

# CSI 5325 Assignment 5

Greg Hamerly

assigned April 2; presentation due April 30; writeup due May 4, 2009

## 1 Expectations

For this assignment, you should work in a group of 3 people. Each group should plan together to distribute the work and turn in *one* report and do *one* presentation. Some people may do different parts of the assignment, but all should contribute equally. This will be the last assignment for the semester, and it carries three times the weight of previous assignments.

The writeups you do for this course should be of high quality; complete but concise. Please structure your writeups well, introducing ideas logically (rather than simply chronologically). Your writeup should be formatted as if you would submit it to a conference or journal. It should include the following sections in the given order:

- Introduction – Describe the problem(s) you are working on at a high level, and give the reader a summary of the results you found.
- Methodology – Explain your data sets, data preprocessing, the algorithms and techniques you apply, the hypotheses you are testing, and how you will evaluate the success of your experiments. Use tables and figures as are helpful.
- Results – Explain the results of the experiments, using graphs and tables and text. Then, analyze your results.
- Conclusion – Summarize the work again, and if applicable, describe further work that you think would build on what you've done.

The writeups are intended for your professor to read and evaluate, so they should be technical in nature. The presentations should be for a broader audience, who you should assume is not familiar with machine learning. We will invite outside members to see the project presentations.

### 1.1 Submitting your work

You should turn in all your work in printed format in class, as well as by email to `hamerly@cs.baylor.edu`. Your email should have attached a zip file which has a single folder that contains all your materials. Please DO include all source code with your emails, but DO NOT print out your source code to hand in (just print and turn in writeup).

### 1.2 Tools for the course

As suggested in assignment 1, your work should be composed in  $\text{\LaTeX}$ , and I suggest looking at MATLAB for programming.

## 2 Automating transcript analysis

For this project, you should use machine learning to automate the process of analyzing instant messaging transcripts. These transcripts are records of conversations that Baylor librarians have with patrons over the internet. The librarians responsible for these transcripts must read them all and tabulate summary statistics about questions like:

- how often do patrons ask *directional* questions, such as ‘where is the bathroom’?
- how often do patrons ask *reference* questions, which require the expertise of a reference librarian?
- how often are the librarians able to give satisfactory answers to the patrons?
- what types of resolutions are given for reference questions?

There are many more possibilities, of course.

There will be several milestones you should accomplish:

- gather data – ask (nicely) that the librarians tag transcripts for you
- label transcripts – choose a representation for the target(s)
- choose a learning algorithm – I recommend trying more than one, and comparing results
- pre-process the data
- train the algorithms
- test on held-out data
- provide an internet-based service that the librarians would be able to use in the future

## 3 Service delivery

You should provide a web-based service that allows the librarians to upload new transcripts, anonymizes them, and automatically makes predictions and tabulates summary statistics.

## 4 Privacy issues

Make sure that your web service uses encryption, and that data is anonymized as much as possible to avoid leaking sensitive information.

## 5 Evaluation criteria

For the presentation, you will be evaluated based on the clarity and quality of your presentation, and its appropriateness for the audience.

The following table will be used for grading your written assignment (these adapted from Charles Elkan’s grading criteria):

Category	Points
Introduction insightful on background and motivation	3
Precise and reproducible description of technical work	3
Sensible implementation decisions (code reuse, choice of PL, etc.)	2
Sufficiently conclusive experiments (large datasets, lesion studies, etc.)	3
Well-designed, well-described, reproducible experiments	3
Understandable presentation of results, preferably graphical	3
Clear, correct analysis of results (including statistical significance)	3
Insightful discussion of all major experimental results	3
Exciting and useful results with general applicability	3
Appropriate organization (logical, not chronological) and easy-to-read plain writing style	3
Correct spelling, grammar, and choice of words	3
<b>Total</b>	<b>32</b>

A lesion study means taking parts out of the system you are using to identify what will cause a system's performance to degrade. For this assignment, you need not perform statistical significance tests, but careful analysis is still important.