**Faculty of Engineering, Environment and Computing 7071CEM**

**Assignment Brief Jan- May 2021**



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Module Title**  Information Retrieval | **Ind/Group:**  Individual | | **Cohort**  Jan- May | **Module Code**  7071CEM |
| **Coursework Title (e.g. CWK1)**  Deferred-CW | | | | **Hand out date:**  21.05.2021 |
| **Lecturer**  Seyed Mousavi | | | | **Due date:**  2.08.2021, 18:00 BST |
| **Estimated Time (hrs):**  40  **Word Limit\*:**  N/A | | **Coursework type:**  Project/Development | | **% of Module Mark**  66.7% |
| **Submission arrangement online via Aula:** x  **File types and method of recording:** pdf or docx  **Mark and Feedback date:** 2 weeks after submission due date  **Mark and Feedback method:** via Aula | | | |  |

|  |
| --- |
| **Module Learning Outcomes Assessed:**  Module Learning Outcomes 1-5:   1. *Demonstrate a sound knowledge of information retrieval principles* 2. *Apply main data structures used in index construction in Python or a similar high- level language* 3. *Implement a typical web crawler and query processor, in Python or a similar high-level language* 4. *Acquire knowledge and skills to apply common machine learning methods for text classification and document clustering* 5. *Build the outline of a minimum viable vertical search engine for text retrieval* |
| **Task:**  Develop a vertical search engine similar to Google Scholar that only retrieves papers/books published by a member of Coventry University. That is, at least one of the co-authors must be from CU. To that end, you crawl the profiles of academic staff at CU available on the CU web site and index their papers in their profiles. The seed page for your crawler, i.e. the first page to crawl, is the following:  <https://pureportal.coventry.ac.uk/en/organisations/coventry-university/persons/>  Your system crawls this page and the links provided for each member of staff there to access their CU profiles (also called “pureportal” profiles). Then, for each profile, it goes through the research outputs (publications), if any, to construct the inverted index using the information about the publications. Make sure you preserve the robots.txt rules. In addition, make sure you do not hit the servers too much or too fast.  Because of low rate of changes to this information, your crawler may be scheduled to look for new information, say, once per week, but it should ideally be able to do so automatically, as a scheduled task.  From the user’s point of view, your system has an interface that is similar to the Google Scholar main page, where the user can type in their queries/keywords about the resources they want to find. Then, your system will display the results, sorted by relevance, in a similar way Google Scholar does. However, only publications with at least one co-author from CU are displayed. You may further specialise your search engine to a specific field, e.g., *computer science*, *mechanical engineering*, *bioinformatics* or whatever you would like.  In addition, whether as a separate program or integrated with search engine, a subject classification functionality is needed. More specifically, the input is a scientific text and the output is its subject among zero or more of the cases: *Health*, *Engineering*, *Business*, *Art*.  You can use any general purpose programming language of your choice although Python is recommended because of its rich library and sample codes developed in the labs.  In case of ambiguity, make reasonable assumptions and/or let me know.  **Requirements and Markings**   |  |  | | --- | --- | | **Requirement** | **Mark%** | | *Fully working crawler component*  Must completely crawl the CU profiles in BFS manner. It should be scheduled to re-crawl to extract new data automatically. Must preserve robots.txt rules and be polite by not hitting the servers too fast. | *25* | | *Construction of Inverted Index*  Construction of the index based on appropriate data structures studied in the module as opposed to naive database tables. The index should be updated incrementally once new data are received from the crawler component. | *15* | | *Fully Working query processor component*  Displaying results relevant to given queries | *25* | | *Fully working subject classification component*  Given the search keyword/text, the system should identify its relevant subject(s), if any, among the predefined subjects. | *20* | | *Overall usability*  Acceptable response time, accuracy of results, nice interface, readable results sorted by relevance, subject classification and anything else that might affect the usability of the product. | *15* |   Based on the above marking scheme, a typical system expected for a mark of 40 or more is a working search engine which accepts users’ queries/keywords and displays relevant results, i.e. publications by CU staff in relevant fields. However, it may not have proper inverted index and subject classification components and may be slow and inaccurate in some cases.  To earn 70 or more, the system is expected to be a working search engine with reasonable accuracy and speed. This ensures that the system contains fully working crawler and query processor components. In addition, it must have at least one, and preferably both, of the other two components, i.e. the inverted index and the text classification components, in fully working status.  Please note that to show that your system meets each of the above-mentioned requirements, your report must provide sufficient evidence including clear description, complete source code, and complete screenshots where applicable. |
| Notes:   1. You are expected to use the  [Coventry University APA](https://eur01.safelinks.protection.outlook.com/?url=https%3A%2F%2Flibguides.coventry.ac.uk%2Fapa&data=04%7C01%7Cad0204%40coventry.ac.uk%7C87c5edec83934362793a08d871e2bae0%7C4b18ab9a37654abeac7c0e0d398afd4f%7C0%7C0%7C637384565890081424%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C1000&sdata=2XH02DM9i%2BMc4hmkkHa2xZ4kdKxQ41ruCCWvvcO3XI0%3D&reserved=0)style for referencing. For support and advice on this students can contact [Centre for Academic Writing (CAW)](https://eur01.safelinks.protection.outlook.com/?url=http%3A%2F%2Fwww.coventry.ac.uk%2Fstudy-at-coventry%2Fstudent-support%2Facademic-support%2Fcentre-for-academic-writing%2F%3Ftheme%3Dmain&data=04%7C01%7Cad0204%40coventry.ac.uk%7C87c5edec83934362793a08d871e2bae0%7C4b18ab9a37654abeac7c0e0d398afd4f%7C0%7C0%7C637384565890091371%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C1000&sdata=dJEnR6LZGKOOuew5A6lthd3u8AhVJ6bP2Bwa5svCZfU%3D&reserved=0)[.](https://eur01.safelinks.protection.outlook.com/?url=http%3A%2F%2Fwww.coventry.ac.uk%2Fstudy-at-coventry%2Fstudent-support%2Facademic-support%2Fcentre-for-academic-writing%2F%3Ftheme%3Dmain&data=04%7C01%7Cad0204%40coventry.ac.uk%7C87c5edec83934362793a08d871e2bae0%7C4b18ab9a37654abeac7c0e0d398afd4f%7C0%7C0%7C637384565890101338%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C1000&sdata=OYZHogfKaGGOYF%2BRUvLmcHXWPvuIDaOWwlAfg8Xm2Xs%3D&reserved=0) 2. Please notify your registry course support team and module leader for disability support. 3. Any student requiring an extension or deferral should follow the university process as outlined [here](https://eur01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fshare.coventry.ac.uk%2Fstudents%2FRegistry%2FPages%2FDeferrals-and-Extension.aspx&data=04%7C01%7Cad0204%40coventry.ac.uk%7C87c5edec83934362793a08d871e2bae0%7C4b18ab9a37654abeac7c0e0d398afd4f%7C0%7C0%7C637384565890101338%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C1000&sdata=E0%2FpGRvSft5reQM2BwpeacSJxTCSdW%2BHcGC88VOPl60%3D&reserved=0)[.](https://eur01.safelinks.protection.outlook.com/?url=https%3A%2F%2Fshare.coventry.ac.uk%2Fstudents%2FRegistry%2FPages%2FDeferrals-and-Extension.aspx&data=04%7C01%7Cad0204%40coventry.ac.uk%7C87c5edec83934362793a08d871e2bae0%7C4b18ab9a37654abeac7c0e0d398afd4f%7C0%7C0%7C637384565890111297%7CUnknown%7CTWFpbGZsb3d8eyJWIjoiMC4wLjAwMDAiLCJQIjoiV2luMzIiLCJBTiI6Ik1haWwiLCJXVCI6Mn0%3D%7C1000&sdata=luUmpj3fSJfJoFMp%2Bwr%2FLpPmoMZdcWManRvJrb2t2wo%3D&reserved=0) 4. The University cannot take responsibility for any coursework lost or corrupted on disks, laptops or personal computer. Students should therefore regularly back-up any work and are advised to save it on the University system. 5. If there are technical or performance issues that prevent students submitting coursework through the online coursework submission system on the day of a coursework deadline, an appropriate extension to the coursework submission deadline will be agreed. This extension will normally be 24 hours or the next working day if the deadline falls on a Friday or over the weekend period. This will be communicated via your Module Leader. 6. You are encouraged to check the originality of your work by using the draft Turnitin links on Aula. 7. Collusion between students (where sections of your work are similar to the work submitted by other students in this or previous module cohorts) is taken extremely seriously and will be reported to the academic conduct panel. This applies to both courseworks and exam answers. 8. A marked difference between your writing style, knowledge and skill level demonstrated in class discussion, any test conditions and that demonstrated in a coursework assignment may result in you having to undertake a Viva Voce in order to prove the coursework assignment is entirely your own work. 9. If you make use of the services of a proof reader in your work you must keep your original version and make it available as a demonstration of your written efforts. 10. You must not submit work for assessment that you have already submitted (partially or in full), either for your current course or for another qualification of this university, with the exception of resits, where for the coursework, you may be asked to rework and improve a previous attempt.   This requirement will be specifically detailed in your assignment brief or specific course or module information. Where earlier work by you is citable, i.e. it has already been published/submitted, you must reference it clearly.  Identical pieces of work submitted concurrently may also be considered to be self-plagiarism. |

**Mark allocation guidelines to students**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 0-39 | 40-49 | 50-59 | 60-69 | 70+ | 80+ |
| Work mainly incomplete and /or weaknesses in most areas | Most elements completed; weaknesses outweigh strengths | Most elements are strong, minor weaknesses | Strengths in all elements | Most work exceeds the standard expected | All work substantially exceeds the standard expected |