**NLP Individual assignment #3**

**Sentiment Analysis of Amazon Product Reviews**

It is increasingly common that Internet users engage in various of online reviews. The availability of these review content offers researchers opportunities to better understand and model online social behavior. In this homework, you will conduct sentiment analysis to gain some understanding about the Amazon product reviews.

**1.Dataset**

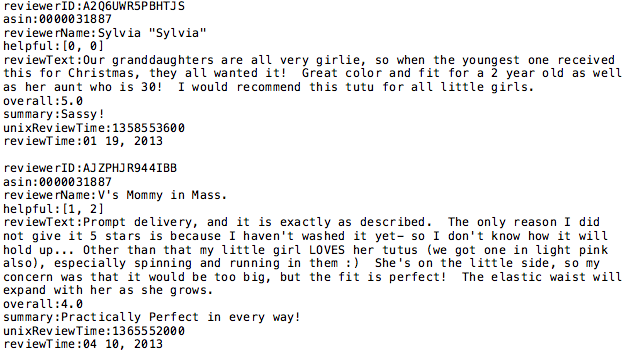
In this problem, you will analyze the review contents from Amazon Product Data provided by Julian McAuley at <http://jmcauley.ucsd.edu/data/amazon/> . This dataset contains product reviews and metadata from Amazon, including 142.8 million reviews spanning May 1996 – July 2014. It includes reviews (ratings, text, helpfulness votes), product metadata (descriptions, category information, price, brand, and image features), and links (also viewed/also bought graphs).

For our tasks, we will use only 5-core subsets of three categories (Baby / Clothing, Shoes and Jewelry / Health and Personal Care). 5-core subsets mean that all users and items in the dataset have at least 5 reviews. Originally, the dataset was a zipped file of json format and the content was arranged in dictionaries. For your convenience, the dataset was modified into text file and is available for download in the Assignment folder in the course web site: Baby.txt and clothing\_shoes\_jewelry.txt. You should choose only one of them for the analysis.

Here are the screenshots of a raw data and a modified review file:



**Figure 1: Screenshot of raw data**

****

**Figure 2: Screenshot of modified data (provided in the assignment)**

Here is the dictionary of the fields provide in the file:

- reviewerID: ID of the reviewer

- asin: ID of the product

- reviewerName: name of the reviewer

- helpful: helpfulness rating of the review, e.g. 2/3

- reviewText: text of the product

- overall: rating of the product

- summary: summary of the review

- unixReviewTime: time of the review (unix time)

- reviewTime: time of the review (raw)

**2. Data Pre-processing (20%)**

You will write a Python code that extracts only reviewTexts. Please submit the sample screenshot of the output (included in your report file).

**3. Sentiment Analysis (80%)**

Based on what we have learned from this class, you will explore the sentiment of the comments at the sentence level. This includes how to process the words and how to conduct the sentiment analysis using classifiers. Ultimately, you will provide two lists of sentences: one is marked as negative and the other as positive.

In your report, please explain in detail the processing techniques that you have applied, the features you used for the classification task, and your experiments. For the data preprocessing/cleaning task, we have learned about several techniques such as tokenization, sentence creation, regular expression processing, stop word filtering, etc. You should describe the techniques you used in this assignment.

For the classification task and the experiments, you should start with the “bag-of-words” features where you collect all the words in the sentence\_polarity corpus and select some number of most frequent words to be the word features. You should use at least NaiveBayes classifier to train and test a classifier on your feature sets, and show its accuracy. In your experiments, in addition to the BOW features, you should use at least two other different sets of features and compare the results. For example, you may take the unigram word features as a baseline and see if the features you designed improve the accuracy of the classification. Here are some of the types of experiments that we have done so far:

• \_Filter by stop words or other pre-processing methods

• \_Representing negation

• \_Using a sentiment lexicon with scores or counts: Subjectivity

**4. Bonus Credit (20%)**

You do not need to work on the following task, but if you do and do well, you will have 20% bonus credit for this assignment.

Choose an additional, more advanced type of task from this list, or propose your own

• \_Using Weka or SciKit Learn classifiers with features produced in NLTK

• \_Using an additional type of lexicon besides Subjectivity

**How to Submit Homework:**

Go to the Blackboard system and the Assignment for Individual Assignment #3. Attach all your files and submit. Your submission should include:

1) your report in a PDF format, including:

* Description of the processing for both part 1(pre-processing) and part 2 (sentiment anlaysis)
* the Table demonstrating two lists of sentences: negative vs positive
* processing screenshots

2) Your Python code