Module 6 Write-up

Natural language processing (NLP) provides many lifts to data analysis, but there are many flaws too, as seen by the NLP performed on the reddit comments. One issue that I noticed was with similar words with sightly different endings. An example of this was that many comments about fantasy football team names and players were spelled differently. Many puns were made using “pitt” because of Kyle Pitts; people using “pitti” versus “pitty” had an issue that could not be fixed by simple NLP. Words ending with “y” benefitted from stemming since the “y” was removed, but “i” at the end was not removed. Similarly, non-common words or proper nouns could be processed when they should not be. Tokenizing by words can also make connotation and meaning difficult to interpret since the layout of words, stems, and combinations of NLP procedures can remove parts of sentences that are important context to the meaning. The words that I chose were related to sports betting, and past tense versus present tense can be a huge factor in sports betting; stemming removes this important factor. NLP can be very helpful and important, but it is easy to see some glaring shortcomings.