CSCI 182 Homework 3
Due Date: Feb 11th, 2022

All assignments MUST have your name, student ID, course name/number at the beginning of your documents.
Your homework MUST be submitted via Camino with the file format and name convention as follows:
HW#_Name.zip (both the python file and the report)
If you have any questions, please don't hesitate to contact me :)

Q1) You are provided with the following datasets:

- olist_order_items_dataset.csv: order_id, order_item_id, product_id, price
- olist_products_dataset.csv: product_id, product_category_name, product_weight_g, product_length_cm, product_height_cm, product_width_cm
- product_category_name_translation.csv: product_category_name, product_category_name_english

Your task is to use these datasets to find association rules among the products.

Note: The transaction is of the form:

Order ID	Order_Item_ID	Product_ID	Price
b8bfa12431142333 a0c84802f9529d87	1	765a8070ece0f138 3d0f5faf913dfb9b	81
b8bfa12431142333 a0c84802f9529d87	2	a41e356c76fab663 34f36de622ecbd3a	99.3

Here, the same order_ID corresponds to the same transaction.
So the transaction would be 765a8070ece0f1383d0f5faf913dfb9b, a41e356c76fab66334f36de622ecbd3a (products bought under the same Order ID makes up one transaction)
You will need to match the product_ID to their name in spanish and match that to the name of the product in English.

The final rules should have the product names in English. You may choose three values for the minsup and minconf as you see fit for the dataset. For the various values

you select, accumulate **5 rules for each case with the highest lift value** in a table that looks like the following for each value that you select:

Example 1:
minsup = x%, minconf = y%

Rule	support	confidence	lift
:			
:			

Example 2:
minsup = a%, minconf = b%

Rule	support	confidence	lift
:			
:			

Q2) You will use **nltk** to explore the Herman Melville novel Moby Dick. Write the python code for answering the following questions:

1. Import the libraries required and Set up Data. Use the link "https://www.gutenberg.org/files/2701/old/moby10b.txt" to access the ebook.
2. Find how many tokens (words and punctuation symbols) are in the text. A token is a linguistic unit such as a word, punctuation mark, or alpha-numeric strings.
3. Find how many of the tokens found in 1.2) are **unique.**
4. Find how many tokens are unique after removing stopwords.
5. What is the lexical diversity of the given text input? (i.e. ratio of unique tokens to the total number of tokens)
6. What percentage of tokens is 'whale' or 'Whale'?
7. What are the 20 most frequently occurring (unique) tokens in the text? What is their frequency?
8. What tokens have a length of greater than 5 and frequency of more than 150?
9. Find the longest word in the text and that word's length.
10. What unique words (not punctuation) have a frequency of more than 2000? What is their frequency?
11. What is the average number of tokens per sentence?