Natural Language Processing

Project Report

Project 1: Corpora construction and tokenization

Group 2

1. Project Requirement

* Crawl text data from website.
* Split text data into sentences and words.

1. Our Solution

* Crawl text data
* Split text data into sentences and words
* Some observations:
  + A sentence usually end with one of following punctuation: full stop, exclamation mark or question mark. After that, a white space is placed. And the first letter in new sentence is upper case.
  + A word contains one or several letters place sequential.
* From above observations, we find ending position of sentence in crawled text data and split the text data into individual sentences. With each sentences we extract words by find group of several letters place sequential.
* For implementation, we use JAVA language and regular expression for detect end of sentence and non-word element in crawled text data.
  + Regex for detect end of sentence: **(?<=[.?!])\\s+(?=[\"0-9\\p{Lu}])**
  + Regex for detect non-word elements: **\\P{L}+**
* See our repo at <https://github.com/minhle2994/NLP-Tokenization> for more details in implementation.
* Result:
  + Our crawled file: 3.8MB text file
  + Number of split sentences: 19880
  + Number of split words: 648852