PR1: Text Clustering

Published Date:
Sept 23, 2021

Due Date:
Oct 7, 2021, 11:59pm

Description:

****
This is an individual assignment.
****
Overview and Assignment Goals:

The objectives of this assignment are the following:

- Implement the ***Bisecting 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 bisecting k-Means clustering algorithm. *You may 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 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.

The train.dat file is a simple CSR sparse matrix containing the features associated with different feature ids and their counts in the input file. (Each line represents a document. Each pair of values within a line represent the term id and its count in that document)

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 challenges to avoid gaming of the system. The private leaderboard will be released after the deadline and evaluates all the entries in the data set.
- Each day, 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 cluster assignments from 1 to 7.

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 Bisecting k-Means clustering algorithm.**

Deliverables:

- Valid submissions (source code and output cluster labels) to the Leader Board website: http://coe-clp.sjsu.edu/ To access the page you need to use the SJSU VPN: https://www.sjsu.edu/it/services/network/vpn/index.php Username: mySJSU email, password: mySJSU password.
- **Canvas Submission of source code and report:**
  - 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. Your approach (pseudocode for Bisecting k-Means)
    3. 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.
    4. Describe, any feature selection/reduction or custom proximity measure you used in this study.

Grading:

Grading for the Assignment will be split on your implementation (50%), report (40%) and ranking results (10%). Extra credit (1% of final grade) will be awarded to the top-3 performing algorithms. Note that extra credit throughout the semester will be tallied outside of Canvas and will be added to the final grade at the end of the semester.

The ranking score will be defined as follows (out of 10%):

0%: If below the baselines
3%: If in low-35% of the final leaderboard
6%: If mid-35% of the final leaderboard
10%: If in top-35% of the final leaderboard

Files:

- *Train Data*
- *Format File*