

Assignment 4:  
Spam filtering  
Spelling and Grammar Correctors  
DUE: Wednesday, February 4, 2004

1. Paul Graham says, “Statistical filters yield fewer false positives because they consider evidence of innocence as well as evidence of guilt.” Based on the discussion in class of how statistical filters work (or on Paul Graham’s page, <http://www.paulgraham.com/wfks.html>), briefly (2-3 sentences) describe in your own words how statistical filters consider evidence of innocence.
2. Go to <http://www.spamassassin.org/tests.html> – in addition to little ninjas, you’ll see a list of rules used by spamassassin, a rule-based filter. Pick three (3) rules, write down the “description of test” and explain to me what this test is looking for in a single sentence.
3. A rule-based filter marks my e-mails as spam or not.
  - (a) Send me a message that you think will get marked as spam (Keep it clean, however!).
  - (b) On your homework, I want an annotated copy of that message, i.e.: 1) a copy of the message and 2) a write-up of why you thought it would be detected as spam—i.e. what particular features made you think it would be spam. Bonus points if it actually is detected as spam. ... And [drum roll] whoever sends me the spam with the highest ranking wins a free can of SPAM luncheon meat (retail value, \$1.69).
4. Take the following text and identify all the misspelled words. For each misspelled word, give:
  - (a) the correct spelling
  - (b) the type of spelling error it is. Choose from the following: run-on, split, keyboard proximity, homophone, Spoonerism, letter substitution, word replacement, lack of knowledge

*I am the word’s greatest spelker, and I thing everyone shood no it. Thon’t you dink so?*
5. 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) [The sequence *ac* has a 1 in italics to help you understand.]

	a	b	c
a	1	1	1
b	1	1	0
c	1	0	1

- (a) According to this chart, out of the nine possible sequences in the table, what two sequences of letters (i.e. two bigrams) are not possible in English?
- (b) Give words for each of the possible bigrams from this table (7 words total)
- (c) There are five (5) misspellings in the following text, all in bold

**Bobb** and his friend Abraham, or “**bae**” for short, were **acberbated bay** their other friend Arbuckle’s **ccat**.

Which of these misspellings will be caught by the bigram array we have? Which misspellings will not be identified?

- (d) Go to [www.spellonline.com](http://www.spellonline.com) and enter the text. For every word it identifies as an error, answer the following.
  - i. Was the word a misspelling?
  - ii. How many potential corrections does it give? (Note that the first word on the list is the misspelled word – do NOT include that in your count.)
  - iii. Is the correct spelling listed among the options?
- (e) Change the nonpositional bigram array into a positional array, namely one which captures the position “start of word.”

6. Calculate the minimum edit distance from each string in column A to each string in column B. Show your work. Insertions, deletions, substitutions, and transpositions all count as 1 each. Note that you are looking for the *minimum* distance.

Example: *hijack* → *hecak*:

- (a) *hijack* → *hejack* (substitution)
- (b) *hejack* → *heack* (deletion)
- (c) *heack* → *hecak* (transposition)

A	B
zardo	zrado
zardo	zalo
zardo	zdrol