HW6: Text Clustering

Published Date:
Apr. 27, 2017, 4:30 p.m.

Deadline Date:
May 9, 2017, 4:30 pm

Description:

Description:
****
This is an individual assignment and deadline is 5/9/2017 4:30 PM.
****
Overview and Assignment Goals:

The objectives of this assignment are the following:

- Implement the K-Means algorithm
- Deal with text data (news records) in document-term sparse matrix format.
- Design a proximity function for text data
    - Think about the Curse of Dimensionality
- Think about best metrics for evaluating clustering solutions.

Detailed Description:

For the purposes of this assignment, you will implement the k-Means or bisecting k-Means clustering algorithm. Please do not use libraries for this portion of your assignment. Additionally, you will gain experience with internal cluster evaluation metrics.

Input data (provided as training data) consists of 8580 text records in sparse format. No labels are provided.

For evaluation purposes (leaderboard ranking), we will use the Normalized Mutual Information Score (NMI), which is an external index metric for evaluating clustering solutions. Essentially, your task is to assign each of the instances in the input data to K clusters identified from 1 to K.

For the leaderboard evaluation, set K to 7. The leaderboard will report the NMI on 50% samples from the dataset.

The train.dat file is a simple CSR sparse matrix containing the features associated with different feature ids in the input file. It differs from previous train.dat files in that it does not contain labels as the first element in each row.

Some things to note:

- The public leaderboard shows results for 50% of randomly chosen test instances only. This is a standard practice in data mining challenge to avoid gaming of the system. The private leaderboard will be released after the deadline and evaluates all the entries in the data set.
- In a 24-hour cycle, you can submit a prediction file up to 5 times.
- The final ranking will always be based on the last submission.
- format.dat shows an example file containing 8580 rows with random class assignments from 1 to 7.
- There are no test.dat files in this assignment.

Rules:
- This is an individual assignment. Discussion of broad level strategies is allowed but any copying of submission files and source codes will result in honor code violation.
- Feel free to use the programming language of your choice for this assignment.
- While you can use libraries and templates for dealing with input data you should implement your own k-Means clustering algorithm.

Deliverables:
- Valid submissions to the Leader Board website: https://coe-cmp.sjsu.edu/ (username is your SJSU ID and your initial password is cmpe239sp17).
- **Canvas Submission of source code and report:**
  - Create a folder called HW6_SJSU-ID
  - Include a 2-4 page, single-spaced report describing details regarding the steps you followed for developing the clustering solution for text data. The report should be in PDF format and the file should be called **report.pdf**. Be sure to include the following in the report:
    1. Name and SJSU ID.
    2. Rank & NMI for your submission (at the time of writing the report).
    3. Your approach (pseudocode for k-Means).
    4. Implement/Use your choice of internal evaluation metric and plot this metric on the y-axis with values for k on the x-axis increasing from 3 to 21 in steps of 2 for the given dataset.
    5. Describe, any feature selection/reduction or custom proximity measure you used in this study.
  - Create a subfolder called src and put all the source code there.
  - Archive your parent folder (.zip or tar.gz) and submit via Canvas for HW6.

Grading:

Grading for the Assignment will be split on your implementation (50%), report (20%) and ranking results (30%).

Files:

- *Train Data:* [Download File](#)
- *Format File:* [Download File](#)