NLP Project

Parsa Kangavari

June 6, 2023

1 Phase 1: Crawling And Cleaning Data

Github Repository Link: Click here.

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. At first we should crawl all URLs from SkySport. So we should crawl urls from a page contains all urls. In this page there will be all urls of all matches. We should choose seasions that we want. In this project we have chosen Primier league, Champions league and Fifa World Cup.

All matches page link: Click here.

Then we have been crawled urls of pages contains reports from this page and save them as a csv file. We have developed a spider that do this. you can see it in MatchURLSpider.py.

Results:

	Unnamed: 0	url
0	0	https://www.skysports.com/football/burnley-vs-bournemouth/373467
1	1	https://www.skysports.com/football/crystal-palace-vs-west-bromwich-albion/373468
2	2	https://www.skysports.com/football/huddersfield-town-vs-arsenal/373469
3	3	https://www.skysports.com/football/liverpool-vs-brighton-and-hove-albion/373470
4	4	https://www.skysports.com/football/manchester-united-vs-watford/373471
5	5	https://www.skysports.com/football/newcastle-united-vs-chelsea/373472
6	6	https://www.skysports.com/football/southampton-vs-manchester-city/373473
7	7	https://www.skysports.com/football/swansea-city-vs-stoke-city/373474
8	8	https://www.skysports.com/football/tottenham-hotspur-vs-leicester-city/373475
9	9	https://www.skysports.com/football/west-ham-united-vs-everton/373476

After that we have crawld reports and player ratings from each urls in dataset above. This informations have been crawld by a spider that you can see in Soc-

cerSpider.py. In this spider we read urls CSV file and then crawl all informations from each url.

Results:

	Unnamed: 0	report	ratings
0	0		
1	1		
2	2		
3	3		
4	4		
5	5		
6	6		
7	7		
8	8		
9	9		

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:

	sent132	sent133	sent134	sent135	sent136	player0	rating0
0	¡PAD¿	¡PAD¿	¡PAD¿	¡PAD¿	¡PAD¿	McCarthy	True
1	¡PAD¿	¡PAD¿	¡PAD¿	¡PAD¿	¡PAD¿	Pope	True
2	¡PAD¿	¡PAD¿	¡PAD¿	¡PAD¿	¡PAD¿	Hennessey	False
3	¡PAD¿	¡PAD¿	¡PAD¿	¡PAD¿	¡PAD¿	Karius	True
4	¡PAD¿	¡PAD¿	¡PAD¿	¡PAD¿	¡PAD¿	Romero	False
5	¡PAD¿	¡PAD¿	¡PAD¿	¡PAD¿	¡PAD¿	Lossl	True
6	¡PAD¿	¡PAD¿	¡PAD¿	¡PAD¿	¡PAD¿	Fabianski	True
7	¡PAD¿	¡PAD¿	¡PAD¿	¡PAD¿	¡PAD¿	Dubravka	True
8	¡PAD¿	¡PAD¿	¡PAD¿	¡PAD¿	¡PAD¿	Lloris	True
9	¡PAD¿	¡PAD¿	¡PAD¿	¡PAD¿	¡PAD¿	Adrian	True
10	¡PAD¿	¡PAD¿	¡PAD¿	¡PAD¿	¡PAD¿	Fabianski	False

You can also see it on : Click here

tokenized by words:

	word2249	word2250	word2251	word2252	word2253	player0	rating0
0	¡PAD¿	¡PAD¿	¡PAD¿	¡PAD¿	¡PAD¿	McCarthy	True
1	¡PAD¿	¡PAD¿	¡PAD¿	¡PAD¿	¡PAD¿	Pope	True
2	¡PAD¿	¡PAD¿	¡PAD¿	¡PAD¿	¡PAD¿	Hennessey	False
3	¡PAD¿	¡PAD¿	¡PAD¿	¡PAD¿	¡PAD¿	Karius	True
4	¡PAD¿	¡PAD¿	¡PAD¿	¡PAD¿	¡PAD¿	Romero	False
5	¡PAD¿	¡PAD¿	¡PAD¿	¡PAD¿	¡PAD¿	Lossl	True
6	¡PAD¿	¡PAD¿	¡PAD¿	¡PAD¿	¡PAD¿	Fabianski	True
7	¡PAD¿	¡PAD¿	¡PAD¿	¡PAD¿	¡PAD¿	Dubravka	True
8	¡PAD¿	¡PAD¿	¡PAD¿	¡PAD¿	¡PAD¿	Lloris	True
9	¡PAD¿	¡PAD¿	¡PAD¿	¡PAD¿	¡PAD¿	Adrian	True
10	¡PAD¿	¡PAD¿	¡PAD¿	¡PAD¿	¡PAD¿	Fabianski	False

1.3 Step 3: Metrics

For this section we should get most regular words, unique words of each reports and ... . At first we should create new datasets that contains of sentences that belongs to positive and negative players. We have done it in PandNSeperation.py. In this script we seperate sentences belong to negative players and positive players and then save them in 2 seperated CSV file; negative and positives.csv. In this datasets, there are seperated sentences.

Results for negative:

	Unnamed: 0	sents
7	7	Solanke clincally finished a scintillating counter involving Salah and Firmino to finally br

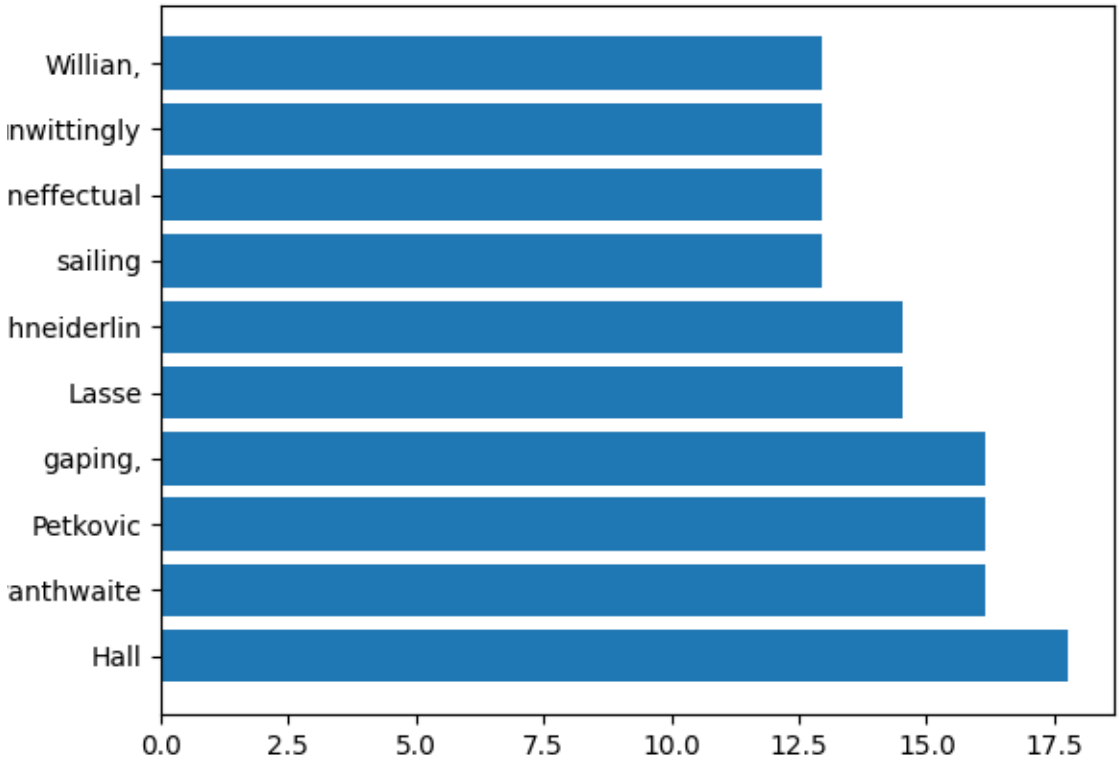
Results for positives:

	Unnamed: 0	sents
7	7	The contest did not really come alive until the 20th minute when England hopeful Nick F

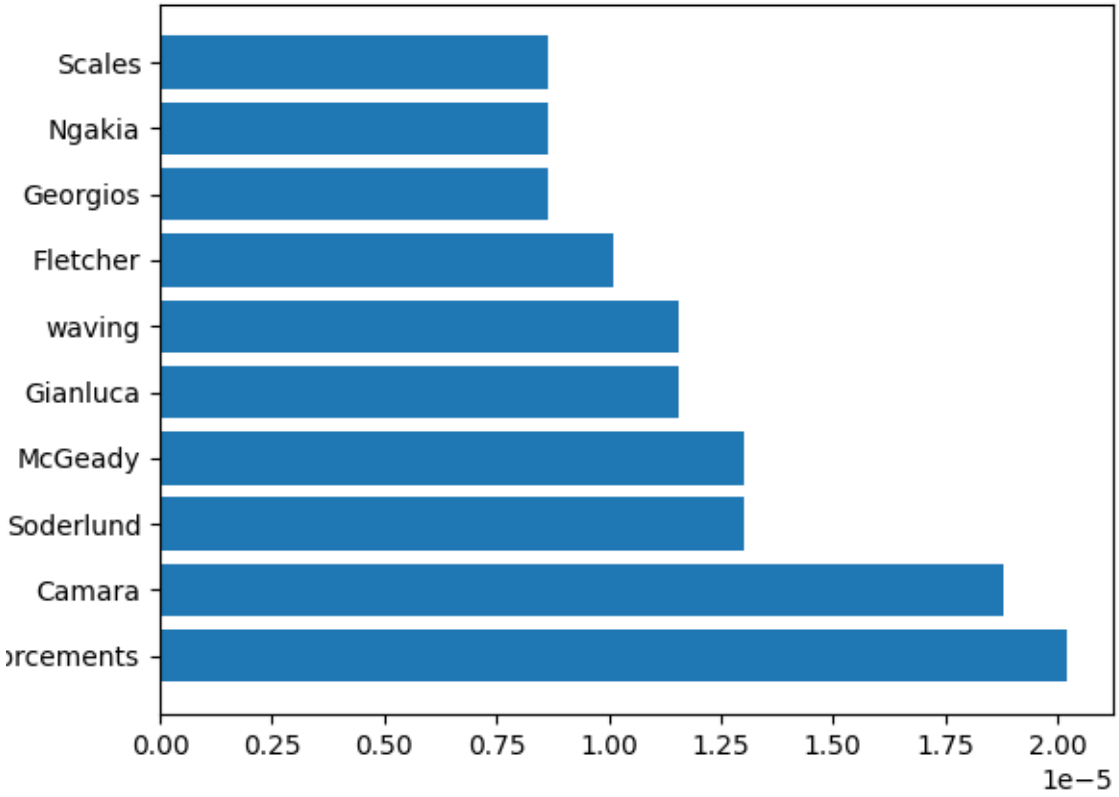
Then we should get metrics from that datas. So we write a script thet get all metrics. In SeperatedDataGetMetrics.py all metrics of seperated datas has been gotten. You can see all metrics as follow:

	Unnamed: 0	keys	values
0	0	Number of positive sents	24616
1	1	Number of negative sents	14597
2	2	Number of all sents	39213
3	3	Number of positive words	780056
4	4	Number of negative words	479749
5	5	Number of all words	1259805
6	6	Number of unique positive words	20253
7	7	Number of unique negative words	16918
8	8	Number of unique all words	22778
9	9	Number of only positive words	5860

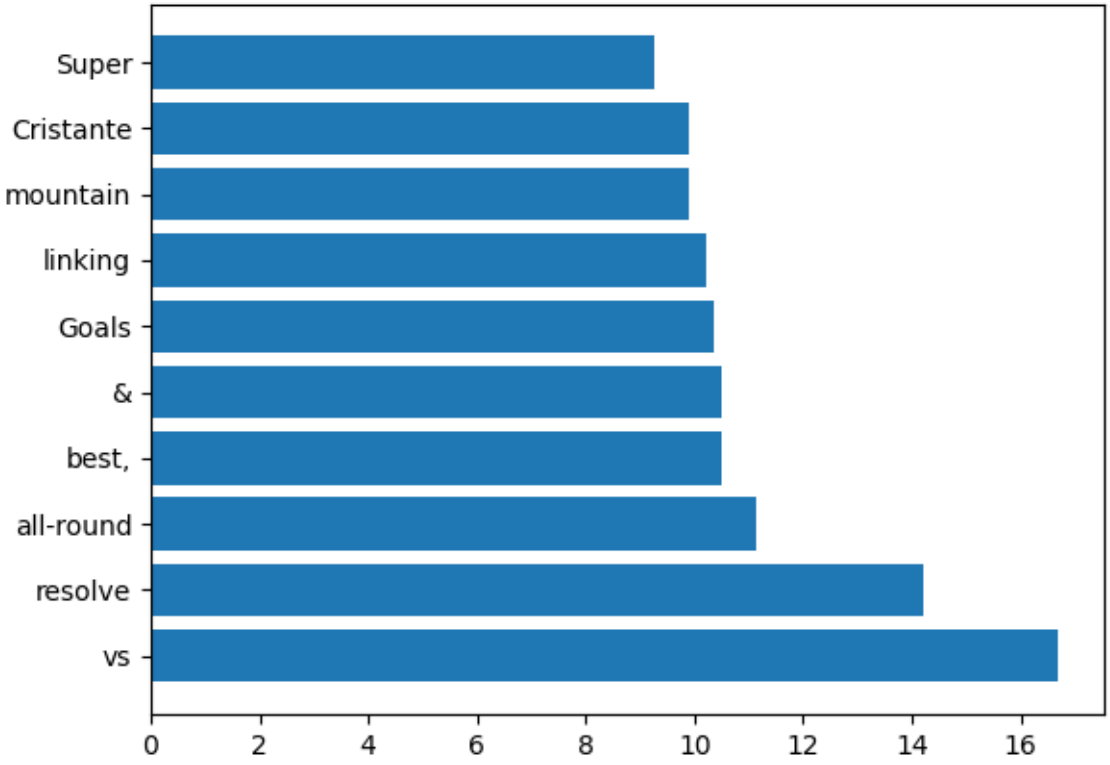
top10CommonNegativeWordsFrequency:



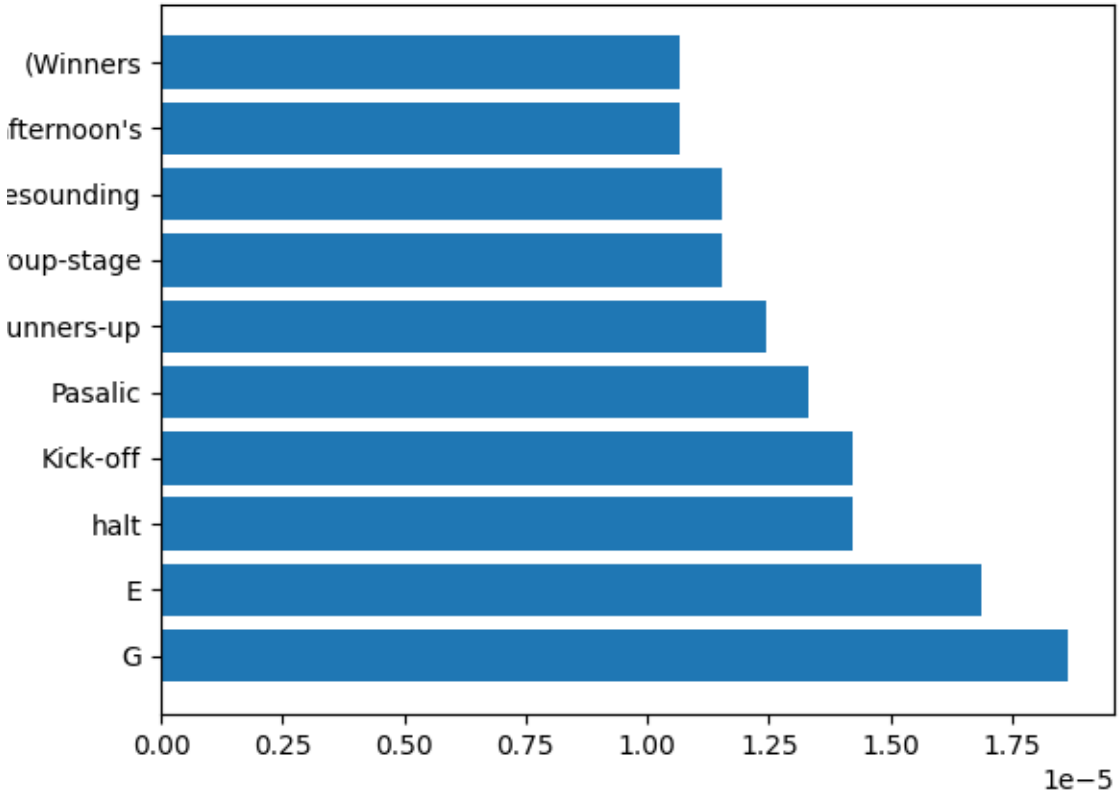
top10CommonNegativeWordsTFIDF:



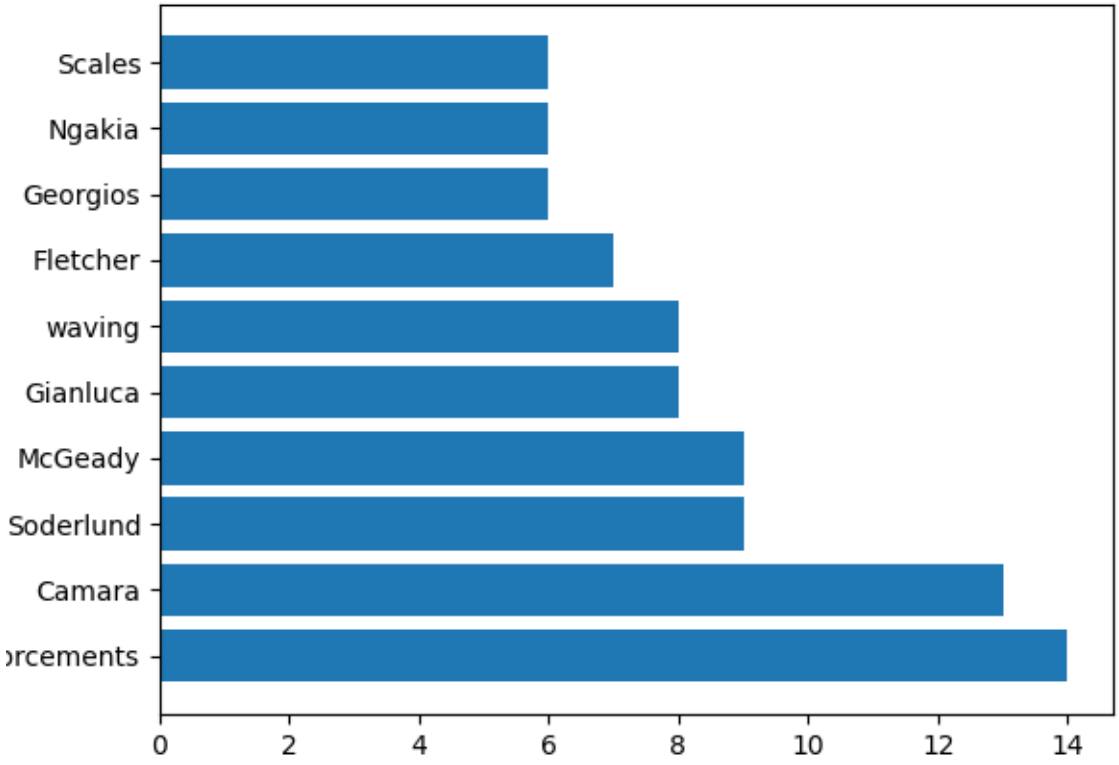
top10CommonPositiveWordsFrequency:



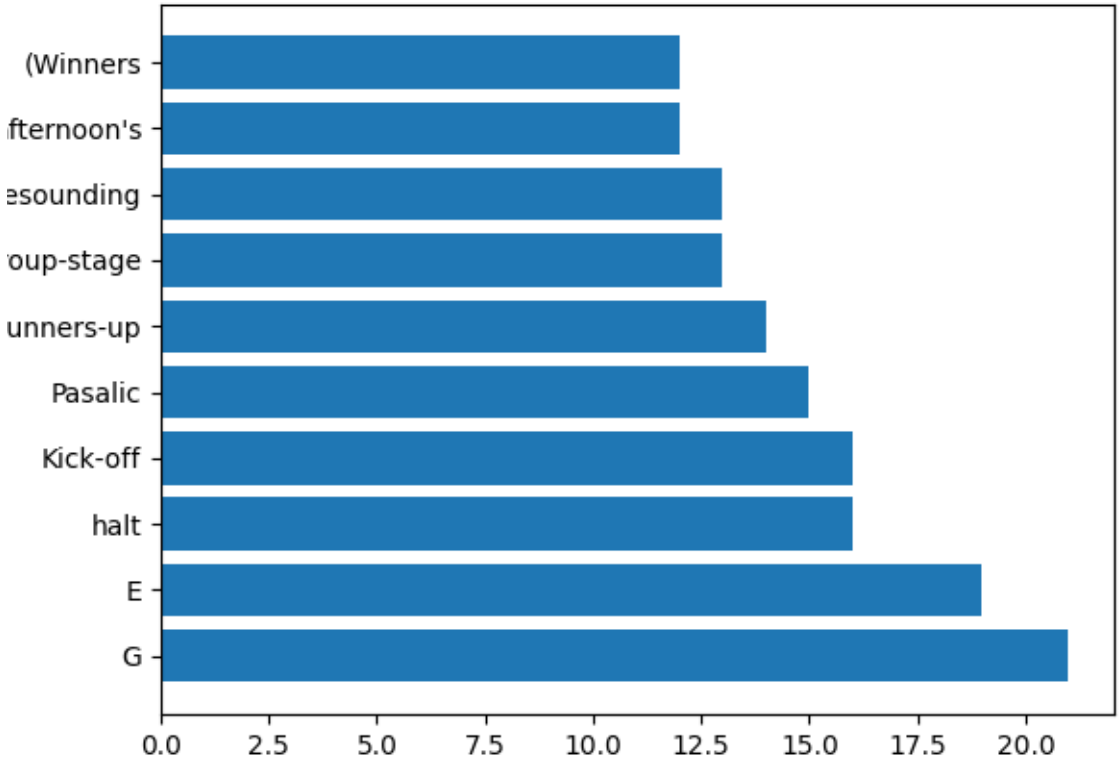
top10CommonPositiveWordsTFIDF



top10NegativeWordsCount



top10PositiveWordsCount



After that we should get metrics from new datasets.

2 Phase 2: Model Implimentation

comming soon!