4.3 Statistical language modeling

In this problem, you will explore some simple statistical models of English text. Download and examine the data files on Piazza for this assignment. (Start with the `readme.txt` file.) These files contain unigram and bigram counts for 500 frequently occurring tokens in English text. These tokens include actual words as well as punctuation symbols and other textual markers. In addition, an "unknown" token is used to represent all words that occur outside this basic vocabulary. For this problem, as usual, you may program in the language of your choice.

(a) Compute the maximum likelihood estimate of the unigram distribution $P_u(w)$ over words w. Print out a table of all the tokens (i.e., words) that start with the letter "A", along with their numerical unigram *probabilities* (not counts). (You do not need to print out the unigram probabilities for all 500 tokens.)

(b) Compute the maximum likelihood estimate of the bigram distribution $P_b(w'|w)$. Print out a table of the five most likely words to follow the word "THE", along with their numerical bigram probabilities.

(c) Consider the sentence **"Last week the stock market fell by one hundred points."** Ignoring punctuation, compute and compare the log-likelihoods of this sentence under the unigram and bigram models:

$$\mathcal{L}_u = \log\Big[P_u(\textbf{last})\, P_u(\textbf{week})\, P_u(\textbf{the}) \ \ldots\ P_u(\textbf{one})\, P_u(\textbf{hundred})\, P_u(\textbf{points})\Big]$$

$$\mathcal{L}_b = \log\Big[P_b(\textbf{last}|\langle\textbf{s}\rangle)\, P_b(\textbf{week}|\textbf{last})\, P_b(\textbf{the}|\textbf{week}) \ \ldots\ P_b(\textbf{hundred}|\textbf{one})\, P_b(\textbf{points}|\textbf{hundred})\Big]$$

In the equation for the bigram log-likelihood, the token $\langle s\rangle$ is used to mark the beginning of a sentence. Which model yields the highest log-likelihood?

(d) Consider the sentence **"The nineteen officials sold fire insurance."** Ignoring punctuation, compute and compare the log-likelihoods of this sentence under the unigram and bigram models:

$$\mathcal{L}_u = \log\Big[P_u(\textbf{the})\, P_u(\textbf{nineteen})\, P_u(\textbf{officials}) \ \ldots\ P_u(\textbf{sold})\, P_u(\textbf{fire})\, P_u(\textbf{insurance})\Big]$$

$$\mathcal{L}_b = \log\Big[P_b(\textbf{the}|\langle\textbf{s}\rangle)\, P_b(\textbf{nineteen}|\textbf{the})\, P_b(\textbf{officials}|\textbf{nineteen}) \ \ldots\ P_b(\textbf{fire}|\textbf{sold})\, P_b(\textbf{insurance}|\textbf{fire})\Big]$$

Which pairs of adjacent words in this sentence are not observed in the training corpus? What effect does this have on the log-likelihood from the bigram model?

(e) Consider the so-called *mixture* model that predicts words from a weighted interpolation of the unigram and bigram models:

$$P_m(w'|w) = (1-\lambda)P_u(w') + \lambda P_b(w'|w),$$

where $\lambda \in [0,1]$ determines how much weight is attached to each prediction. Under this mixture model, the log-likelihood of the sentence from part (d) is given by:

$$\mathcal{L}_m = \log\Big[P_m(\textbf{the}|\langle\textbf{s}\rangle)\, P_m(\textbf{nineteen}|\textbf{the})\, P_m(\textbf{officials}|\textbf{nineteen}) \ \ldots\ P_m(\textbf{fire}|\textbf{sold})\, P_m(\textbf{insurance}|\textbf{fire})\Big].$$

Compute and plot the value of this log-likelihood $\mathcal{L}_m$ as a function of the parameter $\lambda \in [0,1]$. From your results, deduce the optimal value of λ to two significant digits.