**Information Retrieval – Homework1 (Build a web crawler)**

**2) Focused Crawling**

Key word variation process:

* Collected all the words after the “/wiki/” pattern by striping the string (URL) and the anchor text
* For the words collected, special characters were removed (Non-English and numbers) and replaced with a space
* For the given keyword or the set of keywords, checked if the keyword/set of keywords are present in the words collected
* If yes, then the words are split (By removing the keyword from the word), leaving the parts before the keyword and after.
* Using the PyEnchant library, the parts are checked if they are valid English words
  + For example,
  + Reference – Piazza
  + Keyword = “Moon”
  + Words = “fumoon” and “fullmoon”
  + Check if PyEnchant consist of “fumoon” and “fullmoon”
  + “fumoon” will be False and “fullmoon” will be True
  + Split “fumoon” into “fu” and “moon”, check PyEnchant again for both the words
  + Split “fullmoon” into “full” and “moon”, check PyEnchant again for both the words
* If all the above conditions are true, then the words are considered valid (Also, the URL is considered valid to be crawled)
* List of valid URLs are used for the BFS crawl