MO444 - 2018s1 - Activity #3

Prof. Anderson Rocha (anderson.rocha@ic.unicamp.br)

Objective

Given a collection of news headlines, produce meaningful clusters of headlines. In other words, you should cluster them in such a way that each cluster has a common “theme”.

Teams :

This assignment is in teams of 2.

Activities

1. Engineer your features. Here you do not have them for free. You need to think of possible ways for transforming the collected data into meaningful features. For some ideas, read attached papers. If you cannot think of anything, talk to the professor for some ideas.
2. Some ideas for your features:
    a. Term/n-gram frequencies (also re-weighted by inverse document frequency)
    b. Word embeddings — https://fasttext.cc/docs/en/english-vectors.html - This part is optional if you do part 2a.
3. Determine the most adequate number of clusters by plotting the *k*-means cost function over different values of *k*.
4. What is the common “theme” of each cluster? Does the most adequate number of clusters lead to meaningful clusters?
5. Perform the same experiments on some two or three subsets of the data containing only headlines from a single year. Is the number of clusters found earlier still adequate for any given year? Are there recurrent “themes” of clusters? For this activity, you should apply clustering for each year separately. Does this help you with the prediction of the possible number of clusters?
6. Report your results providing both quantitative and qualitative assessments of the clusters found by your solution.
    a. Examples of quantitative assessments: k-means cost function, silhouette coefficient, variance of each final cluster, etc.
    b. Examples of qualitative assessments: most frequent words, most frequent n-grams, word clouds. Suggestion: try with words 2-gram and 3-grams and also char-4-grams. These would give you already very good indications.

Dataset

The dataset contains 1 million news headlines from ABC (Australian Broadcasting Corporation) published over a period of 15 years. It is formatted in CSV, with two columns, described below.

Columns

- publish_date (yyyyMMdd) ranging from 2003-02-19 to 2017-12-31
- headline_text (ASCII; lowercase)

The dataset is available at:

+ https://www.ic.unicamp.br/~rocha/teaching/2018s1/mo444/assignments/news_headlines.zip
+ MD5 (news_headlines.zip) = 9bca6f5ac5c3c572d50c3f332124e547

Deadline

Monday, May 14th in the beginning of the class. There will be no deadline extension.

Submission

Bring your 4-page printed report and submit during class on the deadline day. This activity is PAIRS.

Recommended Readings

[1] https://en.wikipedia.org/wiki/N-gram
[2] https://ieeexplore.ieee.org/document/7555393/