HomeWork-3 Report Template

1. **Crawling [Note down steps implemented for each of the below]**

    a. **URL Canonicalization –**
    if link starts with / or ../, get "base/parent" and join links. Parse the url and remove # and : parts of the link (ports) and then combine all the components of the url again.
    b. **Frontier Management**
    Frontier is implemented on a priority queue. Term list stores keywords that are related to the topic. Calculating the score: + 5 for each word in term list and link <a> tag, +3 for each keyword in title, + 1 for each inlink, + 35 if the link ends in .gov or .edu. (50000 - score) is the ranking for the priority queue since queue takes lower numbers first. Then it continues to get the next element from queue until 10000 docs are reached.
    c. **Politeness Policy**
    before crawling the document, using urlrobot.RobotFileParser(), get "url + /robots.txt" and read it to find the crawl delay. If there is none, sleep for 1 second. If it is given, sleep for however many seconds it defines it as.
    d. **Document Processing**
    using BeautifulSoup, get the text, and using langdetect library, detect if the text is in english. If it is, continue and get the outlinks and inlinks of the document. This is done by finding all instance of <a> and canonicalizing them. Then process the document by getting the title (if it has) and all instances of <p> tags. The doc is stored in the same format as ap89 collection.

2. **Vertical Search**
    a. **Add a Screenshot of your Vertical Search UI**
    b. **Explain briefly how you implemented it.**

3. **Extra Credits Done [Note done what was done for each extra credit]**