COURSEWORK ASSIGNMENT **UNIVERSITY OF EAST ANGLIA**
School of Computing Sciences

MODULE:	**: CMP-5036A Information Retrieval**	
ASSIGNMENT TITLE:	**: Experimental project: search experiments**	
DATE SET	**: Thursday Week 4**	
HAND-IN	**: Slides + flyer Wednesday 13 December (Week 12)**	
	: **Presentation Thursday 14 December (Week 12)**	
MARKS AVAILABLE	**: Spring semester Week 1**	
ASSIGNMENT VALUE	**: 40%**	
SET BY	**: D. J. Smith**	**SIGNED:**
CHECKED BY	**: S. Taylor**	**SIGNED:**

Overview

The aim of this assignment is to construct and evaluate the retrieval performance of a simple search engine, using techniques that have been described in the module. You will compare some different techniques for this and produce a short report and a presentation describing your work and your findings. You are expected to program this in Python.

Aims

- To introduce natural language processing techniques that are used to improve the performance of information retrieval systems
- To understand the main issues and measures used in evaluating the performance of information retrieval and document classification systems
- To give students practical experience of information retrieval systems, experimental work and evaluation
-

Learning outcomes

- Understanding and experience of a range of information retrieval techniques and models and their application, particularly in Web searching
- Understanding of how natural language processing techniques may be applied to information retrieval
- Improved program design and coding skills.
- Improved research and communications skills.

Assessment criteria

You will present your results in a 10 minute presentation. (You are required to be present for the duration of your presentation session.)
Marks will be awarded for:
- the quality of the design of your experiments,
- the quality of the understanding of the topics,
- the understanding you display in response to questions,
- extra work you have done to improve your experimental results,
- the quality of your presentation, slides and flyer.

The breakdown of marks is given on the marking sheet (attached). *Note that the distribution of marks is indicative only and may change*.

Description of the assignment

The assignment is to produce a search engine that works over the https://portal.uea.ac.uk domain. Use the crawler and indexer from Assignment 1 as the starting point.
The tasks are:
1. Construct a simple vector model retrieval system using tf*idf term weighting. The system should:
 a) Have a simple interface to run search queries. This should allow for (i) a single query to be entered from the keyboard, or (ii) a set of queries to be read from a text file.
 b) Return a ranked list of URLs to screen or file. You may also add document snippets to the screen display. The file output should be a CSV file with the same structure as the template provided on Blackboard.
When you have done this, you should (i) evaluate your system using a set of training queries that you have devised and then (ii) attempt to improve your results by investigating the effects of:

2. Using stemmers. You could use the home-grown UEA Lite stemmer, available on Blackboard (the NLTK lemmatizer is similar) and/or an implementation of the Porter algorithm (e.g. in NLTK or at https://pypi.python.org/pypi/stemming/1.0)
3. Giving extra weight to terms occurring in titles, headings, etc.
4. Developing a simple system-centred relevance feedback mechanism.
5. …. anything else that you feel is interesting and appropriate.

After each attempted improvement you should measure its effect by running the system with your training queries. It is extremely important that you always evaluate your system using a set of queries, as performance on a single query is often unrepresentative of the typical performance of the system.

Evaluation

In Week 10 (Thursday 13:00) we will provide a set of 10 test queries. You should run the best version of your system with the test queries and submit the result files via Blackboard before Thursday Week 11 09:00. We will return the pooled relevance judgements by Friday Week 11 12:00, so that you can use them in your flyer and presentation. (If you do not submit evaluation results you will not be given the relevance judgements for the test queries.)

Reference

Manning C., Raghavan P., Schultz H. (2008) Introduction to Information Retrieval, CUP

Required:

1. Presentation and questions

The presentation should highlight the results, assumptions, problems and limitations of your experiments. It should clearly show the techniques you have used and—above all—your experimental results. Graphs showing the performance under various conditions are particularly informative.
In Week 12 you will:
(a) do a presentation of your work and answer questions on it. (You will need to submit your presentation slides the day before the presentation, so that they can be loaded onto the computer being used for the presentations.),
(b) submit a PDF version of the flyer via Blackboard the day before the presentation, so that it is available for the presentation.

2. Flyer

You should produce flyer (2 pages maximum), summarising the techniques used, your experiments and results. It is best presented in ACM SIG alternate style. An example, LaTeX and Word templates for the required style are on Blackboard.

3. Code

You must submit a copy of all the code used in your experiments as a single zipped file, via Blackboard; the filename should be of the form: *yourstudentid*_IRcoursework.zip. The review of this may affect the marks awarded for basic retrieval and additional experiments either positively (e.g. particularly clear, well structured and commented code) or negatively (e.g. code mostly reused from third parties, inconsistent layout, uncommented).

CMP-5036A Marking Sheet: Search experiments

Student name		**No.**	
Marker name			
Baseline retrieval 25%	*Evidence of a basic working vector space / tf*idf retrieval system with correctly calculated ranked results with snippets (from presentation and questions).* *Breakdown:* *4 for a complete crawl (0 if incomplete);* *5 for reasonable tf*idf calculation;* *5 for reasonable vector calculation;* *5 for a ranked list of results (rank no. + URL) for each query (1 if some queries);* *6 for useful snippets in results*		
Core experiments 20%	*Evidence of well-conducted experiments to investigate the impact of stemming, differential weighting of text, and evaluation – p(10), R-precision – against the results provided in Week 11.* *Breakdown (3 marks per experiment lost if no results submitted):* *3 for correctly calculated metrics (0 if no results submitted);* *5 for stemming;* *8 for differential weighting (5 max if just title);* *4 for discussion of possible explanations of differences in performance.*		
Additional experiments 30%	*Evidence of well-conducted, experiments on relevance feedback, ...* *Breakdown (3 marks per experiment lost if no results submitted):* *10 for relevance feedback/query expansion;* *12 for any other interesting experiments (e.g. use of NE) or analyses;* *8 for discussion of possible explanations of differences in performance.*		
Presentation 15%	*Content, delivery and appearance of the presentation, including use of graphs/pictures to present results and ideas.* *Breakdown:* *3 oral delivery, engagement, speed;* *3 slide content, legibility, graphs and diagrams;* *6 for good graphs (e.g. from matplotlib) showing p(n) etc.* *3 presentation structure, timing (max. 1 if stopped for overrun).*		
Flyer 10%	*Content, clarity and accuracy of writing, including use of graphs/pictures to present results and ideas; similarity to an ACM SIGIR short paper.* *Breakdown:* *3 spelling, grammar, references/citations, technical style;* *4 content and structure (including graphs and diagrams);* *3 presentation and layout (0 if not in ACM SIG style).*		
Additional comments			
		Mark	**%**