UNSUPERVISED LEARNING – CASE STUDY: AIRPLANE CRASHES
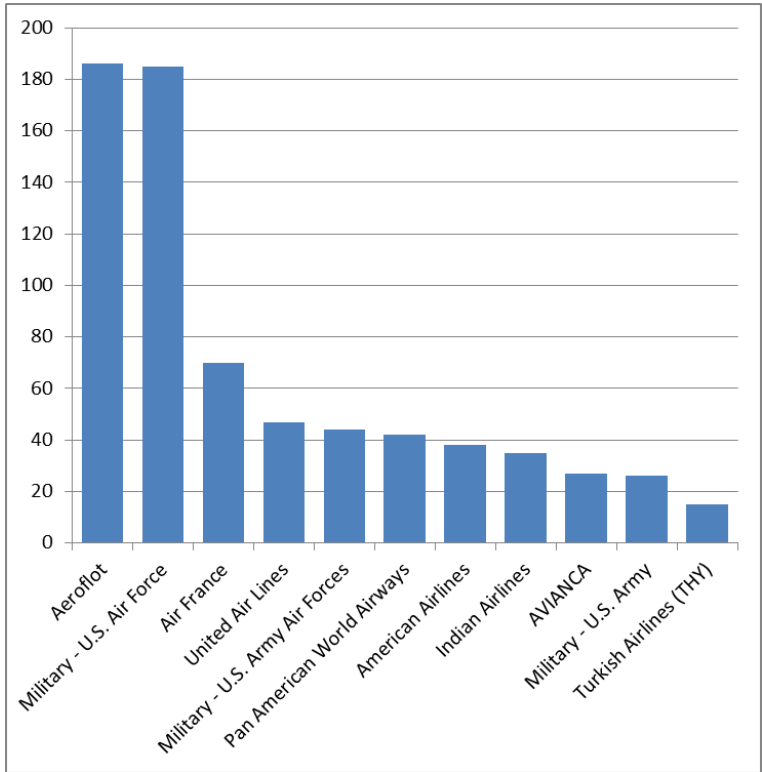
- The data set contains the history of 5268 airplane crashes in the world from 1908 to 2009 (*)

- Fields like Date, Time, Location, Aboard, Fatalities…and Summary

- Objective:
  Extract the main causes of incidents by analysing the 'Summary' field

* Available at:
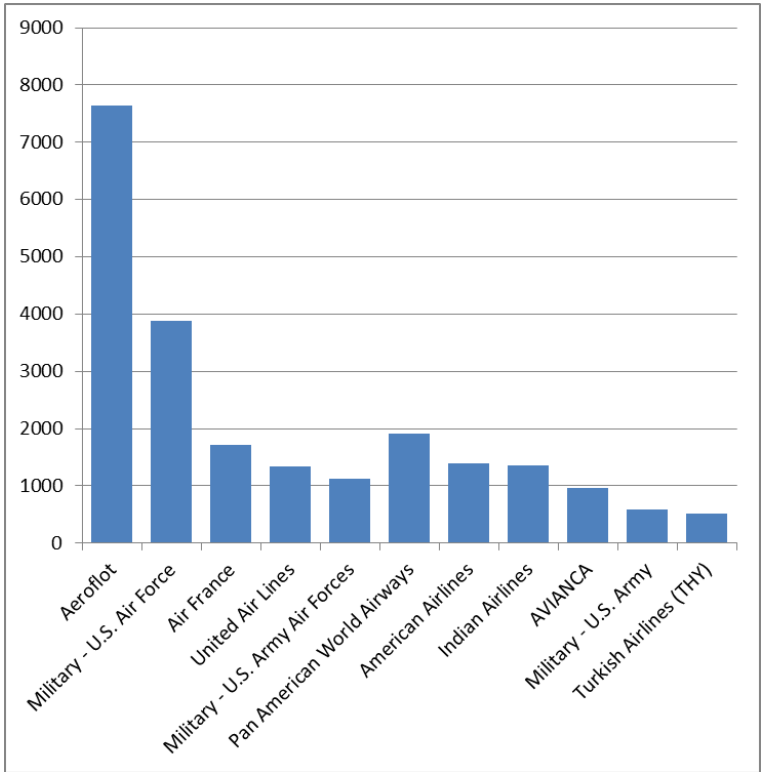https://opendata.socrata.com/Government/Airplane-Crashes-and-Fatalities-Since-1908/q2te-8cvq

Summary
A midair collision occurred between two Black Hawk helicopters. One helicopter was climbing to avoid gunfire when it's rotor struck the other.
All three planes crashed into Devil's Peak and exploded while flying in a V formation in heavy mist.
Crashed 15 ft. below the summit of cloud-obscured Jebel Hilal Mountain on the Israeli-occupied Sinai Peninsula.
Crashed while en route. Twelve Interior Ministers of Poland and Czechoslovakia
The aircraft crashed into a forest, 1.4 miles short of the runway, while attempting to land at Goleniow Airport in rain and poor visibility. Icing. Captain did not turn on dei
The aircraft expierence a midair collision over Corpus Christi Bay while on a training flight.
While attempting to land, the plane fell to the ground from about 500-650 feet and crashed into a wooded area, 2 miles from the airport.
The two aircraft crashed in mid-air in heavy rain and darkness and crashed into the sea.
The aircraft crashed while making an emergency return to the airfield after experiencing engine failure on takeoff.
The plane was making an emergency landing after two previous attempts in heavy snow and strong winds when it clipped treetops and crashed into a field, approxim
Crashed into a hill during a sandstorm.
The plane crashed after the pilot reported a loss of cabin pressure.
Both wing of the cargo plane separated from the fuselage after the aircraft penetrated a thunderstorm.

UNSUPERVISED LEARNING – CASE STUDY: AIRPLANE CRASHES

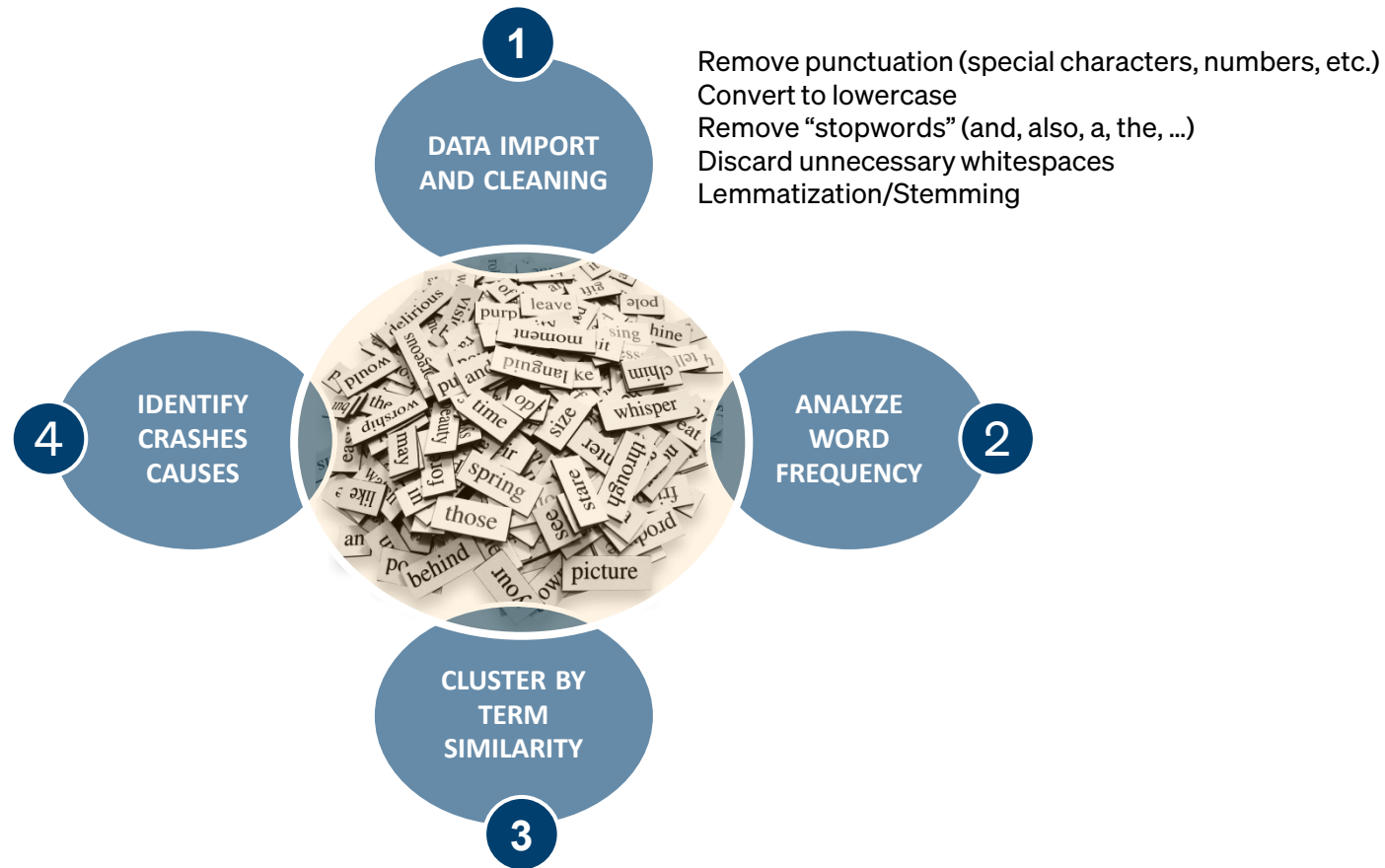One could be interested in the following first simple charts...

Crashes by Operator

Fatalities by Operator

UNSUPERVISED LEARNING – CASE STUDY: AIRPLANE CRASHES

...but, instead, we will analyze the "Summary" field to grasp the main causes of the crashes and categorize them into different clusters.

1 — DATA IMPORT AND CLEANING

Remove punctuation (special characters, numbers, etc.)
Convert to lowercase
Remove "stopwords" (and, also, a, the, ...)
Discard unnecessary whitespaces
Lemmatization/Stemming

2 — ANALYZE WORD FREQUENCY

3 — CLUSTER BY TERM SIMILARITY

4 — IDENTIFY CRASHES CAUSES

The goal of lemmatization and stemming is to reduce inflectional forms of a word to a common base form:

am, are, is ⇒ be

car, cars, car's, cars' ⇒ car

⇓

the boy's cars are different colors
the boy car be differ color

Stemming converts the words of a sentence to its non-changing portions.
Lemmatization converts the words of a sentence to its dictionary form.

```
amusement, amusing, amused => amus
amusement, amusing, amused => amuse
```