LIN570: HW2 – sbd (100pts)

YOUR NAME (UW NetID)

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

For this homework, you are going to perform sentence boundary detection and evaluate its results using a F_1 score. You will also evaluate tokenization results using a F_1 score. All the sample files are under `~/dropbox/19-20/570/hw2/examples`.

Rubric:

2pts `hw.tar.gz` submitted, it should contain following files:

- `file.sbd.system`
- `file.tok.system`
- `f1_score_sbd.sh`
- `f1_score_tok.sh`
- `file.sbd.score`
- `file.tok.score`

2pts `readme.txt` or `readme.pdf` submitted

6pts All files and folders are present in expected locations

10pts Programs run to completion

5pts The output of programs on patas match submitted output

1. (**10pts**) using SPLITTA[1] obtain SBD results:
   - `python2 sbd.py file.txt > file.sbd.system`
2. (**40pts**) 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`
   - Minimum in-line comments should be provided.
3. (**25pts**) Modify your script to calucate a F_1 score `f1_score_tok.sh` for tokenization
   - The command line is:
     - `cat file.sbd.system | ./eng_tokenizer.sh abbrev_list > file.tok.system` (`eng_tokenizer.sh` and `abbrev_list` are from HW1)

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

  - `cat file.tok.system | ./f1_score_tok.sh file.tok.gold > file.tok.score`
- Minimum in-line comments should be provided.

	system	gold	
The	S-SENT	S-SENT	← tp
luxury	O	O	
...	...	...	
year	O	O	
sold	O	O	
1,214	O	O	
cars	O	O	
in	O	O	
the	O	O	
U.S.	S-SENT	O	← fp
Howard	O	S-SENT	← fn
Mosher	O	O	
,			
...			

- See also `conlleval`[2] for the F_1 score used at the CoNLL-2000 shared task data (Chunking).
- parentheses, brackets, etc in the Penn treebank (`file.tok.gold`):

```
# s/(/-LRB-/g
# s/)/-RRB-/g
# s/\[/-LSB-/g
# s/\]/-RSB-/g
# s/{/-LCB-/g
# s/}/-RCB-/g
```

-   - raw:

      ```
      "From the beginning, it took a man with extraordinary qualities to succeed in Mexico," ..
      ```

    - tokenized:

      ```
      `` From the beginning , it took a man with extraordinary qualities to succeed in Mexico ,
      ```

[2] https://www.clips.uantwerpen.be/conll2000/chunking/output.html