MSSP 607: Homework 2

Part 1

1. For each star count, 1-5, what percentage of restaurants receive that score on average? What is the average word count of reviews that give each star count?

Star	Percentage (%)			Average review word count (words)
	Business	Review	average	
1.0	0.5	10.4	5.45	125
1.5	1.5		0.75	
2.0	4.5	9.4	6.95	141
2.5	8.5		4.25	
3.0	16.8	13.9	15.45	138
3.5	24.0		12.0	
4.0	24.2	28.2	26.2	121
4.5	16.5		8.25	
5.0	3.5	38.1	20.8	94

Note: Results relevant to star count 0.0 and 0.5 are omitted in Table[1], due to the fact that no record shows anyone has ever graded any restaurant 0 or 0.5 star (because practically it is impossible to star a restaurant zero or half on yelp).

2. Each restaurant in our dataset is labeled with a number of categories, like "Sports Bars", "Coffee & Tea", "Mexican", or "Delis". How many labels are there? How many restaurants of each category label exist, and what is the mean score for restaurants with that label?

There are 341 labels altogether. The specifics about label count and average score are as in the table label.xlsx. The top 20 most frequently used labels, their count and the mean score of the restaurants with that label are as in Table[2] (top 5 highest average score colored). It is interesting that, shown in the result, bars, coffee & tea, event planning and services, and bakeries tend to receive high score than other kinds of restaurants.

Categories	**Count**	**Ave Score**
Restaurants	3512	3.5
Food	1595	4.31
Pizza	690	3.43
Nightlife	686	3.53
Bars	670	4.42

American (Traditional)	648	3.42
Sandwiches	572	3.53
American (New)	480	3.5
Italian	412	3.48
Coffee & Tea	382	3.64
Breakfast & Brunch	333	3.55
Burgers	329	2.98
Fast Food	328	2.83
Chinese	227	3.44
Salad	213	3.49
Event Planning & Services	201	3.82
Chicken Wings	191	3.14
Grocery	190	3.61
Bakeries	180	3.86

3. What are some common features of restaurants that receive higher-scoring reviews? This can be extracted either from attributes of the restaurant itself or from the review texts.

Firstly, based on attributes of the restaurants themselves, higher-scoring restaurants have some distinct general features as shown in Table[3] (top 5 colored; more information in attributes.xlsx). In this case, for each kind of attribute, if the record shows a dominating customer preferred option, with highest recorded percentage of feature-recognition among all high-scoring restaurants, this option of attribute is considered as a worth-noting feature. (In the cases where most of the restaurants' reviews are missing an attribute, this attribute and its options are not taken into consideration as distinct features). Specifically, high-scoring restaurants are featured with being friendly to groups and kids, offering take-out and catering services, having bike parking spaces and casual attire style, accepting credit cards, etc. Whether or not the restaurant offers delivery, reservation, or outdoor seating does not seem matter to its score.

Attributes	**Option**	**Percentage (%)**
Restaurants Take Out	True	90.33
Business Accepts Credit Cards	True	90.17
Restaurants Good For Groups	True	55.96
Good For Kids	True	55.96
Caters	True	35.06

Bike Parking	True	59.91
Restaurants Delivery	False	56.55
Restaurants Reservations	False	48.1
Outdoor Seating	False	45.43
Restaurants Price Range2	2	47.46
Restaurants Attire	casual	39.6
Business Parking	Street only	34.15

Also, Table^{4} shows some more detailed features about restaurants' business parking, ambience, meal type, music type and best nights. For each kind of features, the most popular 5 options among high-scoring restaurants were recorded.

Attributes	**1st**	**2nd**	**3rd**	**4th**	**5th**
Business Parking	street	lot	garage	valet	validated
Ambience	casual	trendy	hipster	divey	intimate
Good For Meal	lunch	dinner	brunch	breakfast	dessert
Music	Background music	jukebox	live	karaoke	dj
Best Nights	Saturday	Friday	Thursday	Sunday	Tuesday

Another thing to note for high-score restaurants is the high-frequency words that appeared in customer's review texts. Nouns, adjectives and adverbs of concrete meaning (other than functional words such as pronouns, articles, prepositions, conjunctions and so on) may be able to reveal what customers value the most and how customers feel at those high-score restaurants. Among 500 most frequently appeared words (more specific in review_word.xlsx) in reviews, words in Table^{5} are of some reference value. For example, given how frequently nouns such as service, staff, atmosphere, experience, taste, quality and flavor are mentioned, one can conclude that most of the high-score restaurants are at least impressive from one of those aspects.

Nouns	Adjectives/Adverbs	Verbs
Food/chicken/pizza/bar Cheese/ sauce/beer/fries/ Pork/chocolate/fish/steak/wine Drinks/ice cream/bread offee/salad/sandwich **place/service/staff**/friends/ husband/wife/everyone/everything Pittsburgh/**location/parking** lunch/dinner/breakfast/brunch **atmosphere/experience/taste/** **quality/flavor** Thai/Italian/grilled	Good/great/best/nice/super Amazing/perfect/excellent/ Awesome/wonderful **Pretty/delicious/fresh/friendly** **Tasty/cool** Favorite/small/**busy** Hot/sweet/spicy Happy/glad very	Love(ed)/ enjoy(ed) Wait/try/feel recommend

And from how often adjectives such as pretty, delicious, fresh and friendly are used to describe a restaurant, one can read that pretty indoor decoration, deliciousness, freshness and staff's friendliness are potentially what make high score restaurants stand out.

Part 2

1. Individual Page Scraping

1) How did you determine which featured articles were biographies?

I determine which featured articles were biographies by checking if they are under an <h3> headline with key word "biographies" or "Biographies". To be specific, for each line after an <h3> headline labeled with "biographies", I treat it as describing a biography until I run into a new line start with another headline <h2>/<h3>/<h4> label which marks the end of the previous biography section.

2) What percentage of featured articles are biographies?

Number of biographies: 1399

Number of all articles: 5676

Bio percentage: 24.65%

2. Scraping a dataset

1) What percentage of first paragraphs were you able to scrape?

Using my method, I successfully scraped 1289 first paragraphs of entries of featured biographies, which is about 92.14% of all featured biographies.

2) What are the characteristics of the pages you failed to scrape?

As I scrape the first paragraphs (the bio-info paragraph) by assuming that the first non-empty line with more than 100 characters and not starting with "For", "In", "This" or a parenthesis in every text file of a featured biography is the target paragraph, the scraping fails when the API fails somehow, the first paragraph does include the biography's hero's family name and when there are lines describing an extra picture or table on the website such as in the case of Michael Jordan and Tom Pryce.

3. Extracting information from messy content

1) What are the drawbacks of your approach, and what types of content are excluded or missed because of the choices you made?

The problem with my approach is that I regexed the pronouns only with two spaces on each side and did not deal with scenarios where the pronoun follows or is followed by punctuation marks such as in the sentence “I hate him.” or “Do you like her?”. So I probably have missed some pronouns in my calculation.

2) What percentage of biographies uses he/his pronouns, she/her, or they/them pronouns?

	Count	Percentage (%)
Male Pronouns	1169	83.56%
Female Pronouns	204	14.58%
Plural/Non-binary Pronouns	25	1.79%

3) What percentage of pages did your code fail to parse, or have unclear gender? Why?

As the sum of the counts matches the total number of the biographies (except for one failed API request), I think I parsed 100% of the pages.

4. Additional analysis

Define and write a function that will extract one additional quantifiable feature of Wikipedia biographies based on the raw data you scraped. What question did you ask, and why is it interesting? Did you draw any new conclusions based on the feature you found and its distribution in your data? Share any statistics that support your analysis, and include those statistics in your final report.

Question: In which centuries were the heroes of the biographies born? What is the percentage of births of them per century?

This question is interesting as the answer to which reveals the temporal composition of the biographies and the temporal feature of the dataset to see whether the dataset is more contemporary or historical and whether the dataset include more information about ancient, post-classical or modern history. Based on my calculation, figures from the modern era take 85.6% of all the 1139 bio entries from which I was able to extract the hero's birth year, and people born in 20th and 19th century are most highly included in Wikipedia featured biographies. Admittedly, the calculation may not be 100% precise as it is likely that some errors caused by messy formatting are not caught by the program. But the centurial distribution is still of reference value to grasp general features about the dataset.

	Century (AC)	Count	Percentage		Century(BC)	Count	Percentage
modern	21	19	1.67%	ancient	-1	1	0.09%
	20	**404**	**35.47%**		-2	7	0.61%
	19	**421**	**36.96%**		-3	8	0.70%
	18	83	7.29%		-4	4	0.35%
	17	26	2.28%		-5	7	0.61%
	16	22	1.93%		-8	2	0.18%
post-classical	15	8	0.70%		-11	2	0.18%
	14	10	0.88%		-12	2	0.18%
	13	9	0.79%		-13	1	0.09%
	12	18	1.58%		-16	1	0.09%
	11	9	0.79%		-18	1	0.09%
	10	6	0.53%		-19	5	0.44%
	9	8	0.70%		-20	7	0.61%
	8	10	0.88%		-21	1	0.09%
	7	12	1.05%		-24	3	0.26%
	6	5	0.44%		Sum	1139	
	5	4	0.35%				
	4	2	0.18%				
	3	3	0.26%				
	2	5	0.44%				
	1	3	0.26%				

5. Preparing a dataset for sharing

[Technical Documentation] The "p2_5.csv" file contains four columns: index, name, year of birth, and most commonly used pronoun. "Name" column stores the title of each featured biography in Wikipedia; "Year_of_birth" stores each hero's birth year; "Gender_Pronoun" column stores the most frequently used category of gender-implying pronouns in each person's page. Due to differences in each article's page formatting, I failed to extract the protagonist's birth year with precision in some cases, the result of which could be wrong. Also, due to the problem of punctuation marks being interpreted through html and text, a few titles of biographies are not presented in casual readable format. For example, “Gerard K. O&#39;Neill” is actually referring to Gerard K. O’Neill. For access and further use, one can using following codes to read and use “p2_5.csv” file.

```python
df = pd.read_csv('p2_5.csv')
# to learn about the shape of the file
rows, columns = df.shape
print(f"Rows: {rows}\n"
      f"Columns: {columns}")

# to search relevant detail on a particular person
du_fu_info = df.loc[df.Name == "Du Fu"]
print(du_fu_info)

# to search info using row and column index
bronwyn_bancroft_year = df.iloc[0, 2]
print(f"Bronwyn Bancroft's year of birth: {bronwyn_bancroft_year}")

# ro search for male-pronoun users
male_pronoun = df.loc[df.Most_Common_Pronoun == "Male Pronouns"]
print(f"Number of male pronou users: {len(male_pronoun)}")
```

Also, to check if the download was successful, the result to codes above should be as follows.

```
Rows: 1398
Columns: 4

Unnamed: 0    Name      Year of Birth    Most Common Pronoun
292           Du Fu             712.0              Male Pronouns

Bronwyn Bancroft's year of birth: 1958.0

Number of male pronoun users: 1169
```