CSCI 456-01, Large Language Models, Spring 2026
Problem set 1

Exercise 0.1. *This exercise is intended to give you some insight into how perplexity reflects the extent to how well a probability distribution models a set of data. You will compare the perplexities of two distributions when used to describe the frequency of letters in American English.*

Let $P = \{p_1, p_2, \ldots, p_n\}$ be the probabilities of the random events in a (discrete) probability space. The Shannon information of event i

$$I(p_i) = \log \frac{1}{p_i},$$

the Shannon entropy is

$$H(P) = \sum p_i I(p_i) = -\sum p_i \log_2 p_i,$$

and the perplexity is

$$\text{perplexity}(P) = 2^{H(P)}.$$

Let $Q = \{q_1, q_2, \ldots, q_n\}$ be a second probability distribution over the same probability space. The cross-entropy of P and Q is defined to be

$$H(P, Q) = -\sum p_i \log_2 q_i.$$

This is the expected value with respect to the probability P of the information predicted by Q. Note that in general, $H(P, Q) \neq H(Q, P)$.

If P is the true distribution and Q is not, then one can show that

$$H(P) = H(P, P) < H(P, Q).$$

The file `american.txt` *contains a list of over 300,000 words from American English. To simplify matters the words have all been converted to lower case and any apostrophes removed.*

The two probability distributions of letters you will use are an empirical one (though not derived from this dataset),

```
# Realistic letter probabilities, from data at
# https://www3.nd.edu/~busiforc/handouts/cryptography/letterfrequencies.html
actual = {
    'e': 0.111607,  'm': 0.030129,
    'a': 0.084966,  'h': 0.030034,
    'r': 0.075809,  'g': 0.024705,
    'i': 0.075448,  'b': 0.020720,
    'o': 0.071635,  'f': 0.018121,
    't': 0.069509,  'y': 0.017779,
    'n': 0.066544,  'w': 0.012899,
    's': 0.057351,  'k': 0.011016,
    'l': 0.054893,  'v': 0.010074,
    'c': 0.045388,  'x': 0.002902,
    'u': 0.036308,  'z': 0.002722,
    'd': 0.033844,  'j': 0.001965,
    'p': 0.031671,  'q': 0.001962
}
```

and the uniform distribution,

```
uniform = {c : 1/26 for c in actual}
```

We will treat the empirical distribution P as the true(r) one and the uniform distribution Q as an alternative (which is clearly wrong).

What you are to do: *Use the dataset* `american.txt`.

1. *For each letter, compute its observed probability p_i (relative frequency) in the dataset. That is, if character i appears c_i times in the dataset, then*

$$p_i = \frac{c_i}{n},$$

   where n is the total number of letters in the dataset.

2. *Let p_i be the probabilities from the empirical distribution. Compute the entropy $H(P)$ and perplexity* $\text{perplexity}(P)$ *using P.*

3. *Let q_i be the probabilities from the uniform distribution. Compute the cross-entropy $H(P, Q)$ and the perplexity*

$$\text{perplexity}(P, Q) = 2^{H(P,Q)}.$$

4. *How do the entropy $H(P)$ and cross-entropy $H(P, Q)$ compare? How is this reflected in the perplexities?*

What you are to turn in: *Upload to Gradescope a Python file named* `perplexity.py` *that contains a function* `perplexity()`. *This function*

- *takes as its input a string that is the name of a file structured like* `american.txt` *(one wor per line),*
- *opens the specified file and computes $H(P), \text{perplexity}(P), H(P, Q), \text{perplexity}(P, Q)$;*
- *returns $[H(P), \text{perplexity}(P), H(P, Q), \text{perplexity}(P, Q)]$ in the order shown.*

You may find the `Counter` *class from the Python* `collections` *module to be helpful in computing relative frequencies.*