Text Analysis

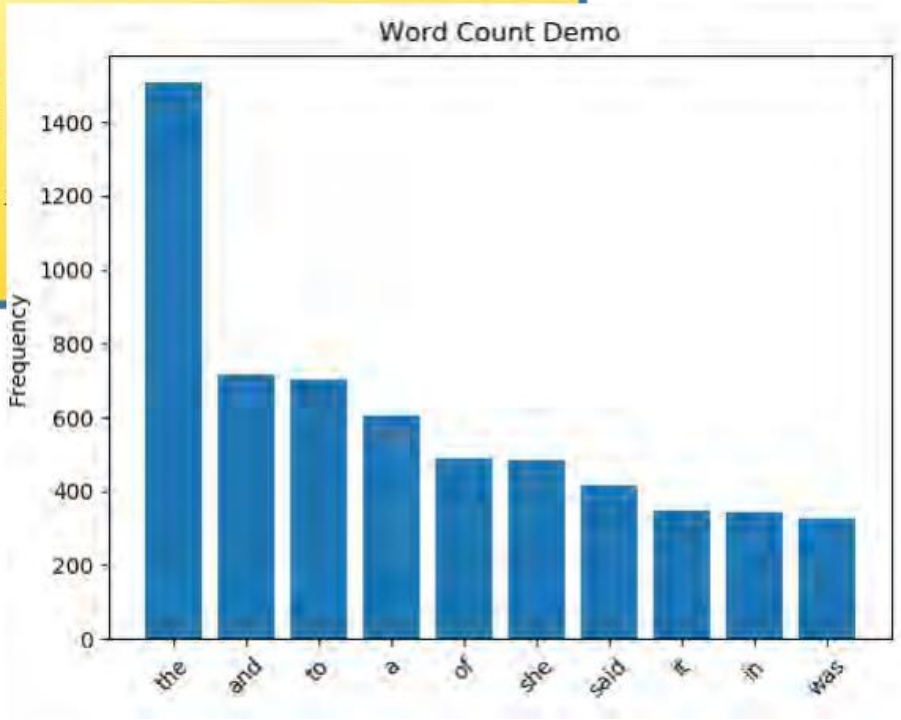
Text Analysis – word frequency

```
# A program to visualise the most common words in a file
from matplotlib import pyplot as plt
from collections import Counter

# IMPORTANT: Make sure book.txt exists in runtime directory
bookFile = open("book.txt","r") # Open the file
text = bookFile.read() # read the file
bookFile.close() # close the file
text_list = text.split() # create a list

# use counter to return the most common words
# format is .... [('the', 1507), ('and', 714), etc
most_common_words = Counter(text_list).most_common(10)

words = [] # an empty list of words
word_count = [] # an empty list of counts

# Build up the lists
for word, count in most_common_words:
    words.append(word) # append the word to the words list
    word_count.append(count)

# Now create and display the chart ….
```

Text Analysis – word frequency

… continued from previous slide

```
# Now create and display the chart ….

# Create the chart
plt.bar(words, word_count)
plt.title("Word Count Demo") # graph title
plt.ylabel("Frequency") # label the y-axis
# put the words on the x-axis
plt.xticks(range(len(words)), words, rotat
plt.show() # display the chart
```



Regular Expressions

A language that enables us to look for patterns in strings

```
import re

text1 = "THERE are 99 RED balloons"
print(re.sub('[0-9]', '', text1)) # remove digits
print(re.sub('[A-Z]', '', text1)) # remove uppercase
print(re.sub('[A-Z0-9]', '', text1)) # remove uppercase and digits
print(re.sub('[^a-z]', '', text1)) # leave lowercase
print(re.sub('[^a-zA-Z ]', '', text1)) # leave letters and spaces
print(re.sub('[^a-zA-Z0-9]', ' ', text1)) # leave letters and digits
print(re.sub(r'\b\w{1,4}\b', '', text1)) # remove words of length 1-3

text1 = "$%^$% joe ^&$%^&"
print(re.sub('[^a-zA-Z0-9]', '', text1))
```

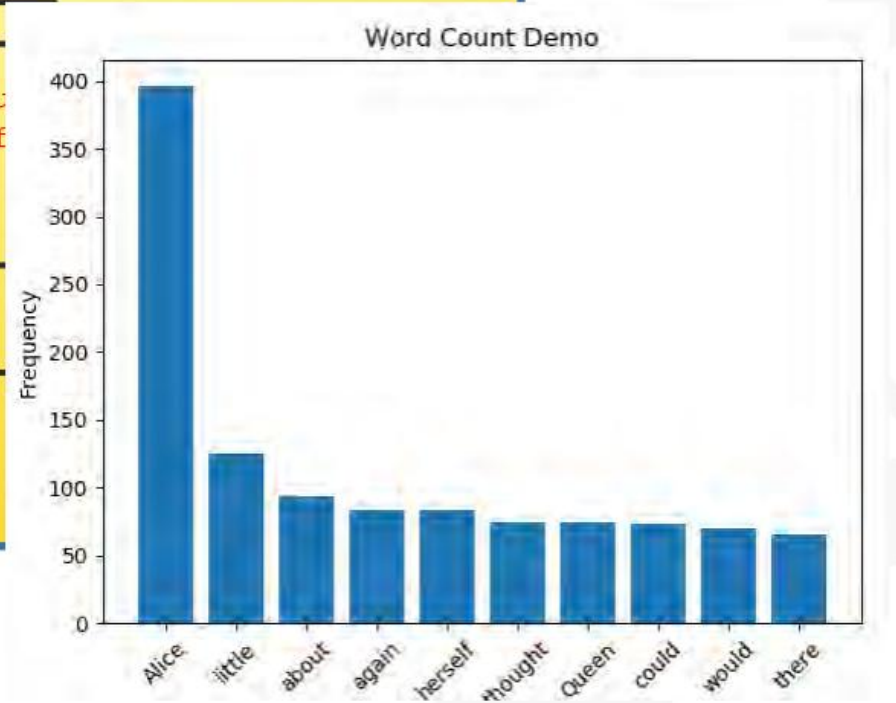
Output

```
THERE are  RED balloons
 are 99  balloons
 are   balloons
areballoons
THERE are  RED balloons
THERE are 99 RED balloons
THERE    balloons


joe
```

Text Analysis – word frequency

Eliminate words of three letters or less … use Regular Expressions

```
# A program to visualise the most common words in a file
from matplotlib import pyplot as plt
from collections import Counter
import re

# IMPORTANT: Make sure book.txt exists in ru
bookFile = open("book.txt","r") # Open the f
text = bookFile.read() # read the file
bookFile.close() # close the file

text = re.sub('[^a-zA-Z0-9 \n]', ' ', text)
text = re.sub(r'\b\w{1,4}\b', '', text)

text_list = text.split() # create a list

# Continue as before …
```



Import the `re` library

Use the `sub` method

Readlines

https://www.w3schools.com/python/ref_file_readlines.asp