Machine Learning and Pattern Recognition MC886/MO444

University of Campinas (UNICAMP), Institute of Computing (IC) Assignment #3, 2018s2, Prof. Sandra Avila

Objective

Discover organizational structures in a dataset with unsupervised learning techniques.

Activities

- 1. Discover the number of groups present in the data or a reliable range of possible values. Do some experiments in this regard.
- 2. Analyze the medoids of some groups (for example, 3 groups) and their closest neighbors in the groups. Do they make sense? Are they talking about the same type of documents?
- 3. Think of possible ways of checking the validity/quality of your clusters.
- 4. Re-do the best experiment above considering the PCA dimensionality reduction. Consider different energies (variance) to cut and reduce dimensionality. What are the conclusions when using PCA in this problem?
- 5. Prepare a 4-page (max.) report with all your findings. It is UP TO YOU to convince the reader that you are proficient on Unsupervised Learning Techniques, and the choices it entails.

Dataset

Health News in Twitter (https://archive.ics.uci.edu/ml/datasets/Health+News+in+Twitter#), from UCI Machine Learning Repository, is a dataset collected in 2015 using Twitter API.

Dataset Information:

- The health news are from BBC ('bbchealth.txt'), CNN ('cnnhealth.txt'), Fox News ('foxnewshealth.txt') and Everyday News ('everydayhealth.txt').
- Each file is related to one Twitter account of a news agency. For example, bbchealth.txt is related to BBC health news. Each line contains tweet id|date and time|tweet. The separator is '|'.
- There are 13,229 health news ('health.txt'). The bag-of-words features vectors (with 1,203 dimensions) representing each new are also available ('bags.csv'), as well the word2vec vectors (with 128 dimensions, 'word2vec.csv'). You should choose one.
- The data is available at: https://www.dropbox.com/s/ahkim9u103v0q9i/health-dataset.zip

Deadline

Tuesday, **November 6th** in the beginning of the class, 7 pm.

Penalty policy for late submission: You are not encouraged to submit your assignment after due date. However, in case you did, your grade will be penalized as follows:

- \bullet November 7th 7pm : grade * 0.75
- November 8th 7pm : grade * 0.5
- November 9th 7pm : grade * 0.25

Submission

On the deadline day, bring your 4-page printed report. The template for report is available at https://www.dropbox.com/s/nc6d89otr8ekvjd/report-model.zip. Please, print on both sides of the page. The report should be written in Portuguese or English.,

Submit a zip file, with the code and the report (PDF file), via Moodle.

This activity is NOT individual, it must be done in pairs (two-person group).