**Title: MIME Diversity in the Text Retrieval Conference (TREC)**

**Group members:**

**1. Executive summary:**

**2. Dataset discerption and download strategy: (suman)**

- How we downloaded the data

- How many files were there in total and how many of them were empty

- Distribution of the mimetypes (with the pie chart)

**3. Methodology and results:**

We need to say which mimetypes we used and why?

**3.1. BFA:**

**3.2. BFC: (simin)**

**3.3. Cross-Correlation: (simin)**

**3.4. FHT: (suman)**

**4.4. Updating Tika Mime repository: (milee)**

**4. Discussion:**

**4.1 Comparison of different algorithms:**

1. How well did BFA work, compared to FHT?
2. Did the D3 interactive visualizations help you understand the byte frequencies, and to identify patterns?

**4.2 Application/octet-stream:**

1. Were you able to discern any new MIME types within the 200, 000 application/octet-stream (“unknown”) types?
2. Why Tika’s detector was unable to discern the MIME types?
3. Was it lack of byte patterns and specificity in the fingerprint?
4. An error in MIME priority precedence?
5. Lack of sensitivity in the ability to specify competing MIME magic priorities and Bytes/offsets?
6. Also include your thoughts about Apache Tika – what was easy about using it? What wasn’t?

**5. Conclusion:**