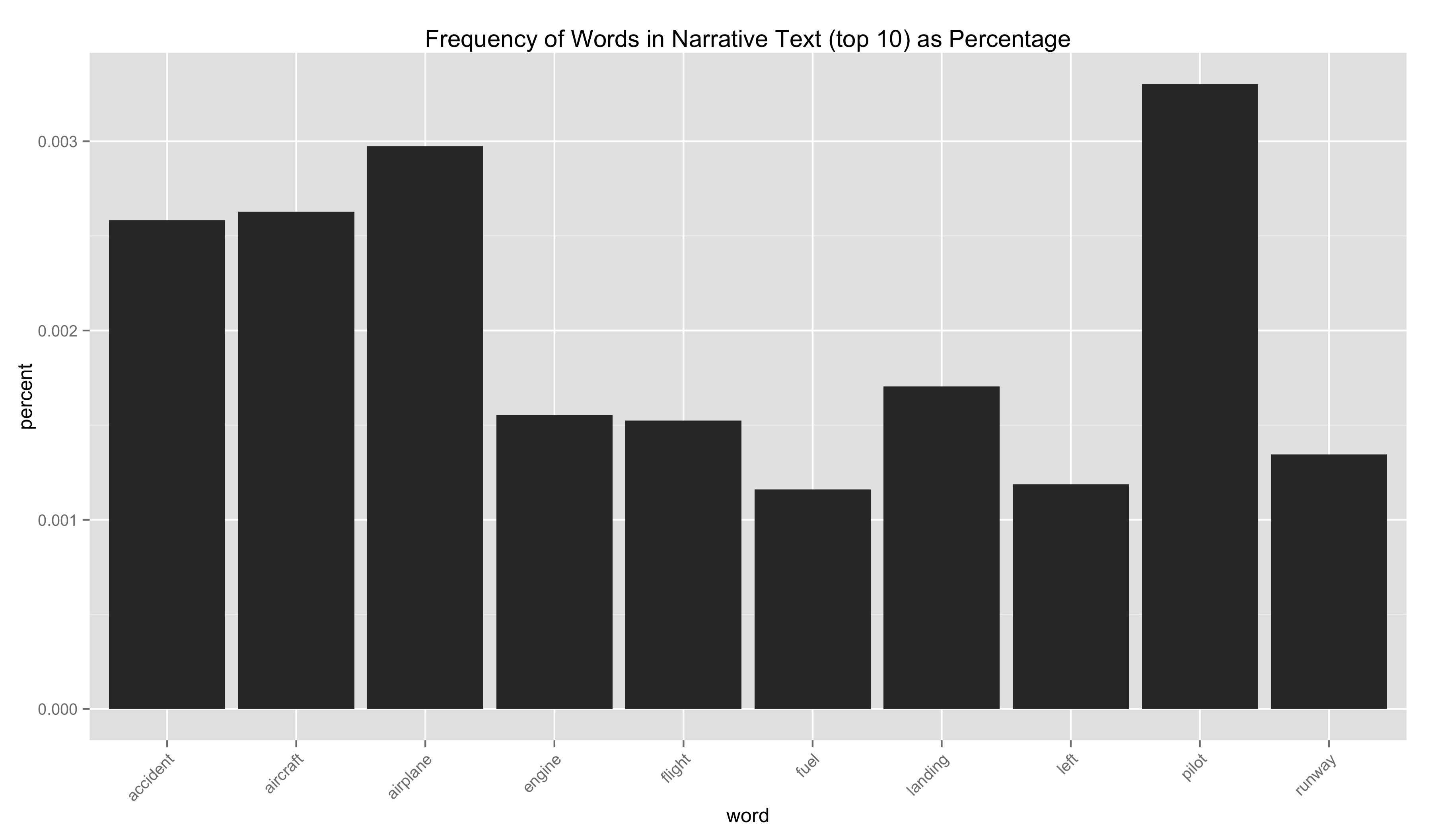


Figure 5

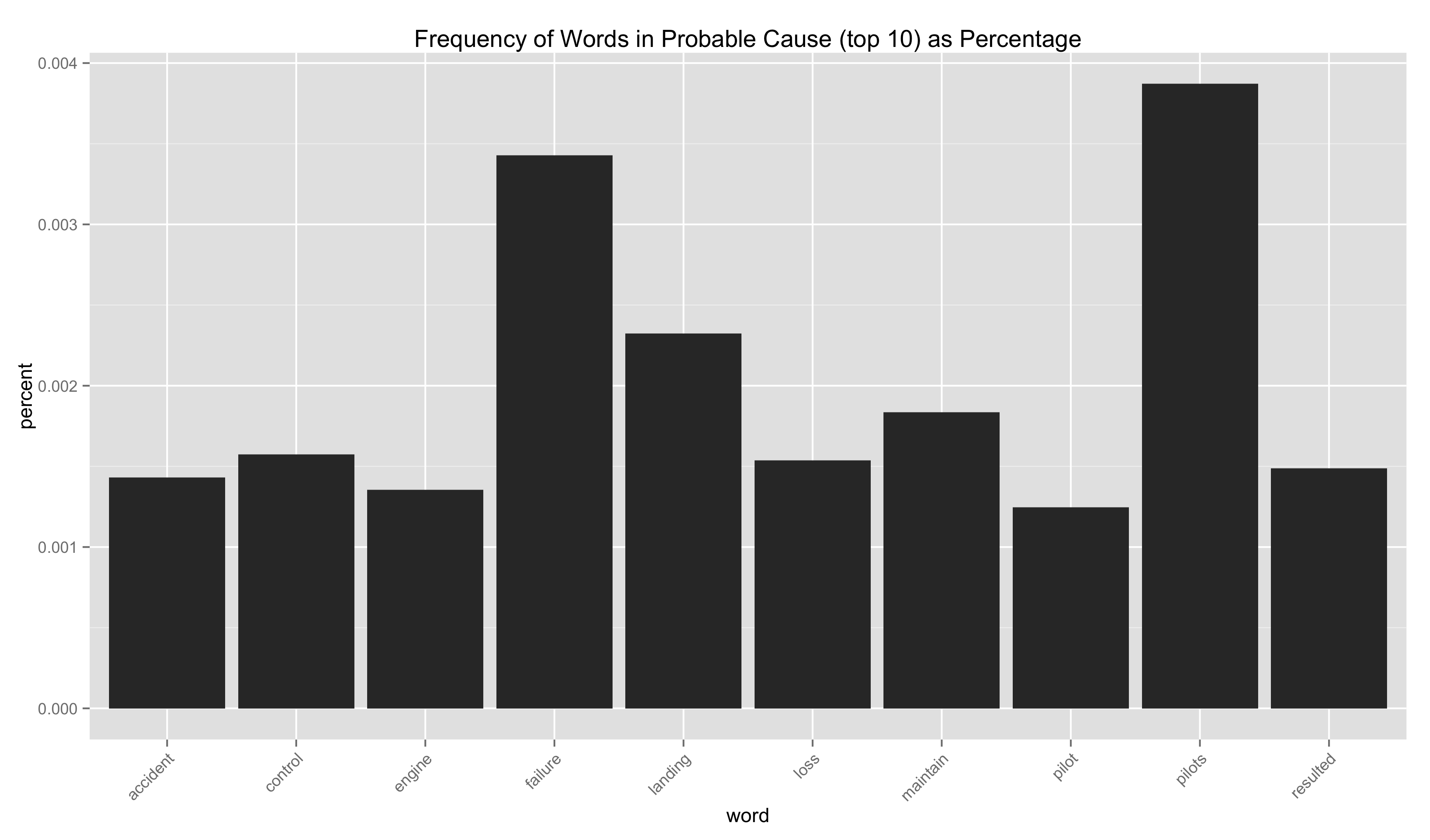
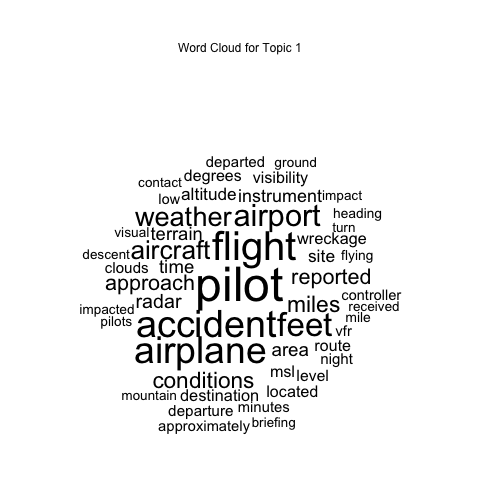
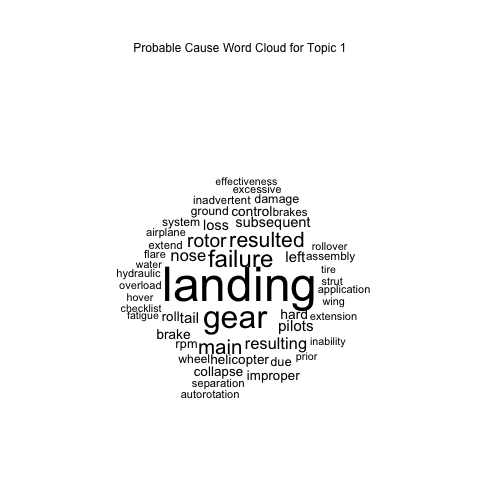


Figure 6 Figure 7



1. A stop-words list from literature analysis was augmented to omit commonly occurring words in the data (e.g. NTSB, source, investigation) in order to remove jargon words present in nearly every record. [↑](#footnote-ref-1)
2. WordCloud visualization for all topics is available upon request (and in the Github fork to my personal account). [↑](#footnote-ref-2)