Java 1.7 encoding:UTF-8

Main Class

* Done
  + Add methods to report everything needed in report
* TODO
  + Add crawler method to traverse each file
  + 1. For each URL in frontier, need to save it to a file, or read it
  + 2. Call Storage class to rip all tokens out of file
  + 3. At the end of calling all URL with crawler, need to summarise report

Storage Class

* Done
  + Added objects for each necessary part of project
* TODO
  + Improve code to:
    - Count longest page
    - Process subdomains
    - Process word frequenicies
    - Process 3-grams
  + Add file writer object to print to file
    - Needed for subdomains
    - Needed for answers.txt
  + Keep full vocabuluary
  + How to store each document? large files?
    - How to parse?
* Anchor text?

TextProcessor

* Done:
  + Results from each file are now cumulative
* TODO:
  + Ignore stop words when processing token list
    - Do this before computing word frequency and 3 grams
  + Modify token list method to also count tokens (required for page length)

Indexer

* Doc id to file indexer
  + How to read and write to a storage file
* Token to file indexer
  + Refers to document id
  + Very basic for now, just each token stores its own list of doc id