HW\_2 Exploring iSchool Faculty Research

The dataset from iSchool SU Experts database is provided to us for analysing the research conducted by “30” different professors in iSchool.

Performed following analysis on the dataset provided

* Vectorization
* Cosine Similarity Measure
* K Means Clustering
* Topic Modelling using LDA

The initial set of analysis include “Vectorization” using TF\_IDF (Term Frequency-Inverse Document frequency) which calculates the frequency of word in the document and it’s importance. Also, removed stop words during this process

A sparse matrix is obtained after applying TF\_IDF vectorization and this matrix is converted to an array for further analysis. Top 20 words in their decreasing frequency from the documents are

[('data', 3.3417241194603253),

('social', 3.224456732001306),

('information', 2.8453581266106114),

('research', 2.817046851834984),

('learning', 2.101340617484771),

('use', 1.9654686012491556),

('online', 1.9551208418005686),

('design', 1.8880682182730133),

('library', 1.8162133423386542),

('work', 1.654923802158007),

('study', 1.6465350783537454),

('paper', 1.6207140293225735),

('media', 1.5953398678879227),

('science', 1.5740156953277389),

('analysis', 1.382128649053957),

('based', 1.3706554776313544),

('model', 1.3509842896062718),

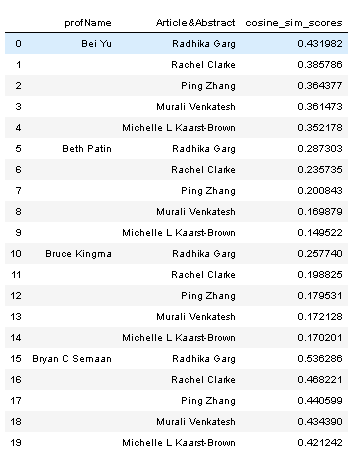
('using', 1.266176569990889),

('new', 1.2659556131159506),

('systems', 1.2412379374376585)]

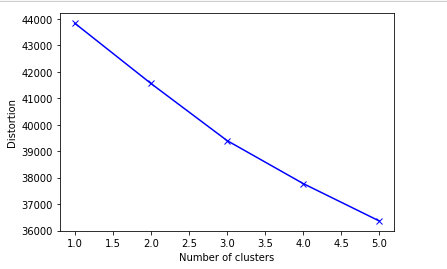
**Cosine similarity Measure:**

Cosine similarity measure is used to calculate the similarity measures between the documents. In our analysis, we found the top 5 faculty members on comparing their documents with each faculty members in the list by sorting them with similarity measures. Sample of faculty members is displayed below

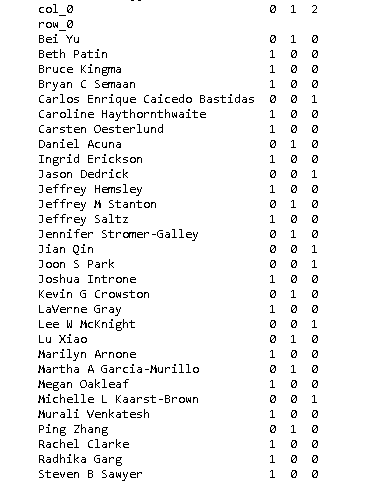


**K Means Clustering:**

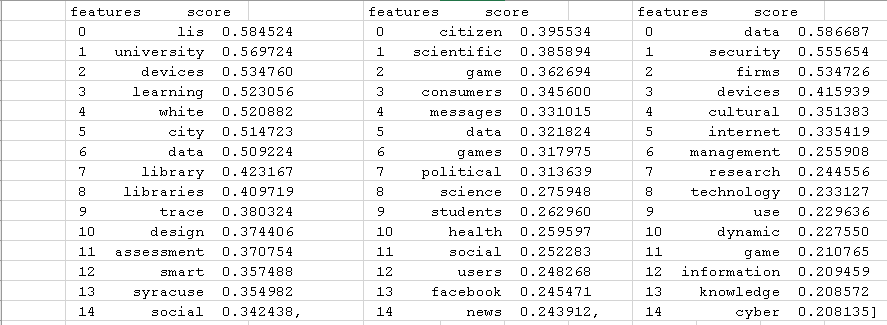
Optical number of clusters are found using elbow method. In this process, graph is plotted against each cluster and the distortion in the cluster. The value of distortion decreases with the increase in the K value. In our analysis, we have considered the K range from1 to 6 and observed that the optimal value of K is at “3” as the error is decreasing slowly after “3” compared to the other K values. Hence, K = 3 is considered as a optimal value for clustering



After, finding the optimal number of clusters, we are going to implement the algorithm by importing kMeans from sklearn. KMeans algorithm doesn’t normalize the data, hence we are using standardscalar method which normalizes the data by equalizing variance along different features as K-means is sensitive to variance in data, and features with larger variance have more emphasis on result. From the below image we can see that the authors are separated into three different clusters. Cluster[0] has 16 authors, cluster[1] has 8 authors and cluster[2] has 6 authors. Majority of the authors belong to Cluster[0].



The common words that are used by the faculty members in each cluster and their TF\_IDF scores are shown below. It is observed that the faculty members that belong to cluster[0] spoke conducted research on the university, city and libraries . Faculty in cluster[1] spoke conducted their research on students, news, messages, people and health. Whereas, the faculty in cluster [2] spoke about the firms and also concerned about the security



**LDA topic Modelling:**

The model was tested with different number of topics and observed that above 10 topics, the topics are redundant and does not have any proper information. Topic 500 features are considered from the from the titles and abstracts column and performed tf\_idf vectorization by removing the stop words.

Topic 10 topics with labels are listed below

**Research on firms**:

data metadata sharing study use science using devices management understanding firms smart systems knowledge method based paper results aims need

**Scientific paper:**

spectrum use research-based data management model new using scientific work paper learning science results communication dynamic systems access analysis

**Cyber Security**:

security cloud data wireless services-based users access computing systems cyber paper game user internet social approach sharing critical model

**Innovation**:

firm’s data smart use value children research devices industry technology using greater countries new developing study global internet innovation paper

**Security:**

social design research data learning online study work metadata new analysis security based media using public sharing model paper scientific

**Research on student academic performance**

learning library value libraries academic assessment literacy student librarians impact research using institutional assessing strategies university outcomes teaching results education

**Unknown**

social research data work online design use study science support analysis paper learning systems technologies citizen technology people knowledge community

**Analysis on Students data**

data science big process teams project paper use projects methodology team results model analytics management framework research learning students study

**Unknown**

science data social research study use model based work analysis support projects management community digital knowledge patterns characteristics practices technology

**Analysing Messages:**

data social research media metadata design analysis use study paper science learning knowledge messages online using model political approach based

**Conclusion:**

From the above analysis, it is observed that most of the faculty members conducted their research on the student’s academic performance, data from the firms, analysed messages and presented their scientific papers. Most of them focused on the cyber security analysis as well

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |