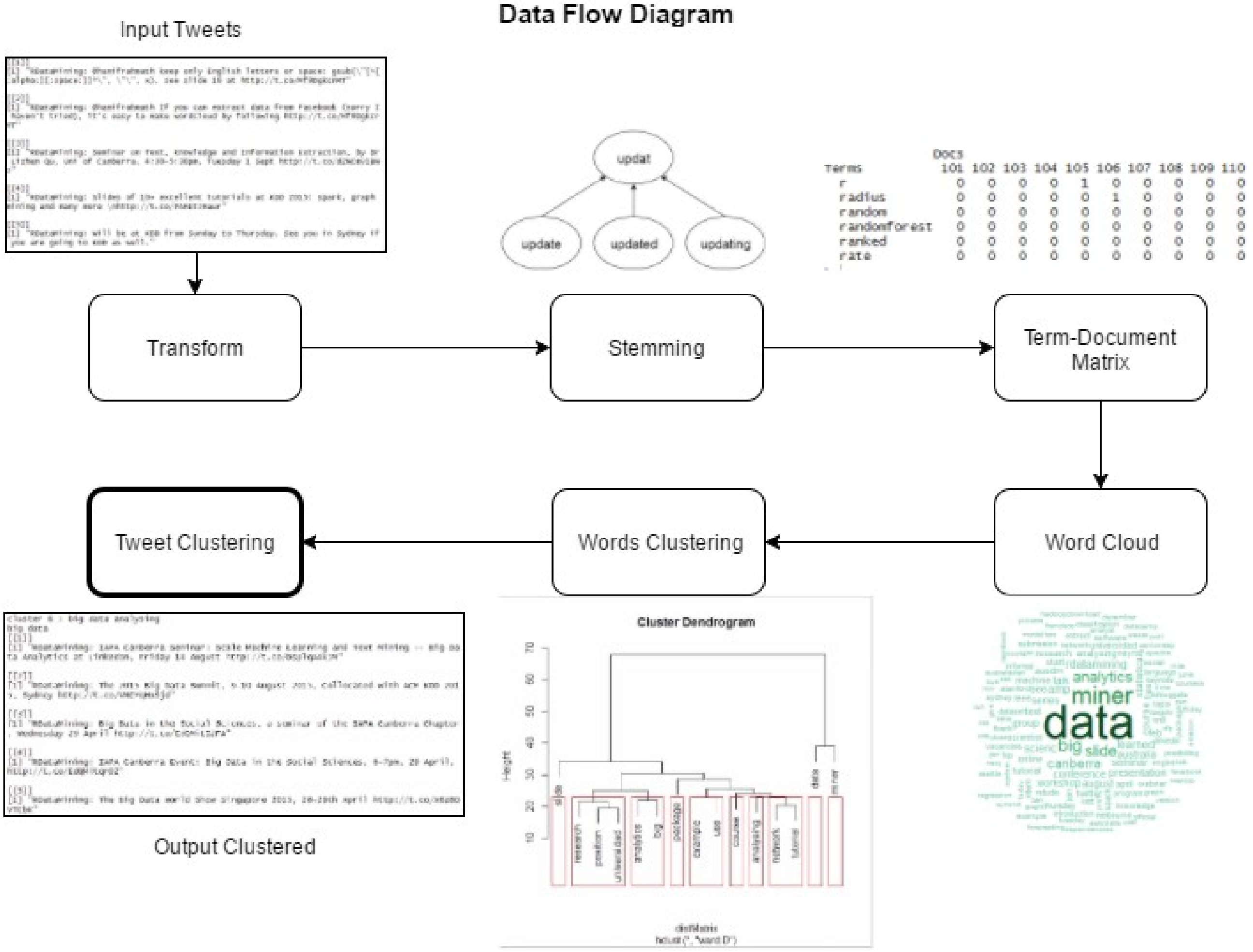
Twitter Text Mining Using Clustered Search

Anjali Malav(UG201310006)
Muttineni Navya(UG201310020)

Problem Statement : Given a search word, the objective is to return results in this case tweets in clustered manner from a large database of tweets.

Introduction : Twitter is a major microblogging service. When a user searched these tweets with a search term, thousands of tweets are displayed. The user has to go through every tweet in order to see if that tweet is of his relevance or not. This burden would be reduced if the tweets were clustered first and then returned to the user topic wise.



Data Flow Diagram

```
[[1]]
[1] "RDataMining: @hanifrahmah keep only English letters or space: gsub("[^[:alpha:][:space:]]*", "", x). see slide 18 at http://t.co/[illegible]"

[[2]]
[1] "RDataMining: @hanifrahmah If you can extract data from Facebook (sorry I haven't tried), it's easy to make wordcloud by following http://t.co/[illegible]"

[[3]]
[1] "RDataMining: Seminar on text, knowledge and information extraction, by Dr Lizhen Qu, Uni of Canberra, 4:30-5:30pm, Tuesday 1 Sept http://t.co/[illegible]"

[[4]]
[1] "RDataMining: Slides of 10+ excellent tutorials at KDD 2015: Spark, graph mining and many more \nhttp://t.co/[illegible]"

[[5]]
[1] "RDataMining: will be at KDD from Sunday to Thursday. See you in Sydney if you are going to KDD as well."
```

```
              Docs
Terms          101 102 103 104 105 106 107 108 109 110
  r              0   0   0   0   1   0   0   0   0   0
  radius         0   0   0   0   0   1   0   0   0   0
  random         0   0   0   0   0   0   0   0   0   0
  randomforest   0   0   0   0   0   0   0   0   0   0
  ranked         0   0   0   0   0   0   0   0   0   0
  rate           0   0   0   0   0   0   0   0   0   0
```

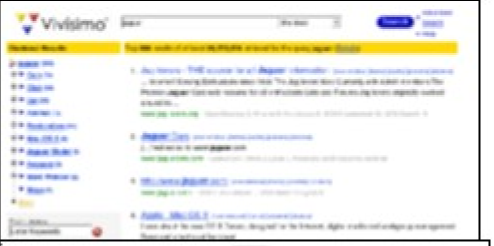
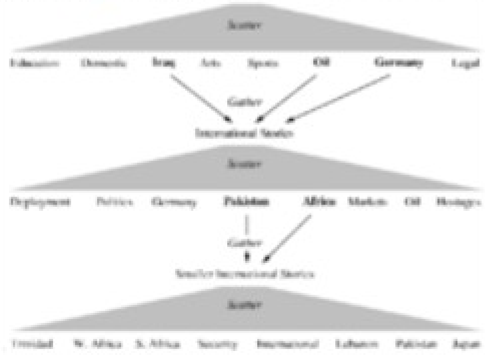
```
Cluster 6 : big data analysing
big data
[[1]]
[1] "RDataMining: IAPA Canberra Seminar: Scale Machine Learning and Text Mining -- Big Data Analytics at LinkedIn, Friday 14 August http://t.co/[illegible]"

[[2]]
[1] "RDataMining: The 2015 Big Data Summit, 9-10 August 2015, Collocated with ACM KDD 2015, Sydney http://t.co/[illegible]"

[[3]]
[1] "RDataMining: Big Data in the Social Sciences, a seminar of the IAPA Canberra Chapter, Wednesday 29 April http://t.co/[illegible]"

[[4]]
[1] "RDataMining: IAPA Canberra Event: Big Data in the Social Sciences, 6-7pm, 29 April, http://t.co/[illegible]"

[[5]]
[1] "RDataMining: The Big Data World Show Singapore 2015, 28-29th April http://t.co/[illegible]"
```

Methodology : The methodology is as given in the above diagram. For clustering of tweets we have used two algorithms, K-means and K-medoids.

Comparison between K-means and K-medoids

1. Number of clusters needs to be specified in advance where as in k-medoids is not required.
2. k-medoids is based on centroids (or medoids) calculating by minimizing the absolute distance between the points and the selected centroid, rather than minimizing the square distance. As a result, it's more robust.
3. K-means is faster than K-medoids.
4. Implementation of k-means is easier than k-medoids which is complex.

Applications :

Application	Benefit	Example
Search result clustering	more effective information presentation to user	
Scatter-Gather	alternative user interface: search without typing	 Scatter Education Domestic Iraq Arts Sports Oil Germany Legal Gather International Stories Scatter Deployment Politics Germany Pakistan Africa Markets Oil Hostages Gather Smaller International Stories Scatter Trinidad W. Africa S. Africa Security International Lebanon Pakistan Japan

Conclusions : The input data i.e the number of tweets being retrieved is small (200-600). Scalability is not handled as working on large data sets needs a lot of computation which would take a lot of time on 1 CPU. This can be handled by computing on more than 1 CPUs. This can be one of the future scopes of this project with others being sentiment analysis and social network analysis.