Clustering Wikipidea Articles using Hierarchical Clustering

Tharindra Galahena (inf0_warri0r)

December 27, 2013

1 About

This project is to cluster set of articles from 'wikipedia' using hierarchical clustering method. The program classify 100 articles from a data set and output results in a set of html documents.

2 Implementation

As I mentioned earlier this program uses 100 articles from a data set downloaded from Middlebury university web site for clustering. First step is to preprocess the data. For that the text of each article is splitted in to words and the stop words will be removed. Only first four sentences of each article is taken by the program to make it easy to run. And the words are stmmed and the count of each distinct word in the text will be calculated. This calculated word list (consist of distinct words in each article and the number of places that each word appear in a article) can be recorded in a 'word_list' file. So you don't have to read the data set each and every time you run the program.

Then a set of words are chosen to use as features. The words I choose are in the text file named selected. The number of each of these selected words in a article will be the feature vector of that article. The next task was to create that feature vectors. And that ends the preprocessing part of the program. After that the clustering algorithm will run and cluster the articles.

For more information on hierarchical clustering go to,
http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/hierarchical.html

3 Running the Program

3.1 Pre-Requirements

The program uses Python on Linux to run. And id doesn't use any additional libraries to run. And also it needs to data set file download it from, http://www.cs.middlebury.edu/~dkauchak/simplification/ (download Version 2.0

document-aligned data set). But you can also use the 'word_list' file provieded with the source code insted of reading from the data set.

3.2 Running

Run the 'main.py' using following command,

$ python main.py

You need to make sure the files the 'selected', 'word_list' (if you are using saved word list), 'stop', 'endings', 'header.html', 'footed.html' and 'pages/index.html' in the same folder as the 'main.py' when you are running the program. After running the program it ask if the articles should be read or it is ok to load the word list from a file. If you want to do it from the beginning you must enter 'y' and press 'enter'. Then it will ask the path of the dataset. This script is designed to read the format of the earlier mentioned data set so download it from the site and extract it. Then give the path to the file 'normal.txt'. This will read and preprocess the first 100 articles in the data set. And also it will save the word list in at text file called 'word_list' as a JSON string. If you have 'word_list' file which you made earlier (or a file I created is also given with the source code) enter 'n' instead of 'y'. Then it will preprocess words and cluster and write the output files in pages folder. In there you can open the 'index.html' see the clusters in each level.

4 Results

The results of the algorithm will be included in the html documents generated by the program. The accuracy of the clustering will be depended on the selected words. So by changing the word list in the 'selected' file you can change the clusters.

5 Author

- Tharindra Galahen
- tcg.galahena@gmail.com
- http://www.inf0warri0r.blogspot.com

6 Licenses