

# Text Analysis of Aviation Safety Data

Predict 453 – Final Project Report

Rahul Sangole  
Mar 17 2018

Northwestern

# Contents

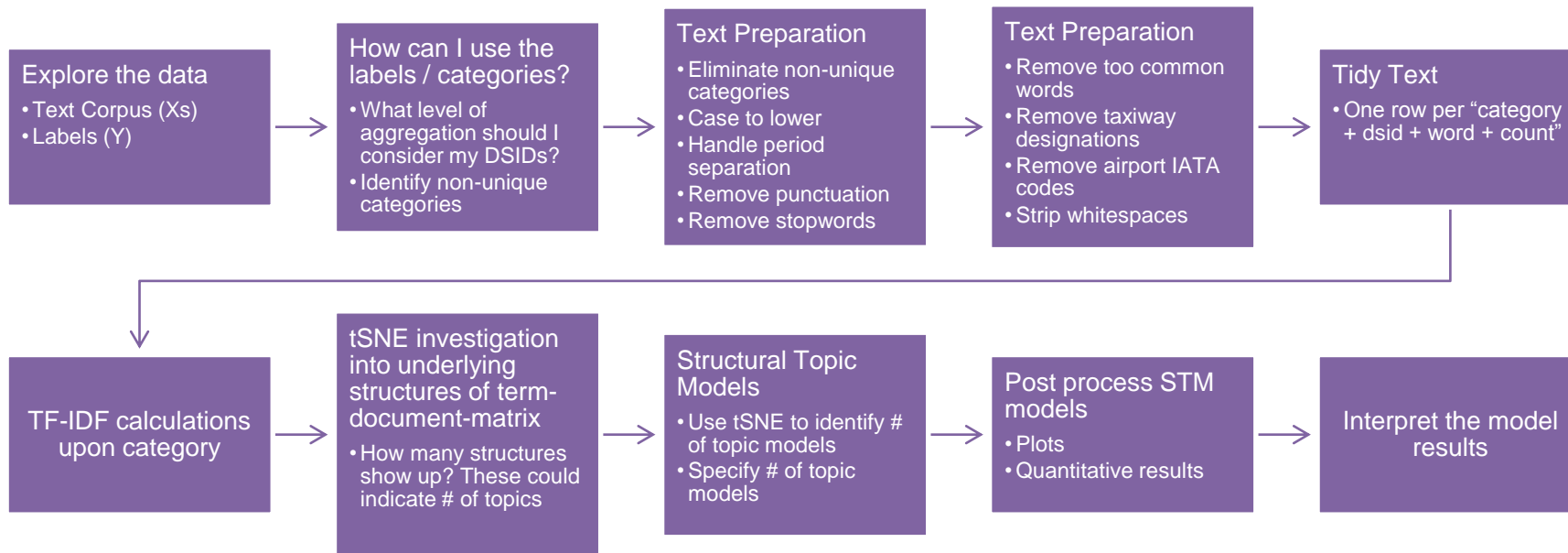
- Problem Statement
- Approach
- Input Data
- Exploratory Data Analysis
- Data Preprocessing
- Modeling
  - Document Term Matrix
  - TF-IDF
  - tSNE
  - Structural Topic Modeling
- Conclusions
- Lessons Learnt
- References

# Data Source

- This is the dataset used for the SIAM 2007 Text Mining competition<sup>1</sup>.
- This competition focused on developing text mining algorithms for document classification.
- The documents in question were aviation safety reports that documented one or more problems that occurred during certain flights.
- The goal was to label the documents with respect to the types of problems that were described. This is a subset of the Aviation Safety Reporting System (ASRS) dataset, which is publicly available.

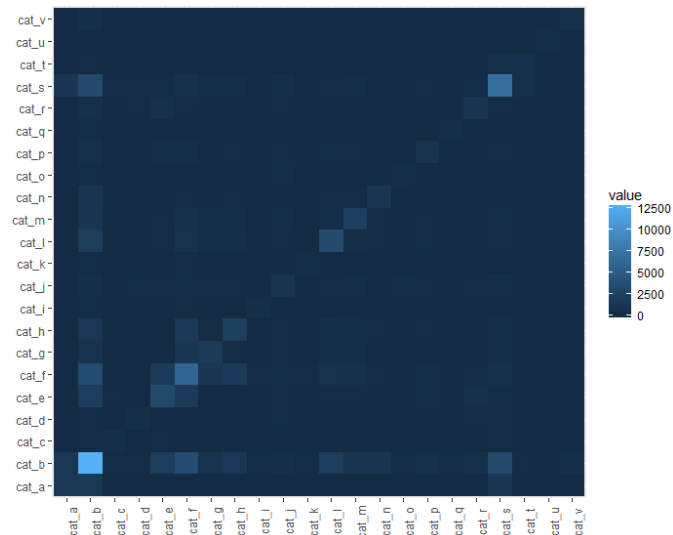
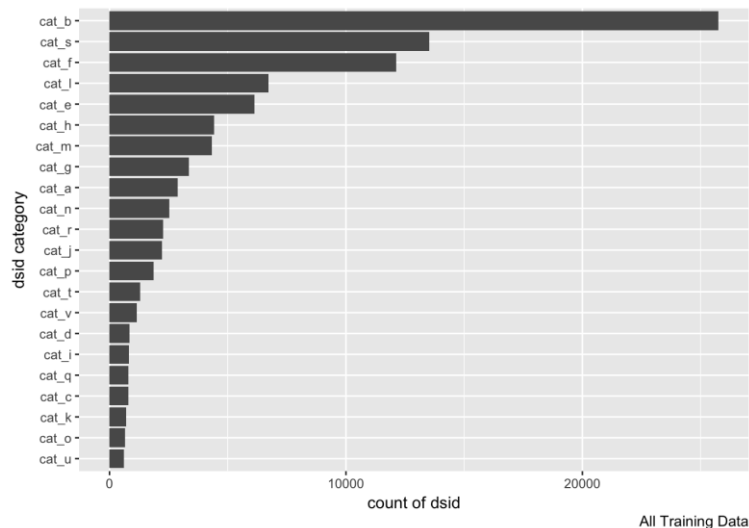
<sup>1</sup> <https://c3.nasa.gov/dashlink/resources/138/>

# Approach



# Input Data

- This is a very large dataset: [21519, 24]
- Each of the 21519 DSI is an aviation safety report
- There are a total of 22 categories. The categories aren't labeled. Does each category correspond to a topic? Or are there multiple topics within each category?
- Each DSID can belong to more than 1 category. [If this was false, all off diagonal cells would be dark].



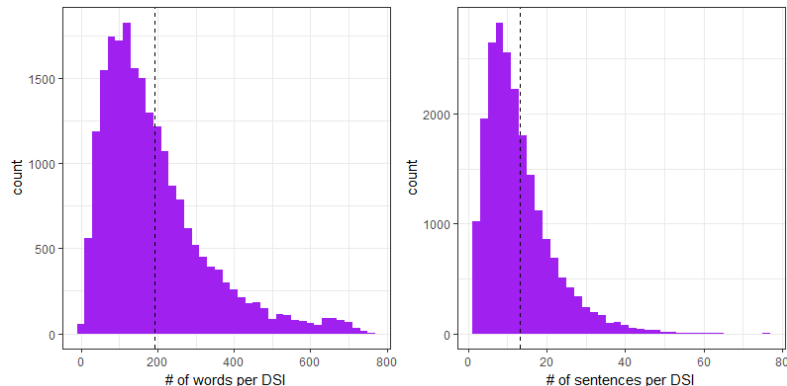
# Exploratory Data Analysis

“pass flightlevel climb TO flightlevel RUDDER RATIO AND  
AILERON LOCKOUT engineindicationandcrewalertingsystemmessage  
occur.near LEVELOFF captain south ALTIMETER fail AND UNWOUND  
TO \_ feet.WE level OFF us firstofficer south ALTIMETER AFTER  
A deviate OF \_ feet.left centralairdatacomputer HAD fail”

“ON taxiout AT bos airport I miss A TURN AND CAME INTO  
CONFLICT WITH AN aircraft \_ .WE BOTH CAME TO A STOP AND I  
realize MY MISTAKE.THERE WAS NEVER ani DANGER OF hit EACH  
OTHER.I WASWAY.humanfactor give OTHER taxiinstruction BY  
ground AND WENT ON MY DURING taxiout THE firstofficer IS  
frequent OUT OF THE LOOP get LOAD DATA AND takeoff  
perform.BOTH set OF eye ARE need DURING TAXI.AFTER EVENT  
OF \_ IT HAS BEEN hard TO FOCUS AND  
CONCENTRATE.firstofficer south HAVE TOLD ME THIS ALSO.WE  
NOW MORE THAN EVER FORCE OURSELVES TO BE mind OF THE TASK  
AT HAND.”

## Interesting artifacts

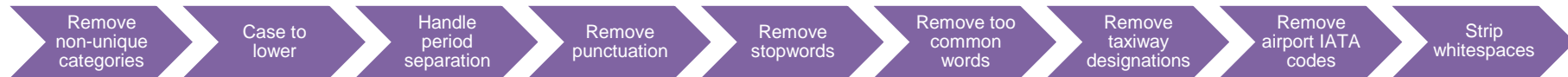
- There is a mixture of lowercase and UPPERCASE words
- Some lowercase words seem to be pre-processed tokens (bi / tri grams)
- Sentences are separated by a period with no adjacent spaces
- Numbers have been replaced by \_
- Stemming seems to have been done
- Airport IATA codes show up in lowercase



# Word Cloud



# Data Processing



approximate \_ ON september \_ captain call ask WHERE hi CLOSEOUT WAS.THE  
loadagent work THE flight inform ME THAT THE FINAL CARGO number WERE  
NOT enter YET AND THERE WAS A CHANCE THE flight MIGHT BE OVER THE  
maximum RAMP weigh.WHEN THE FINAL CARGO number WERE final receive BY  
THE loadagent SHE BECAME suspicion AS TO THE FIGURE...

approximate september captain call closeout loadagent work flight  
inform final cargo number enter yet chance flight might maximum ramp  
weigh final cargo number final receive loadagent became suspicion  
figure...

```
vs <- VectorSource(training_data_uniques$text)
docs <- Corpus(vs)
meta(docs, 'doc_id') <- training_data_uniques$doc_id

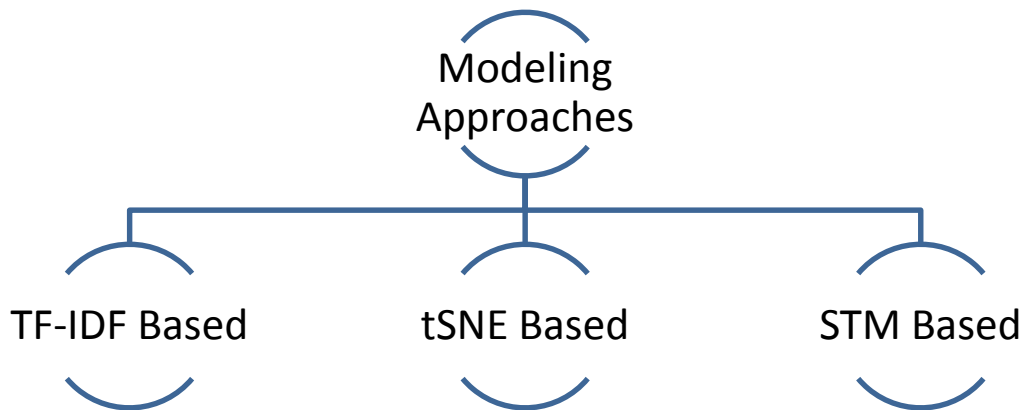
stopwords_w_spaces <- stopwords('english') %>% gsub(pattern = '\\', replacement = ' ')
taxiway_designations <- unlist(c(letters, map(letters, ~ paste0(.x, letters))))
remove_fullstop <- function(x) gsub(pattern = '\\.', replacement = ' ', x = x)
remove_too_common_words <- c('aircraft', 'airport')

docs %>%
  tm_map(content_transformer(tolower)) %>%
  tm_map(content_transformer(remove_fullstop)) %>%
  tm_map(removePunctuation) %>%
  tm_map(removeNumbers) %>%
  tm_map(removeWords, stopwords_w_spaces) %>%
  tm_map(removeWords, remove_too_common_words) %>%
  tm_map(removeWords, taxiway_designations) %>%
  tm_map(removeWords, airport_iata_codes[1:7000]) %>%
  tm_map(removeWords, airport_iata_codes[7000:7800]) %>%
  tm_map(stripWhitespace)

tdm <- TermDocumentMatrix(docs)
```



# Modeling



Using existing *known* categories and TF-IDF, can we identify patterns in the categories?

Using only the term-document-matrix, can we identify any patterns? How many topics exist in the corpus?

Using only the term-document-matrix, can we identify topics in the corpus? How many topics? What are the topics? How do they agree with the *known* categories?

# How I'm defining a DSI

## For TF-IDF Analysis

- All the individual safety reports are grouped into the 22 *known* categories
- Thus, this is equivalent to a total of 22 DSIs in the corpus, each containing thousands of sentences

## For STM Analysis

- Started with the TF-IDF approach
- Switched to defining each safety report as it's own DSI
- Thus, total DSIs = 5949

# Equivalence Classes, Reference Term Vectors, Document Term Matrix

- Manually read through 10s of safety reports
- I could not find any obvious terms which I would group together into ECs
- The RTV is thus the remainder of the terms after pre-processing
- `tidytext` is used to create the RTV which is fed into a `TermDocumentMatrix` generation function
- The definition of 'documents' depends on if I'm doing a TF-IDF or STM

```
> inspect(tdm)
<<TermDocumentMatrix (terms: 13620, documents: 5949)>>
Non-/sparse entries: 355806/80669574
Sparsity           : 100%
Maximal term length: 70
Weighting          : term frequency (tf)
Sample            :

```

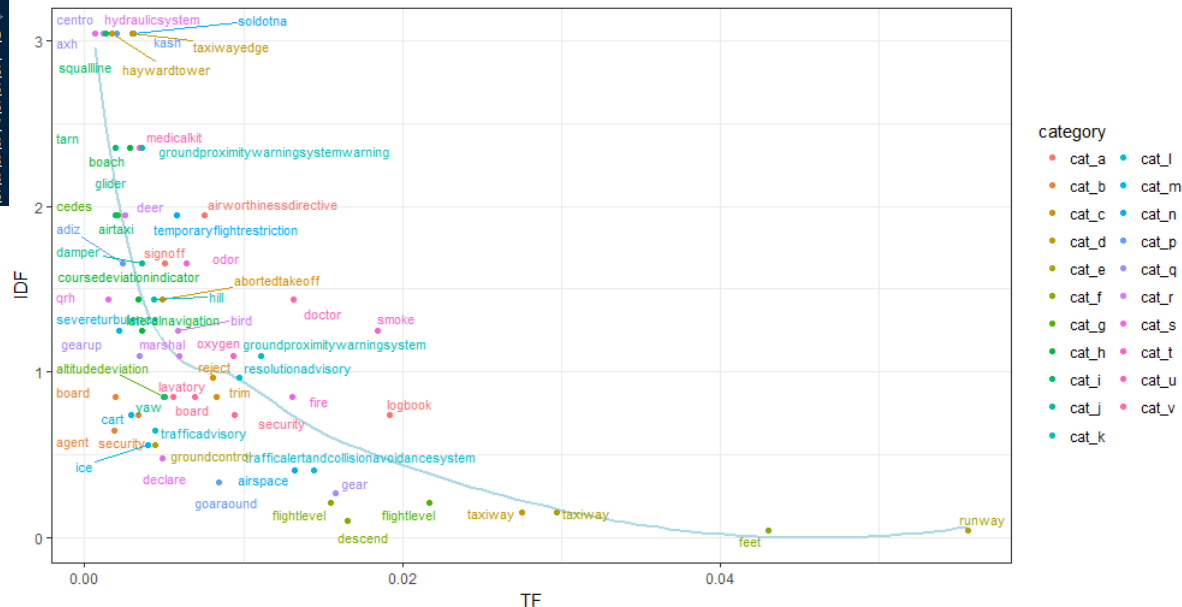
	Docs									
Terms	1520	2733	2959	3039	3758	380	4938	5432	5867	5920
approach	2	0	0	5	4	1	0	0	5	3
call	1	1	0	0	1	0	0	0	4	4
clear	1	0	2	2	5	0	1	7	1	0
control	2	1	4	0	0	0	4	5	1	4
feet	1	0	0	4	7	1	1	13	8	12
flight	2	1	2	2	0	1	3	2	1	1
land	3	2	7	9	0	6	1	1	3	1
passenger	0	0	0	1	0	3	0	1	0	2
report	0	0	0	3	2	1	1	2	1	4
runway	2	3	3	3	0	4	0	1	8	1

# TF-IDF Based Analysis

category	word	n	tf	idf	tf_idf
<fct>	<chr>	<int>	<dbl>	<dbl>	<dbl>
1 cat_f	feet	1323	0.0430	0.0488	0.00210
2 cat_e	runway	876	0.0556	0.0488	0.00271
3 cat_s	declare	751	0.00491	0.480	0.00235
4 cat_l	trafficalertandcollisionavoidancesystem	749	0.0145	0.405	0.00586
5 cat_f	descend	508	0.0165	0.100	0.00165
6 cat_l	resolutionadvisory	503	0.00971	0.965	0.00937
7 cat_f	flightlevel	477	0.0155	0.211	0.00328
8 cat_b	security	475	0.00335	0.742	0.00248
9 cat_e	taxiway	468	0.0297	0.154	0.00458
10 cat_b	board	276	0.00195	0.847	0.00165

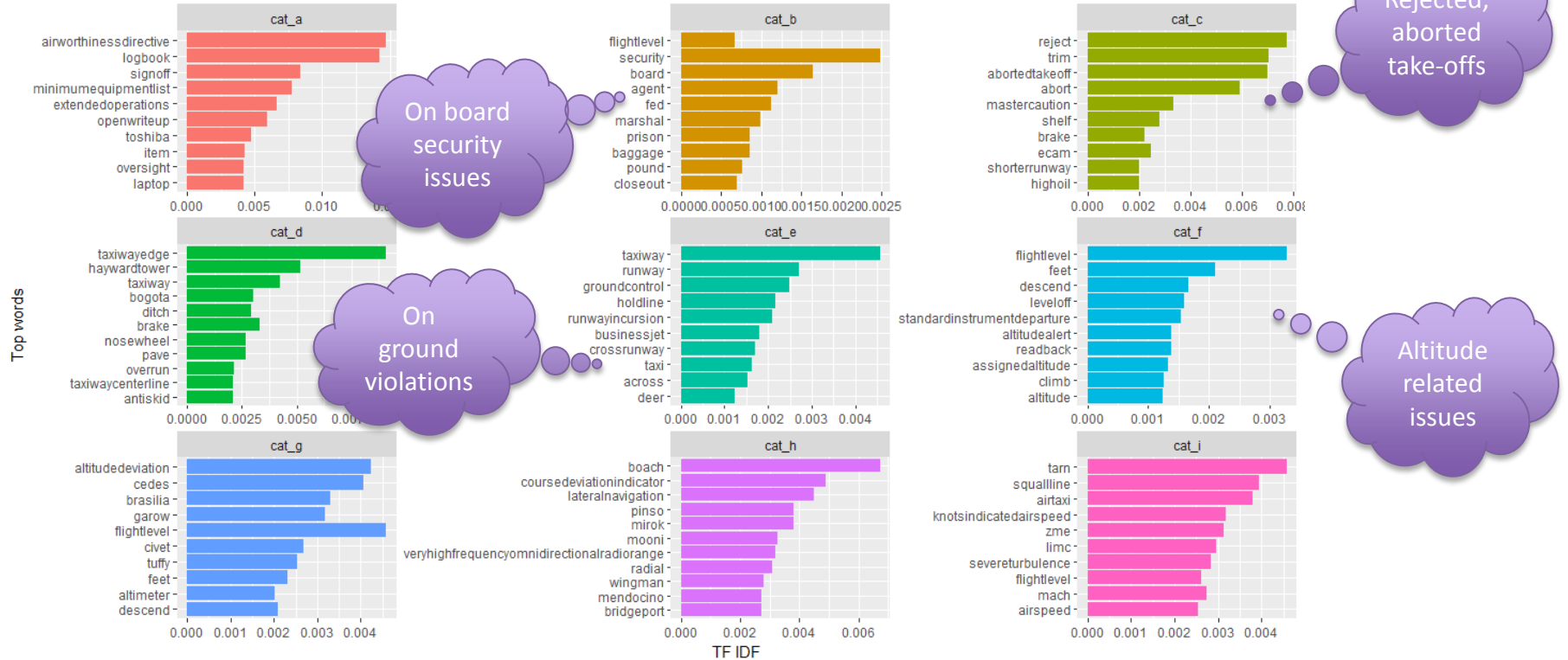
- Some terms are very exclusive to certain categories, with high values of IDF and low values of TF. Ex: *hydraulicsystem* for *cat\_s*
- Others terms like *runway* show up in a lot more documents, yet have high TF values for certain categories like *cat\_d*
- Both these instances result in high TF-IDF values within the respective categories

Top 3 words by tf-idf per category



# TF-IDF Based Analysis

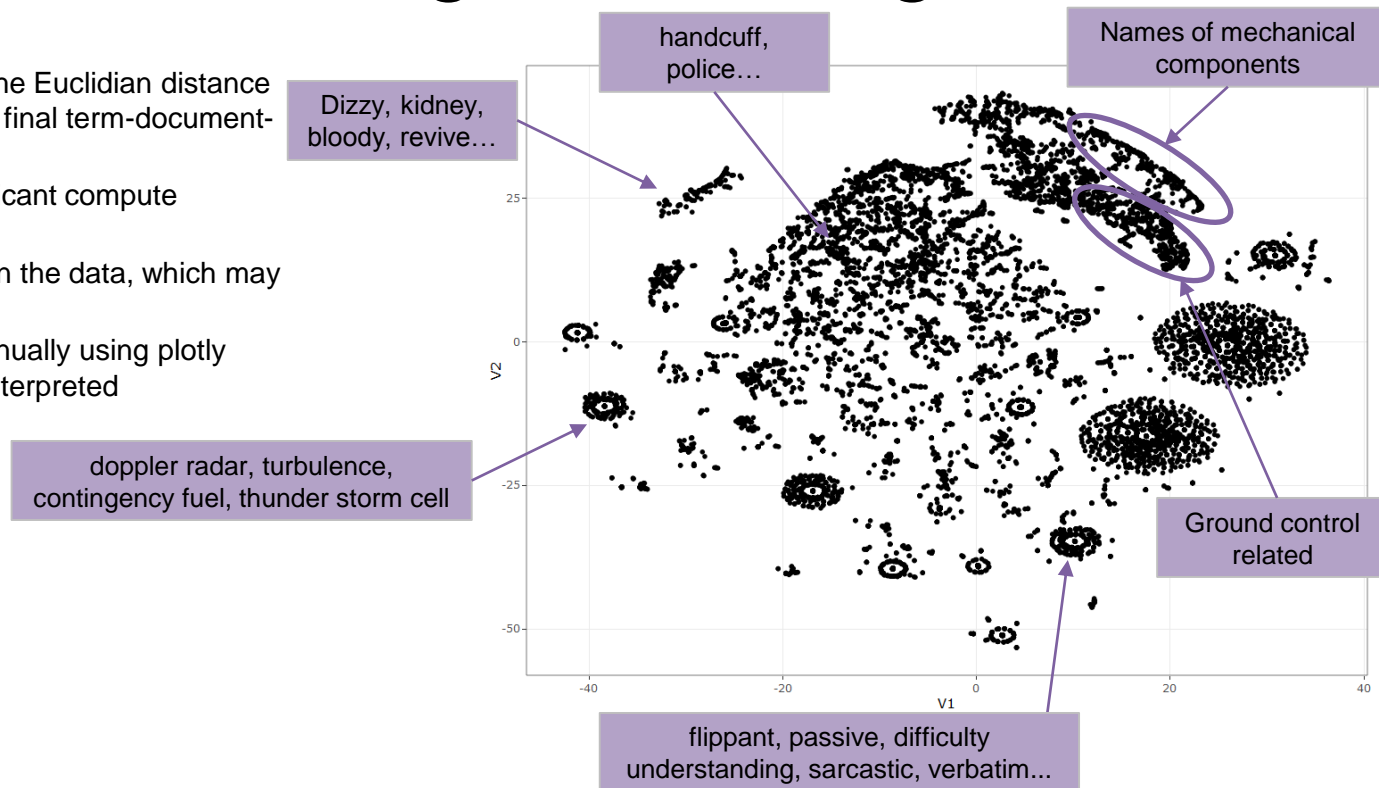
Page 1 of 3



<sup>1</sup> Remainder of the plots in the appendix

# Structural Investigation using tSNE

- tSNE analysis is performed on the Euclidian distance matrix between the words in the final term-document-matrix
- The Rtsne package offers significant compute efficiency vs the tsne package
- We can see very clear clusters in the data, which may point to topics in the corpus
- Investigation of these points manually using plotly reveals patterns which can be interpreted

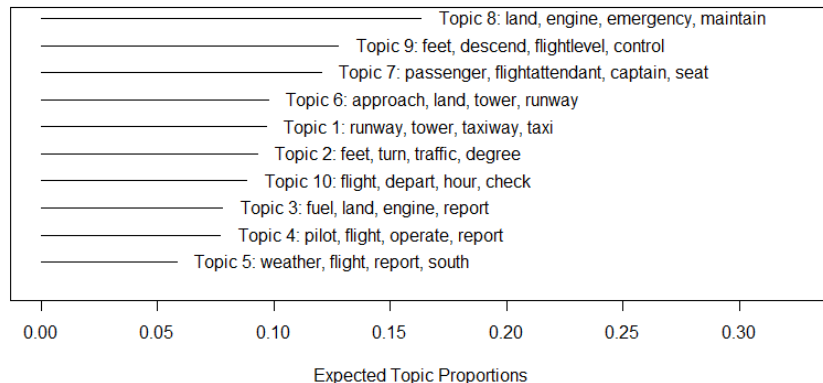


# Structural Topic Modeling

- STMs allow researchers to flexibly estimate a topic model
- STMs are like LDA models, except they extend the capability to include document-level metadata into the topic model. This allows researchers to not only estimate topics but also relationships to document metadata.
- The `stm` package in R is quite powerful with many features which work well with the `tidyverse`, `tm`, `quanteda` and other major R packages
- Using `stm`, there are two main approaches towards deciding how many topics (K) are contained in the corpus:
  - Define the value(s) of K yourself
  - Use `tsne` to perform a dimension reduction on the TDM followed by some geometric calculations to estimate the number of “groupings”
    - This method resulted in 116 topics!
- To begin with, I’m assuming  $K=10$

# STM Results, K=10

Top Topics



- For each topic,
  - Highest probability terms
  - FREX: How exclusive terms are to this topic?
  - Lift: Higher weight to words which appear less frequently in other topics

A topic model with 10 topics, 5949 documents and a 8515 word dictionary.

Topic 1 Top Words:  
 Highest Prob: runway, tower, taxiway, taxi, clear, hold, takeoff  
 FREX: taxiway, crossrunway, groundcontrol, takeoffclearance, taxi, hold, taxiinst  
 Lift: accessroad, airportlight, airportvehicle, apron, blowingsnow, breakaway, c  
 Score: taxiway, runway, taxi, tower, groundcontrol, crossrunway, clearance

Topic 2 Top Words:  
 Highest Prob: feet, turn, traffic, degree, climb, head, trafficalertandcollisions  
 FREX: resolutionadvisory, oclock, target, trafficalertandcollisionavoidancesyste  
 Lift: arrivalcorridor, collisioncourse, computerfailure, conflictresolution, fal  
 Score: trafficalertandcollisionavoidancesystem, resolutionadvisory, traffic, tra

Topic 3 Top Words:  
 Highest Prob: fuel, land, engine, report, left, right, damage  
 FREX: tank, revolutionsperminute, mixture, magneto, propel, fuelleak, deer  
 Lift: accesspanel, airintake, airstrip, amplitude, automotive, auxiliarytank, bo  
 Score: fuel, engine, tank, damage, gear, propel, brake

Topic 4 Top Words:  
 Highest Prob: pilot, flight, operate, report, south, state, part  
 FREX: federalaviationregulation, airman, federalaviationadministration, violate  
 Lift: causeconfusion, coupler, disapproved, eggx, flightdatacenter, furnish, hun  
 Score: temporaryflightrestriction, federalaviationadministration, airman, violat

Topic 5 Top Words:  
 Highest Prob: weather, flight, report, south, encounter, radio, minute  
 FREX: thunderstorm, encounter, windshear, windshield, forecast, navigate, turbul  
 Lift: aerobaticmaneuver, basicvisualflightrules, coldfront, contactor, dispense  
 Score: turbulent, weather, cloud, sector, ice, severeturbulence, encounter

Topic 6 Top Words:  
 Highest Prob: approach, land, tower, runway, call, clear, mile  
 FREX: adiz, finalapproachfix, potomac, instrumentflightrules, glideslope, classb  
 Lift: adiz, angling, arlington, assignedtranspondercode, authoritarian, awo, bet  
 Score: tower, approach, runway, airspace, downwind, adiz, visualflightrules

Topic 7 Top Words:  
 Highest Prob: passenger, flightattendant, captain, seat, cockpit, door, number  
 FREX: galley, doctor, lavatory, paramedic, firstclass, aisle, jetbridge  
 Lift: abrasion, assault, auditor, beverage, bloody, boardingprocess, breath  
 Score: flightattendant, passenger, security, door, board, doctor, agent

Topic 8 Top Words:  
 Highest Prob: land, engine, emergency, maintain, report, normal, flap  
 FREX: qrh, hydraulicsystem, trailingedge, overheat, flap, compressorstall, exhaus  
 Lift: hydraulicsystem, abnormalprocedure, aircyclemachine, allison, amberlight, e  
 Score: engine, flap, emergency, declare, trim, qrh, gear

Topic 9 Top Words:  
 Highest Prob: feet, descend, flightlevel, control, altitude, climb, airtrafficcon  
 FREX: autopilot, flightlevel, altimeter, descend, assignedaltitude, altitude, al  
 Lift: alertwindow, altitudeawareness, altitudecallout, altitudepreselect, altitue  
 Score: flightlevel, feet, descend, climb, altitude, autopilot, clearance

Topic 10 Top Words:  
 Highest Prob: flight, depart, hour, check, time, crew, captain  
 FREX: schedule, load, release, logbook, cargo, paperwork, minimumequipmentlist  
 Lift: accuload, airtrafficcontroldelay, magazines, maintenancediscrepancy, mainte  
 Score: schedule, dispatch, logbook, pound, paperwork, hour, minimumequipmentlist

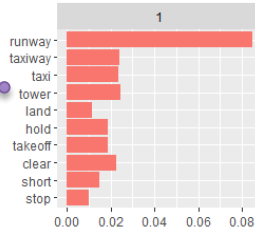


# Word Probabilities Per Topic

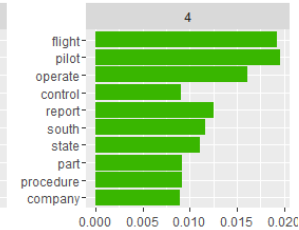
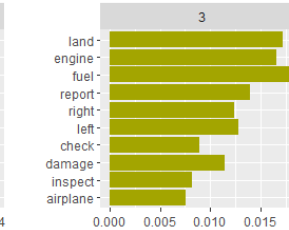
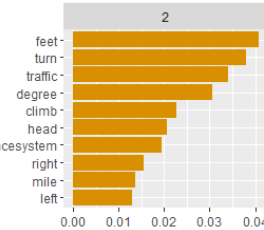
Highest word probabilities for each topic

Different words are associated with different topics

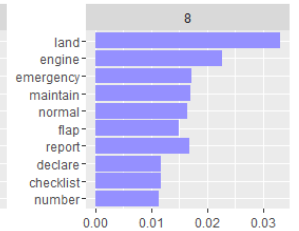
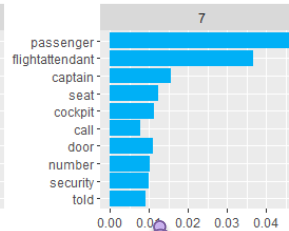
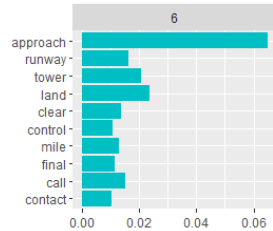
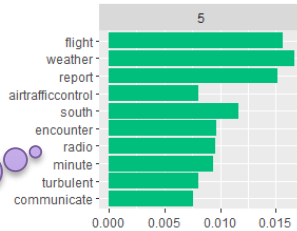
On ground violations



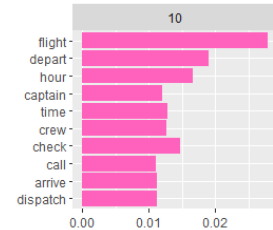
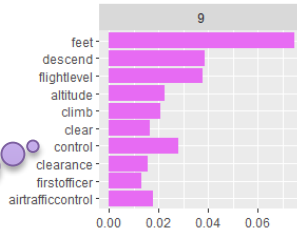
trafficalertandcollisionavoidancesystem



Weather related issues



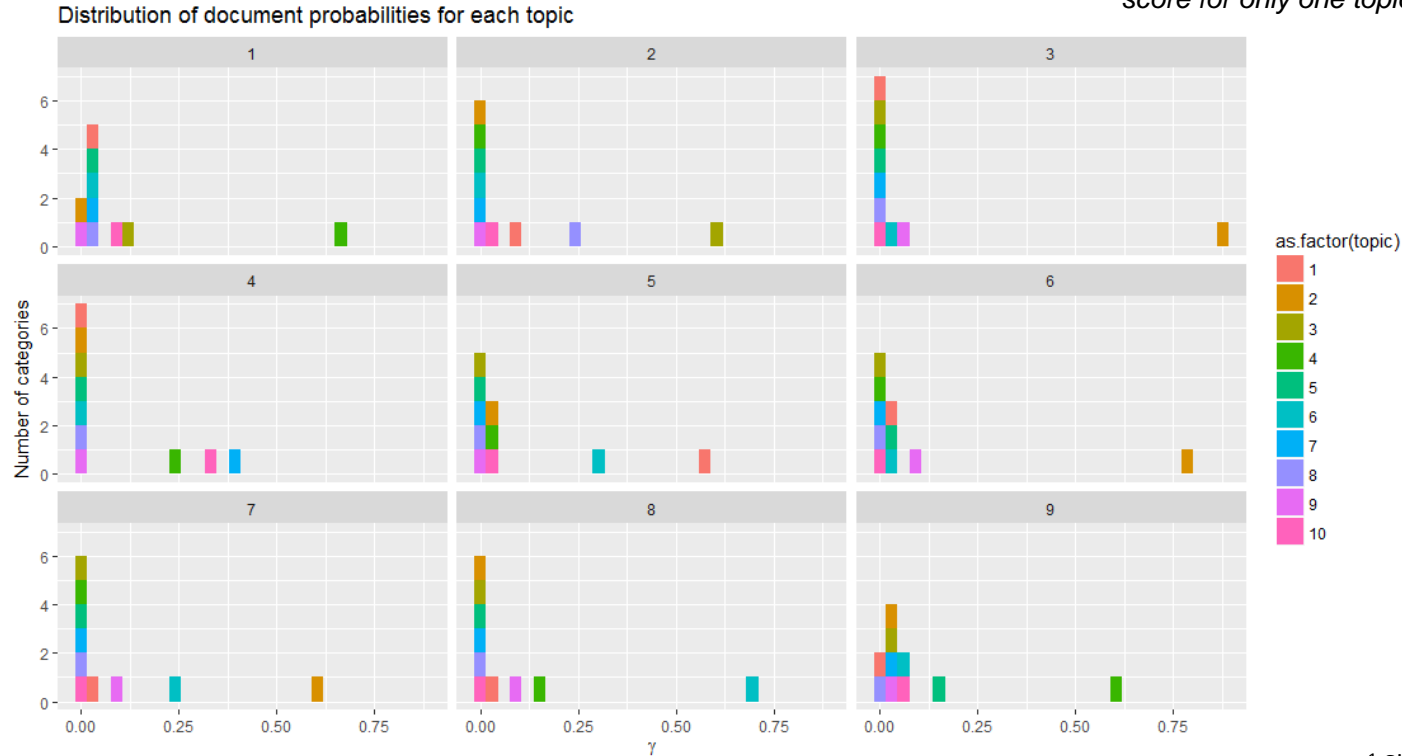
Altitude related issues



On board security issues

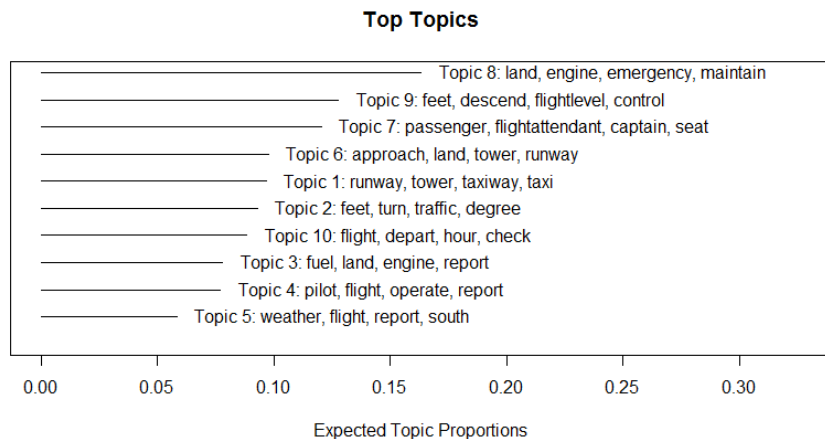
# Document Distribution among Topics

*Each document should have a high gamma score for only one topic*

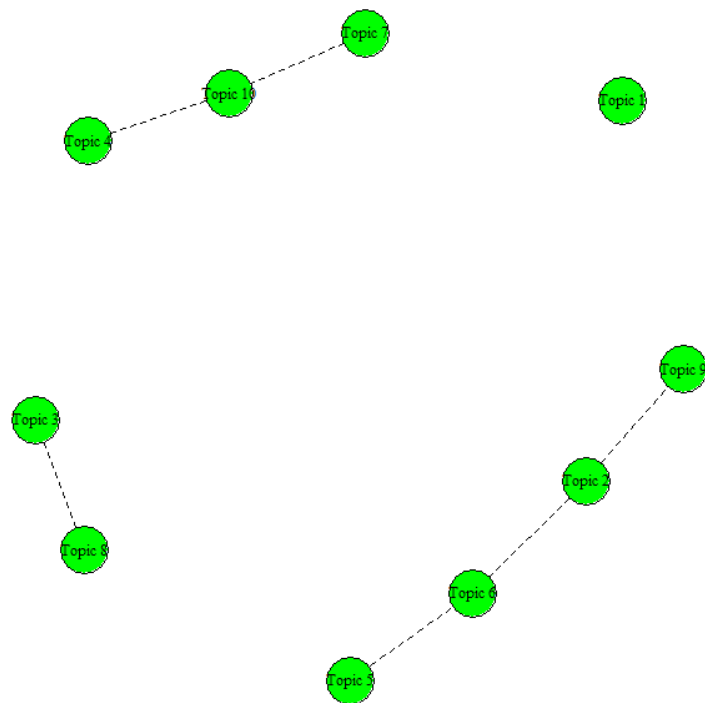


<sup>1</sup> Showing 9 of 5949 DSIs

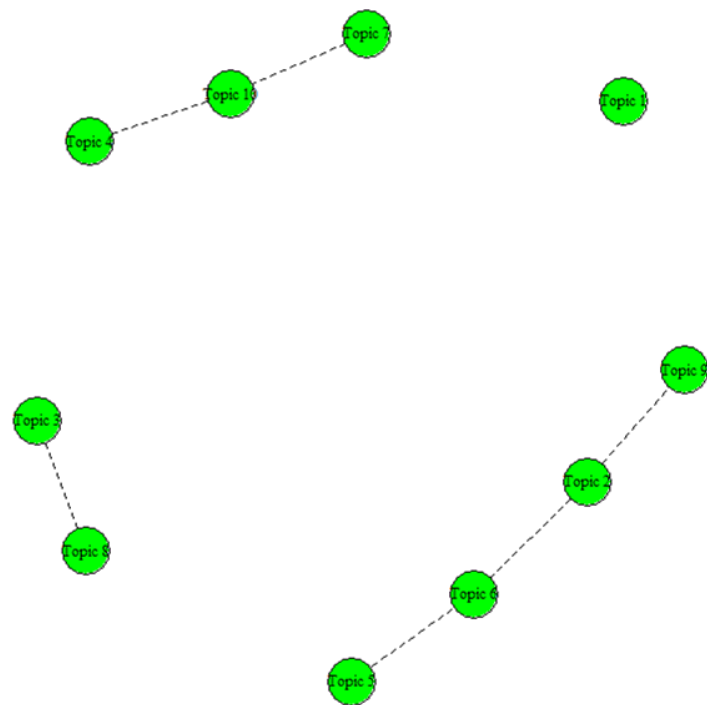
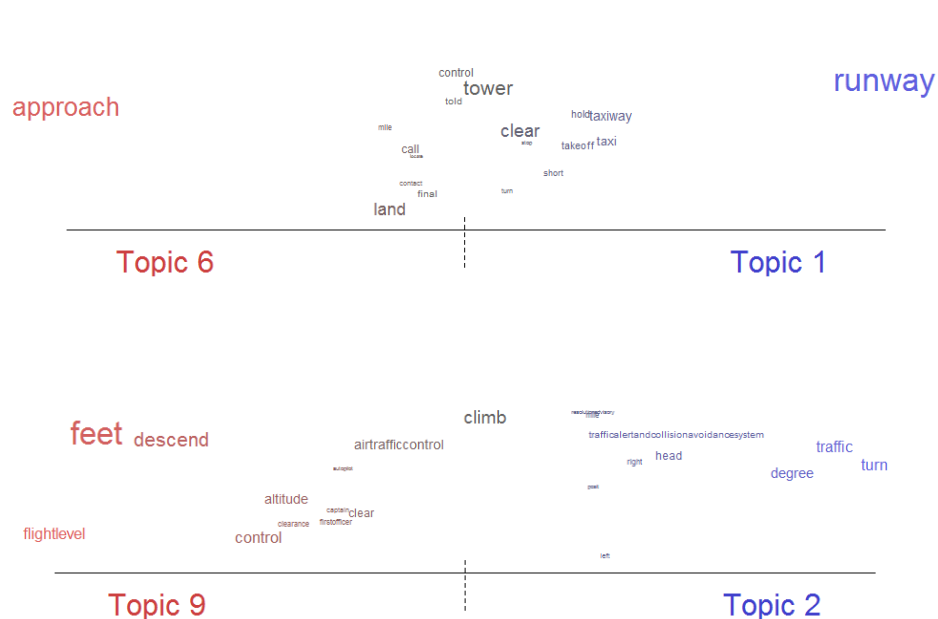
# Topic Model Correlation



- Some topics are closer together, others seem further apart
- For example:
  - Topic 8 and Topic 3 both refer to engine and damage. While Topic 8 references hydraulic systems and flaps, Topic 3 references fuel & fuel leaks
  - Topic 1 is quite different from all the rest – it speaks about runway / taxiway / ground control clearance issues
- We can model this using ‘topic correlations’. Positive correlations indicates that both topics are likely to be discussed in a document.



# Topic Model Correlation



# Lessons Learnt

- Text analytics is a deep subject with many rabbit holes to get lost in
- It's a nascent field with a large number of analytics packages in R developed within the last 5 years
- More non-quantitative work involved than any other course
- Stability of results seem asymptotic and sensitive to pre-processing
- Long way to go

# References

# Appendix

# TF-IDF Based Analysis

Page 2 of 3

