BIRZEIT UNIVERSITY
Electrical and Computer Engineering Department

ENCS539: Special Topics: "Information Retrieval and Web Search"

Major Assignment 2 (Course Project): Instructor: Dr. Adnan H. Yahya,

Selection Deadline (team and topic): October 15, 2021. Due date: January 20, 2021.

Project Discussion: January 22, 2021

This is the start of your main (and last major) assignment for the course. It is a project on one of the topics below, with the last choice being something you wanted to do and never had the chance and is related to Information Retrieval, Web Search and Arabic NLP. The first step is to select a topic by the given deadline. Teams of 2 students is the default. But you can work in groups of 3 but you will be expected to deliver a better product.

The following conditions apply:

- The topic selected cannot be something you already worked on in the past, cannot be another course project and cannot be your graduation project (no double credit principle). In all projects, Arabic NLP elements are strongly encouraged (and expected).
- You can work in groups of 2 students and in certain cases 3, but the amount of work needs to be proportional to the number of partners in the group.
- You need to be able to demonstrate the project to the instructor.
- You need to submit a full report (one per group), 6 pages max, that reports on the theory and results of the project, full with references and citations. The report should be prepared using IEEE style (Templates can be found at: https://www.ieee.org/conferences/publishing/templates.html)
- Timeline: Selection Deadline (team and topic): October 25, 2021. Project and report Due date: January 20, 2021.. Project Discussion: January 22, 2021.

Possible Projects:

1- **Fake Photo/Video detection**: given the title/description of a piece of info, detect if this really is a correct description of the piece.
2- **Search Results Clustering:** Given a collection of ambiguous search results for an ambiguous query, devise a model to cluster these results into the needed subclasses:
3- **Pairing Arabic and English Clauses:** Can you rank and evaluate short sentences in both languages (Arabic and English) so that a search on an Arabic clause finds the corresponding English statement and vice versa. Can you use WordNet/Wikipedia for that? Could be video captions

4- **AutoCompletion system**: at the word and phrase level: based on a language model or collected queries with frequencies: one may try different granularity of prediction. It needs also to learn from correct/incorrect predictions (be adaptive).
5- **Build and evaluate recommender system:** Allow multiple users to access a server and rate items based on their preferences (e.g. books, movies, music, Web pages, etc.). Based on ratings of other **similar users** create dynamic recommendations for the current user of the system. Many different variations of this idea is possible.
6- **Arabic Question Answering:** Given a limited type of questions and a collection of web documents devise an algorithm to answer the query from the most relevant document. The algorithm may very well be based on NLP tasks like NER plus other material (ML or Rule based,…). You may also want to rephrase the query to make sure the user agrees with your interpretation of the query.
7- **Scraping or Harvesting Product descriptions/prices and dates of validity from social media posts like Facebook or similar**: from Arabic documents/websites you need to identify certain types of data and get it into a table.
8- **Scraping or Harvesting Product Descriptions/prices and dates of validity from Web sites of different structure types.** from Arabic websites you need to identify certain types of data and get it into a table. Example product names, prices and validity of these prices.
9- **Authorship analysis through style comparisons:** Given a set of authors and a collection of articles for each author: devise a mechanism to compare unknown documents for similarity in writing style to each of the given authors. Use the mechanism to detect if the styles of two documents are close or distant. The documents can be short or long and the work can be in Arabic or English.

10- **Your Suggested Project: must be with IR and NLP elements and disconnected from anything you do for other courses. It has to be defined through a one page (100 words) abstract, the soonest.**

Good Luck