**Vanessa Williams**

**Project Milestone 2: Identifying Data Sources**

For this milestone, I selected two different data sources for my project, this will help develop a comprehensive analysis by combining both structured metadata and unstructured text data.

**Data Source 1: Doc2Vec Dataset from Kaggle**

* **Link**: [Doc2Vec Pre-trained Model](https://www.kaggle.com/datasets/tayorm/arxiv-papers-metadata)
* **Description**: This dataset provides a pre-trained Doc2Vec model, which is highly useful for capturing the semantic relationships between documents and word vectors. Doc2Vec is an extension of the Word2Vec algorithm, designed specifically for generating vector representations of entire documents, as opposed to individual words. I will use this model to transform large chunks of textual data into fixed-length feature vectors, making it easier to process and analyze textual information.
* **Potential Use**: I will use this dataset to generate vectorized representations of my documents, and will serve as input for machine learning models. By applying clustering and classification algorithms to these vectors, I can gain insights into patterns within the text and explore possible topic groupings.

**Data Source 2: The Open Research Metadata Dataset from CERN**

* **Link**: [Open Research Metadata](https://www.kaggle.com/datasets/japkeeratsingh/arxiv-dataset-trained-doc2vec-model)
* **Description**: This dataset contains structured metadata about various research papers available in open-access repositories. The metadata includes details like authors, titles, keywords, abstracts, and publication dates. While it does not contain the full text of the papers, it provides significant contextual information that will be valuable for understanding trends in research topics and the relationships between different academic fields.
* **Potential Use**: This dataset will allow me to perform exploratory data analysis (EDA) to examine research trends, publication frequency, and the most common keywords across various domains. Combining this structured metadata with vectorized representations from the Doc2Vec model will also enable me to compare and cluster research papers based on their semantic similarity.

**Relationships Between the Data**

Both data sources are highly complementary. The structured metadata from the CERN dataset provides the context necessary for understanding and categorizing research papers, while the Doc2Vec vectors will allow for more in-depth semantic analysis. By integrating these two sources, I will be able to analyze the similarity between papers in both a structured and unstructured way, offering a more holistic view of the data.

**Data Preprocessing and Planned Analysis**

* **Preprocessing**: For both datasets, I will need to perform extensive preprocessing. This will include cleaning text, handling missing values, removing duplicates, and ensuring that the structured metadata and the unstructured text from the papers are compatible for comparison.
* **Feature Engineering**: Using the Doc2Vec model, I will generate document vectors from the textual data. For the metadata, I will focus on extracting key features such as keywords and publication years to facilitate analysis.
* **Exploratory Data Analysis**: I will explore trends in the metadata dataset, such as publication frequency and topic distributions over time.
* **Clustering and Classification**: I plan to apply clustering algorithms to the document vectors to group similar papers. Additionally, classification algorithms can be used to categorize papers based on the metadata features, such as keyword relevance or publication year.

Overall, this combination of structured metadata and unstructured document vectors will provide valuable insights into research trends, document similarity, and topic discovery across academic papers.

### References

Le, Q. V., & Mikolov, T. (2014). Distributed representations of sentences and documents. Proceedings of the 31st International Conference on Machine Learning, PMLR, 32(2), 1188–1196. https://proceedings.mlr.press/v32/le14.html

Zenodo. (2017). Open research metadata. Zenodo. https://zenodo.org/record/1069039