NLP Homework 1

Due Wednesday, February 12, 11:59 pm.

Corpus Statistics and Python Programming

For this assignment, please read Chapter 1 and 2 of NLTK book carefully.

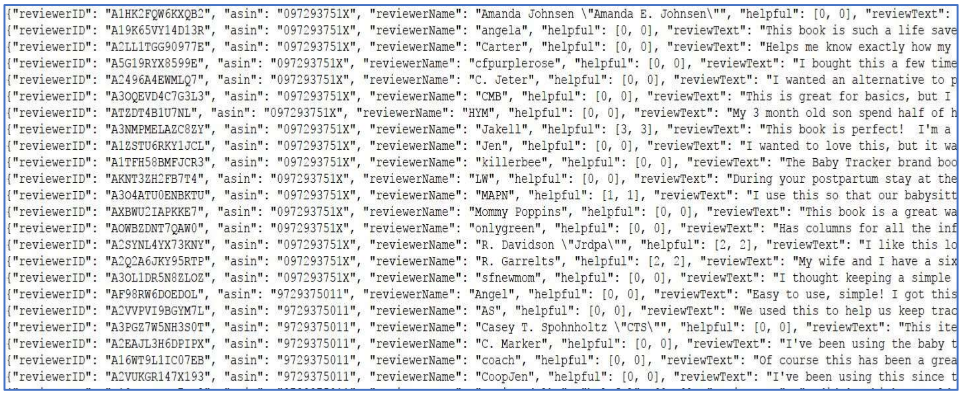
It is increasingly common that Internet users engage in various of online reviews. The availability of these review content offers researchers opportunities to better understand and model online social behavior. As a starting point, you will analyze a subset of Amazon reviews focusing on the word frequency and bigram frequency measures.

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).

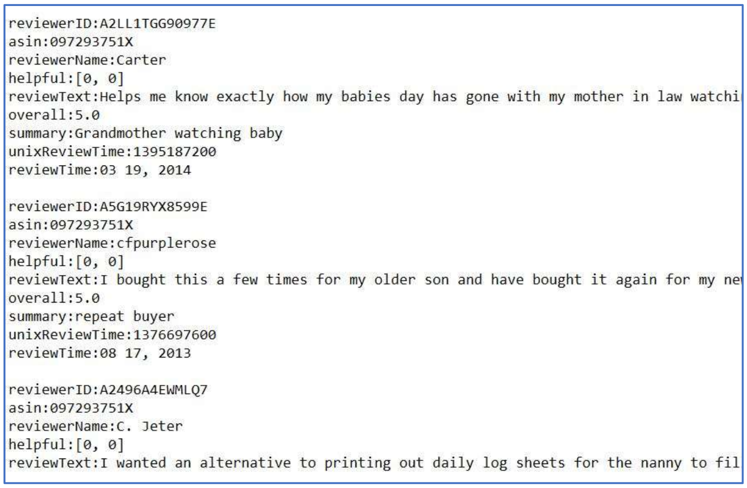
For our tasks, we will use only 5-core subsets of the category “Clothing, Shoes and Jewelry”. 5-core subsets mean that all users and items in the dataset have at least 5 reviews. Originally, the dataset was a zipped file of json format and the content was arranged in dictionaries. For your convenience, the dataset was modified into text file and is available for download in the Assignment folder in the course web site: clothing_shoes_jewelry.txt.

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



```
{"reviewerID": "A1HK2FQW6KXQB2", "asin": "097293751X", "reviewerName": "Amanda Johnsen \"Amanda E. Johnsen\"", "helpful": [0, 0], "reviewText":
{"reviewerID": "A19K65VY14D13R", "asin": "097293751X", "reviewerName": "angela", "helpful": [0, 0], "reviewText": "This book is such a life save
{"reviewerID": "A2LL1TGG90977E", "asin": "097293751X", "reviewerName": "Carter", "helpful": [0, 0], "reviewText": "Helps me know exactly how my
{"reviewerID": "A5G19RYX8599E", "asin": "097293751X", "reviewerName": "cfpurplerose", "helpful": [0, 0], "reviewText": "I bought this a few time
{"reviewerID": "A2496A4EWMLQ7", "asin": "097293751X", "reviewerName": "C. Jeter", "helpful": [0, 0], "reviewText": "I wanted an alternative to p
{"reviewerID": "A3OQEVD4C7G3L3", "asin": "097293751X", "reviewerName": "CMB", "helpful": [0, 0], "reviewText": "This is great for basics, but I
{"reviewerID": "ATZDT4B1U7NL", "asin": "097293751X", "reviewerName": "HYM", "helpful": [0, 0], "reviewText": "My 3 month old son spend half of h
{"reviewerID": "A3NMPMELAZC8ZY", "asin": "097293751X", "reviewerName": "Jakell", "helpful": [3, 3], "reviewText": "This book is perfect!  I'm a
{"reviewerID": "A1ZSTU6RKY1JCL", "asin": "097293751X", "reviewerName": "Jen", "helpful": [0, 0], "reviewText": "I wanted to love this, but it wa
{"reviewerID": "A1TFH58BMFJCR3", "asin": "097293751X", "reviewerName": "killerbee", "helpful": [0, 0], "reviewText": "The Baby Tracker brand boo
{"reviewerID": "AKNT3ZH2FB7T4", "asin": "097293751X", "reviewerName": "LW", "helpful": [0, 0], "reviewText": "During your postpartum stay at the
{"reviewerID": "A3O4ATU0ENBKTU", "asin": "097293751X", "reviewerName": "MAPN", "helpful": [1, 1], "reviewText": "I use this so that our babysitt
{"reviewerID": "AXBWU2IAPKKE7", "asin": "097293751X", "reviewerName": "Mommy Poppins", "helpful": [0, 0], "reviewText": "This book is a great wa
{"reviewerID": "AOWBZDNT7QAW0", "asin": "097293751X", "reviewerName": "onlygreen", "helpful": [0, 0], "reviewText": "Has columns for all the inf
{"reviewerID": "A2SYNL4YX73KNY", "asin": "097293751X", "reviewerName": "R. Davidson \"Jrdpa\"", "helpful": [2, 2], "reviewText": "I like this lo
{"reviewerID": "A2Q2A6JKY95RTP", "asin": "097293751X", "reviewerName": "R. Garrelts", "helpful": [2, 2], "reviewText": "My wife and I have a six
{"reviewerID": "A3OL1DR5N8ZLOZ", "asin": "097293751X", "reviewerName": "sfnewmom", "helpful": [0, 0], "reviewText": "I thought keeping a simple
{"reviewerID": "AF98RW6DOEDOL", "asin": "9729375011", "reviewerName": "Angel", "helpful": [0, 0], "reviewText": "Easy to use, simple! I got this
{"reviewerID": "A2VVPVI9BGYM7L", "asin": "9729375011", "reviewerName": "AS", "helpful": [0, 0], "reviewText": "We used this to help us keep trac
{"reviewerID": "A3PGZ7W5NH3S0T", "asin": "9729375011", "reviewerName": "Casey T. Spohnholtz \"CTS\"", "helpful": [0, 0], "reviewText": "This ite
{"reviewerID": "A2EAJL3H6DPIPX", "asin": "9729375011", "reviewerName": "C. Marker", "helpful": [0, 0], "reviewText": "I've been using the baby t
{"reviewerID": "A16WT9L1IC07EB", "asin": "9729375011", "reviewerName": "coach", "helpful": [0, 0], "reviewText": "Of course this has been a grea
{"reviewerID": "A2VUKGR147X193", "asin": "9729375011", "reviewerName": "CoopJen", "helpful": [0, 0], "reviewText": "I've been using this since t
```

Fig 1. Raw data



```
reviewerID:A2LL1TGG90977E
asin:097293751X
reviewerName:Carter
helpful:[0, 0]
reviewText:Helps me know exactly how my babies day has gone with my mother in law watchi
overall:5.0
summary:Grandmother watching baby
unixReviewTime:1395187200
reviewTime:03 19, 2014

reviewerID:A5G19RYX8599E
asin:097293751X
reviewerName:cfpurplerose
helpful:[0, 0]
reviewText:I bought this a few times for my older son and have bought it again for my ne
overall:5.0
summary:repeat buyer
unixReviewTime:1376697600
reviewTime:08 17, 2013

reviewerID:A2496A4EWMLQ7
asin:097293751X
reviewerName:C. Jeter
helpful:[0, 0]
reviewText:I wanted an alternative to printing out daily log sheets for the nanny to fil
```

Fig 2. Modified review file used for the task

- 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 review texts and create a document to save the extracted text. Please submit the sample screenshot of the output (included in your report file).

Note: you will decide how to process the words, i.e. decide on tokenization and whether to use all lower case, use or modify the stop word list, or lemmatization. Briefly state why you chose the processing options that you did.

3. Data Analysis

To get a rough understanding of what these review texts were about, you will perform the following three tasks on the pre-processed data:

• list the top 50 words by frequency

• list the top 50 bigrams by frequencies, and

- list the top 50 bigrams by their Mutual Information scores (using min frequency 5)

4. Interpretation of the Results

Please explain what you have learned about the reviews, based on the results above. And, please discuss what additional analysis tasks that you think are important to conduct and why.

How to Submit Homework:

Go to the Blackboard system and the Assignment for Homework 1 and submit your report. Your report should include:

1) Description of data pre-processing (with Python processing screenshots in the corresponding section)
2) The results from the analysis tasks (with Python processing screenshots in the corresponding section)
3) Your interpretation of the results and the additional analysis you suggest to perform in the future

Please also upload your Python code besides the report