|  |  |
| --- | --- |
|  | **HASAN KALYONCU UNIVERSITY**  **Computer Engineering Department** **COME 499 Project Proposal Form** |

**Part I. Project Proposer**

|  |  |  |  |
| --- | --- | --- | --- |
| **Name Last-named** | **Assist. Prof. Dr. Saed ALQARALEH** | **E-mail** | **saed.alqaraleh@hku.edu.tr** |

**Part II. Project Information**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Starting Term** | |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | | 2 | 0 | 1 | 9 | / | 2 | 0 | 2 | 0 | |
| **Title of the Project** | **Topic-Specific Web Crawler** |
| **Project Description** | |
| Crawler, which is also known as a spider or robot, is one of the main parts of the search engines. It is a software that visits the websites in a routinely manner. Crawlers are responsible to find new web objects, such as new webpages, multimedia files, articles, etc., and to observe changes in previously indexed web objects.  This project aims at implementing a new efficient Topic-specific crawler that supports the Turkish language and providing a solution to the Web Crawler efficiency problems. The proposal consists of building, first, a watcher module, which can be uploaded to the website's server, prepares a report which contains the address of web pages that have been modified or added to the website. The watcher file allows crawlers to visit only the updated pages and it helps to solve crawlers overlapping and communication problems, and second, a Server module, which will be split into multiple units, where each unit is responsible for performing a specific crawling process. | |
| **Project Justification** | |
| **Novelty** | |
| **New aspects** | Building a Topic-specific crawler that supports the Turkish language |
| **Complexity** | |
| **Challenging problem and issues** | Executing multiple crawling tasks in parallel efficiently is a critical issue for the project. |
| **Related computer science fields and subfields** | Information retrieval, Data mining. |
| **Tools** | Python. |
| **Risk involved** | |
| **Potential problems and alternative solutions** | Implementing multi units may require more duration, to avoid such problem tasks will be divided between the group members, which allows implementing multiple parts at the same time. |
| **Minimum work required** | 4 MONTHS |