**Program no : 1**

Input : Single Topic name , duration

* Example : Sports , 01-06-2017 TO 01-07-2017

Output : all users ( user names ) who tweeted tweets about the topic in that duration in the following format

Ex:

|  |  |  |
| --- | --- | --- |
| SL.NO | USERNAME @ MENTION | USER ID |
| 1 |  |  |
| 2 |  |  |
| ….. |  |  |

**Program no : 2**

Input : all users in program 1

Output : Number of tweets about that topic in descending order

EX:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SL.NO | USER NAME @ MENTION | USER ID | NUMBER OF TWEETS | Log( number of tweets) |
| 1 |  |  |  |  |
| 2 |  |  |  |  |
| ….. |  |  |  |  |

**Program no : 3**

Input : In 2nd programs output top 4 users (users who tweeted more number of tweets in that duration)

Output : Tweets of those 4 users on the topic in that duration

EX:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| SL.NO | USER NAME @ MENTION | USER ID | TWEET ID | YEAR-MONTH-DAY | HOURS: MINUTES: SECONDS | TWEET |
| 1 |  |  |  |  |  |  |
| 2 |  |  |  |  |  |  |

**Program no 4 :** Arrange all the tweets in a time line

**Program no 5 :**

Input : All tweets

Output : TF\* IDF Scores for all the tweets ( consider only english words)

TF = Term Frequency = number of times the word i occurs in the tweet

IDF = Inverse document frequency

= log Total number of tweets in total

----------------------------------------------------------

Number of tweets containing the word i

Ex: if there are 3 tweets

Tweet 1 : I love india

Tweet 2 : India is my country

Tweet 3 : Iam a resident of india

Step 1: Removing all the stop words from the above tweets

Step 1 : Remove all the stop words in the tweets

Stop word list is available in the internet, so we can download it

In the below example, after removing stop words such as ( i,is, my, am, a of etc)

I

Tweet 1 : I love india

iss

Tweet 2 : India is my country

of of

Tweet 3 : I resident india

After removing stop words from all the tweets, the words remaining are

Tweet 1: love india

Tweet 2: india country

Tweet 3: resident india

Step 2: Stemming

Step 2: Stemming

To do stemming there are 2 ways

1. in the internet a package is available to perform stemming, if you type tweets in the box then automatically the tool will give us all the words after stemming

Package name is “ stemming and lemmatization with python NLTK” - STEM PACKAGE

2. Otherwise we can also use “ Porter stemmer “ algorithm . Directly the algorithm is available in the internet for stemming

Ex: stemming means

Suppose 4 words are there, those 4 words would be considered as single word

“assume” ,”assuming”, “assumination”, ”assumes”

Step 3: calculating TF\* IFDF scores for all the words after removing stop words, and after stemming

Ex: in the example there are no stemmed words

Tweet 1: love india

Tweet 2: india country

Tweet 3: resident india

So there are 4 words

Word1= love

Word 2= india

Word 3= country

Word 4 = resident

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| India | | | Love | | | Country | | | Resident | | |
| TF | IDF | TF \* IDF | TF | IDF | TF \* IDF | TF | IDF | TF \* IDF | TF | IDF | TF \* IDF |
| Tweet1 | 1 | Log(3/3) | 1 \* log1 | 1 | Log(3/1) | 1 \* log(3) |  |  |  |  |  |  |
| Tweet 2 |  |  |  |  |  |  |  |  |  |  |  |  |
| Tweet 3 |  |  |  |  |  |  |  |  |  |  |  |  |

In the above example, for tweet 1

TF(LOVE ) = 1

TF ( LOVE) = NUMBER OF TIMES THE WORD “LOVE “ occurs in the tweet

IDF( LOVE) = log( total number of tweets)/ (number of tweets containing the word” love”)

IDF( LOVE) = log(3/1)

So TF \* IDF (LOVE) = 1\* log 3

**PROGRAM NO 6:**

**Represent the tweets in following format**

|  |  |  |  |
| --- | --- | --- | --- |
| Sl no | Tweet | User name @mention | W  ( log(number of tweets made by user)) |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |

**PROGRAM NO 7:**

**Input :** for 3/5 th of the above tweets in a time line **( Program no:4)** apply k-means clustering algorithm

Output : k-clusters are displayed in the following format (3 tables)

Table 1:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| SL.NO | User id | User @ | Tweet id | Tweet | Cluster number |
|  |  |  |  |  | 1 |
|  |  |  |  |  | 1 |
|  |  |  |  |  | 2 |
|  |  |  |  |  | …… |
|  |  |  |  |  | k |

Table 2: In a graph ( OUTGOING EDGE)

Number of tweets made by each user

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Sl.no | Cluster number | Total Number of tweets in the cluster | User 1 | User 2 | User 3 | User 4 |
|  | 1 |  |  |  |  |  |
|  | 2 |  |  |  |  |  |
|  | … |  |  |  |  |  |
|  | …. |  |  |  |  |  |
|  | k |  |  |  |  |  |

**PROGRAM NO 7:**

**Input : Tweets in k clusters by 4 users (individual tweet by each user)**

**Output : Re tweeted by which follower (individual tweets ) in the following formats**

**Table 1:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Cluster number** | **User name** | **Number of tweets posted by each user** | **Tweet id** | **Tweet** | **Re-tweeted by which follower(follower name @ mention)** |
| **C1** | **User 1** | **10** | **19876** | **Tweet 1:**  **India is my country** | **A,b,c,d,e,…..** |
|  | **….** |  |  | **Tweet 2:** | **F,g,h,i.j** |
|  | **….** |  |  | **Tweet 3:** |  |
|  |  |  |  | **Tweet 4:** |  |
|  |  |  |  | **Tweet 5:** |  |
|  |  |  |  | **Tweet 6:** |  |
|  |  |  |  | **Tweet 7:** |  |
|  |  |  |  | **Tweet 8:** |  |
|  |  |  |  | **Tweet 9:** |  |
|  |  |  |  | **Tweet 10:** |  |
|  |  |  |  |  |  |
|  | **User 2:** | **3** | **65436** | **Tweet 1** |  |
|  | **….** |  |  | **Tweet 2:** |  |
|  | **….** |  |  | **Tweet 3:** |  |
|  | **User 3** |  |  |  |  |
|  | **….** |  |  |  |  |
|  | **…..** |  |  |  |  |
|  | **User 4** |  |  |  |  |
|  | **…….** |  |  |  |  |
| **C2** | **User 1** |  |  |  |  |
|  | **….** |  |  |  |  |
|  | **User 2:** |  |  |  |  |
|  | **……..** |  |  |  |  |
|  | **User 3:** |  |  |  |  |
|  | **………** |  |  |  |  |
|  | **User 4:** |  |  |  |  |
|  | **………** |  |  |  |  |
| **…….** |  |  |  |  |  |
| **Cluster k** | **User 1** |  |  |  |  |
|  | **User 2:** |  |  |  |  |

**Table 2: ( first 3 columns are same as in table 1)**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Cluster number** | **User name** | **Number of tweets posted by each user in each cluster** | **Follower name** | **Number of times retweeted** | **Influence score** |
| **C1** | **User1** | **10** | **a** | **8** | **8/10** |
|  |  |  | **b** | **10** | **10/10** |
|  |  |  | **c** | **6** | **6/10** |
|  |  |  | **…..** |  |  |
|  | **User 2** | **8** | **a** | **8** | **8/8** |
|  |  |  | **e** | **1** | **1/8** |
|  |  |  | **……** |  |  |
|  | **User 3** | **10** | **a** | **5** | **5/10** |
|  | **User 4** | **6** | **a** | **6** | **6/6** |
|  |  |  |  |  |  |
| **C2** | **User 1** | **6** |  |  |  |
|  | **User 2** |  |  |  |  |
|  | **User 3** |  |  |  |  |
|  | **User 4** |  |  |  |  |
|  |  |  |  |  |  |
| **Ck ( kth cluster)** | **User1** |  |  |  |  |
|  | **User2** |  |  |  |  |
|  | **User3** |  |  |  |  |

**PROGRAM NO 8 : Weight of an edge**

**Input : (** number of re tweets made by each follower **and** number of tweets

made by each user ) for each cluster

**Output : weight of an edge = Number of re tweets made by follower**

**Total number of tweets posted by user**

**Weight of an edge should be displayed in the following tabular format**

**Example :** in cluster 1 there are 4 users

if user 1 posted 10 tweets (outgoing edge),

then the followers of user1 re tweeted ( incoming edge)

then the weight of edge = (incoming edge)/(outgoing edge)

user 1 posted 10 tweets, follower1 re tweeted 8 times , then weight of an edge=8/10

**Table 1:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Cluster number | User name | Follower name | Edge weight of the follower= (**Number of re tweets made by follower)/( Total number of tweets posted by user)** |  |
| 1 | User1 | a | 8/10 |  |
|  |  | b | 10/10 |  |
|  | User 2 |  |  |  |
| 2 |  |  |  |  |
|  |  |  |  |  |
| 3 |  |  |  |  |
| k |  |  |  |  |

**Table 2 :**

|  |  |  |  |
| --- | --- | --- | --- |
| **Cluster number** | **User name** | **Follower name who has Maximum edge weight from the above table** | **Weight (value)** |
| **1** | **User 1** |  |  |
|  | **User 2** |  |  |
|  | **User 3** |  |  |
|  | **User 4** |  |  |
|  |  |  |  |
| **2** | **User 1** |  |  |
|  | **User 2** |  |  |
|  | **User 3** |  |  |
|  | **User 4** |  |  |
| **…..** |  |  |  |
| **k** | **User 1** |  |  |
|  | **User 2** |  |  |
|  | **User 3** |  |  |
|  | **User 4** |  |  |

**From the above table , we can assume that in cluster 1 , user 1 is influenced by the follower name who has maximum edge weight.**

**PROGRAM NO 9 :**

**INPUT : Edge weights in each cluster by each follower**

**Output : overall influence score from all the clusters**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Exemplar User name** | **a** | **b** | **c** | **d** | **e** | **f** | **g** | **h** | **i** | **…..** |
| **User 1** | **( 8/10 +8/8 +5/10)** |  |  |  |  |  |  |  |  |  |
| **User 2** |  |  |  |  |  |  |  |  |  |  |
| **User 3** |  |  |  |  |  |  |  |  |  |  |
| **User 4** |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |

**PROGRAM NO 10 :**

**Input : Remaining 2/5 th of the tweets apply K-Means clustering algorithm (incremental clustering)**

**Output : display K-Clusters in the following format**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **SL.NO** | **User id** | **User @ mention** | **Tweet id** | **Tweet** | **Cluster number** |
| **1.** |  |  |  |  | **1** |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  | **2** |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  | **K** |
|  |  |  |  |  |  |
|  |  |  |  |  | **K+1** |
|  |  |  |  |  | **…** |
|  |  |  |  |  | **K+2** |