Assignment 4

Spelling & Grammar Correction

Due Friday, October 8

1. Pretend we have a bigram array, as in the given table, where the first letter of the bigram is given in the vertical letters (i.e., down the side), and the second letter is given in the horizontal ones (i.e., across the top)

<table>
<thead>
<tr>
<th></th>
<th>i</th>
<th>j</th>
<th>k</th>
</tr>
</thead>
<tbody>
<tr>
<td>i</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>j</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
<tr>
<td>k</td>
<td>?</td>
<td>?</td>
<td>?</td>
</tr>
</tbody>
</table>

Make this into 3 positional bigram arrays, namely ones which capture the positions:
- start of word
- middle of word
- end of word

For the “middle of word” one, provide a word which justifies each 1 you put in the chart.

2. Do question #3 in chapter 2 of the draft textbook (p. 85).
   - **Bonus:** Do question #4 in chapter 2 of the draft textbook (p. 85).

3. Consider the misspelling forg, and assume our edit distance calculations have insertions, deletions, substitutions, and transpositions. Describe how probabilities are used to rank frog vs. forge. Do you have an intuition as to which should be higher, and why?

4. Pick a search engine of your choice, as long as the search engine provides spelling suggestions. Search with the keywords ghous and ghosts, and do not use quotes.
   (a) Try as many sensible misspellings as you can think of (and vary the order of the words, too). Which ones does the search engine catch and which doesn’t it catch?
   (b) Given what we talked about in class, can you explain why certain variations were flagged as misspellings and others were not? If the search engine always correctly caught your misspellings, it will help to try some even more varied misspellings, in order to find the “boundary” where the correct suggestion is no longer made.

5. Do question #6 in chapter 2 of the draft textbook (p. 86) ... but, instead of evaluating spelling errors, evaluate grammar errors.

6. Write a Python program that takes two strings and says what their shared “prefix” is, if any. For example, bring and behold have a shared prefix of b
• Let’s call the two strings $a$ and $b$ and assume that the first string is always shorter—i.e., $\text{len}(a) \leq \text{len}(b)$

• I (will) have some code available on the website to get you started.