Task: Read theoretically about extracting just pronouns in texts using Perl

Marwan M. Al Omari

Lebanese University

Before starting the search for the way to extract pronouns, I have made some changes in the file that we worked on in classroom. First, adverbs, for example, when they start at the beginning of sentences, the code that we created does not exclude them. Therefore, I added the top used adverbs in English language and the result has changed accordingly. Next, I added also an extended list of words that have not been included in the stopwords.txt. The output's accuracy increased since I made the two steps. However, it is still missing few important things that sill effect the result. For instance, a proper noun may consist of two part as in The White House, and the code will classify it as two separate proper nouns. In addition, the title of text will be also considered as proper nouns according to the code, which is not right! From these points, I start searching for a way just to extract proper nouns for any text.

As the search starts for a way to extract just proper nouns from texts, I come across modules that could be used to help Perl coding, sharpening its accuracy. To let the computer understands the language, parsing modules are important to use. Lingua-LinkParser, Lingua-Stem, link-grammar, Lingua-EN-Tagger and Lingua-EN-NamedEntity. Concerning proper nouns, Lingua-EN-NamedEntity is the one that would help in our goal. "'"Named entities" is the NLP jargon for proper nouns which represent people, places, organizations, and so on. This module provides a very simple way of extracting these from a text" ("CPAN RT", n.d.).

Therefore, I did download the previous modules to start working on the texts, testing its accuracy. However, I am still encountering some problems in how to code the module further than writing **use**, i.e., "Lingua::EN::NamedEntity". Because I did not find a guidance to install or to deal with these modules, I went back working on Stopwords. I did add a dictionary of words. "A list of 109582 English words compiled and corrected in 1991 from lists obtained from the Interociter bulletin board, and "this word list includes inflected forms, such as plural nouns and the -s, -ed and -ing forms of verbs. Thus the number of lexical stems represented in the list is considerably smaller than the total number of words" ("English Wordlists", n.d.). Thus, the result has been sharped to meet at minimum requirements.

In conclusion, the coding files can meet the goal of user but no 100% accurate. By the use of modules that understand the sentence structure as Lingua::NamedEntity, the result would score a higher accuracy.

References

CPAN RT. (n.d.). Retrieved December 30, 2017, from http://search.cpan.org/dist/Lingua-EN-NamedEntity/NamedEntity.pm

English Wordlists. (n.d.). Retrieved January 01, 2018, from http://www-01.sil.org/linguistics/wordlists/english/