Applied Spatial and Temporal Data Analysis
COMP SCI 6001
Manish Jaisinghani

1. Data Collection: The data collection process is completed in two steps as described below:

   a. Website List: I collected the website list by using online website scrapper http://www.xsitemap.com. The scrapper parses all the sub URLs from a given URL and provides a list of the parsed URLs. The URL list has been divided into 4 categories and 25 URLs have been taken from each category:
      - Politics – 25 articles
      - Health – 25 articles
      - World – 25 articles
      - Technology – 25 articles

   b. Website Scrapper: I have written HTML website scrapper in python using beautiful soup and NLTK as packages to parse the HTML packages and remove words from the collected list that do not add to meaning of the article, for example "a, the, he, she, these etc." A complete list of the words which have been remove or not taken into consideration has been attached by the file name "stopwords".

2. Data Processing: The data processing is done in two parts as described below:

   a. Website parsing: A list of URLs is provided in the file by name "**website_list**". The python script parses through each website one by one. The script makes use of beautiful soup, parses the website and pulls out text data eliminating all the symbols, punctuations, styles and scripts. Once the data is collected the data is recorded in a list data structure and then stored making sure that there are no duplicate words in the list. The same procedure is repeated for every URL. While parsing the text of current article frequency of words occurring in the article is also recorded. The final step is just writing this data to a CSV file("**data.csv"**). In the csv file each article is represented as a row and each word is recorded in a column and the final matrix consists of the frequency of words.

   b. Similarity Calculation: The similarity matrix calculation takes place from a single python script(Similarity.py) that calculates Euclidean distance, cosine distance and Jacard distance and records them in three different csv files eudist.csv, cosdist.csv and jacdist.csv respectively. In these files rows and columns both represent the article and the matrix represents the similarity between a row article and a column article.

   c. Sorting: Each type of similarity is sorted in an order where the most similar types of article come first and least similar in the end. The three csv files eudist_Sorted, cosdist_Sorted and jacdist_Sorted represent the sorted files for three different types of similarity matrices.

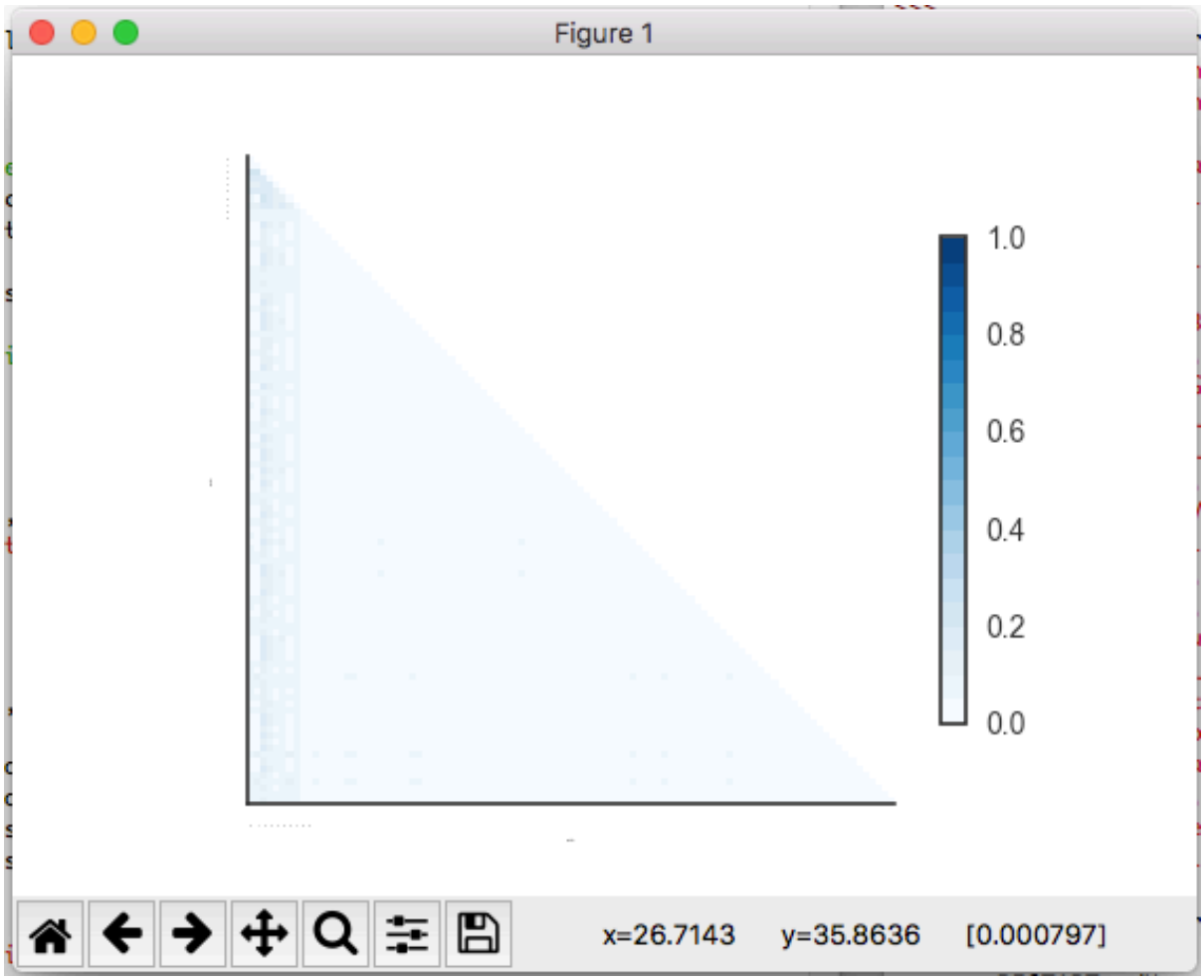
3.Data Analysis: Below is the data analysis for all the type of similarity matrices:

a. Euclidean Distance: Below are the top 5 articles as per Euclidean distance similarity matrix:

- The First column provides similarity between the article compared with itself hence the same is represented as “0”, which suggests that the article is exactly similar.
- The consecutive rows represent the similarity of article in the first column to all the other articles mentioned in consecutive 5 columns.
- The distance matrix has been arranged in lower to higher order where in a lower distance depicts increased similarity, whereas higher distance equals lower similarity.
- Example: As an example let us consider the first row of article, the article which is being compared to all the other articles is “**Tech industry braces for Trump's visa reform - Feb. 5, 2017**”. This article has been taken from Technology category.
- The articles with which above article is being compared with are “Tech industry braces for Trump's visa reform - Feb. 5, 2017”, “Jack Dorsey: 'We benefit from immigration' - Video - Tech”, “Crash dummies are getting 'supersized' to resemble obese drivers - Video – Tech”, “Facebook doubles down on bereavement leave - Feb. 7, 2017”, “'We are entrepreneurs, professors, scientists, artists' - Jan. 29, 2017”. The articles mentioned above are from categories Technology, Technology, Technology, Technology and Technology respectively.
- It can clearly be established that the top most similar articles belong to the same category i.e. Technology and do talk about the same topic of concern.

Article Title	Tech industry braces for Trump's visa reform - Feb. 5, 2017	Jack Dorsey: 'We benefit from immigration' - Video - Tech	Crash dummies are getting 'supersized' to resemble obese drivers - Video - Tech	Facebook doubles down on bereavement leave - Feb. 7, 2017	'We are entrepreneurs, professors, scientists, artists' -
Tech industry braces for Trump's visa reform - Feb. 5, 2017	0	37.01351105	37.94733192	38.19685851	38.45776905
Article Title	Intel CEO touts $7 billion factory investment in Trump	Jack Dorsey: 'We benefit from immigration' - Video - Tech	Crash dummies are getting 'supersized' to resemble obese drivers - Video - Tech	Netflix wants in on toy and licensed merchandise business -	Facebook doubles down on bereavement leave -
Intel CEO touts $7 billion factory investment in Trump meeting - Feb. 8, 2017	0	30.61045573	32.01562119	32.61901286	32.71085447
Article Title	Libya Fast Facts - CNN.com	Oil and Gasoline Fast Facts - CNN.com	Becky Anderson on her heroes: Her fellow journalists - CNN.com	Brexit bill introduced to parliament - CNN.com	Abortion laws around the globe - CNN.com
Libya Fast Facts - CNN.com	0	87.90335602	88.06815543	88.09653796	88.43076388
Article Title	Cervical cancer death rates are much higher than thought - CNN.com	'Automated dermatologist' detects skin cancer with expert accuracy - CNN.com	Becky Anderson on her heroes: Her fellow journalists - CNN.com	Abortion laws around the globe - CNN.com	Boxer battles for title as his father battles cancer - CNN.com
Cervical cancer death rates are much higher than thought - CNN.com	0	72.29107829	73.35529974	73.87151007	74
Article Title	Putting the 'active' in activism: How to get kids to get involved - CNN.com	Becky Anderson on her heroes: Her fellow journalists - CNN.com	Man uses Facebook to find woman who saved his life - CNN.com	Abortion laws around the globe - CNN.com	This is not your typical 'Make America Great Again' ad -
Putting the 'active' in activism: How to get kids to get involved - CNN.com	0	49.8998998	49.95998399	50.51732376	51.51698749

Graphical representation of Euclidean distance matrix:



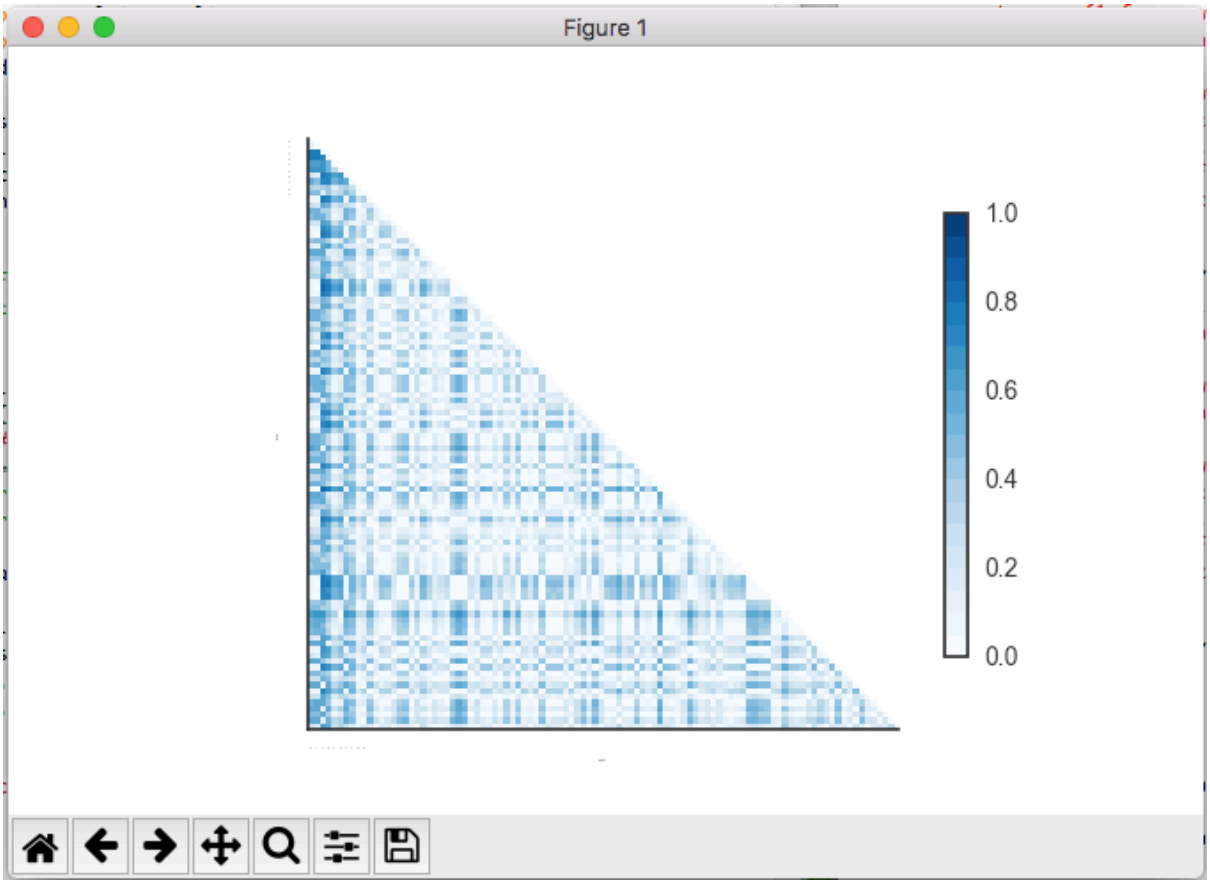
b. Cosine Distance: Below are the top 5 articles as per cosine similarity matrix:
   - The First column provides similarity between the article compared with itself hence the same is represented as "1", which suggests that the article is exactly similar.
   - The consecutive rows represent the similarity of article in the first column to all the other articles mentioned in consecutive 5 columns.
   - The distance matrix has been arranged in higher to lower order where in a higher numeric depicts increased similarity, whereas lower numeric equals lower similarity.
   - Example: As an example let us consider the first row of article, the article which is being compared to all the other articles is "**Tech industry braces for Trump's visa reform - Feb. 5, 2017**". This article has been taken from Technology category.
   - The articles with which above article is being compared with are "Tech industry braces for Trump's visa reform - Feb. 5, 2017", "More than 100 companies join legal fight against Trump travel ban - Feb. 5, 2017", "Intel CEO touts $7 billion factory investment in Trump meeting - Feb. 8, 2017",

"Jack Dorsey: 'We benefit from immigration' - Video – Tech" and "'We are entrepreneurs, professors, scientists, artists' - Jan. 29, 2017". The articles mentioned above are from category Technology, however a new article enters the list of 5 top articles in the cosine distance matrix.

- The similarity between this new entry to top 5 articles is greater as calculated by the cosine rule and the same looks correct as well. Both of the articles are concentrated around the same category and do talk about the same concern.

Article Titles	Tech industry braces for Trump's visa reform - Feb. 5, 2017	More than 100 companies join legal fight against Trump travel ban - Feb. 5, 2017	Intel CEO touts $7 billion factory investment in Trump meeting - Feb. 8, 2017	Jack Dorsey: 'We benefit from immigration' - Video - Tech	'We are entrepreneurs, professors, scientists, artists' - Jan. 29, 2017
Tech industry braces for Trump's visa reform - Feb.	1	0.571	0.527	0.521	0.505
Article Titles	Intel CEO touts $7 billion factory investment in Trump meeting - Feb. 8, 2017	Robot makes money off Trump's tweets and donates it to the ASPCA - Feb. 6, 2017	More than 100 companies join legal fight against Trump travel ban - Feb. 5, 2017	Jack Dorsey: 'We benefit from immigration' - Video - Tech	'We are entrepreneurs, professors, scientists, artists' - Jan. 29, 2017
Intel CEO touts $7 billion factory investment in Trump meeting - Feb. 8,	1	0.607	0.591	0.575	0.544
Article Titles	Libya Fast Facts - CNN.com	Lebanon Fast Facts - CNN.com	Pakistan Fast Facts - CNN.com	Oil and Gasoline Fast Facts - CNN.com	Cuba Fast Facts - CNN.com
Libya Fast Facts - CNN.com	1	0.35	0.307	0.284	0.231
Article Titles	Cervical cancer death rates are much higher than thought - CNN.com	'Automated dermatologist' detects skin cancer with expert accuracy - CNN.com	Can burnt toast and roasted potatoes cause cancer? - CNN.com	Women rush to get IUDs because of Trump - CNN.com	Boxer battles for title as his father battles cancer - CNN.com
Cervical cancer death rates are much higher than thought - CNN.com	1	0.431	0.312	0.281	0.27
Article Titles	Putting the 'active' in activism: How to get kids to get involved - CNN.com	Kids say the darndest things about being president - CNN.com	Dalai Lama's guide to the next four years - CNN.com	Quitting smoking is the hardest resolution to keep - CNN.com	Women rush to get IUDs because of Trump - CNN.com
Putting the 'active' in activism: How to get kids	1	0.516	0.445	0.416	0.414

Below is the graphical representation of the unsorted matrix. We have plotted all the points on x and y axis.



c. Jacard Distance: Below are the top 5 articles as per Jacard distance similarity:

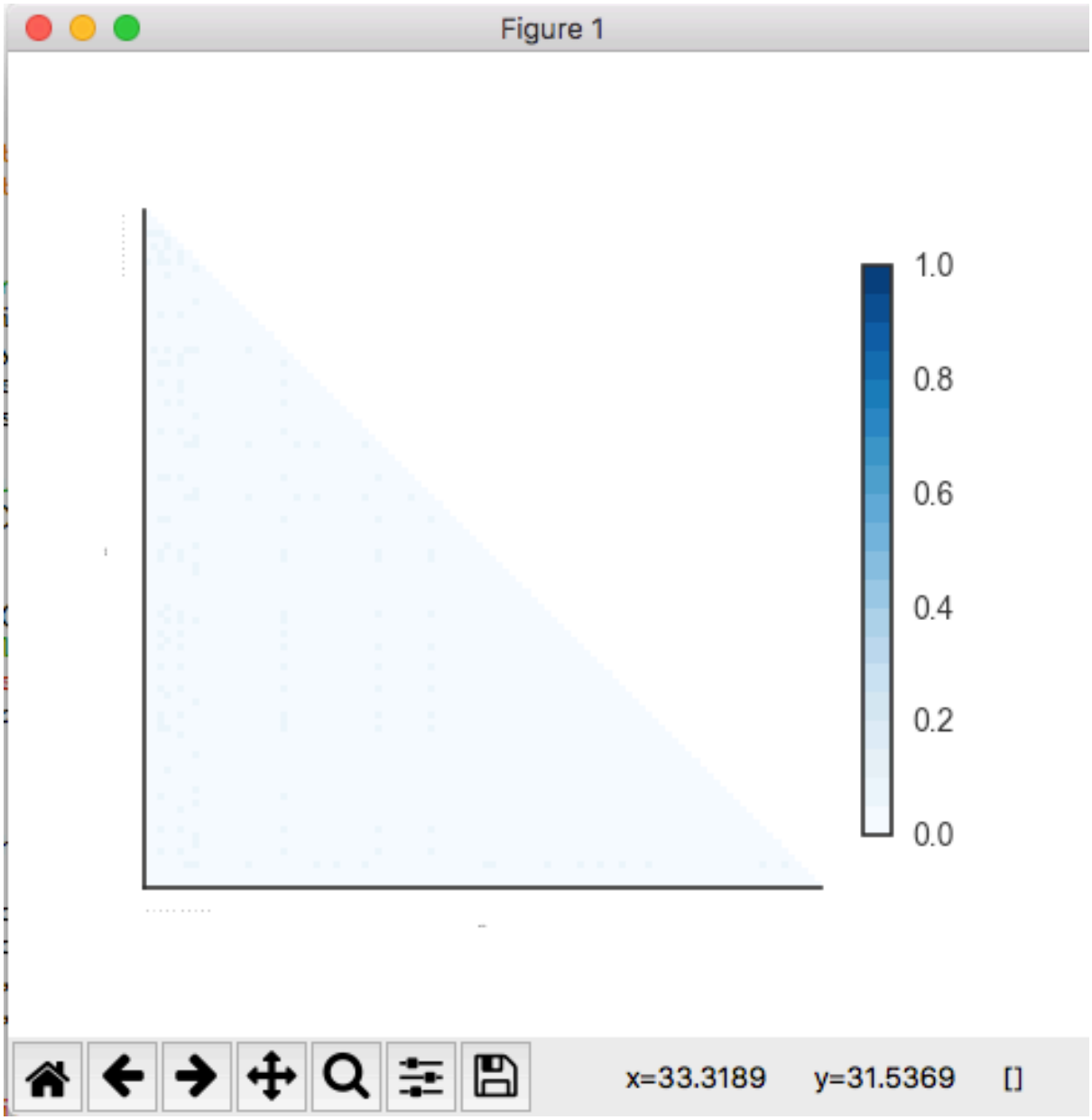
- The First column provides similarity between the article compared with itself hence the same is represented as “1”, which suggests that the article is exactly similar.
- The consecutive rows represent the similarity of article in the first column to all the other articles mentioned in consecutive 5 columns.
- The distance matrix has been arranged in higher to lower order where in a higher numeric depicts increased similarity, whereas lower numeric equals lower similarity.
- Example: As an example let us consider the first row of article, the article which is being compared to all the other articles is “Tech industry braces for Trump's visa reform - Feb. 5, 2017”. This article has been taken from Technology category.
- The articles with which above article is being compared with are “Tech industry braces for Trump's visa reform - Feb. 5, 2017”, “Putting the 'active' in activism: How to get kids to get involved - CNN.com”, “Inside the quirky world of Snapchat - Feb. 3, 2017”, “Trump's talk puts GOP in a jam - CNNPolitics.com” and “Women rush to get IUDs because of Trump -

CNN.com". The articles mentioned above are from category Technology, health, technology, politics and health categories.

- The Jacard distance reports similarity in a different way as compared to cosine and Euclidean distances, in other words we can see we are experiencing information loss in this case.

Article Titles	Tech industry braces for Trump's visa reform - Feb. 5, 2017	Putting the 'active' in activism: How to get kids to get involved - CNN.com	Inside the quirky world of Snapchat - Feb. 3, 2017	Trump's talk puts GOP in a jam - CNNPolitics.com	Women rush to get IUDs because of Trump - CNN.com
Tech industry braces for Trump's visa reform - Feb. 5,	1	0.916666667	0.909090909	0.846153846	0.846153846
Article Titles	Intel CEO touts $7 billion factory investment in Trump meeting - Feb. 8, 2017	Twitter tries new measures in crackdown on harassment - Feb. 7, 2017	Boxer battles for title as his father battles cancer - CNN.com	Inside the quirky world of Snapchat - Feb. 3, 2017	Tech industry braces for Trump's visa reform - Feb. 5, 2017
Intel CEO touts $7 billion factory investment in Trump meeting - Feb. 8, 2017	1	0.909090909	0.909090909	0.909090909	0.833333333
Article Titles	Libya Fast Facts - CNN.com	'Fast radio burst' tracked to a dwarf galaxy 3 billion light-years away - CNN.com	Article 50: UK's path to Brexit, explained - CNN.com	The 6 most scientifically valid methods to quit	Kids say the darndest things about being president - CNN.com
Libya Fast Facts - CNN.com	1	0.782608696	0.777777778	0.75	0.736842105
Article Titles	Cervical cancer death rates are much higher than thought - CNN.com	Wild supercows return to Europe - CNN.com	I rented a guy's $145,000 Tesla Model X - Feb. 11, 2017	Can burnt toast and roasted potatoes cause cancer? - CNN.com	Trump says he'll send in feds if Chicago doesn't fix 'carnage' - CNNPolitics.com
Cervical cancer death rates are much higher than thought - CNN.com	1	0.75	0.733333333	0.722222222	0.6875
Article Titles	Putting the 'active' in activism: How to get kids to get involved - CNN.com	Tech industry braces for Trump's visa reform - Feb. 5, 2017	Inside the quirky world of Snapchat - Feb. 3, 2017	Trump's talk puts GOP in a jam - CNNPolitics.com	Women rush to get IUDs because of Trump - CNN.com
Putting the 'active' in activism: How to get kids to get involved - CNN.com	1	0.916666667	0.833333333	0.785714286	0.785714286

Graphical representation of Jacard distance matrix:



4.Conclusion:

The similarity matrices represented by Euclidean and cosine strategies are pointing out towards similarity in articles in a more similar way, however Jacard seems not to be finding similarity between articles based on categories that were chosen to be parsed.

The graphical representation of all the three matrices highlights the cluster formation. Cluster formation and

visibility is more obvious in cosine distance matrix as compared to other two.

Based on the algorithmic results, my analysis points out that cosine similarity is better towards the information analysis task which was assigned to us.