READ ME file

* Install python 2.7
* Unzip the project folder Task3
* This folder contains 5 python files

1. File\_Convertor.py
2. file\_handler.py
3. Inverted\_Indexer\_unigram.py
4. Inverted\_Indexer\_bigram.py
5. Inverted\_Indexer\_trigram.py

And a txt file i.e URLS.txt

* Run the file\_convertor.py and the parsed articles are stored in separate txt files in **ParsedCorpus** folder. The processing done includes removing punctuation (except hyphen), case folding, and retaining punctuation within digits.
* Unprocessed HTML files are stored in Corpus folder.
* Run Inverted\_Indexer\_unigram.py to get the unigram index and unigram term frequency doc and unigram doc frequency doc.
* Run Inverted\_Indexer\_bigram.py to get the bigram index and bigram term frequency doc and bigram doc frequency doc.
* Run Inverted\_Indexer\_trigram.py to get the trigram index and trigram term frequency doc and trigram doc frequency doc.

Output

|  |  |  |  |
| --- | --- | --- | --- |
|  | Word->(docId, tf) | Word-> (docId1,docId2,…,df) | Word – >tf(in the entire corpus) |
| Unigram | unigram\_index.txt | unigram\_docfrq\_table | unigram\_trmfrq\_table |
| Bigram | bigram\_index.txt | bigram\_docfrq\_table | bigram\_trmfrq\_table |
| Trigram | trigram\_index.txt | trigram\_docfrq\_table | trigram\_trmfrq\_table |
| Format  to read  This list | Term , {doc\_id,term count} | Term | Term count | Term, doc1,doc2,doc3(list of docs containing this term separated by comma), Count of doc containing the term(number after last comma in the line) |