LIN570: HW2 – sbd (100pts)

Yao Yan (1669598)

Due date: 11pm on Oct 15, 2018 (Tuesday)

1. using SPLITTA[1] obtain SBD results:
   - `python2 sbd.py file.txt > file.sbd.system`
2. Implement a script to calculate a F_1 score `f1_score_sbd.sh` for sentence boundary detection
   - The command line is: `cat file.sbd.system | ./f1_score_sbd.sh file.sbd.gold > file.sbd.score`
   - In my solution, the first word in the sentence is represented by "s", the others using "o". The same conversion is applied to both input file and gold file. By comparing the alignment of "s", the F1 score for sentence boundary detection can be calculated.
     e.g.
     "This is sentence1."
     "This is sentence2."
     will be converted to "soosoo"
     "This is sentence1. This is sentence2."
     will be converted to "sooooo"
3. Modify script to calucate a F_1 score `f1_score_tok.sh` for tokenization
   - This method builds on the assumption that both input file and gold file have the same amount characters.
     The starting letter of a token is represented by "t" others by "o"
     e.g.
     input: " 1 , 214 cars in the U . S ."
     will be represented as "tttootooototootttt"
     gold: " 1,214 cars in the U.S."
     will be represented as "tooootooototootooo"
     Then extract the index of t. "tooootooototootooo" will be [0,5,9,11,14] and then create a representation for each token[0-5, 5-9,9-11,11-14]. By converting the input and gold text into [#-#,#-#,...] format and comparing the the converted list, precision and recall for tokenization can be calculated

[1] `https://github.com/lukeorland/splitta`