IR Assignment -3

Group Number 22

ReadMe

Ans 1) We used the 20_newsgroup dataset and took five classes namely, comp.graphics, sci.med, talk.politics.misc, rec.sport.hockey, sci.space.

Dataset Description -

The dataset contained 20 folders having different documents, however restricting to the question we used only 5 folders as mentioned above for computation. Each document had some description followed by the text.

Methodology -

We first loaded the given dataset, each class contained 1000 documents. We preprocessed the dataset using the steps mentioned in the preprocessing steps.

Then we splitted the dataset into different ratios, 80:20, 70:30 and 50:50, using the document ids.

After splitting the data, for training data we calculated the Term Frequency, Class Frequency and Inverse Class Frequency and using these we calculated the TF-ICF scores for each word classwise. Using these TF-ICF scores we extract the top K features by sorting them using TF-ICF scores. After getting the top K features from each class we prepare a final vocabulary. For Naive Bayes we calculated the prior and conditional probability for prediction on Test Data.

Preprocessing Steps -

For the preprocessing, as we noticed that many text documents contained Email Ids so we removed them, also digits, punctuation, special symbols, stop words were removed from the given data.

Assumptions -

As we noticed in each document there was some extra information that was not relevant, so in preprocessing steps we removed those lines.

For example -

Path: cantaloupe.srv.cs.cmu.edu!crabapple.srv.cs.cmu.edu!fs7.ece.cmu.edu!europa.eng.gtefsd.com!emory!wupost!usc!newshub.sdsu.edu!ucssun1.sdsu.edu.sdsu.edu!weston

From: weston@ucssun1.sdsu.edu (weston t)

Newsgroups: comp.graphics

Subject: graphical representation of vector-valued functions

Date: 5 Apr 1993 20:22:28 GMT

Organization: SDSU Computing Services

Lines: 13

Message-ID: <1pq4e4$afc@pandora.sdsu.edu>

.

.

.

Starting 13 lines were removed.

Ans 2)

We have used bitcoin dataset from the following url -

https://snap.stanford.edu/data/soc-sign-bitcoin-alpha.html

Dataset Description –

It represents Weighted Signed Directed Bitcoin Alpha web of trust network.

The format of data is -

Source	Target	Rating	Time

Source represents the node id of the Source node

Target represents the node id of the Target node

Rating represents the rating of the Source node for the Target node

Time represents the rating time. It is measured as seconds since Epoch

Methodology –

- Graph is represented in the form of Adjacency matrix, where 1 denotes the edge from source node to destination node and 0 denotes absence of edge. A preview of adjacency matrix is shown below -

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
1	0	1	0	1	0	0	0	0	1	1	1	0	0	0	1	0
2	1	0	0	1	1	0	1	1	1	1	1	0	0	0	1	0
3	0	1	0	0	1	1	1	1	0	1	1	1	1	0	0	0
4	1	1	0	0	0	0	0	0	1	1	1	0	0	0	0	1
5	0	1	1	0	0	1	0	1	0	0	1	1	1	0	0	0

- Also, the graph is represented as an Edge list. It is shown below -

```
[(7188, 1), (1, 160), (1, 1028), (1, 309), (1, 11), (1, 594), (1, 1316), (1, 1392), (1, 1583)]
```

- Number of nodes are calculated by counting the source set nodes and number of edges are calculated by counting the number of 1's in the adjacency matrix.
- Average In-degree and average out-degree is calculated by counting the number of incoming and outgoing nodes respectively for each node and dividing by the total number of nodes.
- Node with maximum in-degree is the one to which maximum links are incoming and maximum out-degree is one from which maximum links are outgoing.
- In the case of directed graphs, the density of the network is actual connections by possible connections.

- The in-degree and out-degree distributions are plotted.

  Nodes Vs In-degree of nodes curve is plotted

  In-Degree of Nodes Vs Fraction of Nodes with In-Degree curve is plotted

  Nodes Vs Out-degree of nodes curve is plotted

  Out-Degree of Nodes Vs Fraction of Nodes with OutDegree curve is plotted
- Cluster coefficient of each node is calculated which is a measure of the degree to which nodes in a graph tend to cluster together.
- Clustering-coefficient distribution of the network is plotted with respect to fraction of nodes.
- Degree Centrality Measure for each node is calculated for in-degree and out-degree and saved in a text file. Higher the score more central the node is.
- Then, PageRank Scores are calculated for each node.
- Authority and Hub Scores are calculated for each node using Hyperlink-Induced Topic Search (HITS) algorithm.

References -

S. Kumar, F. Spezzano, V.S. Subrahmanian, C. Faloutsos. Edge Weight Prediction in Weighted Signed Networks. IEEE International Conference on Data Mining (ICDM), 2016.

S. Kumar, B. Hooi, D. Makhija, M. Kumar, V.S. Subrahmanian, C. Faloutsos. REV2: Fraudulent User Prediction in Rating Platforms. 11th ACM International Conference on Web Searchand Data Mining (WSDM), 2018.