**Text Analytics Assignment 2**

**Now due: 26th November by 23:59 hours**

**You can choose between two topics for this assignment**

**Choice 1: Building a Crowdsourced Recommendation System**

**High level description:** The objective of this group assignment is to create the building blocks of a crowdsourced recommendation system. This recommendation system should accept user inputs about desired attributes of a product and come up with 3 recommendations.

For this assignment, write a scraper to scrape reviews of craft beer from beeradvocate.com, mainly because the site is easy to scrape compared to many others. However, if you prefer, you may want to scrape another site for a different product – say, musiciansfriend.com for reviews of musical instruments like acoustic drums or electric guitars. The challenge is that many sites (including musiciansfriend.com provide reviews as the number of stars, which are pictures, and harder to process to get the numeric rating – by contrast, on beeradvocate.com, the rating is given as text, which is easier to access).

**Task A.** Extract about 6k reviews of beers from Beeradvocate.com. I would suggest choosing the following link, which shows the top 250 beers sorted by ratings:

<https://www.beeradvocate.com/beer/top-rated/>

The nice feature of the above link is that it is a single-page listing of 250 top-rated beers (avoids the pagination feature, which you need in cases where listings go on for many pages).

Scrape 25 reviews and well as the ratings provided by users for each beer. The way beeradvocate.com organizes reviews is that it provides 25 reviews per page, which means you can avoid the pagination part for the reviews as well. The CSV output file should have 3 columns: product\_name, product\_review, and user\_rating.

**Task B.** Assume that a customer, who will be using this recommendation system, has specified 3 attributes in a beer. One website describes multiple attributes of beer:

<https://www.dummies.com/food-drink/drinks/beer/beer-for-dummies-cheat-sheet/>

* **Aggressive (**Boldly assertive aroma and/or taste)
* **Balanced:** Malt and hops in similar proportions; equal representation of malt sweetness and hop bitterness in the flavor — especially at the finish
* **Complex:** Multidimensional; many flavors and sensations on the palate
* **Crisp:** Highly carbonated; effervescent
* **Fruity:** Flavors reminiscent of various fruits **or Hoppy:** Herbal, earthy, spicy, or citric aromas and flavors of hops o**r Malty:** Grainy, caramel-like; can be sweet or dry
* **Robust:** Rich and full-bodied

Choose three attributes from the above list. If you choose another product (like electric guitars), choose a different set – e.g., action, neck, pickup (for guitars). It is a good idea to perform a word frequency analysis of the data you have collected to make sure that the above attributes are actually mentioned.

**Task C.** Perform a **similarity** analysis with the 3-attribute set and the reviews. See guidelines below on how to do this. From the output file, choose 300 reviews that have the highest similarity scores with the attribute set you have chosen.

**Task D.** Perform sentiment analysis on these 300 reviews and sort them (high to low) by the sentiment scores.

**Task E.** Based on tasks C and D, **recommend 3 beers** (or guitars) to the customer. Note that in task D, multiple reviews may refer to the same beer. Show the average similarity and sentiment scores for each of the 3 recommended products.

**Task F.** How would your recommendations differ if you ignored the similarity and sentiment scores and simply chose the 3 highest rated beers from your entire dataset? To do this, you need to calculate the average rating for each beer mentioned (remember each beer in your data has multiple reviews, and you need to get the average rating for each beer). Would these three beers meet the requirements of the user looking for recommendations? Why or why not? Justify your answer. Use the user\_rating data, similarity and sentiment scores to answer this question.

Your submission (python notebook) should include the following:

1. Names of all team members **inside** the document (only one submission per team)
2. All scripts
3. The average sentiment and similarity scores for the three products you recommended in task E.
4. Your analyses and answer to task F. Make sure you show the ratings, similarity scores and sentiments for the products you recommend in tasks E and F.

**Guidelines**

1. For similarity analysis, use spacy (instead of simple cosine similarity). The script should accept as input a file with the product attributes, and calculate similarity scores (between 0 and 1) between these attributes and each review. That is, the output file should have 3 columns – product\_name, product\_review and similarity\_score.
2. The same beer may appear on multiple rows (in the 300 rows with highest similarity scores) since there are multiple reviews for each beer. Reduce the number of rows by averaging the user\_rating of each beer (can do it in Excel). After this operation, each beer will be present only in one row. The user\_rating column will now have average ratings for each beer.
3. For task F, you do the same averaging, but with ALL rows of data (that is, as many rows as there are reviews).

**Alternative topic for Assignment 2**

**Is a Picture Worth a Thousand Words?**

On Instagram, choose the National Geographic (natgeo) page (do not use hashtags). Write a scraper or use the Web Scraper to extract (i) image URLs (do not extract video URLs, it may end up costing you a lot of money to run analytics on video), (ii) post caption (the text description of a post), (iii) # likes and (iv) # comments. You don’t need actual comments for this assignment. Scrape around 400-500 image posts.

Using the image URLs, obtain **image labels** from Google Vision cloud (you will have to create an account with Google to get your credentials as a json file, though the first $300 are free, which should be more than plenty for this assignment). You will need to write a script to access the Google Vision API. You can also use IBM Watson Vision Analytics (the basic account is free) as an alternative (if you are not a Google fan). Do not use Microsoft image analytics, which I have demonstrated in class to be inferior to Google Vision. Read about Google Vision here: For Google Vision API, look here: <https://cloud.google.com/vision/docs/quickstart>

Task A. Create a metric for **engagement** by using a weighted sum of # likes and # comments. However, first normalize # likes and # comments such that they both have values between 0 and 1. You can scale the # likes by dividing by the maximum # likes (for a post) in your data and do the same for # comments, so that # likes and comments will be in the range [0,1]. Now create an engagement score = .4\*# likes (normalized) + .6\*# comments (normalized). Define High (1) and Low (0) engagement based on whether the engagement score is above or below the median value.

Task B. Run a logistic regression with Engagement (binary) as the dependent variable, and the image labels as independent variables. What is the accuracy (show the confusion matrix)?

What accuracy do you get by using the post caption words as the independent variables instead of image labels? Finally, what accuracy do you get by combining the image labels and post captions and using them as independent variables? What can you conclude from your analysis?

Note: Doing a word frequency analysis and word replacement on the image labels as well as captions will increase the accuracy of prediction. Needless to say, TF-IDF scores should be used.

Task C. Perform topic modeling (LDA) on the image labels. Choose an appropriate number of topics. You may want to start with 5, but adjust the number up or down depending on the word distributions you get. LDA should produce two outputs: (i) A file showing which words load on which topics, and (ii) a file showing topic weights for each image.

Now take the quartiles with highest and lowest engagement scores. What are the differences in the average topic weights of pictures across the two quartiles (e.g., greater proportion of some topics in highest engagement quartile)? Show the main results in a table.

Task D. What advice would you give National Geographic if it wants to increase engagement on its Instagram page based on your findings in Tasks B and C?

**Deliverables:** Create a Python notebook with all code. Also submit all data files. Write the names of all team members **inside** the notebook.