

# Assignment 5

## Spelling & Grammar Correction

Due Monday, November 10, 2008

1. Pretend we have a nonpositional 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)

	i	j	k
i	?	?	?
j	?	?	?
k	?	?	?

- (a) Fill in the chart and provide justification as to why each cell is either 1 or 0, i.e.,:
    - Provide a word for each 1.
    - Provide a brief description of why English doesn't allow such a word.
  - (b) For any 0's, find a website containing such a misspelling.
  - (c) Change the nonpositional bigram array into 3 positional arrays, namely ones which capture the positions:
    - start of word
    - middle of word
    - end of wordNo need to provide justification for these.
  - (d) Why does this method of spell checking work better for Optical Character Recognition spelling errors than for detecting writers' errors?
2. Attempting to type *bling*, I type in *bilng* to a word processor, and I get the following ranked list of corrections:

- |            |            |          |          |
|------------|------------|----------|----------|
| 1) billing | 3) bailing | 5) bang  | 7) bingo |
| 2) belong  | 4) bilge   | 6) binge | 8) blink |

- (a) Show me the (minimum) edit distance for all of these, structuring each answer like the following (which goes from *hijack* to the non-word *hecak*):
  1. *hijack* → *hejack* (substitution)
  2. *hejack* → *heack* (deletion)
  3. *heack* → *hecak* (transposition)
- (b) Focusing on the first 3 suggestions, discuss how probabilistic information would be calculated and applied to provide a ranking.
- (c) Would you want your spell checker to include a word like *bling* in its dictionary? Why or why not?

- (d) Should this list be ranked differently depending upon the immediately surrounding context? Why or why not?
3. Following up on the previous question, we want to compare *bilng* to *blink*.
- (a) Draw a directed graph for this situation.
- Assign weights (0 or 1) to the arcs. To save you some writing, assign them like so: write in the 0s, and assume every unlabeled arc is a 1.
  - Topologically order the nodes. (Note that there is more than one correct way to topologically order them.)
- (b) (Ignoring transpositions,) explain how the least cost of this graph is calculated.
- (c) (**Bonus**): How would you draw an arrow for a transposition in your graph?
4. Pick a search engine of your choice, as long as the search engine provides spelling suggestions. Search with the keywords *ghouls* and *ghosts*.
- (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. I want you to hypothesize how Word's grammar checker works. To start with, come up with a test set of 5 ungrammatical sentences and 5 grammatical but difficult sentences.
- (a) List the test set and the one-phrase description of what is ungrammatical or difficult.
- (b) What is Word's precision and recall in error detection? How many false positives are there?
- (c) Are there any indications of how the grammar checker is working?
6. I told you in class that *I saw the witch with the telescope* was ambiguous.
- (a) Given the phrase structure rules below, draw the two possible trees for the sentence *I saw the witch with the telescope*.
- |                            |                           |
|----------------------------|---------------------------|
| $S \rightarrow NP VP$      | $Det \rightarrow the$     |
| $NP \rightarrow Det N$     | $N \rightarrow witch$     |
| $NP \rightarrow NP PP$     | $N \rightarrow telescope$ |
| $NP \rightarrow Pro$       | $P \rightarrow with$      |
| $VP \rightarrow V NP (PP)$ | $Pro \rightarrow I$       |
| $PP \rightarrow P NP$      | $V \rightarrow saw$       |
- (b) Which is your preferred reading, and why? How could a computer be able to tell that that was the better structure? It might help to compare the sentence to *I saw the witch with the broom*.
7. I'm giving you the skeleton of a python program, available off the course webpage.
- (a) Write the function `fancy`
- (b) Fix the main body of the program, looping until the input is less than 100.