W5 Assignment – Tokenization

Computational Linguistics

John Gamboa

April 22, 2022

1 Tokens and Types

1.1 Definitions

Select the correct option:

a) The sentence

```
my cellphone and your cellphone
```

contains 4 tokens and 5 types

b) The sentence

```
my cellphone and your cellphone
```

contains 5 tokens and 5 types

c) The sentence

```
my cellphone and your cellphone
```

contains 5 tokens but only 4 types

d) The sentence

```
my cellphone and your cellphone
```

contains 5 tokens but only 3 types

answer: _____

1.2 True/False

Answer *TRUE* or *FALSE* for the following assertions:

- The text

  ```
  my Cellphone and your cellphone
  ```

  contains 5 tokens and, debatably, either 4 or 5 types

  answer: _____

- The text

  ```
  my, your, and their cellphone
  ```

  contains 7 tokens and 6 types

  answer: _____

- The text

  ```
  Rosa's cellphone is blue
  ```

  contains, debatably, either 4 tokens and 4 types or 5 tokens and 5 types

  answer: _____

2 Text manipulation / Regular Expressions

2.1 Instructions

Consider the following lines, which are run before the lines in the "Questions" section below.

```
import re
string = "The man and the woman...the the... the dog and the cat"
```

In the gap texts below, write the output of the code immediately preceding it. For example, for the code

```
print('example')
```

you should write

```
example
```

(i.e., without the quotes that denote a string in Python)

Ideally, you should run those lines and try to see why you get each of the outputs. If you have difficulties in understanding the meaning of the Regular Expressions, remember to visit those websites I suggested in the videos.

2.2 Questions

```
if(re.search('man', string):
    print('contains')
```

answer: _____

```
re.sub('man', 'men', string)
```

answer: _____

```
re.sub('the', 'a', string)
```

answer: _____

```
re.sub('[a-z]', 'a', string)
```

answer: _____

```
re.sub('[^a-z]', '_', string)
```

answer: _____

```
re.sub('\W+', '_', string)
```

answer: _____

```
re.sub('(the)', '_\\1_', string)
```

answer: _____

```
re.sub('(the|man)', '_\\1_', string)
```

answer: _____