SAS® GLOBAL FORUM 2018

USERS PROGRAM

Text Analysis and Cluster Analysis of Airplane Crashes from 1908 to 2009

Ritesh Kumar Vangapalli

MS in Business Analytics, Oklahoma State University

April 8 – 11 | Denver, CO

#SASGF

Text Analysis and Cluster Analysis of Airplane Crashes from 1908 to 2009

Ritesh Kumar Vangapalli

MS in Business Analytics, Oklahoma State University

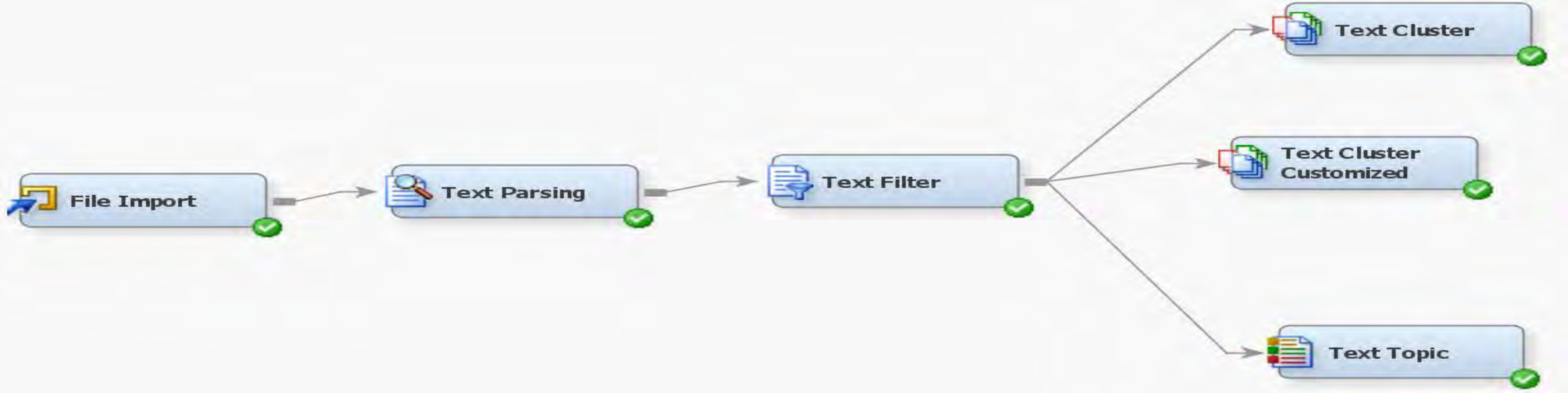
Abstract

A flight in a plane is a profoundly exciting experience. It is flying all around in the air like a feathered creature. The entire thing is crazy and brilliant. But there is also a risk in the flying. Though the fact that instances of the plane crash are not particularly normal, they are very fatal. According to the Telegraph news (United Kingdom), the odds of deaths of a person per total number of passengers flown is 1 to 6 million. Although the year 2015 is considered as the safest year in the aviation history, there are 16 fatal crashes leading to the death of 560 passengers flown. It is followed by 19 fatal crashes leading to the death of 325 passengers flown in the year 2016. Even though the aviation giants took many precautions to control these fatalities incidents are still being reported.

The objective of this paper is to cluster these fatalities into several different segments based on text summary. The text is released by the government after the crash is reported. Finding the major reason associated for these casualties based on this summary is the secondary objective for this paper. I also identified the fatalities by the phase of flight, the cause of fatal airplane crashes and found the number of crashed aircrafts and number of deaths against each category of these segments. Classifying them into different segments based on clustering of the summary of events beforehand helps the aviation giants to take necessary care and precautions which decreases the casualties of airplane crashes and increases the survival rate of these incidents reported. An open dataset by open data from Kaggle containing 5268 airplane crashes with fatalities of 105k is used for this paper. SAS Enterprise Miner and python are also for this analysis

Methodology

Extracted the data from Kaggle and scraped some of the missing events summary from websites using python. After the data cleaning, created new variables for the classification of text into different clusters. This is the methodology used for importing the data, parsing the data using online updated dictionary. Then data is filtered and customized text clustering and text topic building is done.



Project Cycle

- Collecting and Identifying the data
- Cleaning the data/ Removing the unwanted text from the data
- Parsing the data and identifying the most mistaken words
- Filtering the data using the user defined dictionary
- Text Clustering (Customizing into 7 different Clusters)
- Text Topic Building

Data Preparation

- Identified a data set from Kaggle airplane crashes from 1908 to 2009 which have a rows of 6000.
- Scraped the summary data from google using the beautiful soup package on python.
- Also, Identified a dataset from Stanford datasets to validate the some of the records present in the final considered dataset.

Data Filtering

- Repeated punctuation sign normalization
- Lower Casing all the text data
- User defined dictionary
- Identified emoticons and replaced them with words.

Text Analysis and Cluster Analysis of Airplane Crashes from 1908 to 2009

Ritesh Kumar Vangapalli

MS in Business Analytics, Oklahoma State University

Results

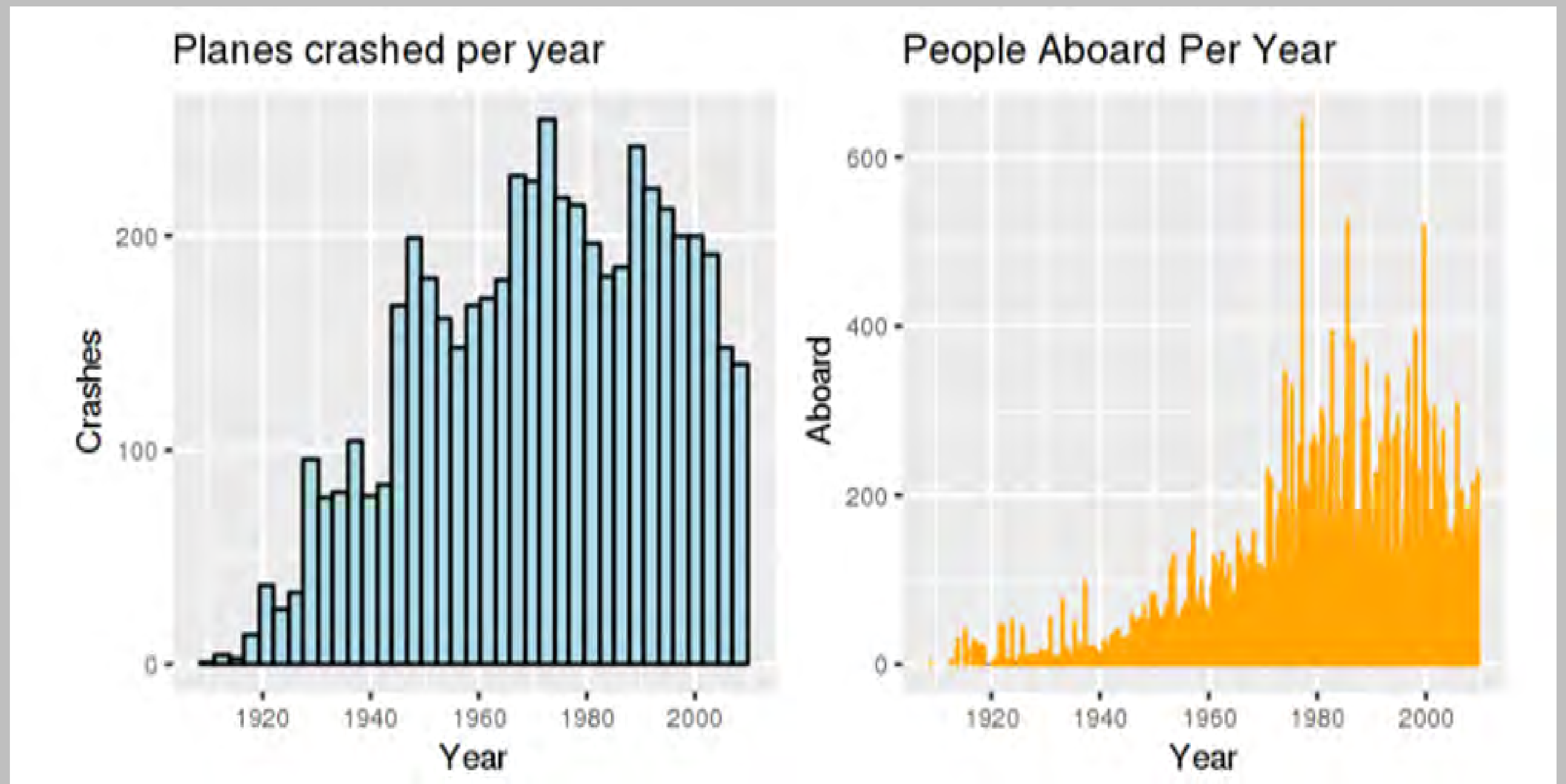
Plane crashes Increased significantly

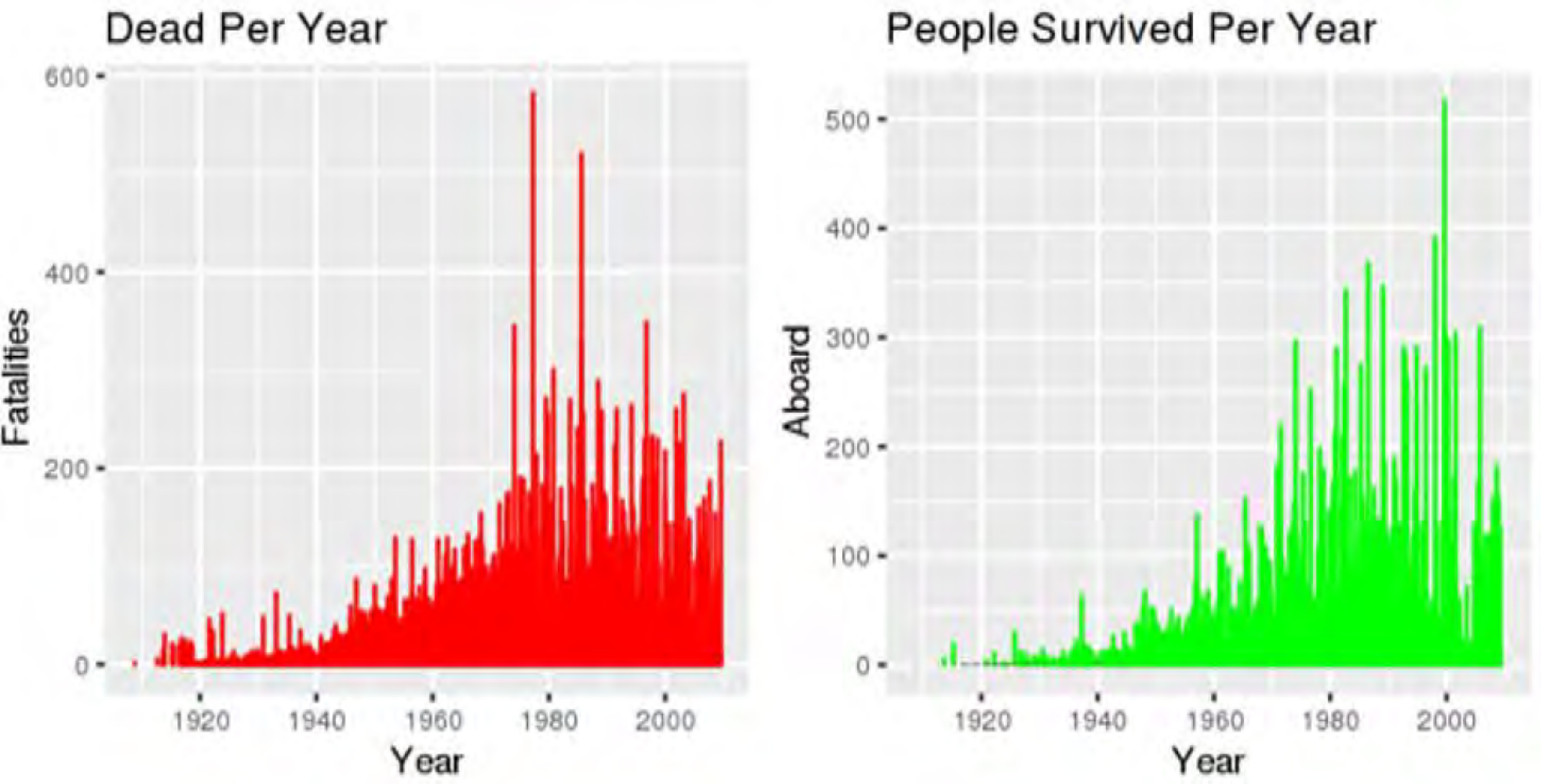
People who are aboard on these fata accidents also increased

Number of people who died on these fatalities increased up to 1990

The people who survived on these accidents also increased as the frequency of these accidents increased in these years

1990 to 2000 is considered as the worst decade for the commercial airliners.

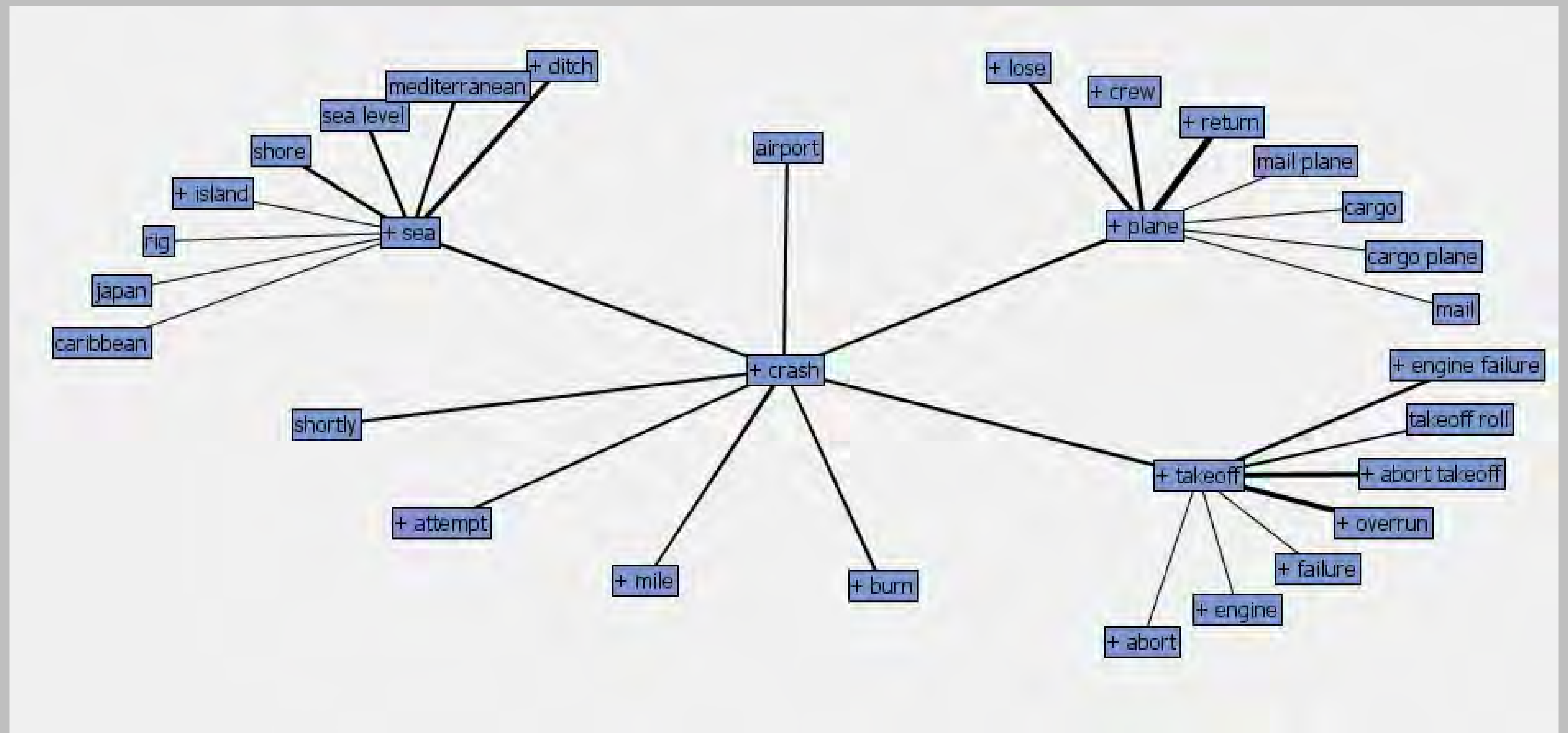




Concept Links

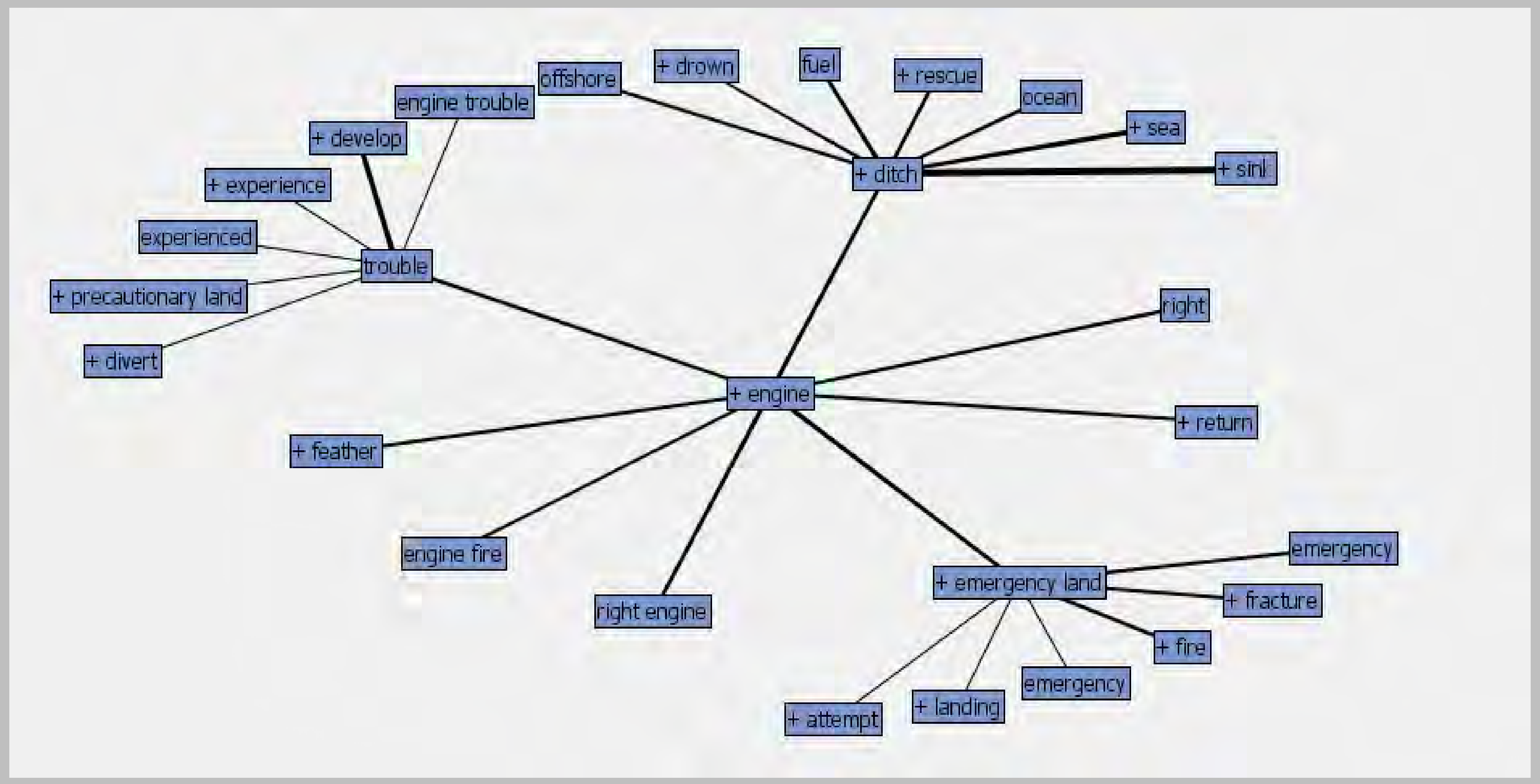
CRASH

- Sea - Island, japan, Caribbean, Mediterranean
- Plane – Mail plane, Cargo, lose, crew
- Take-off – abort, engine failure, overrun, abort takeoff, takeoff roll, engine failure



ENGINE

- Ditch – Offshore, drown, fuel, rescue, ocean, sea, sink
- Emergency land – Landing, fire, fracture, attempt
- Trouble – Engine, divert, experience, precautionary land



Text Analysis and Cluster Analysis of Airplane Crashes from 1908 to 2009

Ritesh Kumar Vangapalli

MS in Business Analytics, Oklahoma State University

Cluster Hierarchy

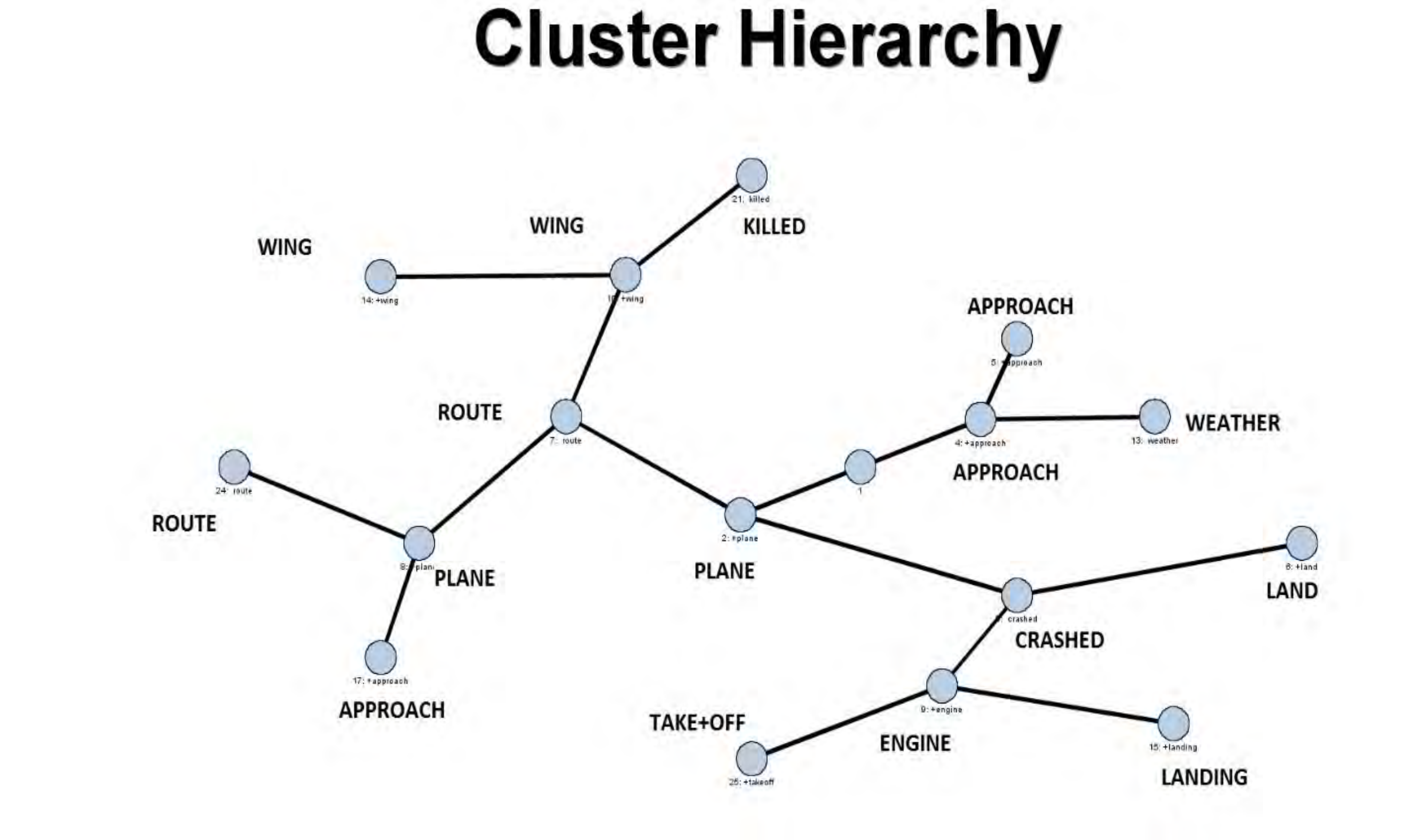
The cluster Hierarchy of the text using the Hierarchical clustering algorithm

The initial nodes are connected to Plane and Approach

Plane is connected to Route, and it is connected to wing and Plane again

Wing is connected to Kill

Approach is connected to Weather.



Cluster1: Explosion and destruction of the fuselage

Cluster2: The plane was hijacked or captured by the rebels

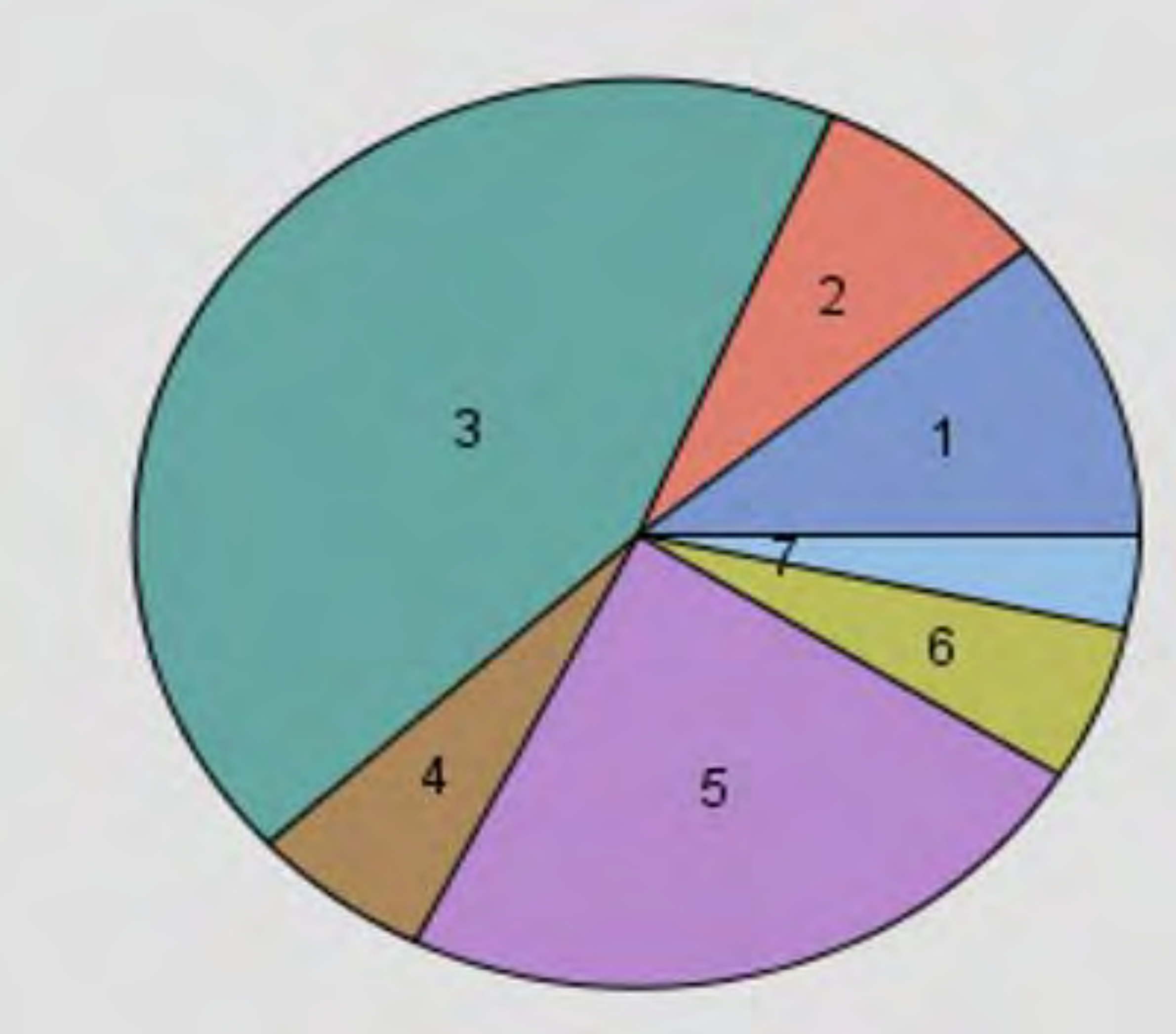
Cluster3: Stalled the engine, Collision with the Mountains

Cluster4: Bad weather conditions: strong wind, snow, ice

Cluster5: Taking off without clearance from ATC. ATC or pilots error

Cluster6: Crash due to manoeuvring

Cluster7: Technical Malfunction



Conclusion

Collecting more about this incidents, having a detailed study of these texts will help them cluster these events and avoid these incidents in the future. This poster briefly explain about classifying the texts into different type of fatalities. Considering a detailed study of these incidents and avoiding the problems which were faced before avoids Fatal incidents in the future.

Acknowledgment

I wish to express my sincere gratitude to Dr. Goutam Chakraborthy for his guidance for accomplishing this paper

.

I sincerely thank Dr. Miriam McGaugh for her constant support and encouragement.

References

https://www.kaggle.com/saurograndi/airplane-crashes-since-1908

https://support.sas.com/resources/papers/proceedings16/SAS6380-2016.pdf

http://www.expertsystem.com/text-mining-research-papers/

http://ieeexplore.ieee.org/document/7339149/?reload=true

https://link.springer.com/article/10.1186/s13173-017-0058-7

https://www.sciencedirect.com/science/article/pii/S2444883417300268

SAS® GLOBAL FORUM 2018

April 8 – 11 | Denver, CO
Colorado Convention Center

Ritesh Kumar Vangapalli
MS in Business Analytics, Oklahoma State University

#SASGF

Paper 2789-2018

Text & Cluster Analysis of Air Plane crashes from 1908 to 2009

Ritesh Kumar Vangapalli, Oklahoma State University

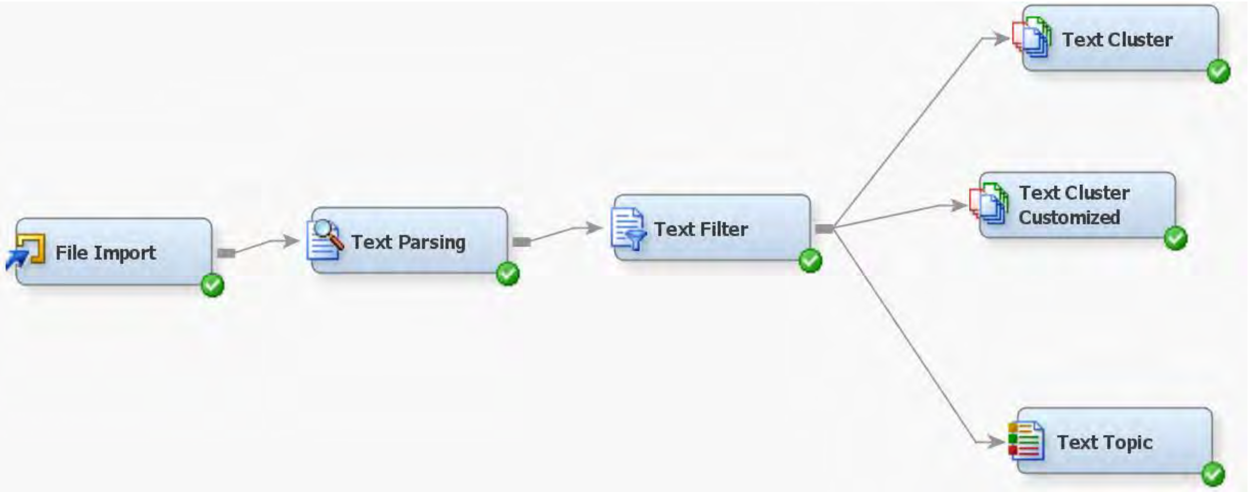
ABSTRACT

A flight in a plane is a profoundly exciting experience. It is flying all around in the air like a feathered creature. The entire thing is crazy and brilliant. But there is also a risk in the flying. Though the fact that instances of the plane crash are not particularly normal, they are very fatal. According to the Telegraph news (United Kingdom), the odds of deaths of a person per total number of passengers flown is 1 to 6 million. Although the year 2015 is considered as the safest year in the aviation history, there are 16 fatal crashes leading to the death of 560 passengers flown. It is followed by 19 fatal crashes leading to the death of 325 passengers flown in the year 2016. Even though the aviation giants took many precautions to control these fatalities incidents are still being reported.

The objective of this paper is to cluster these fatalities into several different segments based on text summary. The text is released by the government after the crash is reported. Finding the major reason associated for these casualties based on this summary is the secondary objective for this paper. I also identified the fatalities by the phase of flight, the cause of fatal airplane crashes and found the number of crashed aircrafts and number of deaths against each category of these segments. Classifying them into different segments based on clustering of the summary of events beforehand helps the aviation giants to take necessary care and precautions which decreases the casualties of airplane crashes and increases the survival rate of these incidents reported. An open dataset by open data from Kaggle containing 5268 airplane crashes with fatalities of 105k is used for this paper. SAS Enterprise Miner and python are also for this analysis.

METHODOLOGY

Extracted the data from Kaggle and scraped some of the missing events summary from websites using python. After the data cleaning, created new variables for the classification of text into different clusters. This is the methodology used for importing the data, parsing the data using online updated dictionary. Then data is filtered, and customized text clustering and text topic building is done.



PROJECT CYCLE

Collecting and Identifying the data
Cleaning the data/ Removing the unwanted text from the data
Parsing the data and identifying the most mistaken words
Filtering the data using the user defined dictionary
Text Clustering (Customizing into 7 different Clusters)
Text Topic Building

DATA PREPARATION

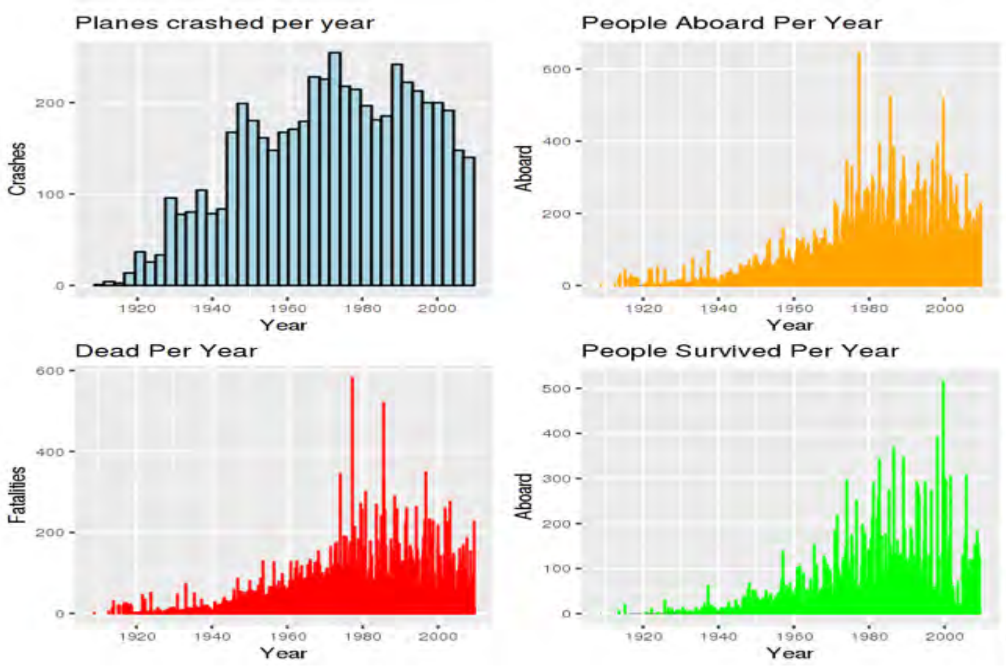
Identified a data set from Kaggle airplane crashes from 1908 to 2009 which have some rows of 6000. Scraped the summary data from google using the beautiful soup package on python. Also, identified a dataset from Stanford datasets to validate the some of the records present in the final considered dataset.

DATA FILTERING

Repeated punctuation sign normalization
Lower Casing all the text data
User defined dictionary
Identified emoticons and replaced them with words

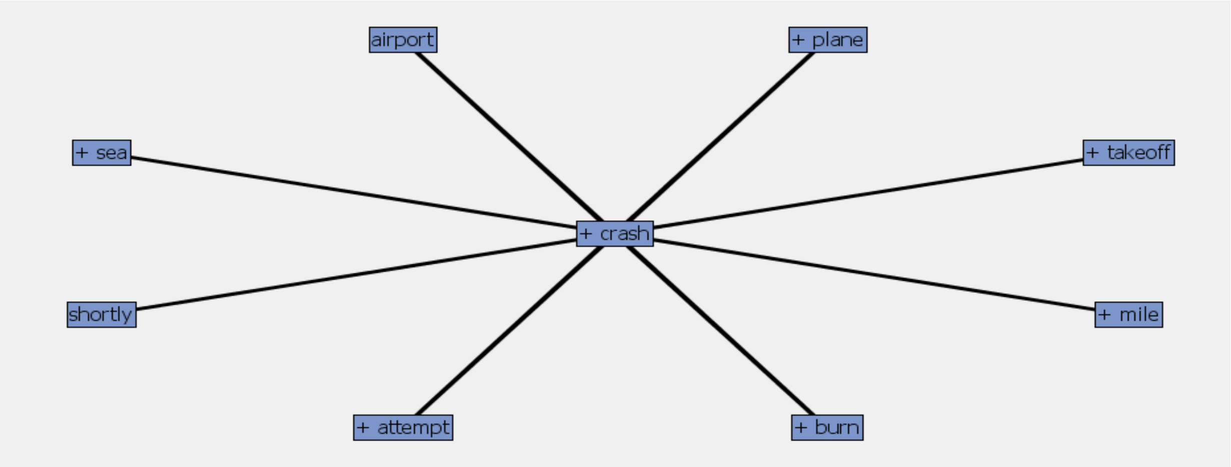
RESULTS

The descriptive statistics of the flights that are crashed per year, the people aboard per year in the flight crash incident, people who are dead per year and the people who survived in the crash. Plane crashes Increased significantly. People who are aboard on these fatal accidents also increased. Number of people who died on these fatalities increased up to 1990. The people who survived on these accidents also increased as the frequency of these accidents increased in these years and 1990 to 2000 is considered as the worst decade for the commercial airliners.

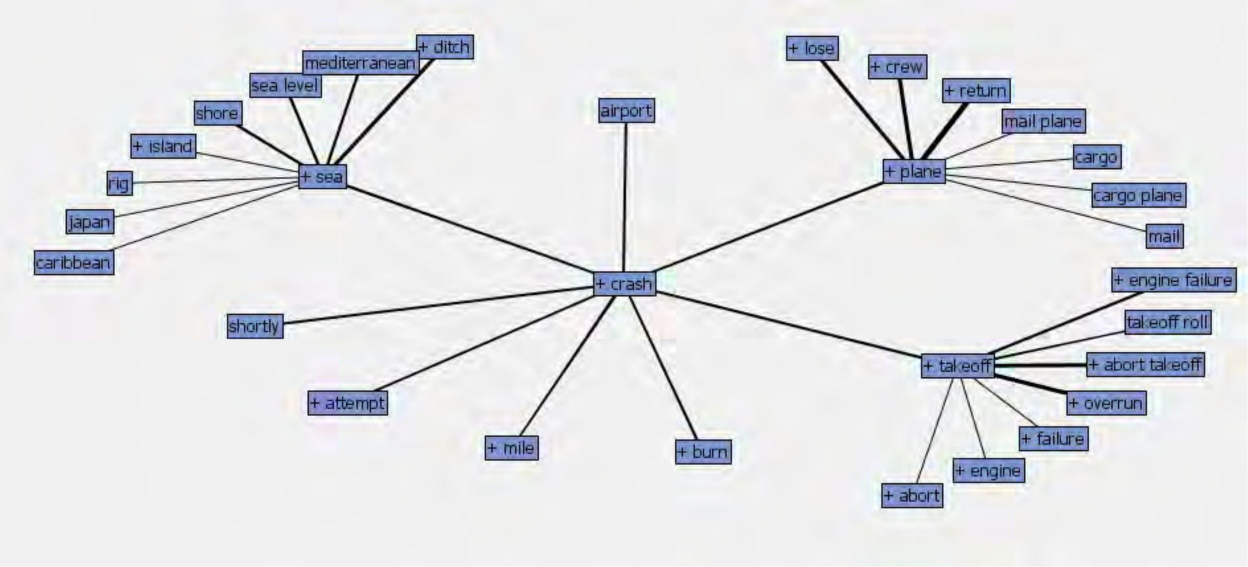


CONCEPT LINKS

The concept links for the keyword **CRASH**.

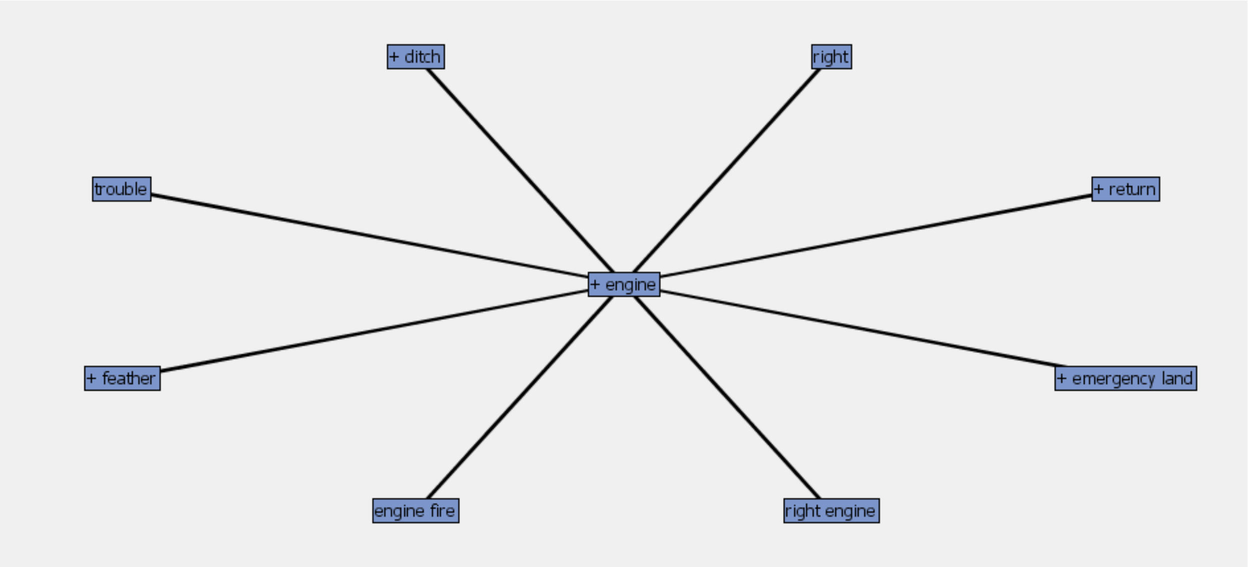


Could extract the concept links of the keywords that are associated with the airplane crash. The word is strongly associated with the keywords attempt, burn, mile, take off etc.

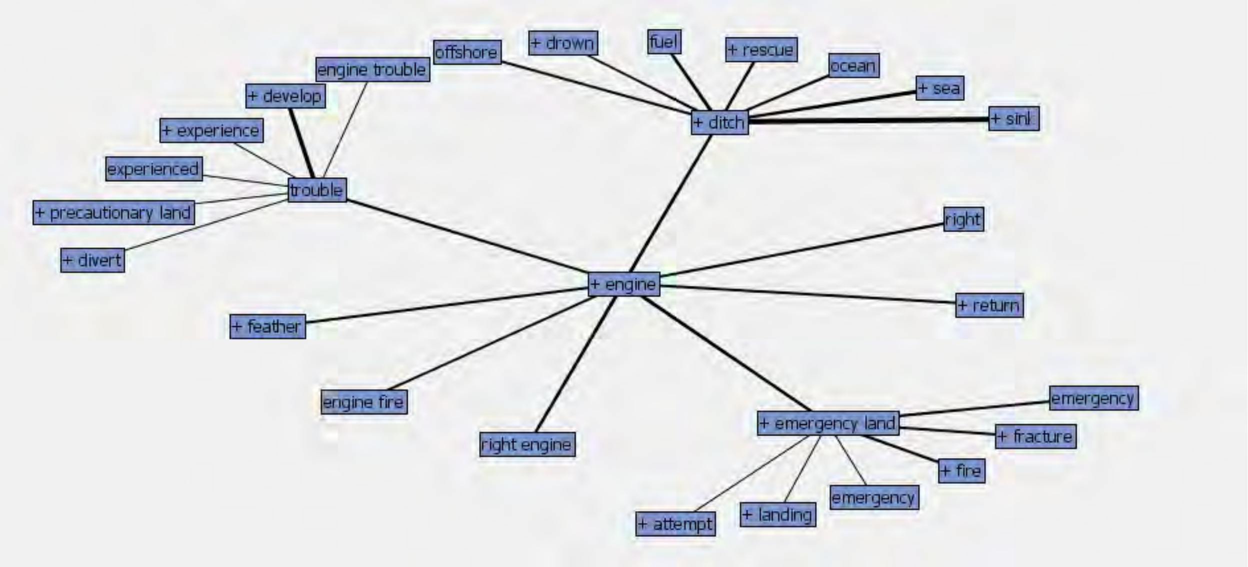


Extending the concept link of **CRASH**, we can see that the words that are connected to crash are interconnected with several different words. Such as
Sea - Island, japan, Caribbean, Mediterranean
Plane – Mail plane, Cargo, lose, crew
Take-off – abort, engine failure, overrun, abort takeoff, takeoff roll, engine failure

The concept links for the word **ENGINE**.

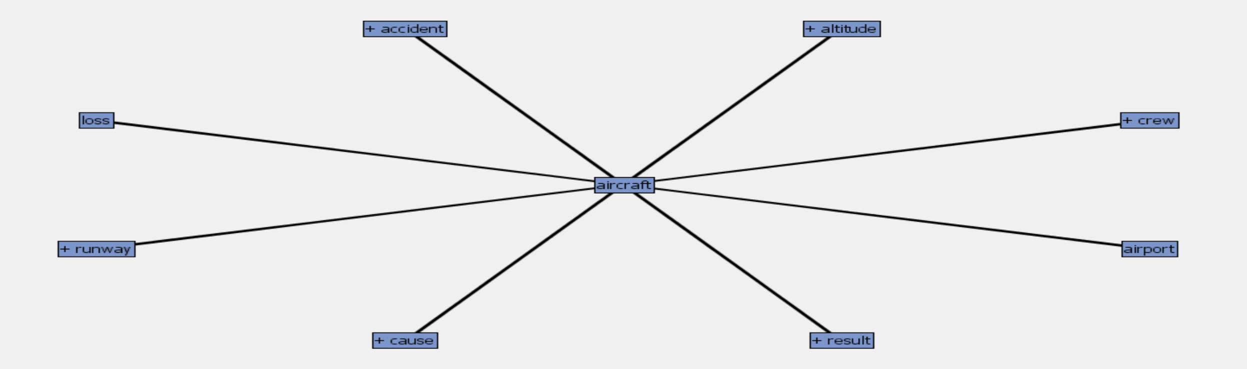


Could extract the concept links of the keywords that are associated with the airplane crash. The word is strongly associated with the keywords ditch, emergency landing, trouble, engine fire etc.



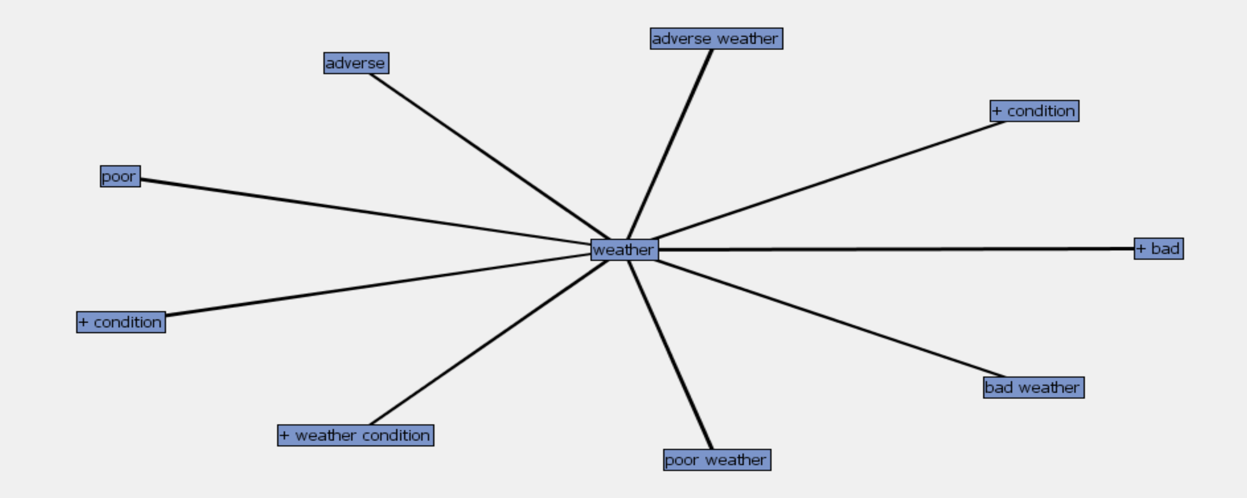
Extending the concept link of **ENGINE**, we can see that the words that are connected to crash are interconnected with several different words. Such as
Ditch – Offshore, drown, fuel, rescue, ocean, sea, sink
Emergency land – Landing, fire, fracture, attempt
Trouble – Engine, divert, experience, precautionary land

The concept link for the word **Aircraft**



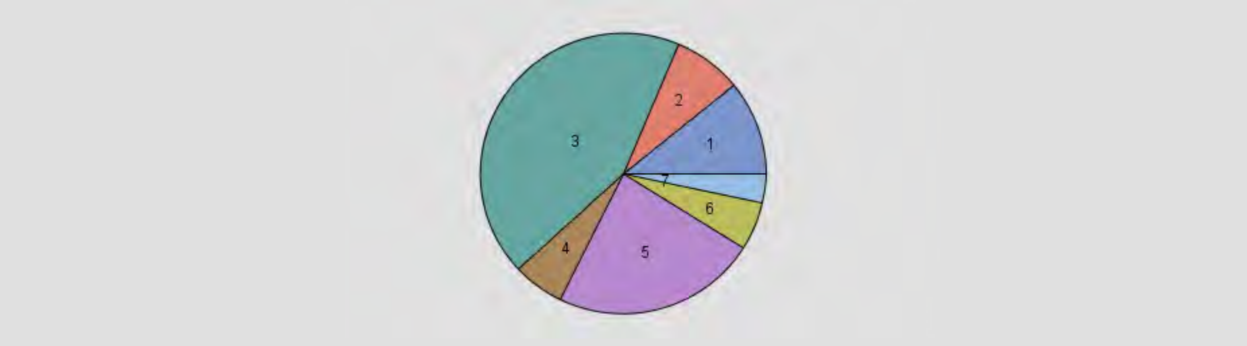
These word is associated with the some of the keywords such as accident, altitude, crew, runaway, cause, result, airport etc. The connecting line between aircraft and airport says that the aircraft has been caught in an incident at the airport itself.

The concept link for the word **Weather**



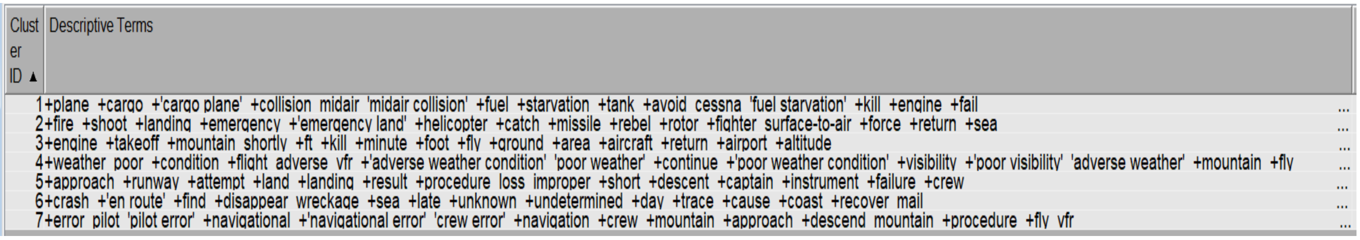
These word is interconnected with condition, poor, adverse weather, poor weather, bad, bad weather, condition. The combination of weather, bad weather and condition says that the condition of the plane was bad to the bad weather.

THE TEXT CLUSTERING OF THE AIRPLANE CRASH INCIDENTS



Identifying the clusters based on the text of the airplane crash summary. Trying to create some good clusters for the different phases of airplane crash and the reason for fatal crash.

CLUSTERS AND THE DESCRIPTIVE TERMS IN THE CLUSTERS



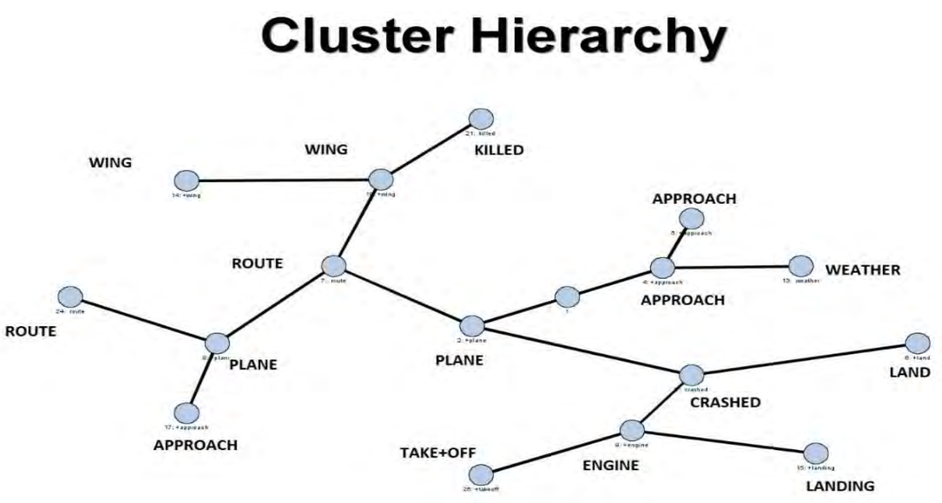
Cluster ID ▲	Descriptive Terms	
1	+plane +cargo +'cargo plane' +collision midair 'midair collision' +fuel +starvation +tank +avoid cessna 'fuel starvation' +kill +engine +fail	...
2	+fire +shoot +landing +emergency +'emergency land' +helicopter +catch +missile +rebel +rotor +fighter surface-to-air +force +return +sea	...
3	+engine +takeoff +mountain shortly +ft +kill +minute +foot +fly +ground +area +aircraft +return +airport +altitude	...
4	+weather poor +condition +flight adverse vfr +'adverse weather condition' 'poor weather' +continue +'poor weather condition' +visibility +'poor visibility' 'adverse weather' +mountain +fly	...
5	+approach +runway +attempt +land +landing +result +procedure loss improper +short +descent +captain +instrument +failure +crew	...
6	+crash +'en route' +find +disappear wreckage +sea +late +unknown +undetermined +day +trace +cause +coast +recover mail	...
7	+error pilot 'pilot error' +navigational +'navigational error' 'crew error' +navigation +crew +mountain +approach +descend mountain +procedure +fly vfr	...

After Analyzing these words, we have categorized cluster into several different categories falling to a subset of the following clusters given below

Cluster1: Explosion and destruction of the fuselage
Cluster2: The plane was hijacked or captured by the rebels
Cluster3: Stalled the engine, Collision with the Mountains
Cluster4: Bad weather conditions: strong wind, snow, ice
Cluster5: Taking off without clearance from ATC. ATC or pilots error
Cluster6: Crash due to maneuvering
Cluster7: Technical Malfunction

We have divided the summary of the text into these different subsets of clustering. So, the main problem of these accidents falls into these categories.

HIERARCHICAL CLUSTERING ALGORITHM



The cluster Hierarchy of the text using the Hierarchical clustering algorithm
The initial nodes are connected to Plane and Approach
Plane is connected to Route, and it is connected to wing and Plane again
Wing is connected to Kill
Approach is connected to Weather

CONCLUSIONS AND FUTURE WORK

Collecting more about these incidents, having a detailed study of these texts will help them cluster these events and avoid these incidents in the future. This poster briefly explains about classifying the texts into different type of fatalities. Considering a detailed study of these incidents and avoiding the problems which were faced before avoids fatal accidents in the future.

REFERENCES

https://www.kaggle.com/saurograndi/airplane-crashes-since-1908
https://support.sas.com/resources/papers/proceedings16/SAS6380-2016.pdf

ACKNOWLEDGMENTS

I wish to express my sincere gratitude to Dr. Goutam Chakraborty for his guidance for accomplishing this paper. I sincerely thank Dr. Miriam McGaugh for her constant support and encouragement.

CONTACT INFORMATION

Ritesh Kumar Vangapalli
rvangap@okstate.edu
LinkedIn: https://www.linkedin.com/in/riteshkumar-vangapalli
Phone: (405)-780-3824