워드 임베딩 기법을 이용한 축구 선수들 간의 친밀도 계산

장정안
국민대학교 소프트웨어융합대학 소프트웨어학부
e-mail: inetty@kookmin.ac.kr

Calculation of intimacy between soccer players using word embedding techniques

Kookmin University

요 약

축구에 있어서 선수들간의 관계는 경기에 있어서 굉장히 많은 영향을 끼친다. 이 프로젝트에서는 축구 관련 뉴스 기사를 크롤링하여 자료를 수집하고 중복되는 기사의 제목의 유사성을 검사하여 전처리 과정을 거친다 그리고, KTL2010 분석기로 형태소 분석을 한 뒤, word2vec 기법을 이용하여 modeling 을 하였다. 이를 통해 축구 선수들의 관계를 파악하고 파악하고 싶은 특정 축구 선수와 관련된 주요 키워드들이 어떤 것이 있는지 시각화한다. 이는 축구 뿐만 아니라 여러 스포츠에서도 팀 내에서의 조화를 도모할 수 있는 방안을 모색할 수 있다. 이러한 결과는 축구 팀의 성적 향상과 관련이 깊고, 더 나아가 축구 선수들 간의 상호작용을 이해하는 데 큰 도움이 될 것이다.

1. 서론

이 연구는 축구 선수들 간의 친밀도를 분석하여 팀 내 조화를 도모하고 축구 팀의 성적 향상을 위한 기반을 마련하는 것을 목적으로 한다. 축구는 개인 능력 뿐 아니라 팀워크와 선수들 간의 조화가 중요한 역할을 한다. 따라서 축구 선수들 간의 친밀도를 파악하고, 이를 기반으로 팀 내에서의 소통과 협업을 촉진하는 것은 축구 팀의 성적 향상과 관련이 깊다. 이를 위해 30 만 문장의 데이터 수집 후 형태소 분석과 word2vec 기법을 이용하여 모델링을 수행하였다. 이를 통해 축구 선수들 간의 단어 벡터를 계산하고, 이를 이용하여 친밀도를 계산할 수 있다. 이러한 분석 결과는 축구 선수들 간의 관계를 파악하는 데 큰 도움이 된다.

과거에는 한 선수의 정보를 알아내고 여러 선수들의 친밀도를 측정하기 위해 주로 설문조사나 인터뷰를 이용하는 방법이 주로 사용되었으나 이러한 방법은 시간과 비용이 많이 들며 선수들이 직접 말하지 않으면 정확한 데이터를 얻는 것도 제한적이다. 따라서 최근에는 직접 조사할 필요없이 여러 기사들의 정보를 수집하는 것에 있어서 수월하게 진행이 가능하다. 크롤링을 통한 데이터 수집은 여러 분야에서도 유용하게 사용되고 있다.

따라서 이번 연구에서는 최신 기술과 여러 기법들을 이용하여, 보다 정확하고 효율적인 분석 방법을 제시하고자 한다.

축구는 선수들 간의 조화와 팀워크가 매우 중요한 스포츠이다. 따라서 축구 선수들 간의 관계를 파악하고 이를 기반으로 한 팀 내 조화는 축구 팀의 성적에 큰 영향을 미친다. 이에 따라 축구 선수들 간의 친밀도를 파악하는 연구는 매우 중요하다. 이번 연구 결과를 토대로 축구뿐만 아니라 다른 팀 스포츠 분야에서도 해당 선수들 간의 친밀도 분석이 가능할 것이며, 이를 토대로 더욱 효과적인 팀 조직과 성적 향상을 이루어낼 수 있을 것이다.

2. 한국어 형태소 분석기

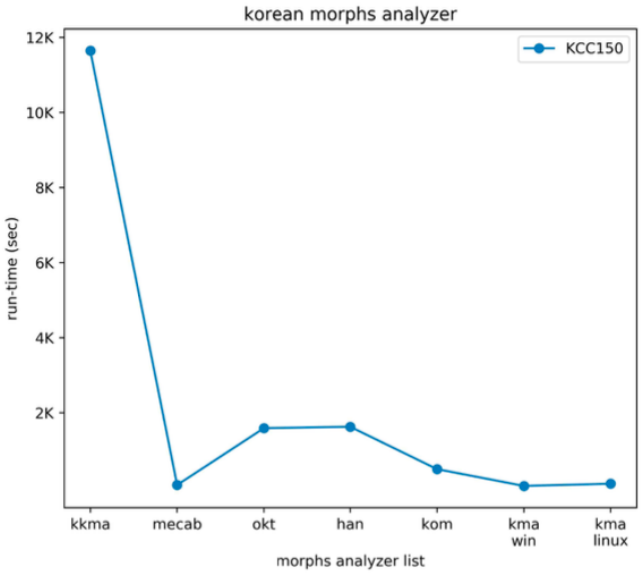
가장 대표적인 한국어 형태소 분석기로는 Konlpy 가 존재한다, 해당 형태소 분석기는 파이썬 패키지 형태로 제공되며, 해당 패키지 내에는 kkma, Mecab, Okt, Hannanum, Komoran 분석기가 존재한다. 이 중에서 Mecab 은 linux 환경에서만 사용이 가능하다.

이외에도 KLT2010(Korean Language Technology)가 존재한다. 해당 분석기는 C 언어로 개발되어 속도가 빠르고 높은 정확도를 특징으로 갖는다. 주 기능으로는 한국어 형태소 분석(KMA), 색인어 추출 및

복합 명사 분해, 형태론적 중의성 해결, 한국어 태깅, 고유명사와 신조어 등 미등록어 분석이다.

KCC150 말뭉치를 형태소 분석하는데 소요된 시간은, 해당 논문, "원혜진, 이현영, 강승식 (2020) 대규모 텍스트 분석을 위한 한국어 형태소 분석기의 실행 성능 비교, 한국컴퓨터종합학술대회 논문집"에 따르면 다음과 같다.

[그림 1] 형태소 분석 성능 비교(KCC150 말뭉치)



[표 1] KCC150 의 형태소 분석 시간 (단위 시간 : 초)

	Kkma	Mecab	Okt	Hannanum	Komoran	KMA -win	KMA -linux
KCC 150	116,430	783	15,916	16,280	5,000	533	1,123

본 연구에서는 대용량 한국어 텍스트를 형태소 분석하기 위해 속도 면에서 뛰어난 Mecab 과 KMA-win 로 형태소 분석을 시도해보았고, 두 분석기의 실행시간은 다음과 같다.

[표 2] Combined.txt(모든 키워드 문장 합친 최종 txt 파일, 문장 수 : 300,737)의 형태소 분석 시간 (단위 시간 : 초)

	Mecab	KMA-win	KMA-linux
combined.txt	23.12	22.57	28.36

3. 방법론

본 연구는 축구와 관련된 키워드를 이용하여 인터넷 뉴스 기사를 크롤링하여 자료를 수집한다. 또한 중복되는 기사들의 제목을 코사인 유사도를 이용하여 비교함으로써 비슷한 내용의 기사 데이터 수집을 방지한다. 수집된 데이터는 형태소 분석을 통해 전처리하고, Word2Vec 모델을 학습시켜 축구 선수들간의 친밀도를 분석하는 것이 이번 연구 실험에 있어서의 기본적인 방법론이다.

자료 수집은 crawl_news_final.py 로 뉴스 기사에 축구 관련 키워드를 입력한 뒤에 원하는 페이지 범위를 입력하면 해당 키워드의 기사들을 수집한다. 자료 수집과정에서 같은 내용에 대해 여러 신문사가 기사를 쓰는 점을 발견하였다. 그런 같은 내용의 기사들을 전부 수집하는 방법은 중복되는 데이터라는 점과 결국 불필요한 데이터들이 많아진다는 점에 있어서 비효율적이다. 그래서 sklearn module 을 통해 CounterVectorizer class 와 cosine_similarity class 를 이용하여 크롤링이 진행되는 동안 수집한 기존 자료의 제목과 이후 수집한 자료의 제목을 토큰 수의 행렬로 변환한다. 그 후에 이들 사이의 코사인 유사도를 계산한다. 본 연구에서는 코사인 유사도 값이 0.7 이상이면 제목을 중복으로 간주하고 다음 기사로 넘어가는 방법을 사용하였다.

수집한 자료를 전처리 과정(불필요한 들여쓰기 제거, 불필요한 내용들 제거)을 거쳐 형태소 분석을 위한 하나의 텍스트파일로 만든다. 본 연구에서는 형태소 분석기로 klt2010 을 사용한 파일로. 형태소 분석 후 Word2Vec 모델로 modeling 을 하였다.

분석에 큰 의미가 없는 단어인 불용어들을 txt 파일로 저장한 뒤에 각 그래프를 가시화 하기 전에 불러와 제거를 하는 작업을 해주었다.

해당 모델을 이용하여 Calculate_affinity.py 를 실행하여 원하는 선수의 이름을 입력 받으면 그 선수와 친밀도가 높은 top = 5 의 선수들과 수치를 출력한다. 좀 더 결과를 가시화하기 위해서 해당 결과를 그래프로 시각화 하였다.

추가적으로 입력 선수와 친밀도가 높은 선수들 이외에도 그 선수와 관련 키워드들 또한 선수 분석에 있어 중요하게 작용될 수 있다고 보고, 선수와 관련된 키워드들을 네트워크 노드 그래프로 시각화도 가능하게 하였다.

4. 실험 결과

'축구', '월드컵', '챔피언스리그' 등 여러 선수의 이름과 정보를 담는 키워드들을 이용하여 검색 및 자료 수집을 했고, 특정 인물의 자료를 수집하면 그 특정 인물을 입력 값으로 주었을 때 더 자세한 결과가 나올 것이라 예상되어 특정 선수를 손흥민 선수로 선정하고 손흥민 선수의 이름을 키워드로 추가 자료 수집을 하였다.

자료 수집에 있어서 유사도 검사를 통해 중복 기사 방지를 하였는데 전처리 전, 후 수집된 문장 수의 결과는 다음과 같다.

[표 2] 대표 상위 3 개 키워드의 전처리 전,후 (단위 : 문장 수)

키워드	전처리 전	전처리 후
축구	73345	40098
월드컵	56750	32018
손흥민	27445	10875

전처리 전과 후를 비교한 결과를 볼 때 같은 키워드와 범위를 입력 값으로 주었음에도 중복된 기사를 수집하지 않아 문장의 수가 적게 수집됨을 알 수 있다.

하나의 텍스트 파일로 합치기 전 키워드 별로 수집한 기사들의 문장 수는 다음과 같다.

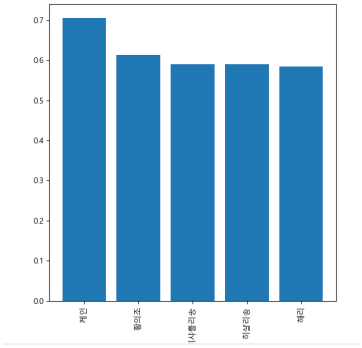
[표 3] 각 키워드 별 수집 문장의 수 (단위 : 문장 수)

키워드	총 문장의 수
축구	40098
epl	41012
라리가	22773
K 리그	26975
분데스리가	29003
축구 국가대표	7116
맨체스터 유나이티드	5454
맨체스터 시티	3698
울버햄튼	9781
토트넘 홋스퍼	2871
아시안컵	8288
SSC 나폴리	7085
베스트 일레븐	5574
북런던 더비	1895
축구 이적시장	10625
프리미어리그	5230
국왕컵	2096
피파랭킹	10043
유로파리그	4694
해트트릭	4393
대한민국 축구	15756
FA 컵	2615
월드컵	17593
챔피언스리그	5194
손흥민	10875
총	300,737

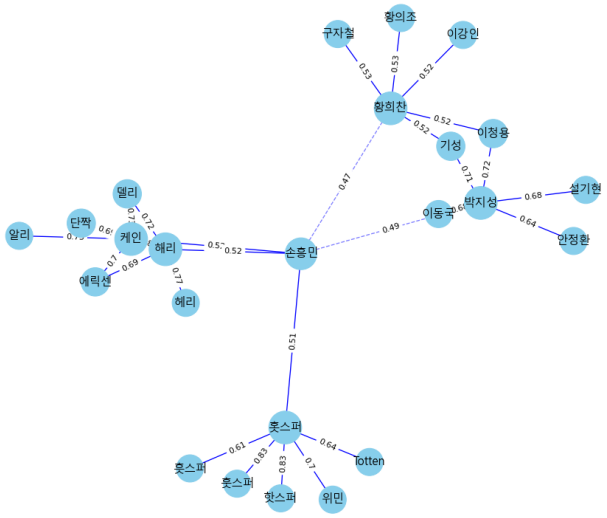
수집한 자료의 문장 수는 총 30 만 737 문장이다.

model 을 가지고 입력 선수와 친밀도가 높은 선수들을 그래프로 시각화한 결과와, 그 선수와 관련된 키워드들을 그래프로 시각화한 결과는 다음과 같다.

[1] 친밀도 top-5 선수들을 수치화 한 그래프



[2] 시각화 한 네트워드 노드 그래프



해당 결과는 '손흥민' 선수를 입력 값으로 주고 지정된 positive 값으로는 친밀도를 위해 '동료' 와 '콤비' 라는 키워드를 지정하였다. 또한 Word2Vec 모델에서 가장 많이 사용되는 negative sampling 기법을 적용하여 학습 시간과 모델 성능을 향상시켰는데, '다툼' 을 negative 값으로 주어 결과를 도출하였다. 결과는 예상한 결과와 비슷하게 나왔다. 같은 팀에서 함께 활약 중이며, 친한 관계인 '해리 케인' 선수와 '히샬리송' 선수가, 국가대표에서 함께 활약 중이며 실제로도 친한 관계에 있는 '황의조' 선수가 높은 수치로 도출되었다. 하지만 '해리'와 '케인'이 각각 형태소로 적용이 되어 각각의 높은 유사도 값으로 해당 결과가 나오는 것을 볼 수 있다. 그리고 관련 키워드를 시각화한 결과도 '손흥민' 선수의 팀, 그리고 국가대표 선수들이 나뉘어져 있음을 가시화하여 확인할 수 있다.

5. 결론

이 연구를 통해서 축구 선수들 간의 친밀도를 분석함으로써 축구 선수들의 관계를 파악할 수 있다. 그리고 수치화 된 데이터를 시각화 함으로써 분석에 있어 용이하게 작용이 가능하다. 이는 축구 선수들의 팀

내 역할 배치나 적응력 등에 대한 통찰력을 기대할 수 있다. 또한 형태소 분석과 워드 임베딩을 통해 구축한 모델은 이후에 다른 종목의 스포츠에 있어서도 응용될 수 있으며, 선수 이적 시장 등 스포츠 산업 분야에서 활용할 수 있는 다양한 연구들을 또한 가능함을 보인다.

본 연구에서는 특정 인물의 이름을 키워드로 자료를 수집한 것은 한 명 뿐이지만, 한 팀에서 이러한 분석을 하려고 하면 그 팀 내의 선수들을 이름을 키워드로 자료를 수집했을 때 해당 팀 내의 선수들간의 친밀도 분석에 있어 더 정확한 데이터를 얻을 수 있을 것이다. 또한 친밀도 분석을 위해 더 최적화된 값을 sampling 기법에 적용할 수 있는 개발이 요청된다.

참고문헌

[1] 원혜진, 이현영, 강승식 (2020) 대규모 텍스트 분석을 위한 한국어 형태소 분석기의 실행 성능 비교, 한국컴퓨터종합학술대회 논문집

[2] 김현준, 박종수, & 이병준. (2016). Word2vec 을 이용한 한국어 형태소 분석기 성능 향상 연구. 정보과학회논문지, 43(1), 7-12.

[3] 최영규 & 장재원. (2017). Word2Vec 기법을 활용한 한국어 문장 유사도 측정 방법에 대한 연구: 문장 내 단어 위치와 유사도 측정 방법의 영향 분석을 중심으로. 정보처리학회논문지 컴퓨터 및 통신시스템 특집, 6(10), 155-162

[4] 이한규. (2015). 스포츠집단의 응집력과 집단효율성. 한국체육학회지, 54(3), 239-247

[5] beausty23. "word2vec 를 사용해보기." 블로그, 2020.1.13, https://beausty23.tistory.com/58.

[6] ChatGPT. (2021). OpenAI.