Capstone 1 proposal

- Amazon Review Analysis for Helpfulness Score Prediction

- Problem statement / Inspiration

Nobody would deny that Amazon is a ecommerce giant. Keeping up with its economical value growth, more people find Amazon to purchase products they want. Though each product has its own description; However, it is natural that users cannot always trust them at face value and are eager to confirm the product is worth to spend money on. You must have thought - “I just want to know this product is trustworthy and this is the one I will not regret purchasing”. This is where helpfulness score comes into play: we need “helpful” reviews that provides information owner did not mention and could not be noticed unless you actually have tried it. Helpfulness score can reveal how useful this review is and thus will make user’s life easier. It will save you from scrolling down many meaning-less reviews. Therefore, trustful helpfulness score prediction system will help the company enhance existing users’ trust and gain new users’. At the same time, it will strengthen their loyalty to Amazon.

- Data (site: http://jmcauley.ucsd.edu/data/amazon/)

All data is provided from Julian McAuley’s Repository of Recommender System Datasets. The data spans may 1996 to July 2014. I will use deduplicated review dataset that is of version that has removed duplicates even if they are written by different users.This accounts for users with multiple accounts or plagiarized reviews. Format is one-review-per-line JSON and example is shown as below:

```
{
  "reviewerID": "A2SUAM1J3GNN3B",
  "asin": "0000013714",
  "reviewerName": "J. McDonald",
  "helpful": [2, 3],
  "reviewText": "I bought this for my husband who plays the
piano.  He is having a wonderful time playing these old hymns.
The music  is at times hard to read because we think the book
was published for singing from more than playing from.  Great
purchase though!",
  "overall": 5.0,
  "summary": "Heavenly Highway Hymns",
  "unixReviewTime": 1252800000,
  "reviewTime": "09 13, 2009"
}
```

Where

1. reviewerID - ID of the reviewer, e.g. A2SUAM1J3GNN3B
2. asin - ID of the product, e.g. 0000013714
3. reviewerName - name of the reviewer
4. helpful - helpfulness rating of the review, e.g. 2/3
5. reviewText - text of the review

6. overall - rating of the product
7. summary - summary of the review
8. unixReviewTime - time of the review (unix time)
9. reviewTime - time of the review (raw)

On the other hand, Metadata includes descriptions, price, sales-rank, brand info, and co-purchasing links and contain 9.4 million products. Example is as follows:

```
{
  "asin": "0000031852",
  "title": "Girls Ballet Tutu Zebra Hot Pink",
  "price": 3.17,
  "imUrl":
"http://ecx.images-amazon.com/images/I/51fAmVkTbyL._SY300_.jpg",
  "related":
  {
    "also_bought": ["B00JHONN1S", "B002BZX8Z6", "B00D2K1M3O",
"0000031909", "B00613WDTQ", "B00D0WDS9A", "B00D0GCI8S",
"0000031895", "B003AVKOP2", "B003AVEU6G", "B003IEDM9Q",
"B002R0FA24", "B00D23MC6W", "B00D2K0PA0", "B00538F5OK"],
    "also_viewed": ["B002BZX8Z6", "B00JHONN1S", "B008F0SU0Y",
"B00D23MC6W", "B00AFDOPDA", "B00E1YRI4C", "B002GZGI4E",
"B003AVKOP2", "B00D9C1WBM", "B00CEV8366", "B00CEUX0D8",
"B0079ME3KU", "B00CEUWY8K", "B004FOEEHC", "0000031895",
"B00BC4GY9Y", "B003XRKA7A", "B00K18LKX2", "B00EM7KAG6"],
    "bought_together": ["B002BZX8Z6"]
  },
  "salesRank": {"Toys & Games": 211836},
  "brand": "Coxlures",
  "categories": [["Sports & Outdoors", "Other Sports", "Dance"]]
}
```

Where

1. asin - ID of the product, e.g. 0000031852
2. title - name of the product
3. price - price in US dollars (at time of crawl)
4. imUrl - url of the product image
5. related - related products (also bought, also viewed, bought together, buy after viewing)
6. salesRank - sales rank information
7. brand - brand name
8. categories - list of categories the product belongs to

For the purpose of this project, we will mostly likely use asin, title, brand and categories.

- Approach:

This project is supervised learning in that provided data has labels (i.e., helpfulness score). I will approach the problem with regression point of view. First thing I will tackle will be reducing the size of data because full dataset is as big as 17.7 GigaBytes excluding metadata that is 3.1 GigaBytes. To decide how, we need to answer a couple of questions:

‘which one is more informational? wide range of products but relatively short period of time analysis or narrow range of products but long period of time analysis’. Both are valid goal; as I want my project be applicable to as many products as possible, I choose the latter. Therefore, this project will only make use of data from 2011 to 2013. Together with time span, due to the fact that this project only focuses on helpfulness score prediction, I will filter the data to utilize reviews with helpfulness scores only. Next, because data is of text, I will need to use bag of word approach so I can analyse which aspects of a review determine how helpful it is. Data format transformation from JSON to DataFrame will be made through Python’s json and Pandas libraries. RMSE (Root Mean Square Erorr) will be used for evaluation metrics because the goal of the project is predict the helpfulness score of a review as correctly as possible. Since data amount is not a problem, I will adopt cross validation method as well to consolidate the generalization of the result.

- Deliverables**:**

1. Codes (notebook) that include:
    a. Data Wrangling
    b. Data Cleaning
    c. Exploratory Data Analysis
    d. Inferential Statistical Analysis
    e. Natural Language Processing
    f. Et cetera
2. Capstone project analysis report
3. Slide deck on the project