

## HOMEWORK #2

CS490M-12/CS580R-1

February 24, 2005

Due date: Mon., March 7, 5:00 PM (hardcopy preferred, but email OK).

See rules for turning in homework in syllabus.

Here are some misspelled words and possible corrections using Kernighan's one-error proposal.

<u>misspelled word</u>	<u>possible corrections</u>
ambitios	ambitious ambitions ambition
adusted	adjusted dusted
dialy	daily diary dials dial dimly dilly

For each of these 3 misspelled words:

- Use the method described in section 5.5 of the text to find the probability of each possible misspelled correction.
- Use a spell checker of your choice (MS Word, Unix ispell, etc.) to generate possible corrections for these words. Is the most probable correction according to Kernighan's algorithm the same as the one suggested by your program? Which additional possible corrections (i.e., non-single-error corrections) does your program generate? Which of Kernighan's possible corrections does your program omit?

You can use any search engine (e.g., Google) to get an approximation for the word frequencies and thus for the word probabilities.

The confusion matrices are available on the web from the ACL archive in Kernighan's original paper, which is recommended reading:

Kernighan, M., Church, K., and Gale, W. (199). A spelling correction program based on a noisy channel model. In COLING '90 (Helsinki), v. 2, pp. 205–211.

The URL is <http://acl.ldc.upenn.edu/C/C90/C90-2036.pdf> , and there is a pointer on the course web site.

Kernighan's original paper omits the unigram and bigram count matrices. They are available on the course web site as ap88-counts-new.txt .