**Twitter Topic Summarisation using**

**Speech Act Classification & Sequential**

**Summarisation**

***DineshV.B 2010103037***

***Daniel Sam Pete.T 2010103065***

***Gopinathan.A 2010103619***

***GUIDE: Dr. A.P.Shanthi***

***DCSE,CEG***

**HOW THE IMPLEMENTATION ACTUALLY WORKS:**

* We would have two seperate programs run in two different softwares:One would extract the json files of tweets collected from the internet .Another program on octave would classify it .Then a summariser module will eliminate the miscellaneous and extract the most necessary information.
* The output will be a display of an extractive summary filtered based on different sub topics.
* At each stage, the output will be shown as a display on the screen
* The demo for a sample number of tweets(200) will be shown during the review.
* For the classification process, the accuracy of the training set classification will be shown in the demo

**MEASUREMENTS:**

* grammaticality – the text should not contain non-textual items (i.e., markers) or punctuation errors or incorrect words

|  |  |  |
| --- | --- | --- |
| Non Textual Items(x) | Total No of words(y) | Grammaticality=(x/y) |
|  |  |  |

* non-redundancy – the text should not contain redundant information

|  |  |  |
| --- | --- | --- |
| No of repeated words(x) | Total no of words(y) | Non-redundancy=(x/y) |
|  |  |  |

* coherence – based on word alignment

Content Evaluation :

* cosine similarity

Check for similarity between each sentence based on Bag Of Words

* Latent Semantic Analysis based similarity measures
* Precision,Recall,Accuracy for Classification of Speech Act

Tp=true positive

Fp=False positive

Fn=False negative

Tn=True negative

* \text{Precision}=\frac{tp}{tp+fp} \, 
* \text{Recall}=\frac{tp}{tp+fn} \, 
* \text{Accuracy}=\frac{tp+tn}{tp+tn+fp+fn} \, 
* Accuracy Value for SVM :60% Accuracy value for Logistic Regression :70% <For Classification>

**COMPARISONS MADE USING THESE MEASUREMENTS:**

* Grammaticality gives a measure of how well the summary is formed.
* Non redundancy gives a measure of how well the summary is informative
* Cosine similarity also gives us a measure of how well the summary is informative and appealing
* Accuracy is a measure of the correctness of the Classification.
* LSA based similarity measures measures semantic based redundancy and eliminates it
* A comparision is made between SVM and Logistic Regression(setting a threshold of 0.005) for classification and the Logistic Regression showed a better accuracy.