Justin Aguila 10/12/2023
AIT-580

M8 Lab 1A – Text Analysis with Python and NLTK

In this assignment, we conducted an in-depth analysis of textual narratives focused on biking accidents. We approached this analysis by utilizing Natural Language Processing (NLP) techniques to analyze textual descriptions of biking injuries.

Through this technique, we aim to unravel insights within the narratives and descriptions from this survey surrounding biking injuries. By using text analytics, we have the unique opportunity to identify not only what types of injuries are most common but also the circumstances. Our method includes keyword frequency analysis, word cloud visualization, and sentiment analysis through NLP libraries.

In the initial phase of our research, we utilized the Natural Language Toolkit (NLTK) to preprocess and analyze a dataset of textual descriptions of biking injuries. Specifically, our code was designed to remove commonly occurring words—known as 'stopwords'—that add little value to the analysis. Following this data cleansing, we identified and displayed the 25 most frequently mentioned types of injuries. This approach allows us to gain valuable insights into the most common hazards facing cyclists, thereby informing potential safety interventions.

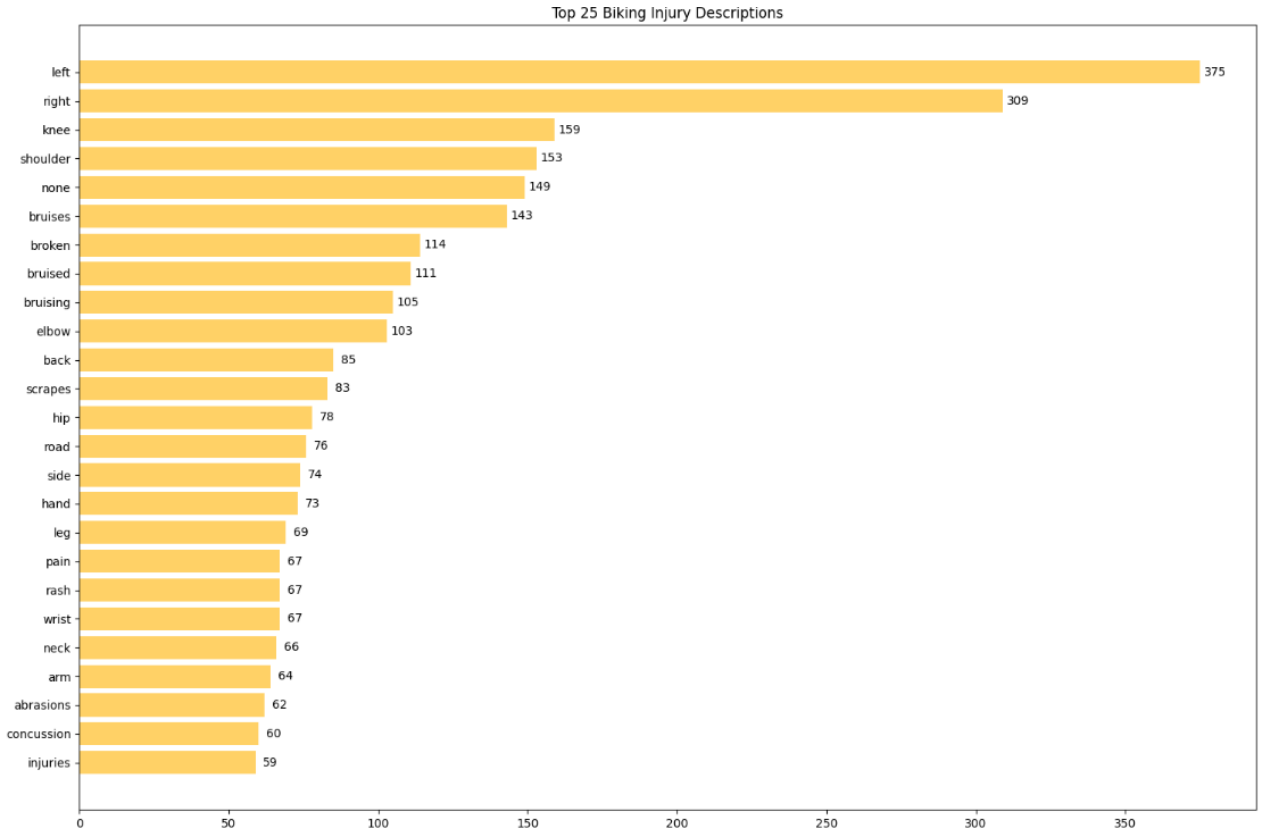

Figure 1. Frequency distribution of the most common areas and location of biking injuries.

Justin Aguila
AIT-580
10/12/2023

As shown in *Figure 1*, the frequency distribution of words describing bike injuries provides an important insight into the common types of injuries and body part locations of injuries. Notably, the bar chart shows that injuries related to the "left" and "right" sides of the body are the most frequently occurring, followed by "knee" and "shoulder" injuries. One thing to note here is that the word “none” also appears among the top terms. This could indicate that many reports may involve near-misses or incidents without injury.

Based on the results of our analysis, it becomes evident that certain body parts, such as the knees, shoulders, and elbows, are more frequently affected in biking accidents. This insight suggests that focusing on protective gear specifically designed for these vulnerable areas could be a highly effective strategy in mitigating the severity of injuries sustained while cycling. Furthermore, the presence of terms like "broken," "bruises," and "abrasions" highlights the types of injuries commonly sustained, potentially informing first aid training and emergency response protocols for biking accidents.

In *Figure 2*, we created an engaging word cloud visualization to offer a more visually intuitive representation of our key findings. This format not only adds an element of visual interest but also serves as a quick reference for identifying the most frequently occurring terms related to biking injuries. Additionally, this visualization serves as a lighthearted yet informative approach to displaying textual data, adding an element of engagement while still providing valuable insights into the most common terms associated with biking injuries.

Figure 2. Wordcloud of the most common areas and body areas of biking injuries.

Similarly, the most prominent words in the word cloud align with those identified in our frequency distribution analysis.

In conclusion, our analysis, leveraging Natural Language Processing techniques, has uncovered the most frequently mentioned types of injuries related to biking accidents within an extensive textual dataset. To reiterate, the data highlights specific body parts such as the "knee" and "shoulder," as well as the frequent mention of the "left" and "right" sides, suggesting these areas as particularly vulnerable in biking incidents. Based on these results, it would be prudent to develop or enhance protective gear aimed at these high-risk areas, and to incorporate these insights into educational programs on bike safety. While our analysis offers valuable information, further research could help to understand the context and severity of these injuries, thereby providing a more comprehensive picture of biking safety needs.