**Text Mining – Final Project**

Your course project will hopefully be a fun investigation of some aspect of text mining and language processing.

**General requirements**

You will conduct an implementation project that is related to something we have studied or will study in this class, or some other topic that is of interest to the text mining community.

An implementation project is a project that *may* have some element of research, meaning that the project investigates something novel; or, a project that involves a significant engineering effort; e.g., implementing an already published algorithm and running it on another dataset; comparing several existing methods and datasets on a particular task, etc. It is possible and recommended to use publicly available resources and tools.

The projects will be done in groups of 1-2 students (size matters – a larger team is expected to do more work).

**Deliverables**

1. **A project proposal** - 1- 2 pages, submitted by April 25 or soon after. It should include the following information:
   * A description of the problem
   * Motivation: why this problem is important/interesting
   * The approach you plan to take to solve this problem
   * Evaluation plan (inc. evaluation measures)
   * A list of at least 3 related papers.
   * A list of possible resources\tools
2. **Presentation** - in the last class(es) of the semester, each team will present their project for 20 minutes, followed by 10 minutes for questions and discussion. Every member of the team must speak for part of the presentation.
3. **A final report** –The report should include 4-6 pages, including references and figures. Format: simply format your paper to look like one of the conference papers we have read. The report (and presentation, to the extent possible) should include the following:

* Problem statement + motivation (i.e., why it is important/interesting)
* Methods: a description of your approach to address the problem
* Experiments: including the description of experimental datasets, how evaluation is conducted, etc.
* Results: including interpretation of results (note that typically, there are previously published results to compare to)
* Related Work: refer to at least 5 works that are related to your work as closely as possible.

**Grading**

All team members will receive the same grade for the presentation and report. The overall project grade will be 50% of the final grade.

You will be graded on these key factors:

1) How clear and understandable is the material that you produced?   
2) How well does your work address the project goals?  
3) Technical correctness (did you understand how the tools you used work? did your experiments and evaluation measures demonstrate the right thing? etc.)  
4) Presentation quality.  
5) Report writing quality.  
6) If the project is creative and/or includes an exceptional effort, it will be taken into account.

**Project Ideas**

You are encouraged to come up with your own project ideas. Otherwise, here are some quick example ideas:

* Twitter Retweets: Collect tweets from news publishers and re-tweets about those tweets. Try to predict if the number of retweets for a particular tweet is going to be high\low. What are useful features in predicting number of retweets? Why?
* Predicting Stock Fluctuations: Gather daily tweets about a particular company (I can show you how to do that) as well as stock price data. Build a model that predicts whether the stock price will go up or down based on previous tweets about the company. This is a very difficult problem. However, the point is to learn about the problem and not necessarily to solve it.
* Detecting Age-Appropriate Language/Content: Find a data set of texts and their age appropriate ratings. Try to build a model that predicts the age-appropriateness of a span of text. Explore different feature representations and discuss what works, what doesn’t, and why.
* Do blog author analysis; for example, can you find out if the blog author is a Democrat or Republican? `left wing' or 'right wing'?
* Many people maintain a homepage. The task is to extract biographical or personal facts of interest from their homepages; for example, extract the person's email address, his or her job title.
* Typo-correction, e.g., based on n-gram statistics and string similarity.
* Visualization of text data (see, for example, <http://benfry.com/linking/>)

Hope you get the idea! (both meanings of this sentence apply..)

**Data resources (very few examples, search for relevant data on the Web!)**

Sentiment analysis: <http://www.cs.cornell.edu/people/pabo/movie-review-data/>

Text classification (20 newsgroups): <http://qwone.com/~jason/20Newsgroups/>

Tweets: http://trec.nist.gov/data/tweets/

**Example tools (outdated)**

MALLET is a Java toolkit for machine learning applied to natural language. It provides facilities for document classification, information extraction, part-of-speech tagging, noun phrase segmentation, general finite state transducers and classification

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. Weka is open source software issued under the GNU General Public License.

NLP tools – parsing, POS tagging, named entity recognizer, word segmentation – see for example <http://nlp.stanford.edu/software/index.shtml>

**Advice**

* Make sure that your project is not TOO ambitious, in the sense that you have some idea about how to reach your goals.
* Check that there are relevant datasets for your work. There are many corpora and labeled datasets available for the evaluation of text processing tasks. Alternatively, you can label your own data. (If there is a similar research paper, check out what data was used there in the evaluation. Experimental data is often freely distributed, either on the Web or by request.)
* Look forward in the list of topics that we will cover, check out the papers and/or the Web, and see if you are interested in one of these topics or applications.
* In general, you are not limited to the topics and methods covered in class, as long as your project has to do with text processing.

Good luck!