Aircraft Crash Analysis by Word Cloud

Gursev Pirge and Meryem Vildan Sarikaya

This article involves the use of Natural Language Processing (NLP), with the target of analyzing the causes of airplane accidents between 1969 and 2009. We used the data set provided by data.world, which is a detailed database about the airplane crashes and gives the opportunity to make an in-depth analysis for anyone interested in the subject. As it was mentioned in the previous paper, the data started from 1908, but we decided to analyze the modern era of flight in order to reflect the effectiveness of the modern-day aerospace safety standards.

Wikipedia's definition is: "Natural language processing (NLP) is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, in particular how to program computers to process and analyze large amounts of natural language data. The result is a computer capable of "understanding" the contents of documents, including the contextual nuances of the language within them. The technology can then accurately extract information and insights contained in the documents as well as categorize and organize the documents themselves."

A tag cloud (word cloud or weighted list in visual design) is a novelty visual representation of text data. Bigger term means greater weight. Word Cloud provides an excellent option to analyze the text data through visualization in the form of tags, or words, where the importance of a word is explained by its frequency.

Some domain knowledge here, Wikipedia's definition for an aviation accident is: "An aviation accident is defined by the Convention on International Civil Aviation Annex 13 as an occurrence associated with the operation of an aircraft, which takes place from the time any person boards the aircraft with the intention of flight until all such persons have disembarked, and in which a) a person is fatally or seriously injured, b) the aircraft sustains significant damage or structural failure, or c) the aircraft goes missing or becomes completely inaccessible."

According to Federal Aviation Administration (FAA), top 10 leading causes of fatal general aviation accidents 2001-2016 are:

1. Loss of Control Inflight,
2. Controlled Flight into Terrain,
3. System Component Failure – Powerplant,
4. Fuel Related,
5. Unknown or Undetermined,
6. System Component Failure – Non-Powerplant,
7. Unintended Flight in IMC (Instrument Meteorological Conditions),
8. Midair Collisions,
9. Low-Altitude Operations,
10. Other.

Another approach summarizes the reasons for airline disasters as follows:

1. Pilot Error – 50%.
2. Mechanical Failure – 20%.

3. Weather – 10%.
4. Sabotage – 10%.
5. Human Error (other) – 10%>

We used Word Clouds with the aim of concentrating on the following issues:

- What are the major causes of the aviation accidents?
- Which operators have the highest accident rates?
- Is the time of the day a parameter?
- Which planes have the worst accident statistics?

Data Cleaning

The first step is to import and clean the data (if needed) using pandas before starting the cluster analysis.

```
df1 = pd.read_csv('Airplane_Crashes_and_Fatalities_Since_1908.csv')
```

```
df1.shape
```

```
(5268, 13)
```

```
df1.sample(5)
```

	Date	Time	Location	Operator	Flight #	Route	Type	Registration	cn/ln	Aboard	Fatalities	Ground	Summary
3985	01/10/1991	18:00	Near Paramo Mucuti, Venezuela	Military - Venezuelan Navy	NaN	Caracas - Merida	CASA 212 Aviocar 200	ARV-0209	264	22.0	21.0	0.0	Crashed into Paramo Mucuti Mountain, 35 miles ...
719	07/11/1945	NaN	Near Kisumu, Kenya	Military - South African Air Force	NaN	NaN	Douglas C-47	6812	NaN	28.0	28.0	0.0	Crashed shortly after take off into Lake Victo...
2207	03/27/1968	10:31	Near Moscow, Russia	Military - Russian Air Force	NaN	NaN	MiG-15 UTI	NaN	NaN	2.0	2.0	0.0	Soviet cosmonaut Yuri Gagarin, 34, the first m...
986	01/11/1949	NaN	Near Pelotas, Brazil	Viacao Aerea Gaucha S.A.	NaN	Porto Alegre - Sao Borja	Lockheed 18 Lodestar	PP-SAC	NaN	12.0	12.0	0.0	The aircraft crashed shortly after taking off....
3531	01/23/1985	10:35	Near Buga, Colombia	AIRES Colombia	NaN	Florencia - Cali	Embraer 110P1 Bandeirante	HK-2638	110341	17.0	17.0	0.0	The aircraft crashed into a mountain at 8,500 ...

5268 accidents were recorded in the database with some nulls. Summary column has 390 NaNs and these rows were dropped.

```
df1.isnull().sum()
```

```
Date              0
Time           2219
Location         20
Operator         18
Flight #       4199
Route          1706
Type             27
Registration    335
cn/In          1228
Aboard           22
Fatalities       12
Ground           22
Summary         390
hour              0
```

The next steps were forming two new columns; ‘year’ and ‘hour’. ‘year’ was extracted from the ‘Date’ column and ‘hour’ was taken from the ‘Time’ column.

```
df1['hour'] = df1['Time'].astype(str).str[0:2]
```

```
df['year'] = df.Date.str.extract(r'([0-9][0-9][0-9][0-9])', expand=True)
```

```
df = df1[['Summary', 'Type', 'Operator', 'Location', 'Date', 'hour']]
```

After that, a new data frame was generated using only the columns, which will be necessary for further analysis.

```
df['year'] = df['year'].astype(str).astype(int)
```

```
df = df[df['year'] >= 1969]
```

```
df.sample(5)
```

	Summary	Type	Operator	Location	Date	hour
866	Went into a turn and lost control, spiraled in...	Lockheed L-049 Constellation	Trans Continental and Western Air	Delaware Bay, New Jersey	05/11/1947	09
5164	Soon after taking off the plane began loosing ...	de Havilland Canada DHC-6 Twin Otter 300	Air Moorea	Off Moorea, French Polynesia	08/09/2007	12
1562	Crashed and exploded on flat terrain easts of ...	Fairchild C-123 Provider	Military - U.S. Air Force	Payette, Idaho	10/09/1958	na
4745	Shortly after departing Perth and after the ai...	Beechcraft SKA 200	Charter - Central Air	Near Burketown, Australia	09/04/2000	15
2539	Crashed in mountainous terrain, 50 miles nort...	NAMC YS-11A-211	VASP	Near Rio de Janeiro, Brazil	04/12/1972	21

Text Mining

Now that the data frame is ready for the further steps, next stage was to prepare the ‘summary’ column for Word Cloud. The function below performs;

- word tokenizing,
- getting rid of special characters and punctuations,
- removing the stop words,
- lemmatizing,
- joining all the words in the summary parts of the reports.

```
def cleaning(data):

    #1. Tokenize
    text_tokens = word_tokenize(data.lower())   #removed the .lower intentionaly to keep NNP s

    #2. Remove Puncs
    tokens_without_punc = [w for w in text_tokens if w.isalpha()]

    #3. Removing Stopwords
    tokens_without_sw = [t for t in tokens_without_punc if t not in stop_words]

    #4. lemma
    text_cleaned = [lem.lemmatize(t) for t in tokens_without_sw]

    #joining
    return " ".join(text_cleaned)
```

```
df['Summary'] = df['Summary'].apply(cleaning)
```

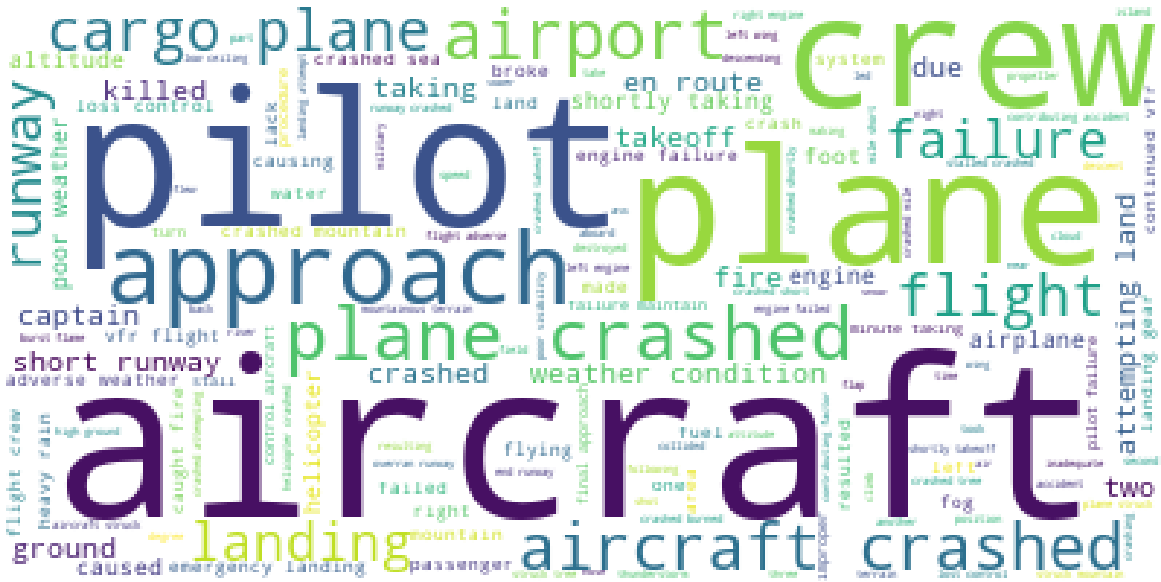
```
df.head()
```

	Summary	Type	Operator	Location	Date	hour	year
1807	crashed approaching land	de Havilland DH-125-1A	Avionas Banamex	Acapulco, Mexico	10/12/1973	na	1973
1965	crashed mountain	Lockheed C-130E Hercules	Military - U.S. Air Force	Fort Smith AFB, Oklahoma	10/15/1973	na	1973
2077	crashed high ground descending	Antonov AN-24	Aeroflot	Near Sukhumi, Georgia, USSR	11/17/1975	na	1975
2243	cargo plane went missing en route	Douglas DC-4	Continental Air Transport	AtlantiOcean	03/08/1969	na	1969
2279	crashed short runway attempting land	Lockheed C-130E Hercules	Military - U.S. Air Force	Ching Chuan Kang AB, Taiwan	03/08/1969	na	1969

Word Cloud – Report Summaries

We know that a word cloud is a collection of words depicted in different sizes. The bigger and bolder the word appears, the more often it is mentioned within a given text and arguably, the more important it is.

Our approach in analyzing the accident report summaries is based on finding the most frequent words and then preparing the Word Cloud in order to get the results in a visual form.



The wordcloud above had the potential to cause misconceptions because the most emphasized words are expected in every accident report but did not tell anything about the causes. The code below shows the most frequent words in the reports, and we removed the first three words:

```
wordcloud.to_file("WC_air_summary3_aircraft-plane-crashed-500words.png")
```

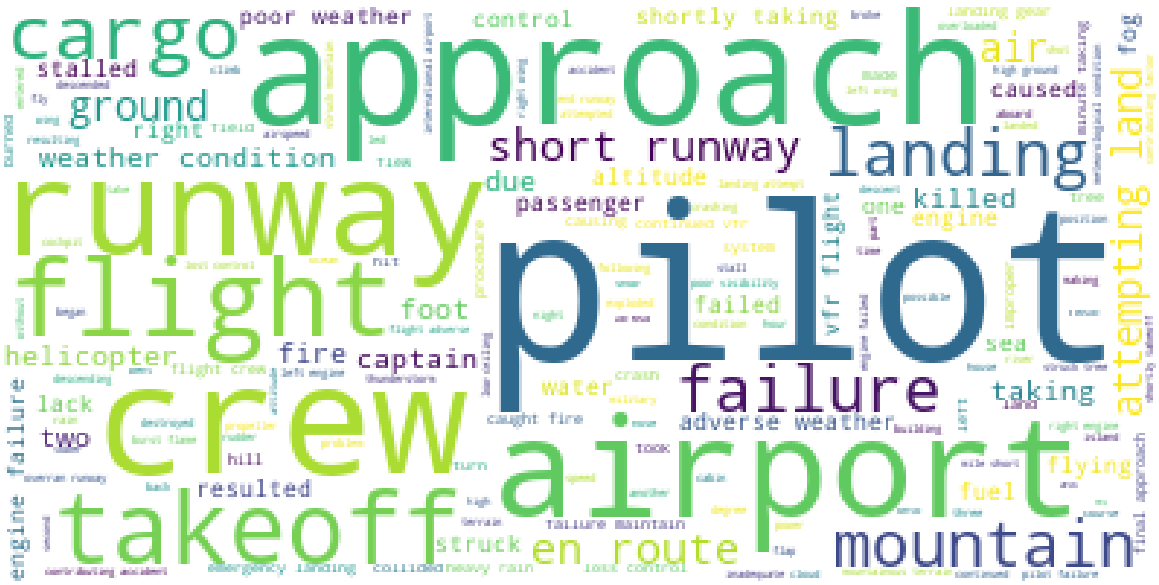
```
<wordcloud.wordcloud.WordCloud at 0x269bbc6ba48>
```

```
from collections import Counter
most_common_words = Counter(" ".join(df["Summary"]).split()).most_common(20)
most_common_words
```

```
[('crashed', 1926),
 ('aircraft', 1515),
 ('plane', 1282),
 ('pilot', 877),
 ('flight', 743),
 ('runway', 692),
 ('engine', 673),
 ('approach', 620),
 ('crew', 615),
 ('failure', 585),
 ('landing', 475),
 ('airport', 442),
 ('mountain', 400),
 ('weather', 392),
 ('condition', 362),
 ('control', 359),
 ('taking', 357),
 ('altitude', 355),
 ('takeoff', 351),
 ('land', 337)]
```

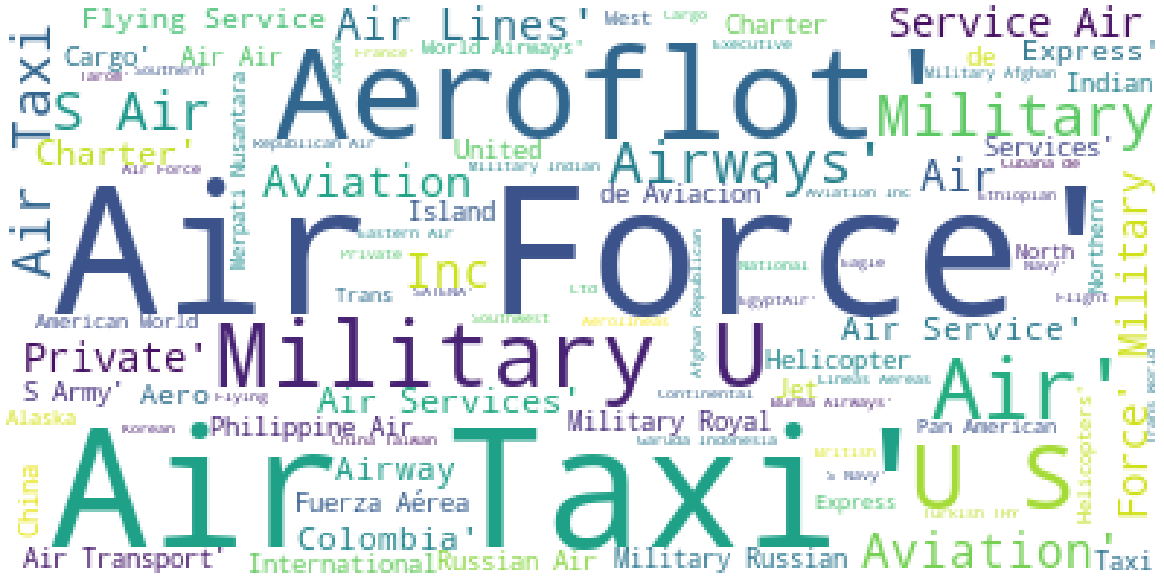
```
reports = reports.replace('crashed', '')
reports = reports.replace('aircraft', '')
reports = reports.replace('plane', '')
```

The new wordcloud below is a more accurate picture of the accident reports:



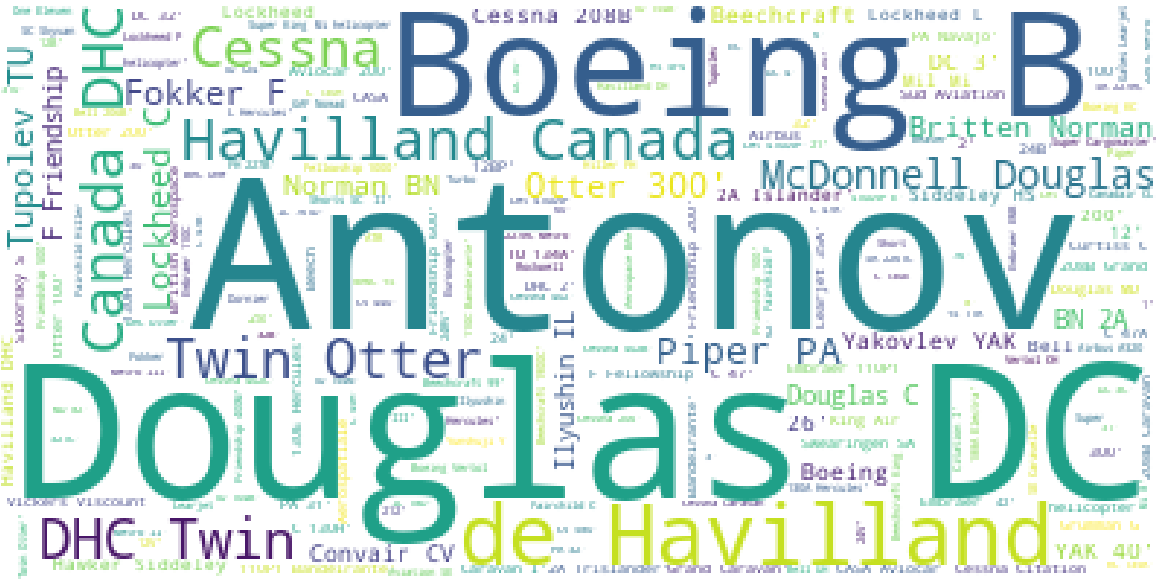
Word Cloud – Operators

Some operators have a long record of accidents. The wordcloud below gives an idea about the riskiest operators in the world. One thing to keep in mind is the potential high risk of military flights.



Word Cloud – Type

The wordcloud below shows the airplane manufacturers. The result may be considered misleading, as some of the companies has been in the market for almost a hundred years and thousands of their planes are flying millions of hours annually.



Word Cloud – Location

The wordcloud below was prepared by using the 'Location' column in the data frame. It should be kept in mind that Russian airplanes have a long history of accidents and covering an expanse of over 6.6 million square miles, Russia is the world's largest country by landmass.

Word Cloud – Hour of the Day

The final wordcloud was prepared with the intention of investigating the riskiest time frame of the day. The figure below gives a poor/insufficient answer to this question.

Conclusion

In this article, we used wordcloud as a data visualization technique and tried to analyze aircraft accidents between 1969 and 2009. The analysis provides evidence that:

- Our results proved to be similar to FAA's and similar sources' claims that put pilot and other human factors at the top of the list, followed by mechanical problems and weather.
- Take-off, approach and landing are the dangerous phases of flights.
- There were a lot of accidents in the largest countries and territories.
- The riskiest hours for accidents are 9 a.m. and 7 p.m.

You can access to this article and similar ones here.

Photo by BOLIEK MEDIA on Unsplash