

NLP Course 2 Week 1 Lesson : Building The Model - Lecture Exercise 01


Estimated Time: 10 minutes

Vocabulary Creation

Create a tiny vocabulary from a tiny corpus
It's time to start small !

Imports and Data

```
In [1]: # imports
import re # regular expression library; for tokenization of words
from collections import Counter # collections library; counter: dict subclass for counting hashable objects
import matplotlib.pyplot as plt # for data visualization
```

```
In [9]: # the tiny corpus of text !
text = 'red pink pink blue blue yellow ORANGE BLUE BLUE PINK wee wee' #

print(text)
print('string length : ',len(text))
```

```
red pink pink blue blue yellow ORANGE BLUE BLUE PINK wee wee
string length : 60
```

Preprocessing

```
In [11]: # convert all letters to lower case
text_lowercase = text.lower()
print(text_lowercase)
print('string length : ',len(text_lowercase))
```

```
red pink pink blue blue yellow orange blue blue pink wee wee
string length : 60
```

```
In [12]: # some regex to tokenize the string to words and return them in a list
words = re.findall(r'\w+', text_lowercase)
print(words)
print('count : ',len(words))
```

```
['red', 'pink', 'pink', 'blue', 'blue', 'yellow', 'orange', 'blue', 'blue', 'pink', 'wee', 'wee']
count : 12
```

