LIN570: HW1 – tokenization (100pts)

Yao Yan (1669598)

Due date: 11pm on Oct 8, 2019 (Tuesday)

For this homework, you are going to write an English tokenizer and its evaluation script from the input text. All the sample files are under `~/dropbox/19-20/570/hw1/examples`.

Rubric:

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

- `eng_tokenizer.sh`
- `abbrev_list`
- `file.tok.system`
- `eng_tokenizer_eval.sh`
- `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. (**25pts**) Implement an English tokenizer using regular expressions and the exception list, `eng_tokenizer.sh`
   - The command line is: `cat file.txt | ./eng_tokenizer.sh abbrev_list > file.tok.system`
   - Minimum in-line comments should be provided.
2. (**10pts**) Calculate LD between *execution* and *intention* by completing the following table of the minimum edit distance algorithm:

	#	I	N	T	E	N	T	I	O	N
#	0	1	2	3	4	5	6	7	8	9
E	1	1	2	3	3	4	5	6	7	8
X	2	2	2	3	4	4	5	6	7	8
E	3	3	3	3	3	4	5	6	7	8
C	4	4	4	4	4	4	5	6	7	8
U	5	5	5	5	5	5	5	6	7	8
T	6	6	6	5	6	6	5	6	7	8
I	7	6	7	6	6	7	6	5	6	7
O	8	7	7	7	7	7	7	6	5	6
N	9	8	7	8	8	7	8	7	6	5

3. (**40pts**) Implement an English tokenizer evaluator using the minimum edit distance algorithm, `eng_tokenizer_eval.sh`

   - The command line is: `cat file.tok.system | ./eng_tokenizer_eval.sh file.tok.gold > file.tok.score`
   - Minimum in-line comments should be provided.