Module: CMP-5036A Information Retrieval
Assignment: 002/102 Individual Experimental Project: Search Experiments

Set by: Farhana Liza: F.Liza@uea.ac.uk
Value: 50% (002) or 55% (102)
Date due: 23 January 2026 14:59
Returned by: Within twenty working days from due date
Submission: Blackboard
Checked by: Muhammad Awais: M.Awais@uea.ac.uk

Learning outcomes

- Explain and experience a range of information retrieval techniques and their applications, particularly for web searching, using Python coding processes
- Develop, demonstrate and analyse information retrieval algorithms, by coding an information retrieval system
- Design, evaluate, and evidence how efficient and effective different information retrieval techniques are, by experimenting with a working system

Specification

Overview

This is an ***individual*** assignment where you will need to design and code a search engine then analyse, evaluate, and discuss its retrieval performance, using techniques described in the module and your own research. You will need to compare multiple different Information Retrieval (IR) techniques/processes and prepare a MP4 video recording of a 10-minute PowerPoint presentation to evidence your research and evaluations. The coding language for the search engine must be Python.

This 10-minute presentation must clearly demonstrate your design and development process. This includes showing how you input and output data, your analytical methods, the use of appropriate IR metrics, evaluations, and critical reflections supported by your research log. You need to concisely explain each retrieval process, how it works, using your experiments to identify and justify how each process helps or hinders IR search engines.

Description

You must download and use one of the three datasets, the *videogames/books/ soccer players* dataset, and a corresponding labelled .csv file for evaluation, saved in the Assessment briefs, under the Assessment folder on Blackboard.

For real world inspiration, have a look at the Internet Games Search Engine at *https://www.igdb.com/ or Online Book Store at https://www.amazon.com/ or Sports Page at https://www.reuters.com/sports/soccer/ .*

You then need to *individually* code and experiment with a Python based search engine that will help you design, experiment, analyse and evaluate IR systems with *precision at top 10* (precision@10) retrieval and standard metric calculations, for the below search questions.

You need to show the evaluations and metrics using visual representations e.g. graphs, tables, diagrams, etc., when discussing your findings, during your 10-minute PowerPoint presentation. Expanding on these training questions will enable you to show a deeper understanding and extra points are allocated for this.

The system should be coded to be domain independent and capable of working with crawled data from any site (assuming UTF-8 character encoding and English language).

Development:

1. Download the *videogames/books/ soccer players html* dataset from Blackboard Assessment folder. Note: you have to develop system for **only** one dataset.
2. Using python code, you need to develop an information retrieval system using at least tf*idf term weighting and more advanced representation. The system should:
    a) Have a simple command line interface to run search queries allowing for:
        i. Single and multiple term queries to be entered from the keyboard
    b) Return a ranked list of URLs and content to the console and a file
3. Evaluate your system for ways to improve your URL ranked list and contents:
    i. Use the below set of **6 training questions** to help you develop, analyse, evaluate and discuss the effectiveness of your coded system (note: you must develop system for only one dataset):
        - For soccer players search engine
            1. Placente, Diego
            2. Wiltord, Sylvain
            3. Carbone, Benito
            4. English players
            5. Fast left sided defender
            6. Defender

        - For books search engine
            1. Copperhead
            2. Star Wars
            3. Love's Labour's Lost
            4. Wild Flowers of Britain & Northern Europe
            5. Books written by William Shakespeare
            6. Uncle Tom's Cabin

        - For videogames search engine
            1. Pokémon Trozei
            2. Tony Hawk’s Downhill Jam
            3. Arcade type games
            4. London Taxi: Rush Hour
            5. Game published by Atari
            6. The Sims 2 Apartment Pets

    ii. Experiment with the use of natural language processing pre-processing techniques including stemming, lemmatisation, stop words etc and analyse how effective each process is for an IR system
    iii. Give extra weight to terms including titles, heading, and the meta-data (e.g., genre) etc.
    iv. Use a simple query expansion mechanism
    v. Identify and experiment with named entities

vi. Anything else (research based method) you think is relevant for improving your retrieval system

Each process should be coded by yourself and you should measure your retrieval system's effectiveness, by using relevant URL and content output order, *precision at 10* (precision@10) and standard metric calculations.

Research Log: When you are designing and developing the system, you need to document your process in a research log, providing a metacognitive account of your development decisions, challenges, and learning. Your research log is a foundational component of this project, serving as a structured and dated chronicle of your entire development process. It must move beyond simple notetaking to provide a critical narrative of your technical and intellectual journey.

Key Required Elements:

- Domain Choice/Dataset Selection Reasoning: Provide a critical rationale for your dataset selection. Your justification could evaluate your personal choice, the chosen domain's structural properties and explain their alignment with the core information retrieval challenges this project aims to address.
- Conceptual Investigation: Document your initial research into core Information Retrieval concepts, such as:
    a. Foundational models (e.g., inverted indexes, Boolean retrieval).
    b. Ranking algorithms (e.g., TF-IDF, BM25).
    c. Advanced techniques explored (e.g., vector space models, query expansion, or deep learning-based approaches).
- Design Rationale: Provide clear justification for your key technical choices, explaining *why* you selected specific:
    a. Programming languages libraries.
    b. Data structures for indexing and storage.
    c. Algorithmic approaches.
- Experimental Record: Maintain a detailed account of your development cycles, including:
    a. Code snippets for different implementations.
    b. Testing methodologies and results (e.g., performance benchmarks, precision/recall scores).
    c. Observations/reflections on the outcomes of each experiment.
- Critical Analysis: Demonstrate your problem-solving skills by analysing:
    a. Failures and dead ends encountered.
    b. How you diagnosed problems and pivoted your strategy.
    c. Reflections on significant challenges and their resolutions.

Relationship to formative assessment

This work builds on the formative work you have completed during lab sessions for this module.

Deliverables

1. **Presentation Demo and Research Log:**

Your MP4 video recording of a 10-minute PowerPoint presentation should contain, at minimum, the following evaluations:

   a. A brief overview of your system and functionality aligned with the marking scheme, coded in python and showing you inputting the relevant search questions and generating outputs.
   b. The results of your experiments, using suitable visual representations e.g., graphs, tables, diagrams, etc.
   c. Discuss and justify the techniques/processes experimented on to demonstrate your reasoning of why they are, or are not, useful for efficient information retrieval systems.
   d. You will need to show your readme file at the end of the presentation
   e. Your presentation must incorporate a curated selection of dated research log entries to critically trace the project's development, evidencing the rationale for key decisions and demonstrating reflective learning throughout the research process. You should reference to the research log as much as possible to showcase your research, development, analysis and critical reasoning skills.

2. **Blackboard submission point upload**

   a. After entering the submission point, submit your source code, a detailed README on how to run the code, a copy of your presentation slides and the completed research log by selecting 'Upload Files'.
   b. Submit your MP4 video recording of a 10-minute PowerPoint presentation to eStream by selecting 'Create submission' and following the instructions in this link: https://my.uea.ac.uk/departments/learning-technology/students/video-and-audio-assignments

You must upload to Blackboard by 14:59 on Friday 23rd January 2026.

The submission point will be released 1 week before the submission date and will show in the Summative Assessment tab on Blackboard. *If you have a Synoptic Project to complete, your submission point is {002}, otherwise you will submit at {102}.*

The submission must contain the presentation PowerPoint slides, presentation recording in MP4, research log, source files for your python system and a README stating how to run the code. Name the submission documents with your student ID (e.g., student_id.mp4, student_id_research_log.doc).

The primary process for marking this assignment is the presentation recording and research log, but the code will be reviewed if the markers deem it necessary.

Resources

https://stackoverflow.com/ is an excellent site for finding information about specific issues relating to various programming languages, including Python. It is important however not to become too reliant on sites such as StackOverflow, which are great for details, but don't give the “big picture”, so it is difficult to get a good understanding of programming in this way.

https://www.w3schools.com/python/python_reference.asp - a website providing reference material on Python and its libraries.

https://docs.python.org/3.7/ - another website providing reference material on Python and its libraries.

This assignment is closely linked to the weekly lab sheets.

If you get stuck, please refer to the relevant lab sheet and ask the Lab Leaders for help during scheduled laboratory hours

Plagiarism, collusion, and contract cheating

The University takes academic integrity very seriously. You must not commit plagiarism, collusion, or contract cheating in your submitted work. Our Policy on Plagiarism, Collusion, and Contract Cheating explains:

- what is meant by the terms ‘plagiarism’, ‘collusion’, and ‘contract cheating’
- how to avoid plagiarism, collusion, and contract cheating
- using a proof reader
- what will happen if we suspect that you have breached the policy.

It is essential that you read this policy and you undertake (or refresh your memory of) our school’s training on this. You can find the policy and related guidance here: https://my.uea.ac.uk/departments/learning-and-teaching/students/academic-cycle/regulations-and-discipline/plagiarism-awareness

The policy allows us to make some rules specific to this assessment. Note that:

> In this assessment, working with others is *not* permitted. All aspects of your submission, including but not limited to: research, design, development and writing, must be your own work according to your own understanding of topics. Please pay careful attention to the definitions of contract cheating, plagiarism and collusion in the policy and ask your module organiser if you are unsure about anything.>*

Marking scheme

See marking scheme on next page:

Marking scheme: Assignment 002/102
Experimental Project – 50% for 002 and 55% for 102

Marking Details	Mark %	Marking Comments
Baseline system Evidence of a working vector space/tf*idf retrieval system with correctly calculated ranked results 2 – reasonable coding style 4 – Ranked list of results (rank number, URL and contents) for each query 3 – Reasonable tf*idf calculation 4 – Reasonable vector calculation 2 – Discussion/justification of differences in performance	15%	
Core Experiments Evidence of experiments that show the effectiveness, or not, of pre-processing (e.g., stemming) techniques and evaluation processes. This should include calculated and visualised metrics and justification of which are most effective and why 6 – Correctly calculated metrics 4 – Stemming, lemmatisation and stop words 5 – Discussion/ justification for differences in performances	15%	
Additional Experiments Evidence of experiments on extra processes e.g. named entities, query expansion and processes you've researched yourself 4 – Query expansion 4 – Named Entity Recognition (NER) experiments 5 – Other interesting experiments or analysis/evaluations 7 – Discussion/ justification for difference in performances	20%	
Presentation Content, structure, oral delivery and appearance of the PowerPoint presentation, should include use of graphs, tables, diagrams to visually support your discussions 5 – Oral delivery, engagement, speed and slide content quality 7 – Visual representations e.g. graphs, diagrams, tables, etc. 5 – Result and discussion 8 – Presentation structure, reference to the research log and timing (0 if more than 10 minutes)	25%	
README and Research Log Content, clarity, accuracy of writing and references, use of graphs, tables etc. To justify your experiments and results 4 – Clear and easy to follow readme 6 – Spelling, grammar, references/citations and technical writing style in research log 15 – The research log will be evaluated on the depth of your critical reflection[1] and the clarity with which it reveals the evolution of both your project and your problem-solving skills.	25%	

Extra Comments: Total Score =

[1] This critical reflection can include IR project's key findings, the problems faced during the IR implementation, the solutions proposed to fix errors, real-world significance of your IR project, how you would approach the project differently if you were to do this again, lessons learnt and skills gained.