# COMP 90042: Project 2: DRAFT

William Webber

April 30, 2014

*Version history:*

- *v 0.1 (30/04/2014) – Draft version*

## 1 Introduction

*This description is a draft. The final version will be posted by Wednesday, 7th May*

The second project is to write a research report on a project of your choice related to the content of the course. The main output of the project will be a research report of between 1500 and 2500 words, describing the outcome of your research. You may also include code, data files, and the like, but assessment will be based upon report itself.

The project and report should contain some combination of the below research elements:

- Literature review of previous work in the area
- Tutorial-style description of methods involved
- Implementation
- Theoretical work
- Experimental setup and experimental results

The combination of the above that is appropriate for your project will depend upon the topic you choose, but there should be at least 2 of the above components. For example, if your project is primarily a literature survey, then you might also implement one of the methods described.

A list of project ideas is given below; but you are also free to choose another project of interest to you. The project ideas below are just ideas, which will need to be elaborated into full project proposals. If you are unclear about what is implied in any of the ideas, please ask me for a fuller explanation.

Whether you pick a project idea from the list below or one of your own, you should submit a project proposal for approval by email to me by the end of **Friday, May 9th**. The project proposal should state in 100-200 words:

- The topic of your research project

- What questions you will investigate as part of the project and how
- Which research elements (from the list above) your project will include

Projects will be assessed upon:

- Quality of presentation (including maths, figures, tables)
- Quality and correctness of writing
- Adequacy of literature citations
- Correctness of implementation (if any)
- Correctness and thoroughness of experimental work (if any)
- Originality and inventiveness of ideas proposed

# 2 Project list

## 2.1 Lecture 1: Terms

- CJK word segmentation
- Automatically learnt stemming
- Automatically learnt stopping
- Noun phrase identification
- Zipfian term distribution

## 2.2 Lecture 2: VSM

- Mahalanobis distance between documents
- Alternative TF*IDF formulations
- Comparing terms in document space

## 2.3 Lecture 3: Information Retrieval

- Index compression
- Document length normalization
- Query term proximity / phraseal detection

## 2.4 Lecture 4: Query expansion

- Improved VSM global term similarity computation
- Detecting terms with multiple senses
- Alternative PRF algorithms (within VSM)

## 2.5 Lecture 5: IR evaluation

- Build test collection around Project 1 dataset
- Detecting answer redundancy
- Significance testing in IR evaluation

## 2.6 Lecture 6: Text clustering

- Term clustering
- Implement internal evalution of clustering algorithms
- External evaluation of clustering algorithms against RCV1v2 collection
- Implement hybrid clustering

## 2.7 Lecture 7: LSA

- Using PCA for text analysis
- Relationship between SVD and PCA, for text
- LSA for suggesting query expansion terms
- Complexity of SVD algorithms
- On-line SVD algorithms
- SVD maintenance by "folding in"
- Effectiveness of LSA search (literature review)

## 2.8 Lecture 8: Text classification

- Decision trees for text classification
- Efficient $k$ nearest neighbours
- Compare effectiveness of multiple classification methods on RCV1v2 dataset
- Active learning for text classification
- Rocchio classification with initial query
- Robustness of classifiers to labelling errors

## 2.9 Lecture 9: Support vector machines

- Efficient methods of calculating SVM, and their complexity
- Linear separability of text classification (with different kernels)
- Effectiveness of different kernels for text classification
- Can we really just "throw features" at an SVM?
- How to intepret SVM models
- Multi-class SVM
- Text kernels in SVM

## 2.10 Lecture 10: Probabilistic IR

- Alternative probabilistic models for IR

## 2.11 Lecture 11: 2-Poisson Model and BM25

- Fit 2-Poisson model to RCV1v2
- Alternative models to 2-Poission model
- Relevance feedback for BM25

## 2.12 Lecture 12: Language models for IR

- $n$-gram language models for IR
- Ponte-Croft vs. Lavrenko language models
- Language models for text classification

## 2.13 Lecture 13: Advanced LM

- Relevance feedback in language models
- Relevance models in LM

## 2.14 Lecture 14: Naive Bayes

- Naive Bayes on top of LSA?
- Feature selection for NB
- TF-IDF features for NB

### 2.15 Lecture 15: Logistic Regression

- Time complexity for full logistic regression
- Online methods for logistic regression
- Regularization in logistic regression
- Relationship between maxent and logistic regression

## 3 Project details

### 3.1 Submission format

Submission is via SVN. Create a directory called *proj2* under the root SVN directory, and add all files under that directory. Please do *not* include large data files ($>$ 5MB).

The report should be submitted in PDF format. You may generate the PDF via LaTeX, MS Word, or another document layout or word processing system of your choice.

### 3.2 Deadline

The due date for this project is **Monday, 2 June, at 11:59pm**. Whatever is committed in Subversion by then will be taken as your project submission.

### 3.3 Individual work

This project is to be completed as individual work, not as a team project. You may discuss high-level questions, but do not share your code or your text with other students.