International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-9 Issue-1, May 2020

K-Means Algorithm for Clustering Afaan Oromo Text Documents using Python Tools

Naol Bakala

Abstract*: With the advancement of technology and proliferation of computers in the country, the amount of Afaan Oromo language news documents produced increasingly which becomes a difficult task for news agencies to organize such huge collection of documents items manually. To solve this problem, researches is conducted using unsupervised machine learning python tools for Afaan Oromo news document clustering with low cost and best quality of clustering solution. In this research work focusing on k-means clustering analysis which produced better results as compared to the other cluster analysis both in terms of time requirement and the quality of the clusters produced.

Keywords*: Afaan Oromo Language, Afaan Oromo Text Document, Kmeans, cluster, Clustering Model.

I. INTRODUCTION

Several Languages categorized under different families in Ethiopia [1] in which Afaan Oromo is one family of Cushitic family widely used in Ethiopia and Horn of Africa. Currently, more than 80 languages are speaking in Ethiopia [2]. Afaan Oromo Language is known language in Ethiopia. More than half of Ethiopia People speaking Afaan Oromo Language in Ethiopia and outside Ethiopia, Afaan Oromo Language spoken by people of Horn of Africa. In small percent people in different parts of world also speak this language. Afaan Oromo uses "Qubee" scripts which has 26 characters taken from latin[2]. Afaan Oromo Language is mother tongue for oromo people and official language for Oromia regional state government t[1]. Since Afaan Oromo language used for different purpose, large amount of text documents is producing in this language. Afaan language are using in primary school, secondary school, court, Journals, journal, and etc are generating today. From this large documents are generating. Producing large amount of text documents resulted in problem of contents of searching and identifying important of group of text documents. To overcome this difficulty of finding important groups of document, researcher designed model that groups document contents according their similarity. To design model Text document clustering Techniques from machine learning techniques, particularly Kmeans algorithm applied to cluster Afaan Oromo Text documents into groups automatically.

Text document clustering is technique of grouping text document into groups based on their internal similarity. Internally similar text documents clustered under the same groups whereas externally similar documents clustered under different clusters [3] It is techniques of that grouping available document into clusters existing collection of text documents into important clusters [4]. Text document clustering applied into different area of Data mining, Text categorization, Natural Language Processing and Information Retrieval [5]. In this research, researcher collected data from Oromia tourism Office, Oromia Broadcast Network, and others offline and online written documents in electronic format. Totally 25 Afaan Oromo Text Documents collected. Each Afaan Oromo Text Documents collected in word format and combined together to develop corpus. After text Preprocessing, cosine similarity, Dimensionality reduction and Dimensionality Reduction Performed on the Corpus, dataset prepared. Dataset loaded into Python and Saved as "AfaanOromoDataSets.csv".

In this study, Kmeans Machine Learning algorithm implemented to cluster Afaan Oromo Text Documents. Internally similar Afaan Oromo text documents clustered under the same groups whereas externally similar documents clustered under different clusters using R. As a result, Afaan Oromo Text document used in experiment grouped into eleven main clusters. Agriculture, Culture, Education, Gada system, Heath, Politics, Religion, and Sport resulted clusters from the experiment. The performance of developed clustering model evaluated depending on the number of instance correctly distributed into clusters. This paper organized as hereunder. In section 1 introduction on entire work discussed. The methodology employed in this study discussed in the section 2. section 3 explains experiment. Finally, section 4 presents Conclusion and Recommendation. The methodology used in this research is the knowledge discovery in text approach by which multi-step process, which includes all the tasks from the gathering of documents to the visualization the extracted information. In this research includes

A. DATA COLLECTION

There is no standard news corpus for Afaan Oromo language. Therefore, researcher collected data from Oromia Tourism Office, Oromia Broadcast Network, Oromia Broadcasting service, Oromia Media Network, Voice of America Radio, Oromia Waqeffatota Association, Oromia Cultural Center and also internet news documents.

Manuscript received on April 02, 2020.
Revised Manuscript received on April 15, 2020.
Manuscript published on May 30, 2020.
* Correspondence Author
Naol Bakala*, Department of Computer Science, Ambo University, Ambo, Ethiopia. naolbakala@gmail.com

B. DOCUMENT PREPROCESSING

One of the basic steps in this research work is preprocessing of the unstructured text documents or news items. Document preprocessing is a very vital role in text clustering, as it is irrelevant and redundant features often degrade the performance of clustering algorithms both in speed and its accuracy and also its tendency to reduce over fitting. Hence, document preprocessing activities are done to increase the performance of clustering algorithms both in speed by removing irrelevant and redundant features of the oromo text documents. Routines were written using Python to tokenize, normalize, remove stop words and stem the documents.

C. TERM FREQUENCY WEIGHTING (TF) FOR INDEXING

Document represented by utilizing document weighting approach. As [6] approved that document representation computed from Term Frequency (TF) and Inverse document Frequency (IDF). Tf(t,cs) computed as sqrt(count(t,cs)) , idf(t,docs) computed as log(D/df(t,docs)) and term weight Wi is computed as TF * IDF[8]. Researcher Utilized to calculated Afaan Oromo Document representation using Python from prepared Afaan Oromo Corpus. In document representation the row contain the name of words reduced to root and the column contain name of documents. The Matrix used as in put for each clustering algorithm. Corpus loaded into python working space after and Python Scripts also used to calculate Term Frequency, Inverse Document Frequency, and Term Frequency * Inverse Document Frequency. Term Frequency and Inverse Document Frequency saved in CSV file format using Python scripts. After Term Frequency and Inverse Document Frequency calculated and saved document index calculated as Term Frequency * Inverse Document Frequency. Term Frequency * Inverse Document Frequency also saved in CSV file format.

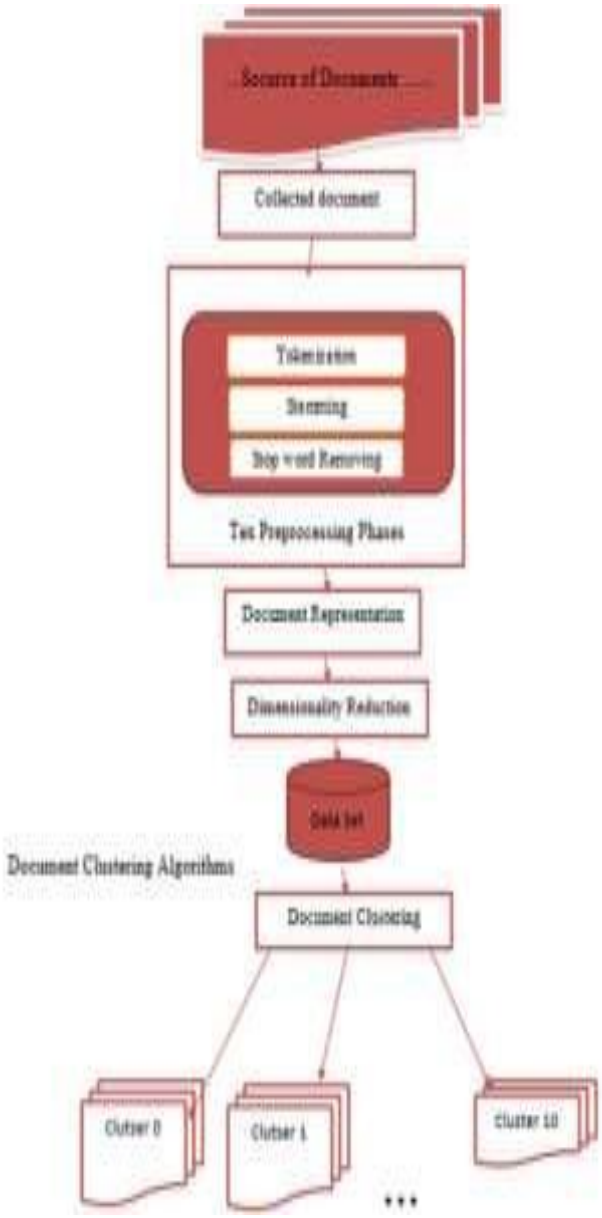
D. AFAAN OROMO TEXT DOCUMENT CLUSTERING ALGORITHMS

In this work, Kmeans algorithm utilized to cluster Afaan Oromo text document. Experiment conducted using Python and matrix we prepared as dataset. The main functions of k-means clustering is

$$J = \sum_{i=1}^{m} \sum_{k=1}^{K} w_{ik} \|x^i - \mu_k\|^2$$

where wik=1 for data point xi if it belongs to cluster *k*; otherwise, wik=0. Also, μk is the centroid of xi's cluster and the Series of steps in Text clustering adopted are 1) Select K points as the initial Centroids, 2) Assign all points to the closest centroid, 3) Recomputed the Centroid of each cluster, repeat 2 and 3 steps until centroid do not change [3]. Researcher assigned different values for K several as steps.

E. MODEL FRAMEWORK



III EXPERIMENT RESULT

Researcher converted data set format into CSV file format that supported by Python for conducting experiment. After Compatible data set loaded into Python, Kmeans clustering algorithm also applied for Afaan Oromo nonfiction text document clustering. Repeated experiment 5 times by assigning 3, 5, 8, 9 and 10 as centroids during clustering experiment in Python. Accuracy of each experiment evaluated by correctly clustered number of instance and incorrectly assigned number of instances. Finally, at centroids values assigned as 10 small number of Clustered assigned incorrectly. At this particular step when centroid 10 was scored as 10.22% as incorrectly assigned and 89.78 assigned correctly. Finally, researcher identified clusters name and the accuracy of the clusters. Clusters scored different accuracy during experiment. Agriculture clusters scored highest accuracy where cluster "Other" scored lowest accuracy

In [239]:
```
%matplotlib inline
from copy import deepcopy
import numpy as np
import pandas as pd
from matplotlib import pyplot as plt
```

In [305]:
```
#Import the dataset
df = pd.read_csv("C:/Users/Mine/Desktop/Publication/Naol publ1-1/AfaanOromoDataSets.csv")
print(df.shape)
df.head()
```

(3094, 24)

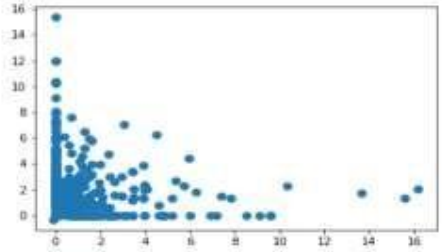
Out[233]:

	1	2	3	4	5	6	7	8	9	10	...	15	16	17	18	19	20	21	22	Class1	24
0	1.00	0.20	0.20	0.20	0.0	0.80	0.40	0.80	0.0	0.60	...	0.00	0.4	0.20	0.60	0.2	0.00	0.80	0.40	Agriculture	1
1	0.79	0.79	0.00	1.58	0.0	1.58	3.15	3.15	0.0	0.79	...	0.00	0.0	0.00	1.58	0.0	0.00	0.00	0.79	Agriculture	1
2	1.70	0.00	0.00	0.00	0.0	0.00	1.70	1.70	0.0	0.00	...	0.00	0.0	0.00	0.00	0.0	0.00	0.00	0.00	Agriculture	1
3	2.40	0.00	0.00	0.00	0.0	0.00	0.00	0.00	0.0	0.00	...	0.00	0.0	0.00	0.00	0.0	0.00	0.00	0.00	Agriculture	1
4	3.41	0.00	0.00	0.00	1.7	0.00	0.00	0.00	0.0	0.00	...	0.00	0.0	1.70	0.00	0.0	0.00	0.00	0.00	Agriculture	1
5	3.98	1.99	0.00	0.00	0.0	0.00	0.00	0.00	0.0	0.00	...	0.00	0.0	0.00	0.00	0.0	0.00	0.00	0.00	Agriculture	1
6	1.00	1.00	0.00	0.40	0.4	1.40	1.40	1.20	0.0	0.00	...	0.80	0.2	0.60	0.80	0.4	0.40	0.60	0.80	Agriculture	1
7	1.59	0.64	0.32	0.32	0.0	1.27	1.27	1.27	0.0	0.32	...	0.32	0.0	0.00	0.64	0.0	0.00	2.23	0.96	Agriculture	1
8	1.99	0.00	0.00	0.00	0.0	0.00	0.00	0.00	0.0	0.00	...	0.00	0.0	0.00	1.99	0.0	0.00	0.00	0.00	Agriculture	1
9	0.45	1.81	0.00	0.00	0.0	0.45	0.45	0.45	0.0	2.26	...	0.90	0.0	0.45	0.90	0.9	0.45	0.45	0.00	Agriculture	1

10 rows × 24 columns

In [306]:
```
f1 = df[["1","2","3","4","5","6","7","8","9","10","11","12","13","14","15","16","17","18","19","20","21","22"]].values
f2 = df[["24"]].values
```

In [307]:
```
plt.scatter(f1[:, 0], f1[:, 1], s=50);
```

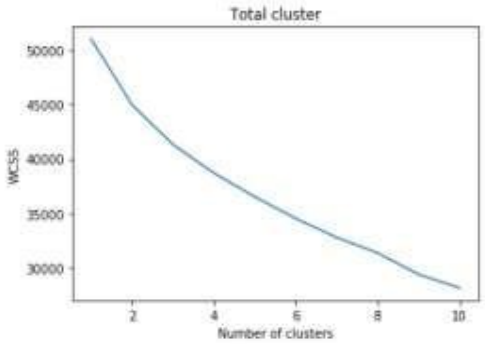


In [308]:
```
def dist(a, b, ax=1):
    return np.linalg.norm(a - b, axis=ax)
```

In [309]:
```
X = np.array(list(zip(f1, f2)))
```

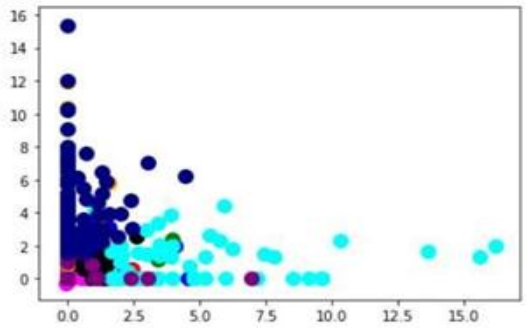
In [310]:
```
from sklearn.cluster import KMeans
wcss =[]
for i in range (1,11):
    kmeans = KMeans(n_clusters = i, init = 'k-means++', max_iter =300, n_init = 10, random_state = 0)
    kmeans.fit(f1)
    wcss.append(kmeans.inertia_)
```

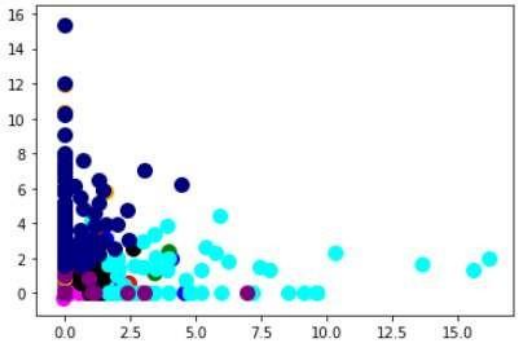
In [311]:
```
plt.plot(range(1,11),wcss)
plt.title('Total cluster')
plt.xlabel('Number of clusters')
plt.ylabel('WCSS')
plt.show()
```



```
In [312]: kmeans=KMeans(n_clusters= 10, init = 'k-means++', max_iter = 300, n_init = 10, random_state = 0)
          Y_Kmeans = kmeans.fit_predict(f1)

In [315]: plt.scatter(f1[Y_Kmeans == 0, 0], f1[Y_Kmeans == 0,1],s = 100, c='red', label = 'Cluster 1')
          plt.scatter(f1[Y_Kmeans == 1, 0], f1[Y_Kmeans == 1,1],s = 100, c='blue', label = 'Cluster 2')
          plt.scatter(f1[Y_Kmeans == 2, 0], f1[Y_Kmeans == 2,1],s = 100, c='green', label = 'Cluster 3')
          plt.scatter(f1[Y_Kmeans == 3, 0], f1[Y_Kmeans == 3,1],s = 100, c='cyan', label = 'Cluster 4')
          plt.scatter(f1[Y_Kmeans == 4, 0], f1[Y_Kmeans == 4,1],s = 100, c='magenta', label = 'Cluster 5')
          plt.scatter(f1[Y_Kmeans == 5, 0], f1[Y_Kmeans == 5,1],s = 100, c='black', label = 'Cluster 6')
          plt.scatter(f1[Y_Kmeans == 6, 0], f1[Y_Kmeans == 6,1],s = 100, c='aqua', label = 'Cluster 7')
          plt.scatter(f1[Y_Kmeans == 7, 0], f1[Y_Kmeans == 7,1],s = 100, c='orange', label = 'Cluster 8')
          plt.scatter(f1[Y_Kmeans == 8, 0], f1[Y_Kmeans == 8,1],s = 100, c='purple', label = 'Cluster 9')
          plt.scatter(f1[Y_Kmeans == 9, 0], f1[Y_Kmeans == 9,1],s = 100, c='navy', label = 'Cluster 10')
          #plt.scatter(kmeans.cluster_centers_[:,0], kmeans.cluster_centers_[:,1], s = 300, c = 'yellow', label = 'Centroids')
```





IV.CONCLUSION AND RECOMMENDATION

In this study, we concluded ten types of clusters in Afaan Oromo News Text documents. The Agriculture, Culture, Education, Gada system, Heath, Politics, Religion, and Sport the main clusters of Afaan Oromo Text documents. Clusters scored different accucy during experiment. Agriculture clusters scored highest accuracy where cluster "Sport" scored lowest accuracy. In addition to this, researcher concluded that Kmeans algorithm scored 89.78% accuracy in Afaan Oromo Text document clustering. In this research, stop words and Punctuation removing algorithms also modified and implemented to remove stop words and Punctuation from corpus. Finally, large size Afaan Oromo text documents corpus also prepared for further researches to that will be conducted in future. Resulted eleven clusters of Afaan Oromo Text documents clusters can be used for Afaan Oromo Text Document classification and Categorization Therefore, It is recommended to use those clusters as main class of Afaan Oromo News Text documents classification and Categorization in future.

AUTHORS PROFILE

Naol Bakala, M.Sc is Senior Lecturer and former Head of Department of Computer and now working as Administrative Vice Director in Ambo University Institute of Technology. Ethiopia. His major research interest lies in Natural Language Process

REFERENCES

1 G. M. Wegari, "OroRoots: Rule-Based Root Generation System for Afaan Oromo," International Journal of Scientific & Engineering Research, vol. 8, no. 3, p. 1, 2017.
2 N. Bakala, "Information Retrieval System By Using Vector Space Model," INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH, vol.
3 VOLUME 8 , no. ISSUE 10, p. 2, 2019.
4 N. S. E. T. S. S. Anthon Roberto Tampulon, "Document clustering Using Combination of K-means and Single Linkage Clustering Algorithms," Journal of Engieering anf Applied Sciences, no. 13 (Special Issue 3): 31883192, , p. 3, 2018 .
5 K. D. G. Hannah Grace, "Experimental Estimation of Number of Clusters Based on Cluster Quality," Journal of mathematics and computer science 12 (2014), 304315, vol. 12, 2014.
6 N. Bakala, "A two Steps approach for Afan Oromo nonfiction text categorization," International Journal of Scientific Research in Computer Science, Engineering and Information Technology, vol. 3, no. 1/1661, p. 2, 2018.
7 D. P. A. V. Prafulla Bafna, "Document Clustering: TFIDF approach," International Conference on Electrical, Electronics, and Optimization Techniques, p. 1, 2016.
8 S. S. M. S. Hariharan, "EXPERIMENTS ON DOCUMENT CLUSTERING IN TAMIL LANGUAGE," ARPN Journal of Engineering and Applied Sciences, vol. 13, no. 10, p. 1, 2018.
9 C. S. M. J. Rakesh Chandra Balabantaray*, "Document Clustering using K-Means and K-Medoids," InternationalJournal of Knowledge Based Computer System, vol. 1, no. 1, 2013.
10 N.Kannaiya Raja, S. Suresh, NLP: Text Summarization By Frequency And Sentence Position Methods