# Computational Thinking MSCI:3020

**Homework Assignment #3**

**Due:** Friday, November 15th, 2019 @ 11:59 pm

**Problem: Webscraping Yelp! Data (5 points)**

Part A: Webscraping (3 points)

* For this assignment, you are going to create a function that scrapes the first 80 Yelp! reviews from the 2 most reviewed restaurants in Iowa City – Pullman Diner and La Regia Taqueria.
* To do this, use the BeautifulSoup and requests python libraries similar to examples from Lecture 7 and Lab 9.
* The links to the restaurant Yelp! pages are:
  + <https://www.yelp.com/biz/pullman-bar-and-diner-iowa-city>
  + <https://www.yelp.com/biz/la-regia-taqueria-iowa-city>
* There are 20 reviews per page, so when you are scraping the pages, use a loop that will change the name of the webpage on everything but the first page of reviews, using a structure as shown below.
  + <https://www.yelp.com/biz/pullman-bar-and-diner-iowa-city?start=20>
  + <https://www.yelp.com/biz/pullman-bar-and-diner-iowa-city?start=40>

Hint – Use f-strings and if/else to get the desired webpage each time through the loop.

* Use time.sleep() to pause your code for 2 seconds between each scrape of the Yelp! website.
* Begin with a blank string variable all\_reviews = '' for example.
* Now, on each webpage use requests and BeautifulSoup to obtain and parse the HTML code.
* There are 20 reviews per page, so you want to use the find\_all function from BeautifulSoup.
* Hint – you are looking for paragraph elements (‘p’) and the class is

'lemon--p\_\_373c0\_\_3Qnnj text\_\_373c0\_\_2pB8f comment\_\_373c0\_\_3EKjH text-color--normal\_\_373c0\_\_K\_MKN text-align--left\_\_373c0\_\_2pnx\_'

* Finally, loop through each of the paragraphs and add the text from each paragraph to your existing all\_reviews string. Make sure to use .text when getting the text from each ‘p’ element.
* The function should return your all\_reviews string.

Part B: Text Analytics (2 points)

* Use the existing functions from Lecture 6, remove\_punc, doc\_to\_frequencies and top\_k to analyze the text from both Pullman Diner and La Regia.
* You can replace the stop\_words in the doc\_to\_frequencies function with a more comprehensive list as shown below:

stop\_words = ['iowa','city','food','good','great','a', 'about', 'above', 'after', 'again', 'against', 'all', 'am', 'an', 'and', 'any', 'are', "aren't", 'as', 'at', 'be', 'because', 'been', 'before', 'being', 'below', 'between', 'both', 'but', 'by', "can't", 'cannot', 'could', "couldn't", 'did', "didn't", 'do', 'does', "doesn't", 'doing', "don't", 'down', 'during', 'each', 'few', 'for', 'from', 'further', 'had', "hadn't", 'has', "hasn't", 'have', "haven't", 'having', 'he', "he'd", "he'll", "he's", 'her', 'here', "here's", 'hers', 'herself', 'him', 'himself', 'his', 'how', "how's", 'i', "i'd", "i'll", "i'm", "i've", 'if', 'in', 'into', 'is', "isn't", 'it', "it's", 'its', 'itself', "let's", 'me', 'more', 'most', "mustn't", 'my', 'myself', 'no', 'nor', 'not', 'of', 'off', 'on', 'once', 'only', 'or', 'other', 'ought', 'our', 'ours', 'ourselves', 'out', 'over', 'own', 'same', "shan't", 'she', "she'd", "she'll", "she's", 'should', "shouldn't", 'so', 'some', 'such', 'than', 'that', "that's", 'the', 'their', 'theirs', 'them', 'themselves', 'then', 'there', "there's", 'these', 'they', "they'd", "they'll", "they're", "they've", 'this', 'those', 'through', 'to', 'too', 'under', 'until', 'up', 'very', 'was', "wasn't", 'we', "we'd", "we'll", "we're", "we've", 'were', "weren't", 'what', "what's", 'when', "when's", 'where', "where's", 'which', 'while', 'who', "who's", 'whom', 'why', "why's", 'with', "won't", 'would', "wouldn't", 'you', "you'd", "you'll", "you're", "you've", 'your', 'yours', 'yourself', 'yourselves']

* Use the top\_k function to display the top 10 words from each restaurant.
* Based on this list of Top 10 words, what is one item you would suggest ordering from each of the two restaurants? Put this answer as a comment in your code.