Assignment 3

# Clustering and segment profiling

**Group – 6 (CS 553 – W1)**

Shreesh Chavan - 10440695

Kaustubh Karlekar - 10444193

Divya Goalla - 10433125

**Deliverable 1**

**A screenshot of a video game

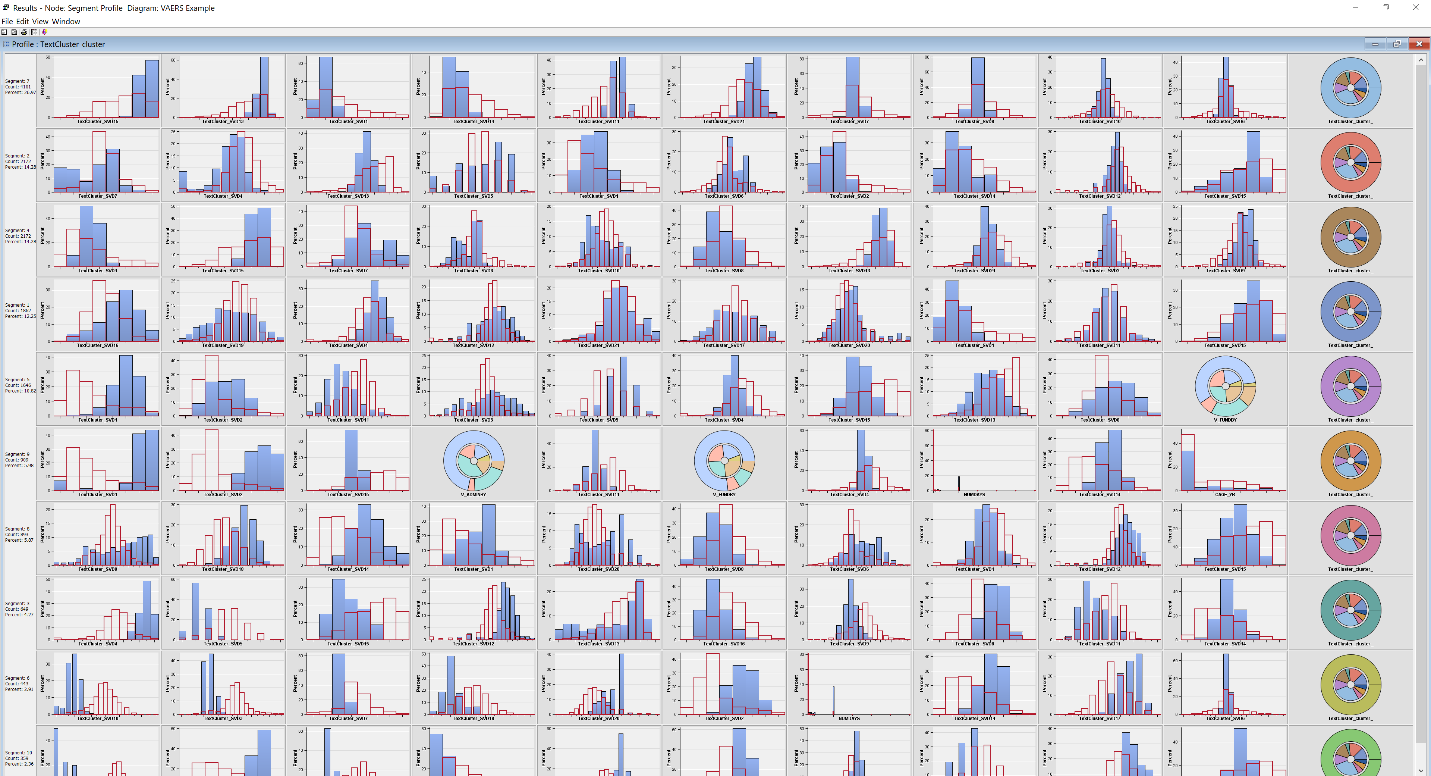
Description automatically generated**

**A screenshot of a social media post

Description automatically generated**

**A screenshot of a video game

Description automatically generated**

****

**Deliverable 2**

**Variable Summary**

**Measurement Frequency**

**Role Level Count**

**ID INTERVAL 2**

**ID NOMINAL 1**

**INPUT INTERVAL 24**

**INPUT NOMINAL 6**

**REJECTED INTERVAL 10**

**REJECTED NOMINAL 14**

**REJECTED UNARY 8**

**SEGMENT NOMINAL 1**

**TARGET BINARY 1**

**TEXT NOMINAL 3**

**TIMEID INTERVAL 4**

**Frequencies: TextCluster\_cluster\_**

**Percent of**

**Segment Frequency Total**

**Segment Variable Value Count Frequency**

**TextCluster\_cluster\_ 7 4101 26.9696**

**TextCluster\_cluster\_ 2 2172 14.2838**

**TextCluster\_cluster\_ 4 2172 14.2838**

**TextCluster\_cluster\_ 1 1862 12.2452**

**TextCluster\_cluster\_ 5 1646 10.8247**

**TextCluster\_cluster\_ 9 909 5.9779**

**TextCluster\_cluster\_ 8 893 5.8727**

**TextCluster\_cluster\_ 3 649 4.2681**

**TextCluster\_cluster\_ 6 443 2.9133**

**TextCluster\_cluster\_ 10 359 2.3609**

**Variable: TextCluster\_cluster\_ Segment: 7 Count: 4101**

**Decision Tree Importance Profiles**

**Variable Worth Rank**

**TextCluster\_SVD15 0.28363 1**

**TextCluster\_SVD13 0.16726 2**

**TextCluster\_SVD1 0.16082 3**

**TextCluster\_SVD14 0.13792 4**

**TextCluster\_SVD11 0.13303 5**

**TextCluster\_SVD21 0.11728 6**

**TextCluster\_SVD7 0.10990 7**

**TextCluster\_SVD8 0.09002 8**

**TextCluster\_SVD10 0.08591 9**

**TextCluster\_SVD6 0.07749 10**

**Variable: TextCluster\_cluster\_ Segment: 2 Count: 2172**

**Decision Tree Importance Profiles**

**Variable Worth Rank**

**TextCluster\_SVD7 0.070153 1**

**TextCluster\_SVD4 0.051778 2**

**TextCluster\_SVD13 0.044028 3**

**TextCluster\_SVD5 0.039335 4**

**TextCluster\_SVD1 0.034757 5**

**TextCluster\_SVD6 0.033625 6**

**TextCluster\_SVD2 0.032431 7**

**TextCluster\_SVD14 0.026521 8**

**TextCluster\_SVD12 0.025789 9**

**TextCluster\_SVD15 0.023657 10**

**Variable: TextCluster\_cluster\_ Segment: 4 Count: 2172**

**Decision Tree Importance Profiles**

**Variable Worth Rank**

**TextCluster\_SVD1 0.053269 1**

**TextCluster\_SVD15 0.048182 2**

**TextCluster\_SVD7 0.047825 3**

**TextCluster\_SVD6 0.041601 4**

**TextCluster\_SVD10 0.034541 5**

**TextCluster\_SVD8 0.033411 6**

**TextCluster\_SVD13 0.027815 7**

**TextCluster\_SVD21 0.022615 8**

**TextCluster\_SVD3 0.022269 9**

**TextCluster\_SVD9 0.022242 10**

**Variable: TextCluster\_cluster\_ Segment: 1 Count: 1862**

**Decision Tree Importance Profiles**

**Variable Worth Rank**

**TextCluster\_SVD16 0.053189 1**

**TextCluster\_SVD19 0.041045 2**

**TextCluster\_SVD4 0.018146 3**

**TextCluster\_SVD12 0.016521 4**

**TextCluster\_SVD21 0.015273 5**

**TextCluster\_SVD17 0.014643 6**

**TextCluster\_SVD20 0.012141 7**

**TextCluster\_SVD1 0.012068 8**

**TextCluster\_SVD11 0.010569 9**

**TextCluster\_SVD15 0.009175 10**

**Variable: TextCluster\_cluster\_ Segment: 5 Count: 1646**

**Decision Tree Importance Profiles**

**Variable Worth Rank**

**TextCluster\_SVD1 0.10167 1**

**TextCluster\_SVD2 0.03926 2**

**TextCluster\_SVD11 0.03669 3**

**TextCluster\_SVD3 0.03495 4**

**TextCluster\_SVD5 0.02942 5**

**TextCluster\_SVD4 0.02822 6**

**TextCluster\_SVD15 0.02490 7**

**TextCluster\_SVD13 0.02369 8**

**TextCluster\_SVD8 0.02279 9**

**V\_FUNDBY 0.01860 10**

**Variable: TextCluster\_cluster\_ Segment: 9 Count: 909**

**Decision Tree Importance Profiles**

**Variable Worth Rank**

**TextCluster\_SVD1 0.076140 1**

**TextCluster\_SVD2 0.075904 2**

**TextCluster\_SVD15 0.022953 3**

**V\_ADMINBY 0.014411 4**

**TextCluster\_SVD11 0.013453 5**

**V\_FUNDBY 0.012841 6**

**TextCluster\_SVD4 0.012767 7**

**NUMDAYS 0.011299 8**

**TextCluster\_SVD14 0.010439 9**

**CAGE\_YR 0.009841 10**

**Variable: TextCluster\_cluster\_ Segment: 8 Count: 893**

**Decision Tree Importance Profiles**

**Variable Worth Rank**

**TextCluster\_SVD9 0.032538 1**

**TextCluster\_SVD18 0.022942 2**

**TextCluster\_SVD14 0.015305 3**

**TextCluster\_SVD1 0.015045 4**

**TextCluster\_SVD20 0.012622 5**

**TextCluster\_SVD8 0.010848 6**

**TextCluster\_SVD6 0.008651 7**

**TextCluster\_SVD4 0.006312 8**

**TextCluster\_SVD12 0.005317 9**

**TextCluster\_SVD15 0.005041 10**

**Variable: TextCluster\_cluster\_ Segment: 3 Count: 649**

**Decision Tree Importance Profiles**

**Variable Worth Rank**

**TextCluster\_SVD4 0.051558 1**

**TextCluster\_SVD5 0.035521 2**

**TextCluster\_SVD15 0.011854 3**

**TextCluster\_SVD12 0.011585 4**

**TextCluster\_SVD13 0.009953 5**

**TextCluster\_SVD16 0.009779 6**

**TextCluster\_SVD9 0.009264 7**

**TextCluster\_SVD8 0.005398 8**

**TextCluster\_SVD11 0.004511 9**

**TextCluster\_SVD14 0.003488 10**

**Variable: TextCluster\_cluster\_ Segment: 6 Count: 443**

**Decision Tree Importance Profiles**

**Variable Worth Rank**

**TextCluster\_SVD10 0.040626 1**

**TextCluster\_SVD3 0.028547 2**

**TextCluster\_SVD7 0.016912 3**

**TextCluster\_SVD18 0.009858 4**

**TextCluster\_SVD20 0.009088 5**

**TextCluster\_SVD2 0.008485 6**

**NUMDAYS 0.006772 7**

**TextCluster\_SVD14 0.003926 8**

**TextCluster\_SVD17 0.003541 9**

**TextCluster\_SVD6 0.002414 10**

**Variable: TextCluster\_cluster\_ Segment: 10 Count: 359**

**Decision Tree Importance Profiles**

**Variable Worth Rank**

**TextCluster\_SVD12 0.040553 1**

**TextCluster\_SVD14 0.017096 2**

**TextCluster\_SVD11 0.009994 3**

**TextCluster\_SVD1 0.008085 4**

**TextCluster\_SVD10 0.006419 5**

**TextCluster\_SVD8 0.003672 6**

**TextCluster\_SVD6 0.003012 7**

**TextCluster\_SVD5 0.002629 8**

**TextCluster\_SVD4 0.002548 9**

**TextCluster\_SVD15 0.002067 10**

**\*------------------------------------------------------------\***

**\* Score Output**

**\*------------------------------------------------------------\***

**\*------------------------------------------------------------\***

**\* Report Output**

**\*------------------------------------------------------------\***

**Export table EMWS1.Prof\_TEST**

**A screenshot of a social media post

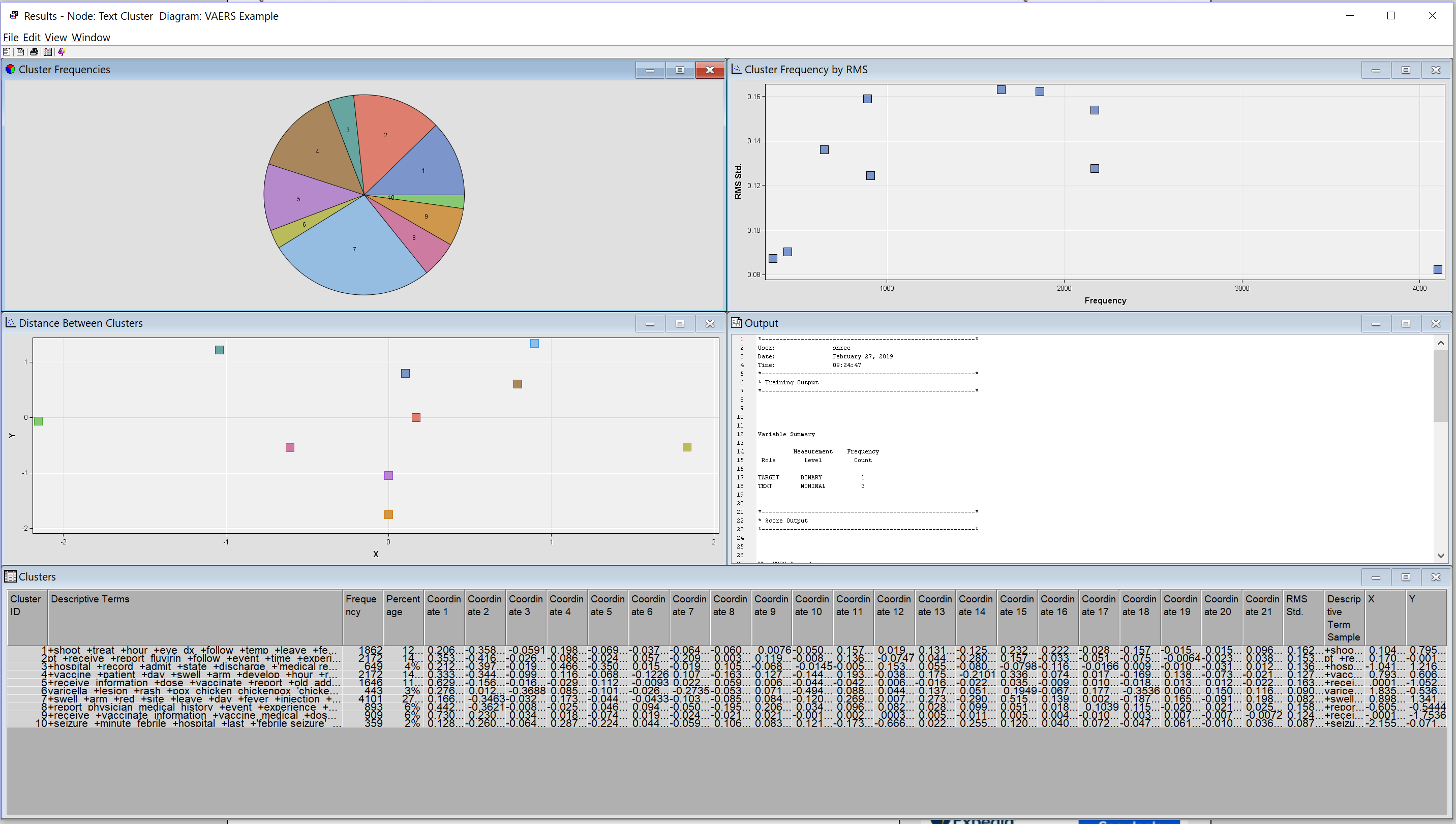
Description automatically generated**

**Pivot Table**

**A screenshot of a cell phone

Description automatically generated**

**Deliverable 3**

**Text cluster with default settings**

**Text cluster with hierarchical clustering algorithm**

**A screenshot of a cell phone

Description automatically generated**

**Text Cluster with Maximum Number & Hierarchical Clustering**

**A picture containing screenshot

Description automatically generated**

**Text Cluster with Exact Number**

**A screenshot of a computer

Description automatically generated**

**Differences between the above mentioned combinations**

|  |  |  |  |
| --- | --- | --- | --- |
| **Default** | **Hierarchical** | **Hierarchical + Maximum** | **Exact** |
| No hierarchy, separate clusters | High no. of levels of clusters | Low no. of levels of clusters | No hierarchy, separate clusters |
| Slight increase in rms std. as the cluster frequency increases, its difficult to find a trend by observation since the frequency RMS graph is spread out | As cluster frequency increases, RMS std decrease. | No. of clusters are less, its difficult to find a trend by observation since the frequency RMS graph is spread out | Slight decline in rms std. as the cluster frequency increases, yet overall stagnant graph. |
| No. of clusters = 10 | No. of clusters = 41 | No. of clusters = 10 | No. of clusters = 39 |

**Setting variable SERIOUS for attributes Use and Report as YES**

**A screenshot of a cell phone

Description automatically generated**

**Result of Text Cluster after changing setting**

**A screenshot of a computer

Description automatically generated**

**Deliverable 4**

Cluster labeling in natural language processing and information retrieval is problem of picking human readable and descriptive labels for the clusters produced by document clustering algorithm. Cluster labeling algorithms examine the contents of documents per cluster to find a labeling that summarize the topic of each cluster and distinguish the clusters from each other.

Schemes for labeling clusters are:-

* **Differential cluster labeling**

It labels a cluster by comparing term distributions across clusters which are used for feature selection in document classification such as mutual information and Chi-Squared selection.

Terms in low frequency are not that good in representing the whole cluster and can be ignored in labeling a cluster. We will get best results for differential cluster labeling by ignoring these rare terms and using a differential test.

1. **Pointwise mutual information**: In this mutual information measures the degree of dependence of two random variables.
2. **Chi-Squared selection**: This test is use to calculate the probability that the occurrence of an event matches the initial expectations. It can be used to determine if two events are statistically independent.

* **Cluster-Internal labeling**

It selects the label that depends only on content of cluster of interest. It does not compare with other clusters. It uses variety of methods such as

1. **Centroid labels**: Most frequently used model in field of information retrieval. It represents documents as vectors. Binary vectors have values of 1 if the term is present within a particular document and 0 if is absent. For a particular cluster of documents we can calculate the centroid by finding arithmetic mean of all document vectors.
2. **Contextualized centroid labels**: This technique overcome the limitation of centroid labels and it is also cost effective and simple technique.
3. **Title labels**: Its an alternative to centroid labeling. It finds smallest Euclidean distance to the centroid and uses it as label for cluster. It provides additional information that would not be present in list.
4. **External knowledge labels**: Cluster labelling can be done indirectly using external knowledge such as pre-categorized knowledge. These can be used to retrieve K-nearest documents.
5. **Combining several cluster labels**: This technique is used to obtain better labels. Linear regression is an example of this.

**Characteristics of the cluster:**

Some of the essential characteristics that a cluster should possess are:

* The dissimilarities within-cluster should be small.
* The dissimilarities between-cluster should be large.
* The number of clusters should be low.
* Clusters should be stable.

A label for a cluster can summarize very large numbers of documents and distinguish the clusters from each other. Labels for the clusters are produced by a clustering algorithm by examining the contents of the documents per cluster. A cluster’s key words help in identifying the meaning to those labels. A label uses the words to express the meaningful result of the clustering as documents are clustered by similarity of words. Hence, it is very important in labeling the clusters with right words that gives a reasonable meaning.

**Topics**

Topic analysis helps in discovering abstract topics from tweets and webpages. This helps in searching and retrieval of documents based on topic. We also use topic modeling for number of documents to extract topics and mine phrases to provide better functionality to user.

**Extraction of topics from document**

* Each document is modeled as a multinomial distribution of topics and each topic is modeled as a multinomial distribution of words.
* LDA assumes that the every chunk of text we feed into it will contain words that are somehow related. Therefore choosing the right corpus of data is crucial
* It also assumes documents are produced from a mixture of topics. Those topics then generate words based on their probability distribution.