NLP Project

Parsa Kangavari

June 2, 2023

1 Phase 1: Crawling And Cleaning Data

1.1 Step 1: Data Crawling

At first we should find websites that include match reports and player's ratings of a match in a single pages. The best website for this is SkySport. Then we should find and store urls of that pages. This urls has been crawled from a webpage that contains all matchs. For crawling we use Scrapy framework.

Results:

After that we scrap reports and player ratings from each urls in dataset above. Results:

1.2 Setp 2: Cleaning Data

For this section we have to clean datas that we crawled before. We should tokenize reports of matchs by sentences and words. For first one we split report by dots and for second one we split them by spaces. Then we should clean player ratings. player ratings are in range 0 to 10. We should set ratings true if they are bigger than 6 and set false otherwise.

tokenized by sentences:

tokenized by words:

1.3 Step 3: Metrics

For this section we should get most regular words, unique words of each reports and ... .

Results:

2 Phase 2: Model Implimentation