CS6670 Topics in Data Mining

Assignment: Opinion Mining

Marks: 60

Start Date: 20-08-2017

Due Date: 11-09-2017

Overview

In this project, you will learn how to find aspects and opinions in product reviews. You will also learn how to perform basic text pre-processing before you can do opinion mining. We are providing you a digital camera dataset in two forms: tagged and untagged (see readme). Reference paper "Mining and summarizing customer reviews".

Getting Started

The coding is based on Java & python 2.7. Download and install python 2.7: sudo apt-get install python2.7. For windows, you can get it from here.

You need to use the Opinion Finder tool to do pre-processing steps (you also try out the GATE tool).

For basic text pre-processing using python see here. Some more info about text mining see here. Natural language processing with Python and NLTK p.1 (youtube).

For association rule mining using FP-Growth algorithm

https://github.com/enaeseth/python-fp-growth (recommended)

http://www.borgelt.net/fpgrowth.html
FP-growth algorithm (youtube)

As mentioned in the class, you can also consider using the TextBlob tool. Following are some relevant links.

http://rwet.decontextualize.com/book/textblob/

http://textblob.readthedocs.io/en/dev/install.html

http://textblob.readthedocs.io/en/dev/quickstart.html

https://textblob.readthedocs.io/en/dev/

Your assignment will be graded based on the result of Opinion Finder. It is optional if you want to give a comparison of the result using Opinion Finder vs TextBlob in the report.

Exercises

Exercise 1 [5 points] Take the untagged dataset and perform POS tagging using the Opinion Finder tool.

Exercise 2 [10 points] From the POS tagged dataset extract all the Nouns. Call this the list of candidate aspects. (Hint: Need to write a Python script)

Note: While extracting aspects considered the lemma of the nouns instead of the nouns. Lemma will remove unnecessary duplicates. For instance, "picture" and "pictures" are treated as a different noun by Opinion Finder although they represent the same feature "picture". Lemmas are used to solve this problem

Exercise 3 [10 points] Select all the frequent aspects using Association Rule Mining algorithm with minimum 0.5% and 1% support. (Submit the top-10 most frequent aspects in decreasing order of support.). What difference do you see in using 0.5% vs 1% minimum support?

Note: You need to consider each sentence as a transaction and not the whole review. Same feature might be mentioned multiple number of times in a single review. As a result, the count of frequent features may be more than the total number of reviews.

If two or more nouns are occurring consecutively, then you should treat it as follows: for a bigram "battery life", pass three nouns to the association rule mining algorithm: "battery", "life" and "battery life". With this approach if "battery life" is a frequently occurring phrase then that will remain after association rule mining whereas less frequently occurring phrases such as "battery duration" will only contribute to the feature "battery".

Exercise 4 [15 points] Use the identified frequent aspects in Exercise 3 to find candidate opinion words. These are the adjectives or adverbs which are near the frequent aspects identified in Exercise 3 (within five words on both sides of a frequent aspect). (Submit the list of top-10 most frequently used opinion words for both 0.5% and 1% support.)

Note: If the same opinion word is near to one or more frequent aspect in same sentence, then you should count it only once.

Exercises 5 [15 points] Compute precision, recall and F1 score for the aspects identified obtained step in Exercise 3. Use the tagged dataset to know the ground-truth aspects.

Exercise 6 [5 points] Compute the number of positive and negative opinion words you obtained in Exercise 4. You need to use the provided list of standard opinion words in English.

What to Submit

Your answers to Exercise 3, 4, 5 and 6 in the form of a report. Also submit your code.