**REPORT**

**2A Crawling Topic, Seeds, Related words**

|  |  |
| --- | --- |
| **Topic** | Apple Corporation  Looking for pages related to Apple company |
| **SEEDS** | https://en.wikipedia.org/wiki/Apple\_Inc. |
| https://en.wikipedia.org/wiki/App\_Store\_(iOS) |
| **Related Words** | steve jobs |
| ipod |
| ipad |
| iphone |
| siri |
| iTunes |
| Macintosh |
| Tim Cook |
| icloud |
| apple |

**2B Crawler implementation**

|  |  |  |  |
| --- | --- | --- | --- |
| FUNCTION NAME | INPUT | OUTPUT | INFORMATION |
| getURLTEXT | String URLseed | StringBuilder URLText | Returns the source code of the given URL in form of text |
| extractURLS | StringBuilder URLText | LIST<String> URLS | Returns list of all the URLs present in the URLText |
| hasRelatedTerms | String URLText  StringArray relatedTerms  String URL | Boolean result | Returns true if at least 2 of the related words are present in the URLText |
| getURL | NA | String URL | Returns the top Queue element if it is not present in the Visited URL list |
| addURL | LIST<String> URLs | void | Adds the list of URLS to the queue |
| isMarked | String URL | Boolean result | Returns true if the URL is present in Visited URLs List else returns false |
| addtoList | String URL | Void | Add the given URL to visited URLs list |
| addRelatedwords | String relatedWord | Void | Add the given word to related words list |
| getRelatedWords |  | String Array | Returns the list of related words in form of Array |
| loadPropertyfile | NA | Void | Loads the Seeds and Related words from the Config.properties file to the list. |
| markedcount | NA | Integer count | Returns the number of visited URLS |
| writetoFile | NA | Void | Creates a new file with name VisitedURLS(currentTimeStamp).txt and writes all the crawled URLS to the file. |
| saveCrawledPage | String URLText | Void | Creates a new HTML page by writing the content of URLText to the file. |

**Implementation Algorithm:**

Step1 : Load property file **loadPropertyfile()** **NOTE**: Two seed URL are added to the Queue

Step2 : IF **markedcount()** < 500 continue ELSE go to Step7

Step3 : *seedURL* = **getURL()**

Step4 : *urlText* = **getURLTEXT(***seedURL***)**

Step5 : **hasRelatedTerms(** *urlText* ***,*** *relatedWordsList***)**

Step6 : if **Step5** = **True** then Add seedURL to list **addtolist(***seedurl***)** and **SaveCrawledPage()**

Step7 : Add the extracted URLs to the queue **addURL(***urlText***)**

Step8 : Write the crawled URLs to the file **writetoFile()**

**Number of Pages Crawled:**

The crawler successfully crawled 500 pages.

**URLS of all crawled pages:**

Please find the URLs of all crawled pages in the attached text file.

