# Assignment 9: NLP/NLTK
# Bhargavanarasimhan Lakshminarasimhan – G01093525

***Using Python and the NLTK, extract a word-count summary and graph of the most frequent types of biking injuries described in the dataset.***

## i) Before Data cleansing

**Frequency distribution summary**
[('.', 1043), (',', 923), ('and', 788), ("''", 586), ('my', 453), ('on', 434), ('I', 403), ('the', 388), ('left', 348), ('to', 338), ('right', 297), ('``', 242), ('a', 235), ('of', 200), ('in', 171), ('was', 159), ('knee', 155), ('shoulder', 147), ('bruises', 117), ('(', 113)]

**Graph for most frequent types of biking injuries**

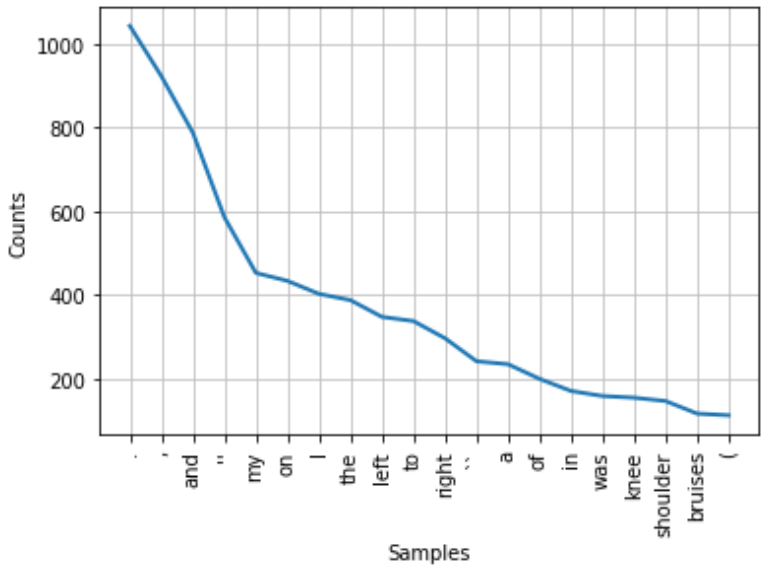


***Fig i: Bike Injury descriptions (Before cleansing)***

From the above figure, we can see that lots of punctuations occur frequently and this graph isn't particularly helpful.

**<u>After removing stop words, tokenizing the words.</u>**

**The word count summary is**
<FreqDist with 1747 samples and 8893 outcomes>
[('left', 375), ('right', 312), ('knee', 159), ('shoulder', 156), ('none', 151), ('bruises', 147), ('bruised', 119), ('broken', 115), ('bruising', 109), ('elbow', 104), ('back', 87), ('scrapes', 87), ('hip', 85), ('road', 78), ('side', 77), ('hand', 77), ('rash', 70), ('leg', 70), ('pain', 69), ('neck', 68), ('wrist', 68), ('arm', 67), ('abrasions', 65), ('concussion', 61), ('sore', 60)]

**Graph for 25 common injuries**

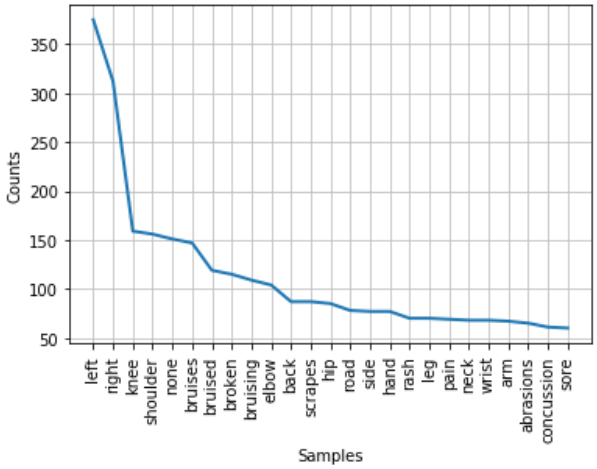


***Fig ii: Bike Injury descriptions (After cleansing)***

From above graph, we can see the types of injuries and the parts affected. This graph can be further cleansed by using Word lemmatization technique where the words like bruises, bruising, bruised will be combined together into one common issue.

### *ii) Display several injury description sentences using the sent_tokenize() function (10 points)*

```
#Use sent_tokenize to display several injury description sentences
sentences = nltk.sent_tokenize(lines)
print(sentences[0])
print(sentences[1])
print(sentences[5])
```

The injury description sentences are
"Concussion, head and neck pain"
"Concussion, dislocated shoulder, broken rib."
"I have deep tissue damage to my left side, and multiple bruises over my body.
If I had not been wearing a helmet my injuries would be much more severe.

### *iii) Describe how this analysis might be used to improve bike safety (10 points)*

This analysis of Bike injury is very explanative. The Bike injury.txt has all the common list of accidents that has happened during biking events/leisure biking. After analyzing it using NLTK package, we drew a graph mentioning the frequent injuries.

From the fig ii,

1) we can see the words like left, right knee, shoulder, leg, neck, wrist. They clearly state the common parts of the body where the injuries might occur. So, for biking, these areas are to be protected with safety gears in order to avoid injuries.

2) Next, the words like bruises, scratch, abrasions, broken elbow, concussions are also mentioned frequently. These state the types of injuries that occur. Using these words, the bikers along with doctors and health care people in hospitals or during events organization can be better equipped to deal these injuries and can arrange their facilities accordingly.