Report

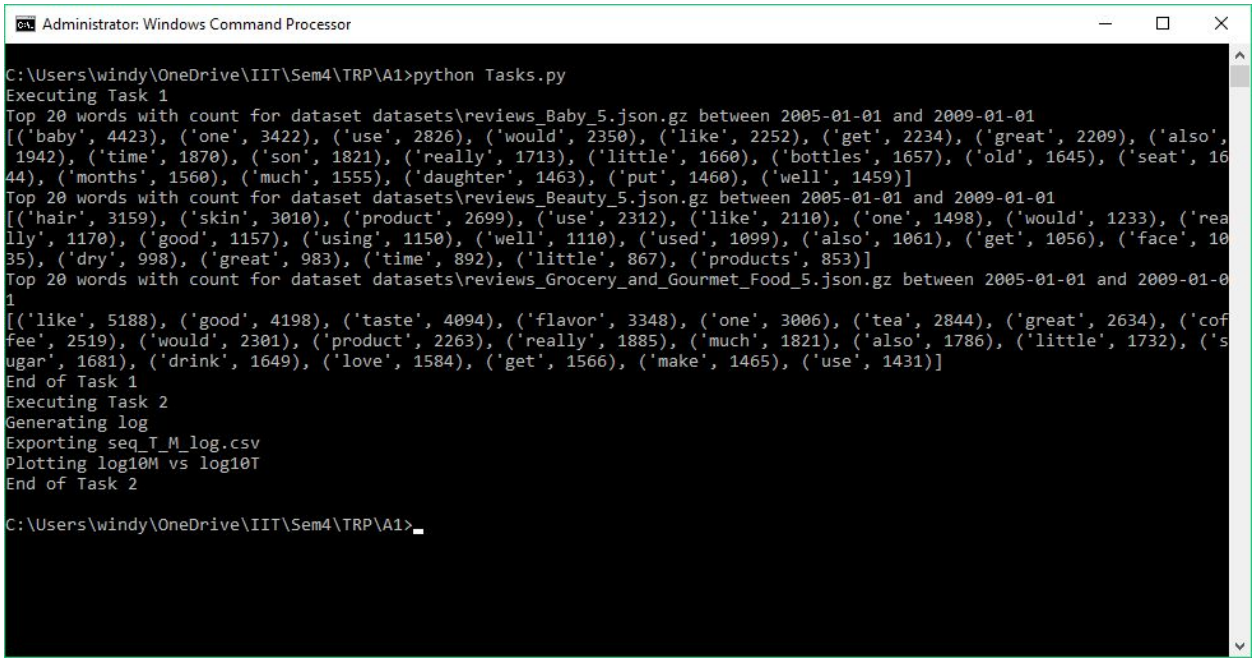
Datasets:

Dataset	Reviews Count	Tasks Used In
reviews_Beauty_5.json.gz	198,502	Task 1 & Task 2
reviews_Baby_5.json.gz	160,792	Task 1
reviews_Grocery_and_Gourmet_Food_5.json.gz	151,254	Task 1

Place the .gz files in directory named *'dataset'*

Run the command: python Tasks.py

Overall Output:



Task 1:

Top 20 words by count per review

Top 20 words with count for dataset datasets\reviews_Baby_5.json.gz between 2005-01-01 and 2009-01-01

Word	Count
baby	4423
one	3422
use	2826
would	2350
like	2252
get	2234
great	2209
also	1942
time	1870
son	1821
really	1713
little	1660
bottles	1657
old	1645
seat	1644
months	1560
much	1555
daughter	1463
put	1460
well	1459

Inference:

For Baby dataset we observe that the top 20 words includes words such as *baby, one, old, months* that represent the age of the target consumer of which reviewed by parent/guardian and also most frequent words includes *son, daughter, little* describing the gender and other aesthetics of the target consumer i.e, Babies.

Top 20 words with count for dataset datasets\reviews_Beauty_5.json.gz between 2005-01-01 and 2009-01-01

Word	Count
hair	3159
skin	3010
product	2699
use	2312
like	2110
one	1498
would	1233
really	1170
good	1157
using	1150
well	1110
used	1099
also	1061
get	1056
face	1035
dry	998
great	983
time	892
little	867
products	853

Inference:

For Beauty dataset we observe that the top 20 words includes words such as *skin, face, hair* and condition of beauty like *good, well, dry* which are more relative to the reviewers' choice of

beauty product being reviewed and the state of their beauty probably after the use of the product.

Top 20 words with count for dataset datasets\reviews_Grocery_and_Gourmet_Food_5.json.gz between 2005-01-01 and 2009-01-01

like	5188
good	4198
taste	4094
flavor	3348
one	3006
tea	2844
great	2634
coffee	2519
would	2301
product	2263
really	1885
much	1821
also	1786
little	1732
sugar	1681
drink	1649
love	1584
get	1566
make	1465
use	1431

Inference:

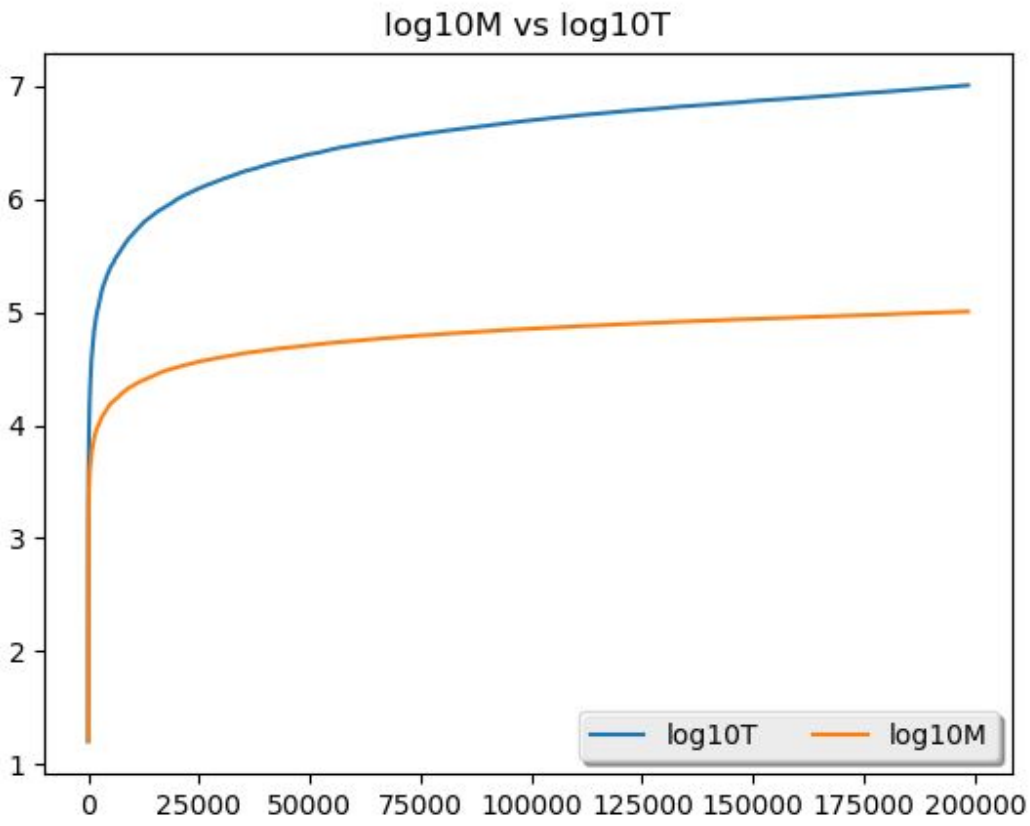
For Grocery and Gourmet Food dataset we observe that the top 20 words includes words such as *like*, *taste*, *flavour* and most frequently reviewed product categories like *tea*, *sugar* and *coffee* which are more relative to the user preferences and category respectively.

Task 2:

The ReviewSeqNo, T, M are logged to a *seq_T_M_log.csv* file located in the directory named *output*.

Dataset used:

reviews_Beauty_5.json.gz



*fig(a): Plot showing **Review Count** on x-axis and $\log_{10}^{T}$ vs $\log_{10}^{M}$ on y-axis*

Inference:

From *fig(a)* we infer that the total words T with $\log_{10}^{T}$ value shoots up as the reviews reckon to around 6k to 8k, further onwards the change in $\log_{10}^{T}$ slowly keeps rises till the total number of reviews are read.
We also observe that the distinct words M with $\log_{10}^{M}$ values too shoots up as the reviews reckon to 5k to 6k range, there is a moderate rise till 50k review count range and the graph of $\log_{10}^{M}$ looks stable as there are hardly new words being added to the distinct set of words collection M