# READ ME file

* Install python 2.7
* Copy the project folder HW3
* Separate files for unigram, bigram and trigram

1. Task2\_unigram for unigram tables
2. Task2\_bigram for bigram tables
3. Task2\_trigram for trigram tables

* Run the Task1.py and the parsed articles are stored in separate txt files in **Parsed** folder. The processing done includes removing punctuation (except hyphen), case folding, and retaining punctuation within digits.
* Next the text is split into terms and the term frequency of the term for a particular document is obtained. It is stored in “unigram\_inverted\_list.txt” in the following format: Word-> (df, tf)
* Along with this a “unigram\_df\_table.txt” file is created which has the list of documents that has that particular word and also the number of documents containing the word
* “unigram\_tf\_table.txt” is created that has the terms and the total frequency of that particular term in the entire corpus.
* For bigram and trigram, the entire process is repeated. The only difference being we add an extra step that is making bigram and trigram.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Word-> (docId, tf) | Word-> (docId1, docId2,…, df) | Word – > tf (in the entire corpus) |
| Unigram | unigram\_inverted\_list.txt | unigram\_df\_table.txt | unigram\_tf\_table.txt |
| Bigram | bigram\_inverted\_list.txt | bigram\_df\_table.txt | bigram\_tf\_table.txt |
| Trigram | trigram\_inverted\_list.txt | trigram\_df\_table.txt | trigram\_tf\_table.txt |