DSCI591 ACRSent Notes

The book dataset is very large. One file (of 3 possible files) is over 10 million lines long. All of the work below will need to be done in batches instead of all at once due to space and time constraints.

Book Data Cleaning Issues

4/17/2021

1. Need to run read\_csv with bad\_error\_lines = False to remove the rows that have less or more items than number of columns defined
2. Some dates are missing
   1. Plan is to remove all rows that have missing dates
3. There are nans in other columns such as the review body and verified purchase
   1. How do we want to handle these? Remove them?
   2. No review body will probably just need to be removed entirely since it adds nothing to the NLP
   3. NOTE: the nans seem to be fairly rare and considering the size of the data set I don’t think removing them is even slightly an issue
4. There are emojis in the review body (😄). What do we do? Possibilities:
   1. The NLP can handle it as it is so leave them in (might not need to deal with it till next term in this case)
   2. Change emojis to words (Python emoji package) for example thumbs up emoji becomes 'thumbs up' or 'cool'
   3. Remove all rows that have an emoji in the review body (how many are there though?)
5. Some review bodies are quite short (e.g. ‘Ok’, ‘Yes’, ‘meh’, ‘No’). Should we keep these? Will we know if they are useful until we run a model?
6. Sometimes the data is in the wrong columns. Found a spot where we’re expecting ‘Books’ for the product category but instead got what looks like a review.
   1. Drop these rows
7. For comparing reviews to top 100 list we will have to match book titles. Will there be issues with this? Will have to try and see.
8. Going to need to write a plain python program that will load the data in batches, organize by title, then remove rows for specific books that have less than a certain number of reviews.
   1. Idea is that books with only 5 or less reviews will not be in the top 100 and hard to judge that book based on a small number of reviews.

Book Data Modeling Issues

4/17/2021

1. There are ‘helpful votes’ and ‘total votes’. Not sure if we should use this or not
   1. Helpful votes is the option where you can indicate if the review was helpful to you or not.
   2. Not sure what total votes is

Book Data Exploratory Statistics (EDA)

4/17/2021

1. What is happening over time? Assume the number of reviews increasing. Make a graph to show this
2. What is the average length of review? For the whole set? Over time? Make a graph
3. Average star rating for the whole dataset? Over time? Make a graph.
4. Average number of helpful votes? Over time? Make a graph
5. Understand and visualize the distribution of number of reviews per book
   1. Most books will have a small number of reviews
   2. A few books will have a very large number of reviews
   3. How many books have more than 100 reviews?

5/3/2021

Some EDA

* Original file from Amazon (amazon\_reviews\_us\_Books\_v1\_00.tsv) contained 10,319,090 reviews. It was 6.7GB.
* After the first pass the data became 7,450,412 reviews so we threw away about 3 million reviews for having various data issues. By far the most common issue was we threw away reviews that were not verified purchases. It became 2.66GB.
* We did our second pass twice: once with every book and once with only books that contain 30 reviews or more. Our second pass created a default dictionary that holds every book and the number of reviews associated with that book.
  + With all books there are 1.65 million books
  + When filtered for 30 reviews or more there are 34,194 books in total.

Some SDA: We loaded the two default dictionaries into pandas.

* Looking at all books
  + We find the average number of reviews is 4.5
  + The median number of reviews is 1. Most books have 1 review
  + The 75th percentile jumps to 3 reviews.
* Looking at books with at least 30 reviews
  + We find the average number of reviews is 78
  + The median number of reviews is 48.
  + The 75th percentile jumps to 76 reviews.