Modifications done for Hindi Index and Search

1. Replace NLTK Stopwords with a custom list of Stopwords
    - A third party list was used, which can be further enhanced

2. Replace NLTK Stemmer with custom code for stemming of tokens

3. Change the way offsets are getting computed and stored - Instead of first storing the data in a variable and writing it to a file in a batch mode, write each line and then get the offset from f.tell() where f is the filestream pointer. This makes the process of file writing a little slower, however, since UTF-8 encoding results in variable bytes per character, it solves the problem.

4. Remove the ASCII coding and decoding done to remove characters from other languages for English Wikipedia -for both storing title of the page and the text processing. However, to limit the characters to Hindi language only – a check on Unicode range was done to remove characters from Chinese, Arabic and other languages.

5. Performance optimizations can be done –
    - Since the number of pages was small 0.284 million pages (as compared to English Wikipedia which was more than 21.3 million pages), pageStats is not split into smaller files. This can be done for incremental gains in performance.
    - Vocabulary has not been split into smaller files – first character based splitting was done for English Wikipedia