TEAM NOISE

Determining Noise and Actuals in the Sightings/Incidence Database

Partner: Federal Aviation Administration (FAA)

Instructor: Professor Brett Berlin

GEORGE MASON UNIVERSITY | Data Analytics Engineering

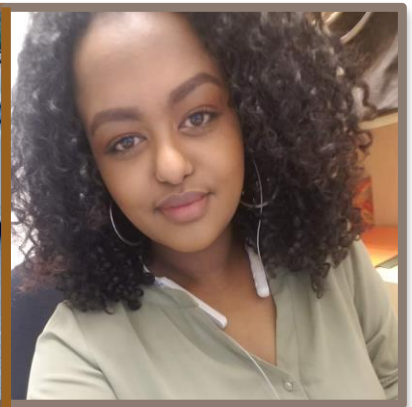
TEAM NOISE - MEMBERS

Samuel Razia
Scrum Master

John Brzezinski
Product Owner

Joey Rainey
GitHub Lead & Developer

Lahari Tadepalli
Developer



Hermella Tessema
Developer

Federal Aviation Administration

Name	Position
Sherri Shearon	**Chief Data Officer**
Michael Lukas	Aviation Safety
Inderbir Singh	Aviation Safety
Matthew O'Meara	Aviation Safety
Dipasis Bhadra	Aviation Policy and Plan
Samuel Pascoe	Aviation Policy and Plan

Mission

- Continuing mission is to provide the safest, most efficient aerospace system in the world.

Vision

- Strive to reach the next level of safety and efficiency and to demonstrate global leadership in how we safely integrate new users and technologies into our aviation system. We are accountable to the American public and our aviation stakeholders.

TEAM NOISE – PROBLEM STATEMENT

Problem Context:

Domain of problem: Air transportation

Importance of problem:

- Safe integration on manned and unmanned aircraft
- Being able to determine the extent of noise over actuals
- Validation across different databases

Problem Statement:

- The goal is to validate the information from the FAA UAS Sightings compared to other databases.
- A credible database is the NASA Aviation Safety Reporting System (ASRS) which reports confidential reports of aviation traffic.

TEAM NOISE – DATASETS

Dataset Name:	FAA UAS Sightings Report	ASRS DB
Dataset Owner:	FAA	NASA
Dataset Type:	Open Source	Open Source
Dataset Size:	1.76MB	8.19MB
Dataset License:	N/A	N/A
Dataset Location:	Internet	Internet
Dataset Access:	https://www.faa.gov/uas/resources/public_records/uas_sightings_report/	https://akama.arc.nasa.gov/ASRSDBOnline/QueryWizard_Filter.aspx
Dataset Restrictions:	No Restrictions	No Restrictions
Dataset Time Range:	Q42014 - Q22021	Q42014 - Q22021
Dataset Collection Process:	Pilots reporting to Air Traffic Control Facilities	Anyone reporting to Air Traffic Control Facilities
Analytic/Algorithm that will use on dataset:	RegEx / NLP / Anomaly Detection	RegEx / NLP / Anomaly Detection

TEAM NOISE – DATASET QUALITY

FAA UAS Sightings Report

Completeness – The dataset has no null values. Each record is complete.

Consistency– The dataset defers between the year in terms of how they recorded data. Information was merged into one area after around 2019.

Uniqueness– The dataset has few duplicate reporting with slight change in language., no unique IDs so will have to manually parse through.

Conformity– Dataset had few areas that were not standardized in dates but was able to through some cleanup.

Accuracy– Values are accurate as the narrative for each matches the location

Aviation Safety Reporting System (ASRS) DB

Completeness – The dataset has MANY null values. Records need to be parsed through to see which information is relevant to us and removing the areas that have too many null values.

Consistency– The dataset is not consistent in their reporting as many times even when they have a location in the narrative, the DB reports it default as 'ZZZ' Airport tower with 'US' as the state. Many records do not tell the full story easily but can generally derive an understanding from the narrative.

Uniqueness– The dataset is unique as it has an identifier for the ASRS reports. Need to look more deeply at narratives.

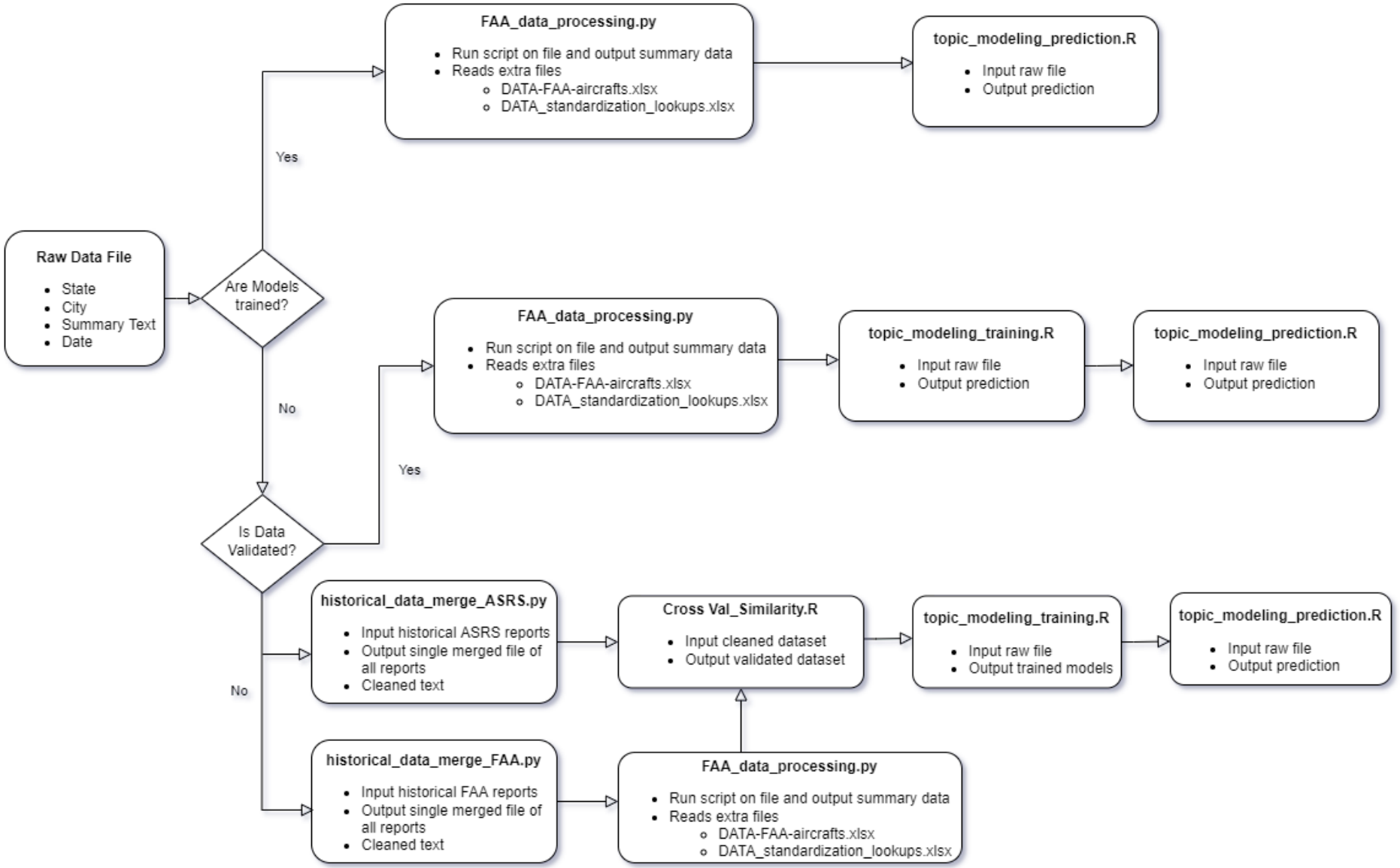
Conformity– Dataset has multiple fields that could be considered duplicates. Looking into tackling that in our data clean up – but our goal is to create a word corpus from the FAA dataset before tackling this issue.

Accuracy– Accuracy of the dataset is slim since it has many default values being used in records.

SPRINT WORKS

Sprint 1 Problem Definition	Sprint 2 Dataset Analysis	Sprint 3 Algorithms	Sprint 4 Visualizations & Modeling	Sprint 5 Delivery
▪ Focus on understanding the data set and correlating it to the problem statement. ▪ Discuss with partners what their goal is for the project. ▪ Learn the key FAA concepts that are important to the partners to extract from the dataset.	• Access the data quality of the FAA & ASRS datasets. • Clean out the database and identify patterns. Look into how we can incorporate algorithms into the database.	• Provided a cleaned dataset to be cross validated with the internal FAA Surveillance database. • Understand NLP techniques and how they work.	• Have our data validated and a working model started with data given • Perform cross validations with the ASRS & FAA Surveillance datasets	• Prepare a clean script for cleaning internal datasets. • Provide a predictive model for detecting potential matching components.

WORKFLOW FOR PROJECT



WORD EMBEDDING

Word embeddings are vector representations of a particular word.

text2vec package

sim2(x, y, method) - calculates similarity between each row of matrix x and each row of matrix y using given method.

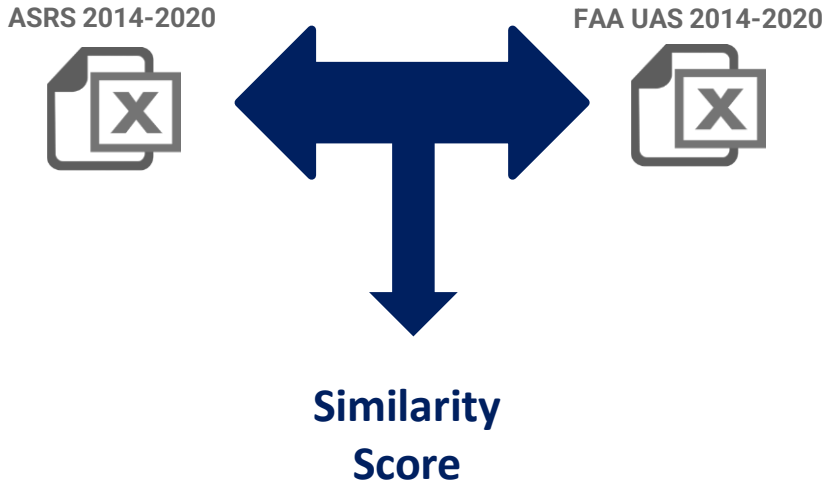
$$similarity(doc_1, doc_2) = cos(\theta) = \frac{doc_1 doc_2}{|doc_1||doc_2|}$$

$$distance(doc_1, doc_2) = 1 - similarity(doc_1, doc_2)$$

How we are utilizing word embedding.

1. Clean out database, tokenize, and clear out any stop words necessary.
2. Vectorize each record in both database
3. Run Cosine distance function to calculate similarity of word matrixes.
4. Produce a similarity score, set a threshold for what we believe to be a hit.

Further cleaning to produce better scores



ASRS CROSS VALIDATION – FIRST ROUND

ASRS 2014-2020

FAA UAS 2014-2020

FAA_Narrative00	Narrative_Hit%	ASRS_Narrative
C303 FLYING TOWER PATTERN RY34. RIGHT BASE REPORTED TOWER "UAV" BELOW HIM. NO AVOIDANCE MANEUVER WAS REQUIRED.	0.433124577	C303 VFR TRAFFIC PATTERN; RIGHT BASE RUNWAY 34 APPROXIMATELY 3;000 FEET; REPORTED SIGHTING 'UAV' APPROXIMATELY 500-1;000 FEET BELOW HIM. THIS WAS REPORTED CONTROL TOWER TOWER FREQUENCY. TOWER CONTROLLER REPORTED THIS OCIC. OCIC FAILED LOG INCIDENT OR FILE AN MOR. NO RECORD OF CONTACTIONNG ANYONE ELSE; INCLUDING LAW ENFORCEMENT. TRAINING 'DRONE' PROCEDURES/REPORTING REQUIREMENTS.
E170 DEPARTED RWY 6L FAILS INTERSECTION. AIRCRAFT WAS SWITCHED ZOB AND REPORTED DRONE SIGHTING 50100 FEET ABOVE THEIR 6000 FOOT ALTITUDE, APPROXIMATELY 1215 MILES NE OF CLE. PILOT STATED NO EVASIVE ACTION WAS TAKEN AND NO DESCRIPTION OF ANY KIND WAS AVAILABLE. PILOT REPORTED DRONE WAS ONLY VISIBLE FOR SPLIT SECOND. ALL CONCERNED NOTIFIED.	0.253184842	WHILE BASE LEG FOR ILS 1 CAK; NOTICED BLUE DRONE OFF LEFT SIDE OF MY WING. 3;000 MSL AND DRONE APPEARED BE APPROXIMATELY 3;200 MSL. NOTIFIED ATC IMMEDIATELY AND EVEN THEY FOUND IT AS PRIMARY TARGET THEIR RADAR. BEING SAID; THIS DRONE WAS WITH CLASS C AIRSPACE. NO DAMAGE OR INJURIES. THIS IS ONE TIME EVENT MAY BECOME MORE FREQUENT EVENT WITH DRONE POPULARITY INCREASING.
UNAUTHORIZED DRONE ACTIVITY INBOUND ARRIVAL FOR RWY 08L LC1 119.1 REPORTED 3 MILE FINAL 500 FEET LEFT SIDE OF FLIGHT. FULT COUNTY AND DEN NOTIFIED.	0.142539329	WAS PILOT MONITORING OUR WAY IN ATL. APPROXIMATELY 1;500 OR SO FEET FINAL IN [RUNWAY] 8L WHEN NOTICED LARGE WHITE DRONE AIRCRAFT OFF OUR LOWER LEFT SIDE MOVING WEST BOUND APPROXIMATELY 500 FT. BELOW US. BEST GROUND REFERENCE COULD MAKE WAS LARGE MANUFACTURING WAREHOUSE WITH 'BJ'S' LOGO PAINTED ROOF. IT WAS WHITE AND HAD FOUR PROPS. NOTIFIED ATC OF SUCH WHICH THEY PASSED PILOT REPORTS AND RELAYED IN ATIS. AFTER LANDING; TOWER REQUESTED PHONE CALL FOR FOLLOW UP. [AIRLINER] TWO AIRCRAFT BACK ALSO SAW DRONE CIRCLING AN APARTMENT COMPLEX. CALLED ATC GATE WHICH THEY TOOK DOWN BASIC INFORM AND ASKED IF TOOK EVASIVE ACTIONON; WHICH HAD NOT. **EVENT WAS CAUSED BY AN INDIVIDUAL OR INDIVIDUALS ILLEGALLY OPERATING UAS WITH 5NM OF AN AIRFIELD.** MIGHT SUGGEST SOME GUIDANCE MEMO OR CFM/FOM PROCEDURE OPERATIONS THESE CIRCUMSTANCES AS DRONE ENCOUNTERS ARE BECOMING MORE AND MORE FREQUENT.
EROC ADVISED , CESSNC310, OBSERVED UAS WHILE 3 MILES WEST OF RIC 2,000 FEET. NO EVASIVE ACTION REPORTED. RICHMOND PD NOTIFIED NO PHONE PROVIDED.	0.056613852	GOT LOC FAILED RESEARCH AIRSPACE. WAS RELYING MANUFACTURE'S GEO FENCING GUIDE ME WITH AIRSPACE. VIOL WAS DISCOVERED BY COMPANY UAS COORDINATOR. VIOL WAS DISCUSSED AND CORRECTIVE ACTION FOR FURTHER USE IS UNDERSTOOD. DISCUSSED AND NOW UNDERSTAND DIFFERENCE BETWEEN GEO FENCING AND NAS. GOING FURTHER; WILL LOOK UP AIRSPACE BEFORE FLIGHTS. WILL NOT FLY CONTROLLED AIRSPACE WITHOUT PROPER AUTHORIZATION.

PREDICTIVE MODELING WITH TOPIC MODELING

- **Name:** Topic Modeling
- **What is it**: Topic modeling is a type of modeling that allows for classifying documents to find natural groupings and extract hidden topics from large documents or text.
- **How it works:** Using column analytics – identify key words from word corpus in 'Summary' columns. Utilize this to find topic relevance and create clusters. Try to match the clusters to others.
- **Data Inputs:** using original descriptive data; also look at broken down data (after parsing and cleaning) see if it helps.
- **Data Outputs:** Classification of if it is noise or actual.

Use of Stop Words

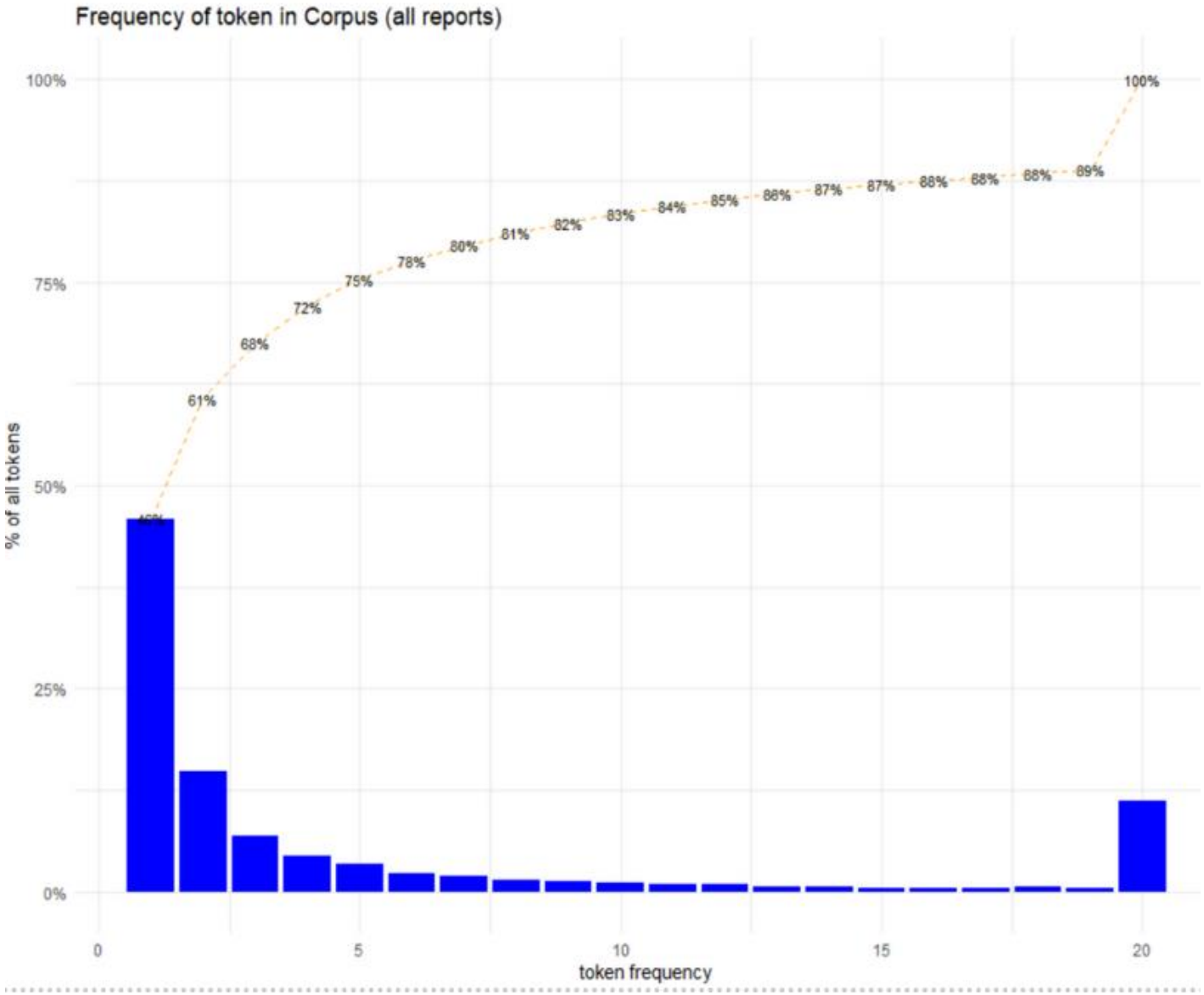
```
# stop word list
stopWords.list <- c(
    # first round top 50
    "the","a","reported","at","was","of","and","to","no", "evasive","action",
    "notified","advised","on", "while","taken","in","not","off",
    "from","that","for","it", "he", "approximately","his","with",
    #second round top 50
    "were","did","as","by","or","they","an","be",
    #random found in work
    "report", "receive","received","observed","observe","drone","uas", "pilot",
    "aircraft", "about", 'airport', 'feet', 'ft', 'miles', 'mi','mile','him',
    'had', 'acft','o'clock", 'arpt', 'they', 'them', 'her', 'she', 'he', 'said',
    'is', 'if', 'notification', 'have', 'but','stated','aviation','approx',
    'information','their', 'when', 'called', 'just'
)

# stop word dataframe
stop.words <- data.frame(word=stopWords.list, stringsAsFactors=F)

# removing the stop words
faa.stopped <- faa %>%
    left_join(y=stop.words, by = 'word', match='all') %>%
    filter(is.na(stopword))
```

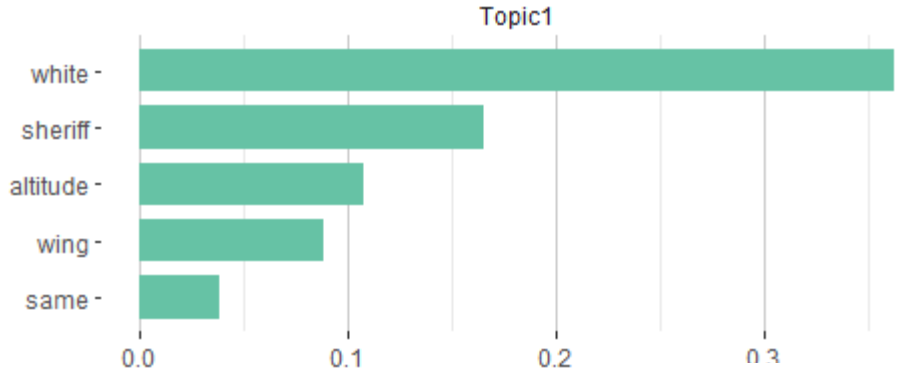
MOST FREQUENT WORDS

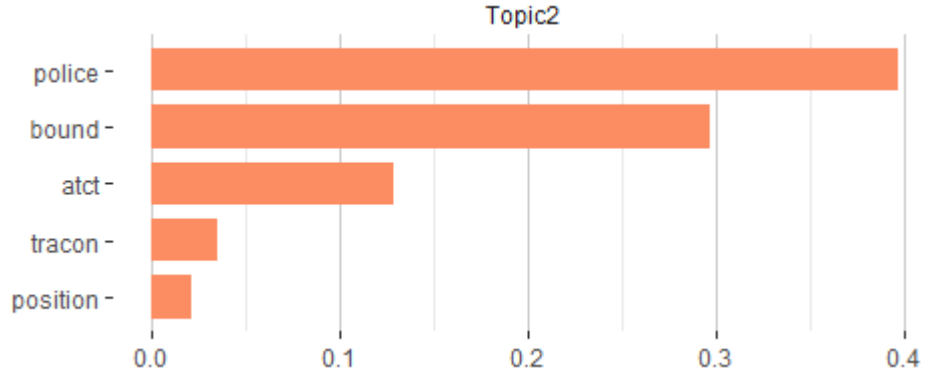
FREQUENCY OF TOKEN IN CORPUS

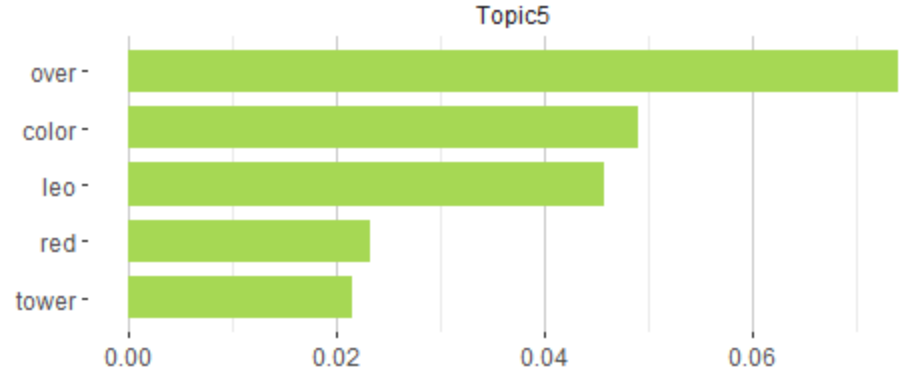


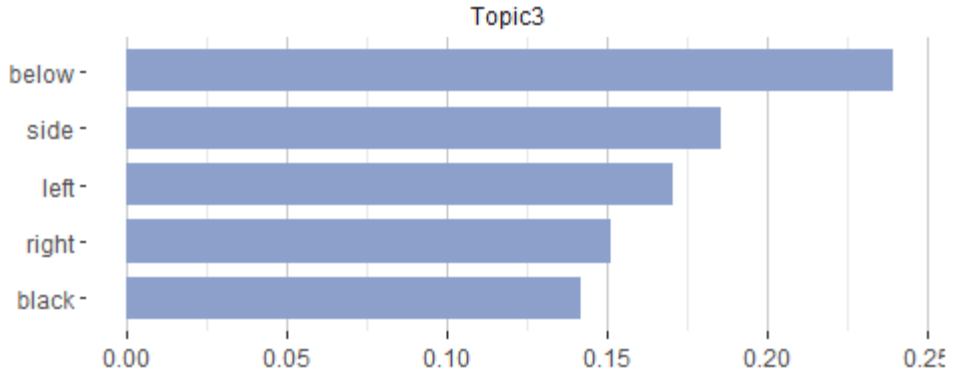
- Similar to stop words, aim to remove words that add minimal information
- Words that only appear in a few reports contribute little to building a corpus
- Removing these words lets the high info words have a greater impact

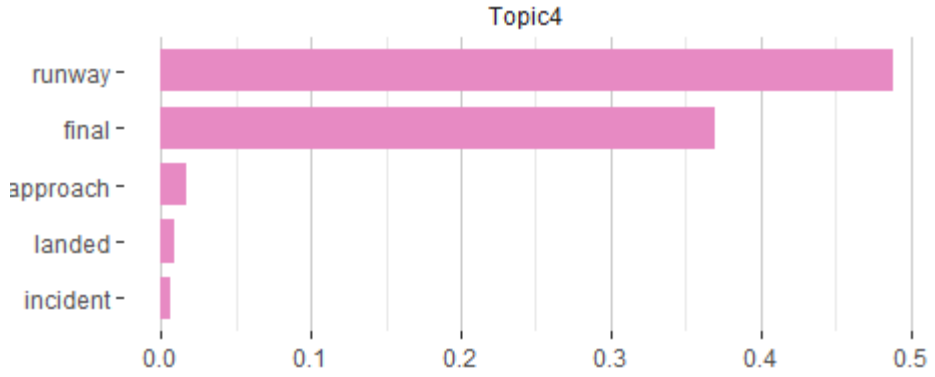
TOPIC MODELING



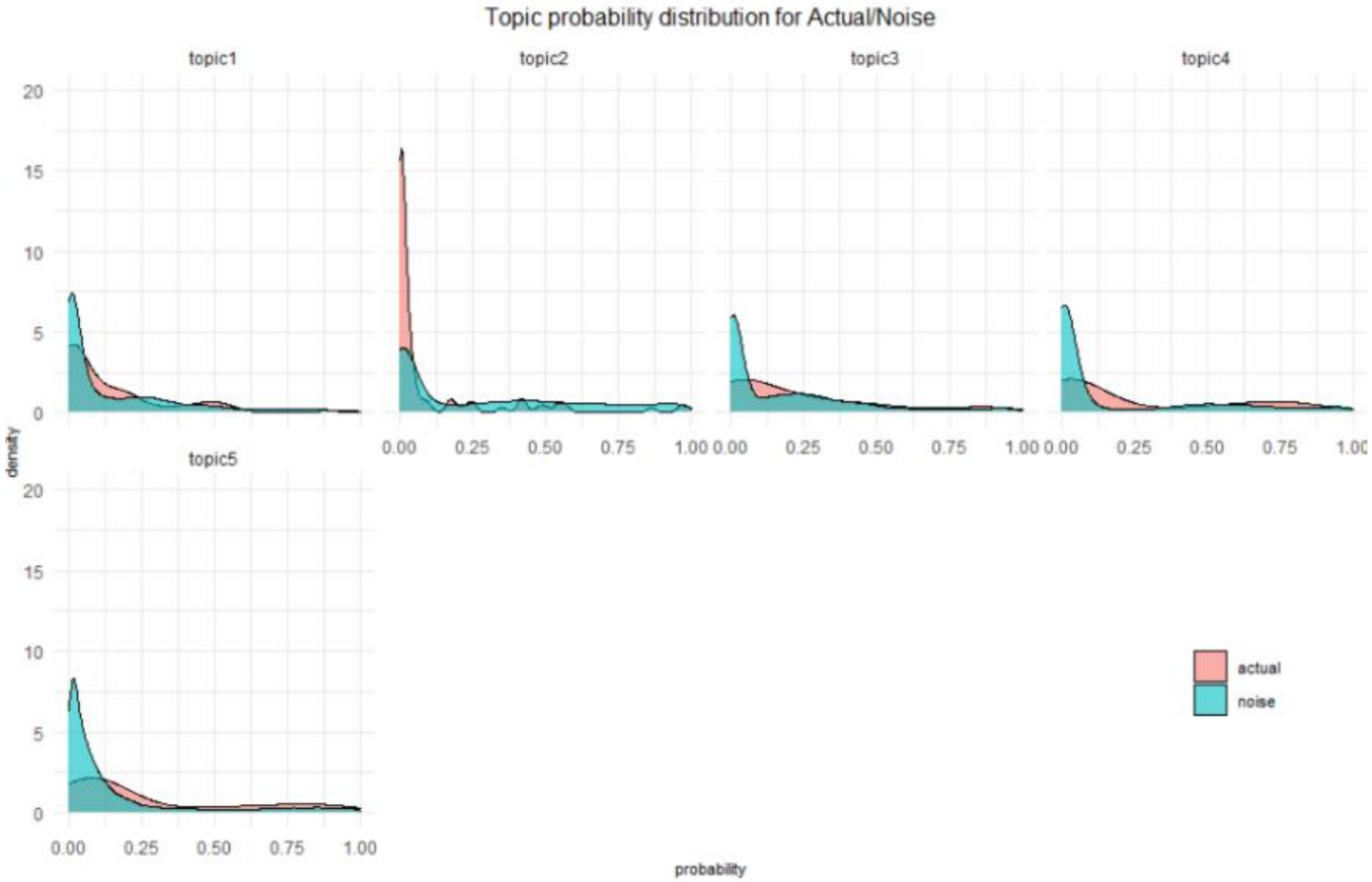




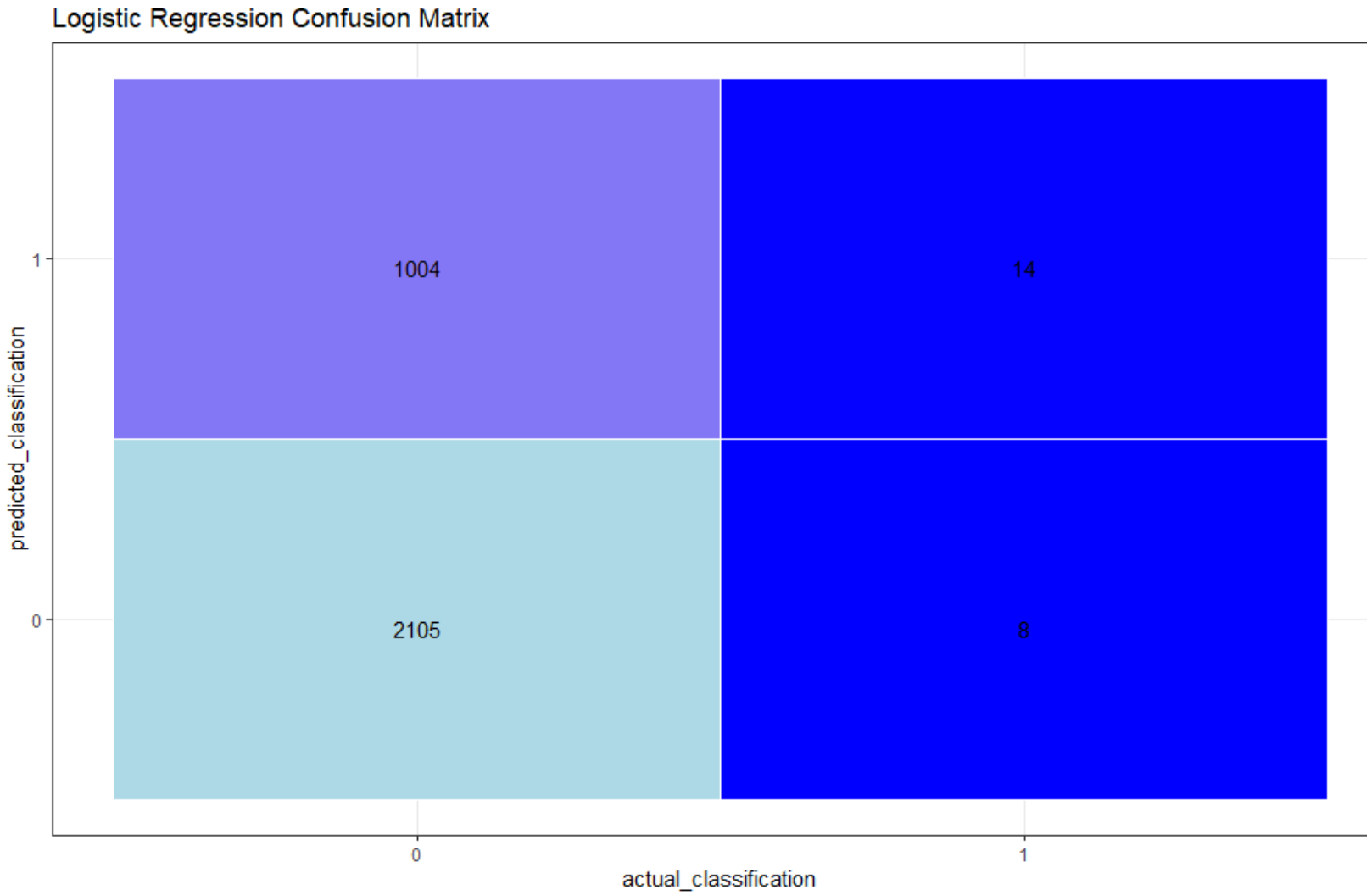




PROBABLITY DISTRIBUTION FOR ACTUAL VS NOISE



CONFUSION MATRIX



- True Positive Rate = 0.63
- False Positive rate = 0.32

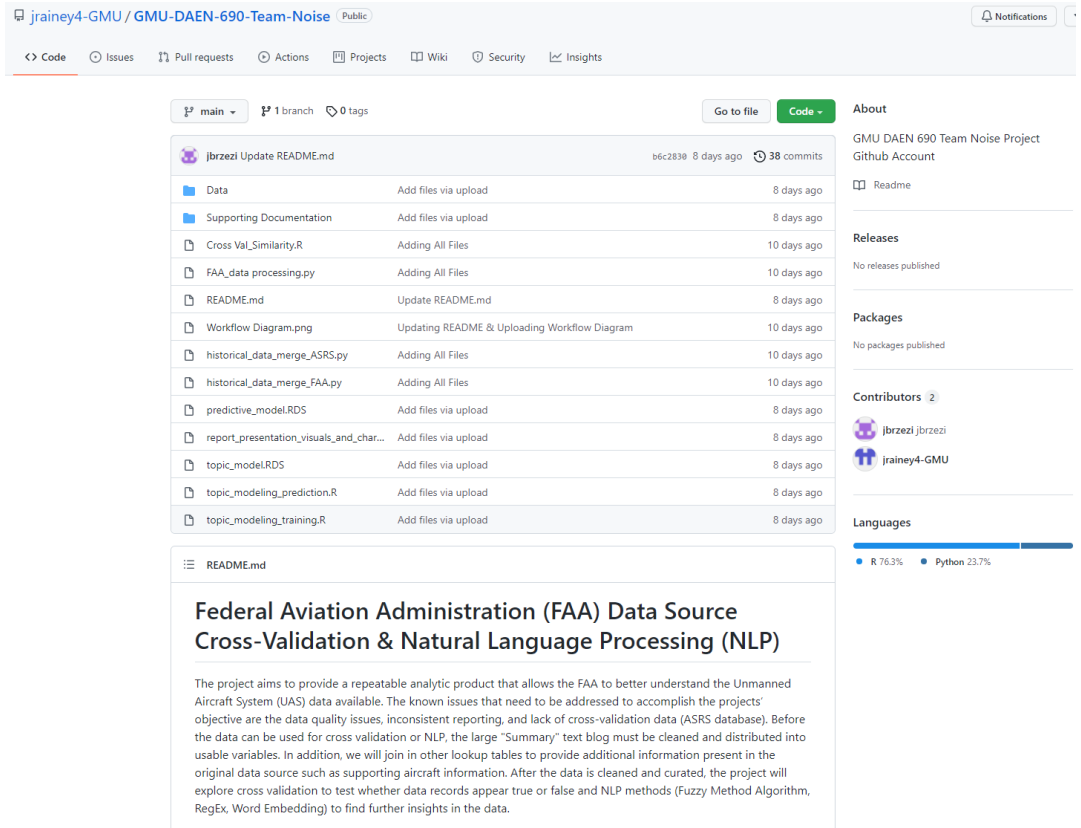
SUMMARY

The team was able to:

- Streamline process for cleaning, standardizing and extracting data from our data sources
- Cross-validate FAA UAS data against ASRS data
- Build predictive model to classify new reports

Files can be found here:

- https://github.com/jrainey4-GMU/GMU-DAEN-690-Team-Noise



NEXT STEPS/RECOMMENDATIONS

- Validating new reports with actual confirmed readings from internal FAA data
- If true actuals can’t be retrieved:
  - Using multiple UAS data bases with the similarity matching – ASRS, Cities (San Francisco, New York, etc..)
  - More sophisticated similarity matching
- More advanced predictive models that are popular for NLP

QUESTIONS?

Q & A

REFERENCES

1. "A Brief History of the FAA | Federal Aviation Administration," Faa.gov, 2021. https://www.faa.gov/about/history/brief_history.
2. "Unmanned Aircraft Systems (UAS)," Faa.gov, Sep. 10, 2019. https://www.faa.gov/uas/.
3. Lizotte, Katherine (FAA), "FAA Aerospace Forecast Fiscal Year 2016-2036," 2016. [Online]. Available: https://www.faa.gov/data_research/aviation/aerospace_forecasts/media/unmanned_aircraft_systems.pdf.
4. "UAS by the Numbers," Faa.gov, Sep. 30, 2019. https://www.faa.gov/uas/resources/by_the_numbers/.
5. B. I. Intelligence, "Drone market outlook: industry growth trends, market stats and forecast," Business Insider, Feb. 04, 2021. https://www.businessinsider.com/drone-industry-analysis-market-trends-growth-forecasts.
6. "Global Consumer Drones Market Report 2021: Hobbyist & Gaming; Aerial Photography; Others - Forecast to 2030 - ResearchAndMarkets.com," www.businesswire.com, Jun. 21, 2021. https://www.businesswire.com/news/home/20210621005402/en/Global-Consumer-Drones-Market-Report-2021-Hobbyist-Gaming-Aerial-Photography-Others---Forecast-to-2030---ResearchAndMarkets.com. (accessed Nov. 29, 2021).
7. M. D. Shear and M. S. Schmidt, "White House Drone Crash Described as a U.S. Worker's Drunken Lark," The New York Times, Jan. 27, 2015.
8. M. Gajewski, "Drone strikes commercial aircraft in Quebec: Garneau," CTVNews, Oct. 15, 2017. https://www.ctvnews.ca/canada/drone-strikes-commercial-aircraft-in-quebec-garneau-1.3633035..
9. R. Niles, "172 Substantially Damaged By Police Drone - AVweb," AVweb, Aug. 21, 2021. https://www.avweb.com/aviation-news/172-substantially-damaged-by-police-drone/. (accessed Nov. 29, 2021).
10. P. Gregg, "Risk in the Sky?," Udayton.edu, Oct. 22, 2018. https://udayton.edu/udri/news/18-09-13-risk-in-the-sky.php.
11. A. May, "Drones can do serious damage to airplanes, video shows," USA TODAY, Oct. 17, 2018. https://www.usatoday.com/story/travel/nation-now/2018/10/17/drones-crashing-into-airplanes-quadcopters-damage-video/1657112002/.
12. "5 Biggest Drone Incidents at Airports in 2020," www.robinradar.com, Dec. 23, 2020. https://www.robinradar.com/press/blog/5-biggest-drone-incidents-at-airports-in-2020.
13. "Small Unmanned Aircraft Systems (UAS) Regulations (Part 107) | Federal Aviation Administration," www.faa.gov, Oct. 06, 2020. https://www.faa.gov/newsroom/small-unmanned-aircraft-systems-uas-regulations-part-107?newsId=22615.
14. "UAS Remote Identification Overview," www.faa.gov, Oct. 13, 2021. https://www.faa.gov/uas/getting_started/remote_id/.
15. "UAS Sightings Report," Faa.gov, Jul. 22, 2019. https://www.faa.gov/uas/resources/public_records/uas_sightings_report/.
16. U. S. G. A. Office, "Small Unmanned Aircraft Systems: FAA Should Improve Its Management of Safety Risks," www.gao.gov, May 24, 2014. https://www.gao.gov/products/gao-18-110 (accessed Nov. 10, 2021).
17. "Simple, Consistent Wrappers for Common String Operations," stringr.tidyverse.org. https://stringr.tidyverse.org/ (accessed Nov. 30, 2021).
18. "10 minutes to pandas — pandas 1.1.0 documentation," pandas.pydata.org. https://pandas.pydata.org/pandas-docs/stable/user_guide/10min.html.
19. S. Prabhakaran, "Cosine Similarity – Understanding the math and how it works (with python codes)," Machine Learning Plus, Oct. 22, 2018. https://www.machinelearningplus.com/nlp/cosine-similarity/.
20. R. Khandelwal, "Word Embeddings for NLP," Medium, Dec. 28, 2019. https://towardsdatascience.com/word-embeddings-for-nlp-5b72991e01d4.
21. S. Li, "Natural Language Processing for Fuzzy String Matching with Python," Medium, Dec. 06, 2018. https://towardsdatascience.com/natural-language-processing-for-fuzzy-string-matching-with-python-6632b7824c49.
22. S. Li, "Topic Modeling and Latent Dirichlet Allocation (LDA) in Python," Medium, Jun. 01, 2018. https://towardsdatascience.com/topic-modeling-and-latent-dirichlet-allocation-in-python-9bf156893c24#:~:text=Topic%20modeling%20is%20a%20type.
23. M. Raza, "Importance of NLP in Data Science," TheLeanProgrammer, Jun. 12, 2021. https://medium.com/theleanprogrammer/importance-of-nlp-in-data-science-3e2fd2424b2d (accessed Nov. 07, 2021).
24. A. Sieg, "Text Similarities : Estimate the degree of similarity between two texts," Medium, Nov. 13, 2019. https://medium.com/@adriensieg/text-similarities-da019229c894.
25. D. Q. Nykamp, "The dot product - Math Insight," Mathinsight.org, 2021. https://mathinsight.org/dot_product (accessed Nov. 10, 2021).
26. N. Seth, "Latent Dirichlet Allocation (LDA) | LDA using Gensim," Analytics Vidhya, Jun. 27, 2021. https://www.analyticsvidhya.com/blog/2021/06/topic-modeling-and-latent-dirichlet-allocationlda-using-gensim-and-sklearn-part-1/ (accessed Dec. 01, 2021).
27. S. Chakravarthy, "Tokenization for Natural Language Processing," Medium, Jul. 10, 2020. https://towardsdatascience.com/tokenization-for-natural-language-processing-a179a891bad4.
28. S. Srinidhi, "Understanding Word N-grams and N-gram Probability in Natural Language Processing," Medium, Jan. 09, 2020. https://towardsdatascience.com/understanding-word-n-grams-and-n-gram-probability-in-natural-language-processing-9d9eef0fa058.
29. N. Seth, "Topic Modeling and Latent Dirichlet Allocation (LDA) using Gensim," Analytics Vidhya, Jun. 28, 2021. https://www.analyticsvidhya.com/blog/2021/06/part-2-topic-modeling-and-latent-dirichlet-allocation-lda-using-gensim-and-sklearn/ (accessed Nov. 02, 2021)