Text Document Clustering

—

Using Spectral Clustering & Particle Swarm Optimization

Table of Contents

What is Text Document Clustering?

- Document clustering is a gathering of textual content documents into groups or clusters
- Aim is to cluster the documents, which are internally logical but considerably different from each other
- Used in Information Retrieval, Information Extraction and Document Organization
- Google’s search engine is probably the best and most widely known example of Text Document Clustering

Motivation of Document Clustering

- To create structure of text data
- The number of available articles is large
- A large number of articles are added each day
- To link similar documents and remove duplicate documents
- The recommendations has to be generated and updated in real time
- Articles corresponding to same news are added from different sources
- Automatically group related document based on their content
- By clustering the articles we could reduce our domain of search for recommendations which leads in improved time efficiency to a great extent

Existing Algorithms in Literature

- Topic models
- Spectral Clustering
- Genetic Algorithms
- Semi-Supervised Clustering
- Meta Heuristic Optimization
- Particle Swarm Optimization(PSO)
- K-means clustering(gives local optimum)
- Swarm Intelligence(gives global optimum)
- Combination of Nystrom Spectral Clustering & Genetic Algorithm

Steps in Document Clustering

- Document preprocessing
  - Tokenization
  - Stemming
  - Lemmatization
- Document Representation
  - Word Count vector
  - TF-IDF vector
- Maximum Likelihood Estimation
  - Bernoulli distribution
  - Gaussian distribution
  - Poisson distribution

- Similarity measures
  - Cosine Similarity
  - Euclidean distance
- Spectral Clustering Algorithm
  - Similarity graph
  - Graph Laplacian Matrix
    - Un Normalized (D - W)
    - Normalized (D^−1/2*(D-W)* D^−1/2)
    - Normalized with Random Walks (D^−1/2*(D-W))
  - Eigen Vectors
- Particle Swarm Optimization

Spectral Clustering with PSO(SCPSO)

- It combines Spectral Clustering & Particle Swarm Optimization
- Improve the accuracy of Text Document Clustering
- Reduces global Convergence, computational complexity
- Can handle large number of objective function
- More flexible
- Can find clusters of arbitrary shapes, under realistic separations
- Able to cluster points which are not necessarily vectors

Spectral Clustering with PSO(SCPSO)

- The SCPSO is a spectral based clustering method which uses the particle swarm optimization algorithm.
- It has the following three important steps:
  - **Similarity graph generation**
  - **Particle swarm optimization**
  - **Clustering methods**
- Dataset that we have used is Reuters which contains text documents of different categories.

Flow of Clustering is as follows:

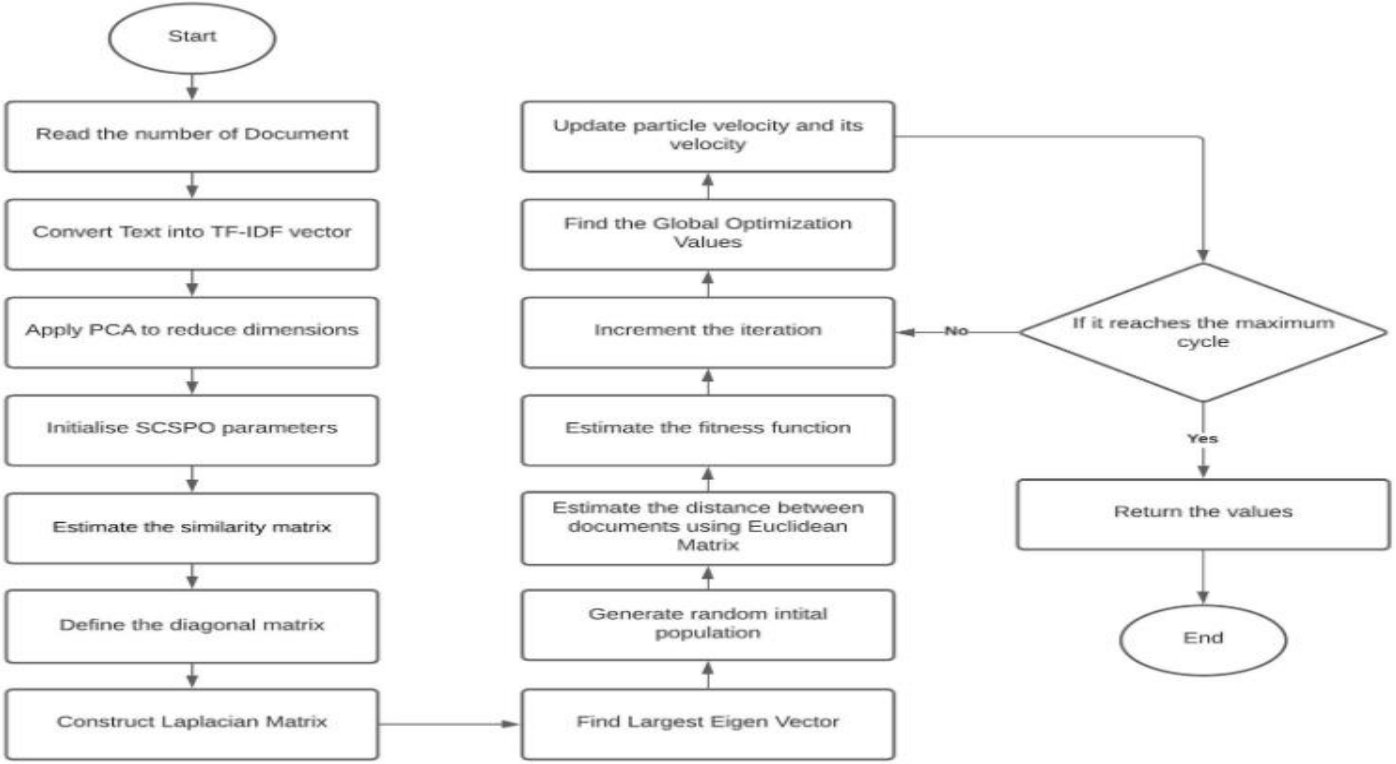


Fig 4.4 Flow chart for Text Document Clustering using while using PCA as Dimensionality Reduction and Spectral Clustering for clustering with Particle Swarm Optimization

Our Ideas and Contribution

We first implemented the SCPSO algorithms mentioned in paper and tried to achieve the accuracy as that of paper. In paper there is no mention of PCA so we tried the following:

- Computed **PCA** of **TF-IDF** feature vector for doing **Dimensionality Reduction**
- Tried PCA for values such as 2, 100, 200, 250, 280, 350, 500, 750, 900
- Best score was found when number of features = 270
- Tried SCPSO algorithm on Affinity matrix with **Gaussian Kernel**
- Tried SCPSO algorithm on Affinity matrix with **Euclidean Distance**

- For making visualisation possible(as there are 1000 features), we did **PCA** with 2 features to visualise the word Embeddings in 2D space
- We have used **Adjusted Rand Index(ARI)** to check how efficient results we have obtained
- Implemented bar plots for visualising the **ARI** of various clustering methods we tried
- Created **Word Embedding** of the input Text Document for visualising the text of Document in 2D space using **Gensim** library

Implementation Details

- Experimental Setup
  - 2.00 GHz Intel CPU with 2GB of RAM & running on windows 10
- Dataset
  - Reuters 21578
- Algorithms
  - SCPSO with K-means
- Idea incorporated
  - Applied PCA to TF-IDF feature vector
  - Applied PCA on Affinity Matrix with Euclidean Distance & Gaussian Kernel
- Performance measure
  - Adjusted rand Index(ARI)

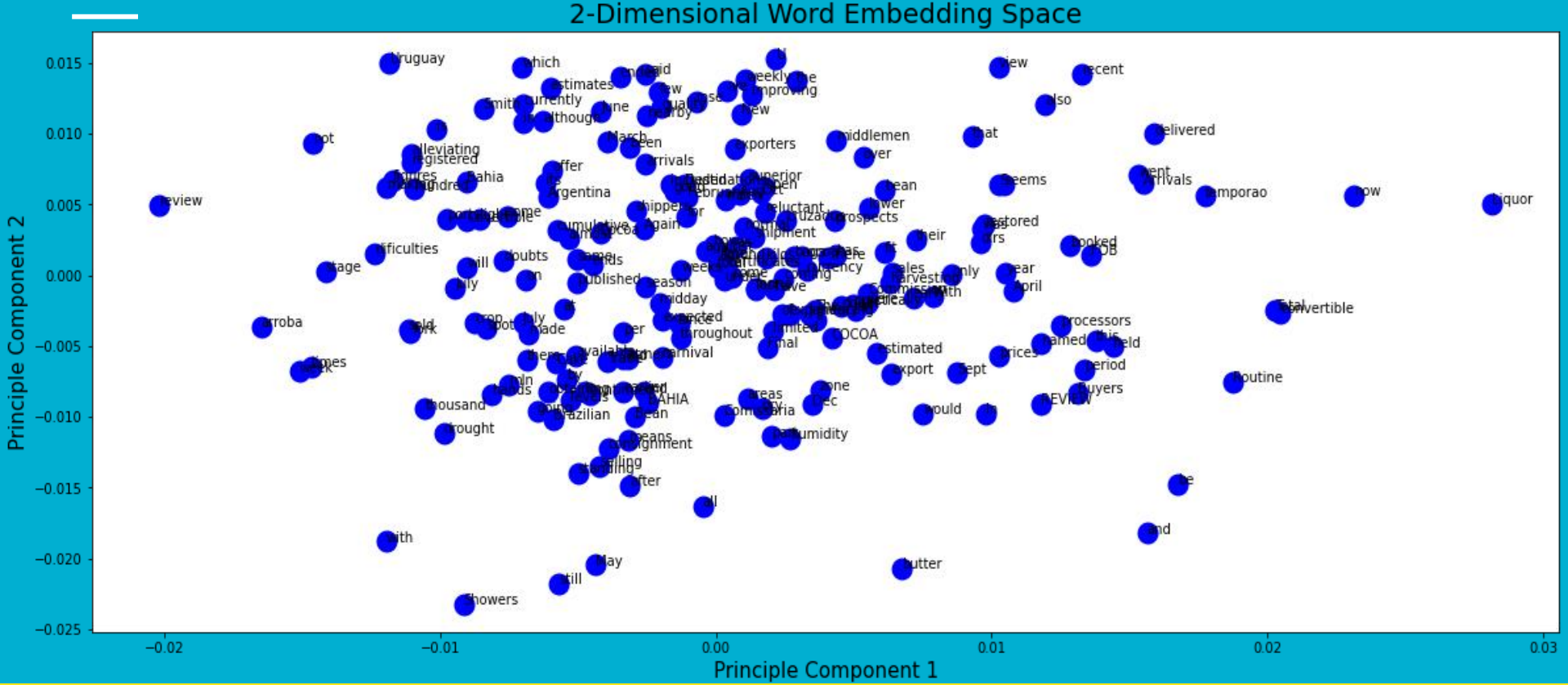
We have used **Reuters** dataset, used **NLTK** library for importing stopwords, converted text to lowercase containing only letters. Created a vocabulary of words that will be used. We created **TF-IDF** vectors from the text. We have then done the visualisation of the text as embeddings, word embeddings give us the idea of words which have similar meanings or representation. Then we have used the **Spectral Clustering(SC)** with **Particle Swarm Optimization(PSO)** on the cleaned data. We have fit the data in the model and then predicted the data. Then we have used **Adjusted Rand Index(ARI)** which is giving us the measure of how well the clusters have been formed.

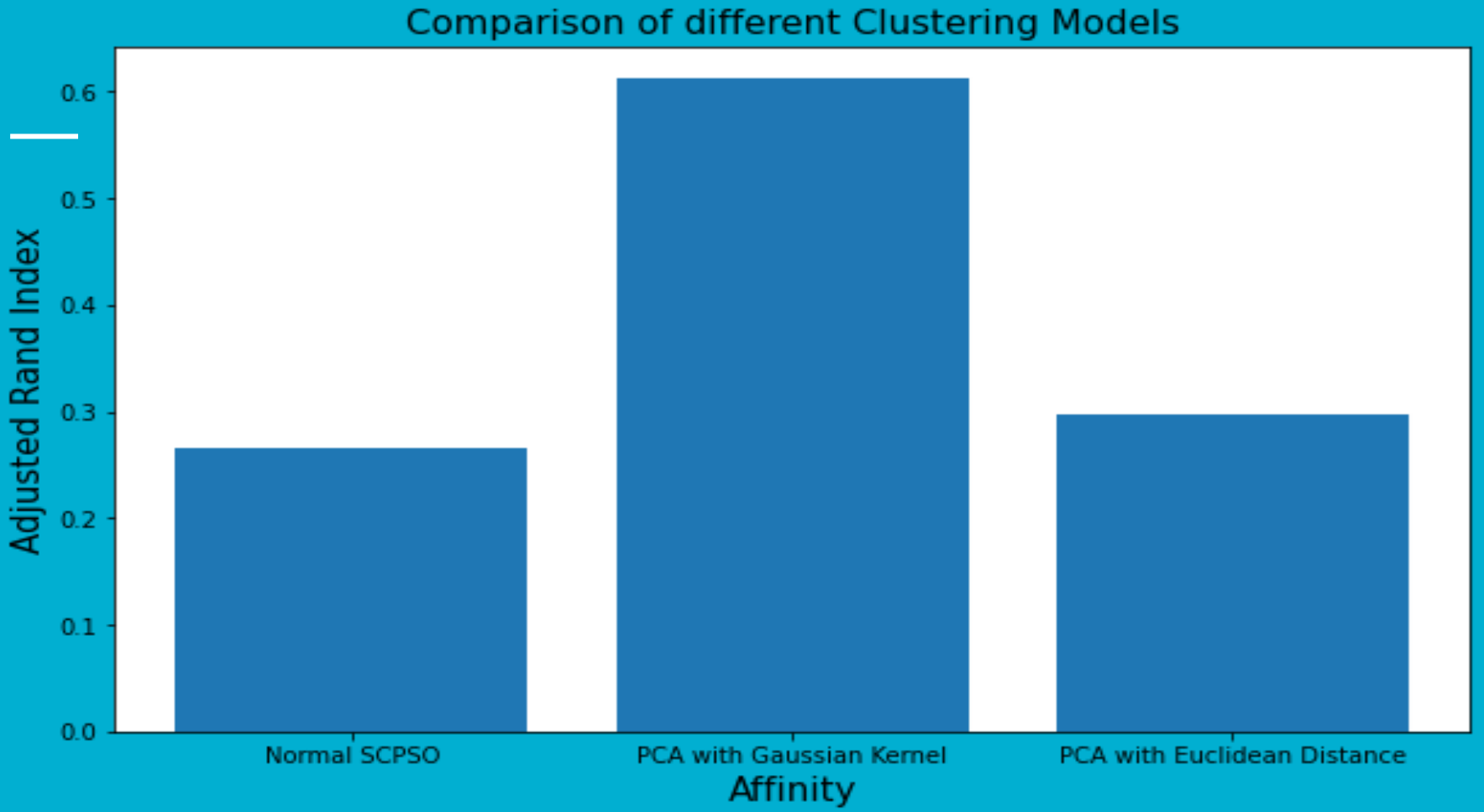
Then we have tried out to apply new ideas and make changes. We have used **Principal Component Analysis(PCA)** on the dataset which will reduce the number of features while the information of the data is retained and the training becomes more meaningful and easier. The first idea is to use **Euclidean Distance** on the affinity matrix for the **Spectral Clustering**. The second idea is to **Gaussian Kernel** on the affinity matrix for the **Spectral Clustering**. For each of these ideas we get **Adjusted Rand Index(ARI)**. At last we compare the score for all the models that we have obtained.

Language & Frameworks used

- **Python** is used as scripting language
- **Jupyter Lab** is the editor used for writing the code & visualising the graph
- Following libraries are used:
  - NLTK - for Natural Language Processing
  - Numpy - for mathematical operation
  - Pandas - for dataframe operation
  - Matplotlib - for plotting & drawing graphs
  - Sklearn - for SCPSO, PCA & other ML algorithms
  - Gensim - for Visualising Word Embeddings in 2D space

Experiments & Results




Comparison of different Clustering Models
Adjusted Rand Index
0.0
0.1
0.2
0.3
0.4
0.5
0.6
Normal SCPSO
PCA with Gaussian Kernel
PCA with Euclidean Distance
Affinity

Conclusion

- SCPSO get a solution for graph partition hence useful for creating clusters.
- Document Clustering problem is an open issue for researchers in the area of text mining and information retrieval.
- Spectral Clustering algorithm has better results when compared to the existing clustering methods such as K-means.
- Spectral Clustering with Particle Swarm Optimization (SCPSO) improves the document clustering accuracy and it leads the result on the way to an optimal solution.

- The problems that can be solved by **K-means** can also be solved by **Spectral Clustering** but not the other way around.
- We observed that applying **PCA** reduces the number of dimensions in the data while retaining most of the information, hence training has become more meaningful.
- On comparing all the models that have been built and tested, the model that has **PCA** applied on the data using **Gaussian Kernel** is giving the best result overall.

Future Direction

- The method can be implemented on a multi-core CPU.
- Hybrid clustering can be applied in order to increase the robustness of the algorithms.
- Instead of PCA, other Dimensionality Reduction algorithms such LDA can be applied.
- To attain the accomplished results of text document clustering, distinct intentions may be introduced.
- This work will be motivated by the enhancements that can be applied to the Spectral & Optimization algorithms.

Thank You !!!