Homework 5 - Extra Credit - 75 points

Due: May 12. 11:59PM

Note: I will only grade this homework if you do not have an A.

DASK

1. **Rewrite/Refactor Homework 3, Question 7, In Dask - 25 points**

   Datasets: Restaurants_in_Durham_County_NC.csv
   durham-nc-foreclosure-2006-2016.json

   Solve:

   For each restaurant (‘Restaurants_in_Durham_County_NC.csv) with “status”=“ACTIVE” and “"rpt_area_desc"="Food Service", show the number of foreclosures (‘durham-nc-foreclosure-2006-2016’) within a radius of 1 mile of the restaurant’s coordinates.

   Note: Use any assumption for the shape of Earth… Or you can use the Haversine distance. https://pypi.org/project/haversine/

2. **Rewrite HW1 Q2, Language Models, in Dask – 50 points**

   Input: hw1dir1.zip (provided in class website)
   Solve: conditional probability of w2 given w1, P(w2|w1)

From HW1:

A language models LM describes the probability of words appearing in a sentence or corpus.

A unigram LM models the probability of a single word appearing in the corpus, but an n-gram LM models the probability of the n_th word appearing given the words n-1, n-2, … .

As an example, given the following corpus: “The Cat in the Hat is the best cat in the hat”, a unigram LM language model would be: (using fractions for clarity). Let P(w) be the probability of w:

P(the) = 4/12
P(cat) = 2/12
P(in) = 2/12
P(hat) = 2/12
P(is) = 1/12
P(best) = 1/12

For unigrams, the probability of ‘cat’ appearing anywhere in the corpus is 2/12 using maximum likelihood estimation MLE (a.k.a. word count) - note this is a very simplistic model – the closed universe model.

A bigram (n-gram, n=2) LM:

P(the cat) = 1/8
P(cat in) = 2/8
P(in the) = 2/8
P(the hat) = 2/8
P(hat is) = 1/8
P(is the) = 1/8
P(the best) = 1/8
P(best cat) = 1/8

P(B given A) = P(A and B) / P(A)

Let’s approximate this using the closed-corpus assumption (no unseen words exist, so no smoothing for those statisticians in class):

P(cat|the) = P(the cat) / P(the) = (1/8) / (4/12) = 0.375

- punctuation does NOT count; so the words is ‘(1991)’ and ‘1991’are the same.
  You must parse your input: replace all characters not in this set: [a-z, A-Z, 0-9] with spaces.
- all text should be normalized to lowercase
- Ignore lines with less than 3 words.
- Input should be lines of text (separated by new line and/or carriage return)
- Write your own code in your language of choice, but your code **MUST BE PYTHON/DASK.**