

# Assignment 4

Text classification/Spam filtering

Due Monday, October 27, 2008

1. Select two articles in the *Indiana Daily Student* (<http://www.idsnews.com/>) written by different authors.
  - (a) Give the title, author's name, and date of each article.
  - (b) What are some prominent stylistic differences between the two authors? Describe at least 3 differences.
  - (c) How could these differences be detected automatically?
2. Read the paper by Paul Clough on plagiarism detection, *Old and new challenges in automatic plagiarism detection*, found at: [http://ir.shef.ac.uk/cloughie/papers/pas\\_plagiarism.pdf](http://ir.shef.ac.uk/cloughie/papers/pas_plagiarism.pdf)
  - (a) In one paragraph, summarize the technique in section 4.1
  - (b) In one paragraph, summarize the technique in section 4.2
  - (c) Section 5 presents ideas for future work (some of which has now been done); what is your favorite idea and why?
3.
  - (a) Assume you have a training corpus of 5000 emails, 3000 of which are spam. We find the word *medicines* in 250 spam messages and in 100 non-spam messages. What is the probability  $P(\text{spam}|\text{medicines})$ ?
  - (b) "Statistical filters yield fewer false positives because they consider evidence of innocence as well as evidence of guilt" (Paul Graham). Based on in-class discussion of how statistical filters work (and/or on the various articles on Paul Graham's page, <http://www.paulgraham.com/antispam.html>), describe in your own words how statistical filters consider evidence of innocence (2-3 sentences).
4. Rule-based filters and statistical filters could be used together, in what we will call a *hybrid* filter. Describe how such a hybrid filter could work. That is, how would both kinds of information be used? How would you propose to integrate statistical knowledge with rule-based knowledge? Consult [http://spamassassin.apache.org/tests\\_3\\_1\\_x.html](http://spamassassin.apache.org/tests_3_1_x.html) for a list of real rules used by a filter.
5.
  - (a) Send me an email message, the contents of which are spam, but are disguised in such a way as to "fool" the filter into thinking it's non-spam.
  - (b) Write up a description of why you thought this would be seen as non-spam.