Stemming a great deal of language-dependent linguistic

ASSIGN BUSTER

Stemming techniques are used to find the root/stem of a word.

Stemming converts words to their stems, which incorporates a great deal of language-dependent linguistic knowledge. Behind stemming, the hypothesis is that words with the same stem or word root mostly describe same or relatively close concepts in text and so words can be conflated by using stems. For example, words; affect, affected, affection, affects all can be stemmed from the word ‘ affect’. affect affected affection affects Stemming affect Figure 12: Stemming Process (Vijayarani, Ilamathi, and Nithya 2015)There are many kinds of algorithms used for stemming purposes such as; Lovins stemmer, porter’s stemmer, Paice/Husk stemmer, Dawson stemmer, N-gram stemmer, HMM stemmer, YASS stemmer, and many others (Vijayarani, Ilamathi, and Nithya 2015).

In this research, we used Martin Porter’s Stemmer algorithm (Porter, M. F., 1980) (Porter, M. F., 2001), which is the most commonly used algorithm for English language proposed in the ’80s.

Many modifications and enhancements have been made and suggested on the basic algorithm. It is based on the idea that the suffixes in the English language (approximately 1200) are mostly made up of a grouping of smaller and simpler suffixes. It has five steps, and within each step, rules are applied until one of them passes the conditions. If a rule is accepted, the suffix is removed consequently, and the next step is performed. The resultant stem at the end of the fifth step is returned. The rule looks like the following (Vijayarani, Ilamathi, and Nithya 2015): For example, a rule (m > 0) AAD EE means “ if the words have at least one vowel and consonant

plus AAD ending, change the ending to AA.” So, “ argued” becomes “ argue” while “ feed” remains unchanged.

This algorithm has about 60 rules and very easy to understand. Porter designed a detailed framework of stemming which is known as ‘ Snowball.’ The primary purpose of the framework is to allow programmers to develop their stemmers for other character sets or language. The “ tm” package provides the stemDocument() function to get to a word’s root. This function either takes in a character vector and returns a character vector, or takes in a plain text document and returns a plain text document.

Stemming uses an algorithm that removes common word endings for English words, such as “ es,” “ ed,” and “ s.”· Strip white space After applying all the above mentioned pre-processing techniques, our corpus would have many white spaces which need to be eliminated from the corpus. This is also a part of text preprocessing in which tm_map () function with stripWhitespace () removes the extra white space in the corpus. Figure 13 presents the sample screenshot of the terms/words after pre-processing stages. There a total of 21, 241 terms obtained from 853 documents after applying preprocessing stages.