**Datasets**

Podcasts: (transcribe folder) <https://nam02.safelinks.protection.outlook.com/?url=https%3A%2F%2Frutgers.box.com%2Fs%2Fxdgzu9p54dsn81g8ofm08g4uegcr2y9n&data=05%7C01%7Crm1667%40scarletmail.rutgers.edu%7Ccc503d7b29814dea6a4f08dba8bc18b1%7Cb92d2b234d35447093ff69aca6632ffe%7C1%7C0%7C638289298032978620%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=rf%2B5YrMCRuFY%2Bdh16E3o2mmr2NTJrfPCpEHSD%2FXmlRc%3D&reserved=0>

Email: <https://nam02.safelinks.protection.outlook.com/?url=https%3A%2F%2Fdocs.google.com%2Fdocument%2Fd%2F1zGrmTyGxRPrrwX4LXumJmnd_-oak_KA5vXZzlkD7CPc%2Fedit&data=05%7C01%7Crm1667%40scarletmail.rutgers.edu%7Ccc503d7b29814dea6a4f08dba8bc18b1%7Cb92d2b234d35447093ff69aca6632ffe%7C1%7C0%7C638289298032978620%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=l%2F2SWWfraAWxL5QY2A%2FN87M5cRsgR2GppMM4w%2F3PBIA%3D&reserved=0>

TV data: <https://nam02.safelinks.protection.outlook.com/?url=https%3A%2F%2Fgithub.com%2Fnotnews%2Farchive_news_cc%23data&data=05%7C01%7Crm1667%40scarletmail.rutgers.edu%7Ccc503d7b29814dea6a4f08dba8bc18b1%7Cb92d2b234d35447093ff69aca6632ffe%7C1%7C0%7C638289298032978620%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=hVUg2q8Vs9UDal753XEjz%2Bq8Zi5wGVlnqtMf87G%2F1wo%3D&reserved=0>

Research paper related to the project <https://nam02.safelinks.protection.outlook.com/?url=https%3A%2F%2Farxiv.org%2Fpdf%2F2308.02068.pdf&data=05%7C01%7Crm1667%40scarletmail.rutgers.edu%7Ccc503d7b29814dea6a4f08dba8bc18b1%7Cb92d2b234d35447093ff69aca6632ffe%7C1%7C0%7C638289298032978620%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C3000%7C%7C%7C&sdata=4TqYcdxvYRAM%2B1b39R43SZ8vixVedmldw9C5CqfYLmY%3D&reserved=0>

Github link for the project:

<https://github.com/RohitMacherla3/narrative-detection-nlp-topicmodeling>

**Week 1: Sep 1 - Sep 11**

1. **Load the Datasets**

**Emails:**

Total Emails count - 186618

2022 Emails count - **80100**

**Podcasts:**

Total Episodes - 10245

2022 Episodes - **6949**

**TV Data:**

Total shows - 917221 (32 GB)

2022 Shows of selected 5 broadcasters - **43284** (1.7GB)

Commands used to load

1. cat archive-cc-2022.csv.gz{aa,ab,ac,ad,ae,af} > combined-archive-cc-2022.csv.gz
2. gunzip combined-archive-cc-2022.csv.gz
3. **Clean and perform EDA**
4. **Read the Research Paper**

**Week 2: Sep 12 - Sep 20**

1. **Standardize the data**

Columns were fixed for each source

1. **Find data consistency over months**

Data was available for 12 months for both Email and Podcast data. However, data was missing for the months November and December for TV data.

1. **Perform Pre-processing**

**Next action items**

1. **Understand LLM usage from the paper**
2. **Get Embeddings for the data**
3. **Run the DP-Means Clustering to identify narratives.**

**Resources**

We haven’t finalized the release code for the narrative detection, but the code is largely based off of<https://github.com/BGU-CS-VIL/pdc-dp-means> except with a modification to utilize cosine similarity by altering (<https://github.com/BGU-CS-VIL/pdc-dp-means/blob/main/paper_code/cluster/_k_means_lloyd.pyx)>. Hope that helps!