HW 1 Rubric

1. Running program pipeline/code **(50pts)**

1.1 Additional features about the data (15pts, task 4)

1.2 Identify at least one other datasets with at least one feature (10pts, task 5)

1.2.1 For each additional dataset beyond minimum 2pts, with 1pt for each additional feature beyond minimum (at most 10pts)

1.3 Applying Tika Similarity (15pts)

2. Readme **(5pts)**

3. Report **(45pts)**

3.1 Structure **(5pts)**

3.2 Satisfactory answers to the following questions **(35pts)**

3.2.1 For each feature you add, be prepared to discuss what types of queries it will allow you to answer and also how you computed the feature?

3.2.2 Compare and contrast clusters from Jaccard, Cosine Distance, and Edit Similarity – do you see any differences? Why? What similarity metrics produced more (in your opinion) accurate measurements? Why?

3.2.3 How to the resultant clusters generated highlight the features you extracted?

3.2.4 what you noticed about the dataset as you completed the tasks.

3.2.5 What questions did your new joined datasets allow you to answer about the Bik et al papers previously unanswered?

3.2.6 You should also clearly explain which datasets you used to join the problematic papers data and how you extracted the new features from each dataset.

3.3 Summary **(5pts)**