FINAL PROJECT - SUBMISSION 2

TONY ANDRYS

I have downloaded and decompressed the Amazon review dataset. Regarding this data, there are two points data worth noting in this submission: **size** and **format**.

DATA SIZE

The review data is available as 28 files, each file containing the reviews for a specific product category. The categories and their sizes are enumerated below:

Product Category	Review Count	Uncompressed Filesize
Amazon Instant Video	717,651	839.8 MB
Arts	27,980	19.6 MB
Automotive	188,728	129.8 MB
Baby	184,887	42 MB
Beauty	252,056	171.9 MB
Books	12,886,488	14390 MB
Cell Phones & Accessories	78,930	69.4 MB
Clothing & Accessories	581,933	372.9 MB
Electronics	1,241,778	1110 MB
Gourmet Foods	154,635	106.3 MB
Health	428,781	313.1 MB
Home & Kitchen	991,794	757.6 MB
Industrial & Scientific	137,042	85.7 MB
Jewelry	58,621	34.9 MB
Kindle Store	160,793	190.6 MB
Movies & TV	7,850,072	8770 MB
Musical Instruments	85,405	68.1 MB
Music	6,396,350	6510 MB
Office Products	138,084	107.7 MB
Patio	206,250	158.7 MB
Pet Supplies	217,170	164.6 MB
Shoes	389,877	240.4 MB
Software	95,084	95.8 MB
Sports & Outdoors	510,991	368.3 MB
Tools & Home Improvement	409,449	321.2 MB
Toys & Games	435,996	314.7 MB
Video Games	463,669	485.2 MB
Watches	68,356	51.5 MB
TOTAL:	**35,358,850**	**36.31 GB**

DATA FORMAT

- **Data Files**
  - Each file contains a variable amount of reviews for a specific category.
  - Each review is separated by new/blank lines (“\n”).

- **Reviews**
  - Each review has 10 properties, one per line.
  - The data for each property is encoded as a key value pair, where the key and value are separated by a colon.
    - General form —> keyText: valueText

#	Property	Description	Notes
1	**product/productId**	Alphanumeric product identifier	Standardized - Amazon Standard Identification Number. Maps to URL: http://amazon.com/dp/<product/productId>
2	**product/title**	Product title	As printed on product’s webpage.
3	**product/price**	Price in dollars	Should convert to a float to include cents.
4	**review/userId**	Alphanumeric user identifier	Maps to URL: http://www.amazon.com/gp/cdp/member-reviews/<review/userId>
5	**review/helpfulness**	“Helpfulness” score of this review	Amazon users can mark reviews published by others as “Helpful” or “Not Helpful”.
6	**review/profileName**	Human readable profile name	Appears above each review authored by the user.
7	**review/score**	Review score	Valid score range: 0.0, 0.5, 1.0, …, 4.5, 5.0
8	**review/time**	Date & Time a review was published	Unix time (ex: 1280188800)
9	**review/summary**	Summary of review	This is the “header” of the review, shown in larger font than the review text.
10	**review/text**	Full text of review	Always on one line, which is really nice.

Sample Review (From Industrial & Scientific):

```
product/productId: B000796XXM
product/title: Uranium Ore
product/price: 39.95
review/userId: A16A2VQQ59I01D
review/profileName: Mark Snakelord
review/helpfulness: 34/35
review/score: 3.0
review/time: 1280188800
review/summary: I named him Jimmy
review/text: Yeah it's nuclear, but the snap on tin lid provides all the
peace of mind I would ever need. Sometimes I open the lid and peek inside at
my little pile of Uranium Ore (ONLY DO THIS FOR SHORT PERIODS OF TIME). When
I am feeling frisky I put on my lace gloves (for protection) and pet my
little Uranium Ore (I named him Jimmy). Again, you can feel safe because the
tin lid snaps into place so no nuclear junk can get out.
```