토픽 모델링을 활용한 사용자 게시글 의견 분석 및 이슈 트래킹

분석지원팀 지원자 서수민

INDEX

과제 설명

과제1. 유저들의 의견을 참고하여 게임 운영에 도움이 되는 인사이트를 도출해보고자 합니다.

유저의 의견을 분석해주시고 데이터 처리하는 방법에 관한 설명 자료와 분석 결과를 제출해주세요.

분석을 통해 도출한 유저들의 의견 중 게임 운영에 도움이 될 만한 주제가 있다면 분석 방법과 기대 효과를 설명해주세요

1. 데이터 수집

수집 기간 : 2018.xx.xx ~ 2021.09.11

수집 데이터 : 피파온라인4 인벤(이슈/토론/버그 게시판), 피파온라인4 (커뮤니티 추천 게시글)

수집 도구 : Python Selenium, Request, BeautifulSoup

설명 : 피파 온라인4 내의 사용자 의견을 분석하기 위하여 크롤링 수행, 총 21,862건의 데이터를 수집

피파온라인4 인벤(이슈/토론/버그 게시판)

	day	title	content	view	hit
0	2021-09-11	겜 렉이 갑자기 엄청 심한데	업댓하고 나서 2판정도 이주 스무스하게 잘 돌아갔었는데 3판째 끝나고 갑자기 이적시...	36	0
1	2021-09-11	5백 2볼란치 는 비매너다?	진선에서 이런 채팅을 하시는 분이 있어서 한번 이야기 해보고 싶네요.해당 경기 내용...	161	0
2	2021-09-11	패키지 추천	9000fc로추석 풍년 + 볼토끼 3개추석 연쇄 패키지 2개중에 구매하려는데 어떤게...	79	0
3	2021-09-11	20A 메시 로패하면 오버롤 오르나요?	20A 메시 은카 갖고있습니다. 로스터패치하면 오버롤 오르나요?	69	0
4	2021-09-11	알파	19시즌 잉글랜드 몰백 금카 있는데 알파 당연히 불가능이겠죠?	100	0
...	...	...	...	...	...
8497	2018-09-25	만약에 모드리치가 발롱받으면	토티 모드리치가 제일 좋아져야 하는거라고 생각하는데 어떤가요?발롱이면 진짜 최고의 ...	2714	2
8498	2018-09-25	게임중 채팅차단은 상대가 차단하면 같이 안되게 하면 됩니다.	인벤에도 메인에 올라왔네요?피파온라인3 에서 1:1 2:2 3:3에서게임이긴사람이든...	1523	0
8499	2018-09-25	추석버닝 추천	이렇게 떳는데 뭐받죠??	2734	0
8500	2018-09-25	**전술, 포메관련 토론합시다**	두달간 전술수치랑 포메 관련해서 여러가지 시험해봤지만 명확한 답을 내리지 못하고 있...	2724	0
8501	2018-09-25	말디니 짝으로 tt라모스 3카 vs tt홈멜스 3카	뭐가좋을까요 ? 피린이에게 도움의 손길을ㅠㅠㅠ	1100	0

8502 rows × 5 columns

피파온라인4(커뮤니티 추천 게시글)

	day	title	content	view	hit
0	2021-09-09	뇌없고 무능한 개발진들아 ~ 욕안했다 글삭제하지마라	ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ 이게 게임이냐?ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ패스는 했다하면 ...	154	4
1	2021-09-09	피파에 현질하기 vs 롤에 현질하기	ㅇㅇ?	276	0
2	2021-09-09	얘들아 지금 현질 유도 기간이다	제갈 개 그지같은거 일부러 현질하게 만드는거니까 참고하자 얘들아개 ㅈ망 그래픽 쪼가...	235	1
3	2021-09-09	도G드래곤-무전	도G드래곤-무전	316	0
4	2021-09-09	개 ㅈ망게임 공이 키퍼를 그냥 지나감?	아주그냥 캐**는거에만 눈이 뱔갛지?	184	4
...	...	...	...	...	...
13355	2019-06-07	제갈 구리다는 애들 잘들어라	현질하셈 자본주의 사회에서 뭘 더바라냐ㅉㅉ 투자도 안하고 꼭 제갈만 구리다고 징징대...	1,607	5
13356	2019-06-07	님들 근데 왜 팀들이 다 똑같아요?	선수카드는 몇백장인거같은데 쓰는선수는 다 똑같네.. 글리트 앙리 호나우두.. 호날두...	1,528	4
13357	2019-06-07	12시 땡하자마자 팅겨버렸네...	신데렐라가 이런 기분인가?	946	4
13358	2019-06-06	인생강화 6천으로 개꿀	2칸의기적	1,790	5
13359	2019-06-06	프로 월클 구간에 호골비말 쉰내충 안쓰는 사람 잇냐?	시부레 무슨 10판하면 10판다 호골비말 호골비말 홍철 손흥민ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ...	126	0

13360 rows × 5 columns

2. 데이터 가공

설명 : 크롤링한 사용자 의견 데이터를 토픽 모델링의 입력 형태로 가공하기 위해 데이터 가공 실시

1) 분석 기간을 2020.01.01 ~ 로 통일 : 크롤링시 두 수집 장소에 대한 수집기간이 다르기 때문에 교집합으로 통일
2) Title + Content Concat : 제목과 내용을 연결하여 하나의 변수로 구축
3) 문자, 숫자를 제외한 불필요 문자 제거: 특수 문자 및 자모음을 제거하기 위해 수행

1) 분석 기간 선정 2020.01~ 2021.09.11

day	title	content
2021-09-11 00:00:00	5백 2볼란치 는 비매너다?	친선에서 이런 채팅을 하시는 분이 있어서 한번 이야기 해보고 싶네요.해당 경기 내용...
2021-09-11 00:00:00	패키지 추천	9000fc로추석 풍년 + 별토끼 3개추석 연쇄 패키지 2개중에 구매하려는데 어떤게...
2021-09-11 00:00:00	20A 메시 로패하면 오버롤 오르나요?	20A 메시 은카 갖고있습니다. 로스터패치하면 오버롤 오르나요?
...	...	...
2020-01-01 00:00:00	강화확률	강화확률 0.1이냐 ** 2카에서 3카 풀이 터지네 개 **
2020-01-01 00:00:00	지금 토티 시즌 사도 될까요	지금 사도 될까요?사게 되면 메시,날두, 라모스 정도일 것 같아요그리고 셋다 가격 ...
2020-01-01 00:00:00	토트넘 티비금카팀 선수고민중 입니다	모두 아시겠지만 티비금카가 많이 풀렸어요티비금카중 토트넘스쿼드가 가성비가 좋아서 토...
2020-01-01 00:00:00	아래 ICON 개봉 결과다..	썩을 발데라마나 렐러스 안나온걸 다행으로 알아야하나? 드로그바나 지단 바랬는데.. 썩을...

2) Title + Content 결합

day	title_body
2021-09-11 00:00:00	5백 2볼란치 는 비매너다? 친선에서 이런 채팅을 하시는 분이 있어서 한번 이야기 ...
2021-09-11 00:00:00	패키지 추천 9000fc로추석 풍년 + 별토끼 3개추석 연쇄 패키지 2개중에 구매하...
2021-09-11 00:00:00	20A 메시 로패하면 오버롤 오르나요? 20A 메시 은카 갖고있습니다. 로스터패치하...
...	...
2020-01-01 00:00:00	강화확률 강화확률 0.1이냐 ** 2카에서 3카 풀이 터지네 개 **
2020-01-01 00:00:00	지금 토티 시즌 사도 될까요 지금 사도 될까요?사게 되면 메시,날두, 라모스 정도일...
2020-01-01 00:00:00	토트넘 티비금카팀 선수고민중 입니다 모두 아시겠지만 티비금카가 많이 풀렸어요티비금카...
2020-01-01 00:00:00	아래 ICON 개봉 결과다.. 썩을 발데라마나 렐러스 안나온걸 다행으로 알아야하나?...

3) 불필요한 문자 제거

text_pre
5백 2볼란치 는 비매너다 친선에서 이런 채팅을 하시는 분이 있어서 한번 이야기 해...
패키지 추천 9000 로 추석 풍년 별토끼 3개 추석 연쇄 패키지 2개중에 구매하려는데...
20 메시 로패하면 오버롤 오르나요 20 메시 은카 갖고있습니다 로스터패치하면 오버...
...
강화확률 강화확률 0 1이냐 2카에서 3카 풀이 터지네 개
지금 토티 시즌 사도 될까요 지금 사도 될까요 사게 되면 메시 날두 라모스 정도일 ...
토트넘 티비금카팀 선수고민중 입니다 모두 아시겠지만 티비금카가 많이 풀렸어요티비금카...
아래 개봉 결과다 썩을 발데라마나 렐러스 안나온걸 다행으로 알아야하나 드로그바나지...

2. 데이터 가공

4) L-Tokenizer 학습을 통한 도메인 명사 추출 : 토픽 모델링 수행 시, 문장을 토큰으로 분절(Parsing)해야 한다. 이때, 일반적인 Tokenizer를 사용할 경우 분석에 중요한 단어가 분절될 수 있다. 따라서 의미있는 토큰을 구축하기 위해서는 명사를 온전한 형태로 분절할 필요가 존재하며, 이를 위해 L-Tokenizer를 활용

*L-Tokenzier: 한국어 어절은 L+[R] 구조로 “Cohesion Score”를 통해 단어의 경계를 학습하여 분절하는 비지도 학습 기반 Tokenizer

Mecab Tokenizer : ['5', '백', '2', '볼란치', '는', '비', '매너', '다', '친선', '에서', '이런', '채팅', '을', '하', '시', '는', '분', '이', '있', '어서', '한', '번', '이야기', '해', '보', '고', '싶', '네요', '해당', '경기', '내용', '포메이션']

L-Tokenizer : ['5백', '2볼란치', '는', '비매너', '다', '친선', '에서', '이런', '채팅', '을', '하시', '는', '분이', '있어서', '한번', '이야기', '해보', '고', '싶네요', '해당', '경기', '내용', '포메이션']

L-Tokenizer 학습 후, 명사 추출 리스트(20,826 건의 명사 추출)

Nouns List	
고민입니다	일렉트로닉아츠코리아
실축존재	아틀레티코마드리드
매출상승가능성	3000마일리지
특수시즌	월드클래스문지기
레알첼시	닌텐도스위치엔진
은카성공	문애란의뮤직서핑
언제나옴	6차넥스트필드전
질문합니다	피파하는월급루팡
쿠폰입니다	패드립과섹드립등
소모됩니다	위닝일레븐10
팀평부탁드립니다	생각했지만이번
회사입니다	게임시작하기전
1차쿠폰	축구온라인게임
…	…

2. 데이터 가공

5) L-Tokenizer를 통해 추출된 명사들을 Mecab 사용자 사전에 추가하여 텍스트 분절 수행 최종 분석 데이터 추출(Token_list)

Nouns_Count
(게임, 4647)
(선수, 4240)
(넥슨, 2550)
(유저, 2545)
(피파, 2416)
(팀, 2360)
(생각, 2349)
(보정, 1891)
(시즌, 1808)
(사람, 1713)
(정도, 1514)
(패스, 1499)
…

token_list
['5백', '2볼란치', '비매너', '친선', '채팅', '이야기', '해당', '경기', '내용', '포메이션'],
['패키지', '추천', '9000', '로추', '풍년', '토끼', '3개추석', '연쇄', '패키지', '구매'],
['20', '메시', '로패', '오버', '메시', '은카', '로스터패치', '오버'],
['알파', '19시즌', '잉글랜드', '풀백', '금카', '알파', '가능'],
['로랑', '코시엘니', '은카'],
['컨뽕', '컨뽕'],
['겜','렉','2판','정도','3판','이적','시장','공경','할때','렉이','진짜','칠정','게임','문제','컴터','문제', '컨터', '사양', '발로', '란트', '피파', '화질','렉',
['볼타', '매칭', '혹시', '피파', '관계자분들', '게시판', '모니터링', '매칭시스템', '개선', '고객센터', '문의', '답변', '모드', '시작'],
['요한', '크루이프', '1카', '이거', '고민입니다'],
…

분절된 텍스트 데이터는 토픽 모델링을 위한 2가지 방법으로 활용

1) LDA Topic Model : Mecab 사용자 사전에 도메인 명사를 추가하여 분절 후, LDA 모델 수행

2) FIFA-BERTopic model : Mecab 사용자 사전에 도메인 명사를 추가하여 분절 후, BERT 추가 사전 학습을 위한 Token ID로 변환

3. 토픽 모델링

토픽 모델링에 대한 3가지 모델 구축 : LDA, BERTopic, FIFA-BERTopic

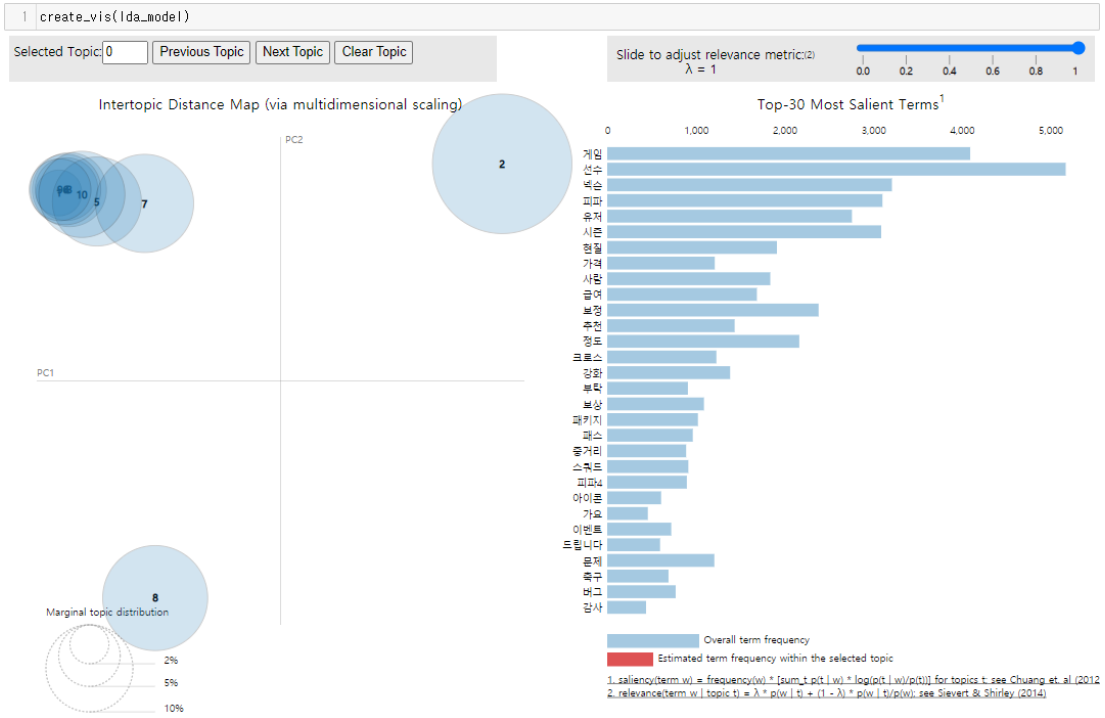
1) LDA(Latent Dirichlet Allocation) : 토픽 모델링 분야에서 널리 사용되는 대표적인 모델로 토픽에 대한 단어 분포를 기반으로 주어진 문서들에 대한 토픽을 할당하는 방법

2) BERTopic : 사전 학습 언어모델인 BERT를 활용하여 고품질의 문장 임베딩(Embedding)을 추출한 후, 밀집 클러스터링을 통해 군집화를 수행하여 토픽을 구성하는 방법, 군집된 토픽에 대한 단어 중요도를 통해 쉽게 해석 가능한 인사이트를 제공

3) FIFA-BERTopic : 기존의 사전 학습된 BERT의 경우, 피파 온라인에서 자주 등장하는 핵심 단어 정보를 가지고 있지 않다. 따라서 "추가 사전학습"을 통해 도메인 단어들을 추가 학습하여 단어의 의미를 파악한 BERT를 구축

3. 토픽 모델링

1) LDA

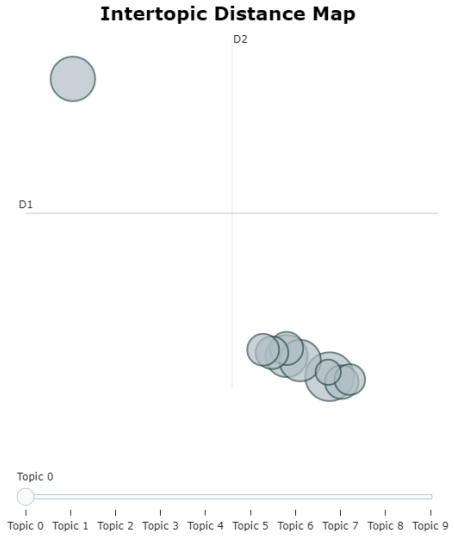


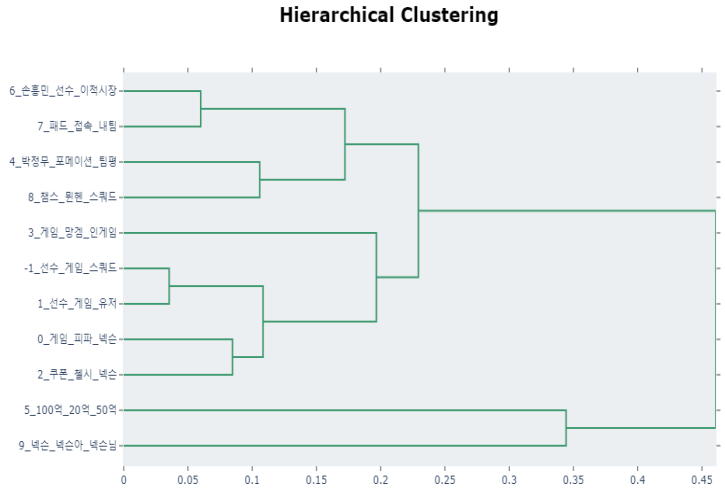
uninterpretable	시즌 카드	uninterpretable	카드팩	점검 & 이벤트	스쿼드	구입	캐시 충전	uninterpretable	게임조작
가요	선수	티어	가격	보상	부탁	급여	게임	매물	크로스
감사	시즌	19	아이콘	패키지	드립니다	추천	넥슨	챔스	패스
첼시	보정	댓글	가능	스쿼드	피파온라인4	강화	피파	전술	중거리
박정무	정도	스텟	20토티	이벤트	수수료	피파4	유저	20	축구
점검	문제	결과	금카	업데이트	의견	버그	현질	과금	굴리트
접속	상대	서버	토티	카드	친구	확률	사람	방법	키퍼
포메이션	시세	호날두	은카	귀속	골키퍼	5카	시작	참고	에이전트
피시방	패치	고민	오버롤	버닝	호나우두	3카	적폐	날두	개인
2억	경기	신규	기준	아이폰	7카	구매	무과금	맨유	적용
드록바	출시	10억	19토티	프로	검색	월클	확인	레알	메시

3. 토픽 모델링

2) BERTopic

BERTopic을 위해 사용된 사전 학습 언어모델은 오픈 소스로 제공하는 KcBERT를 사용, KcBERT는 뉴스 기사의 댓글과 대댓글을 수집하여 학습한 모델로, 30,000개의 단어와 15GB(8,900만 개)의 뉴스 기사를 사전 학습한 모델



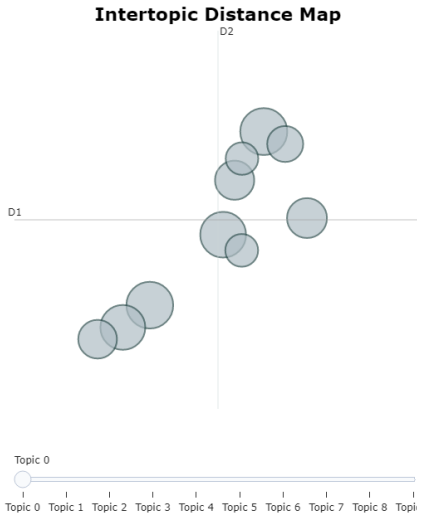


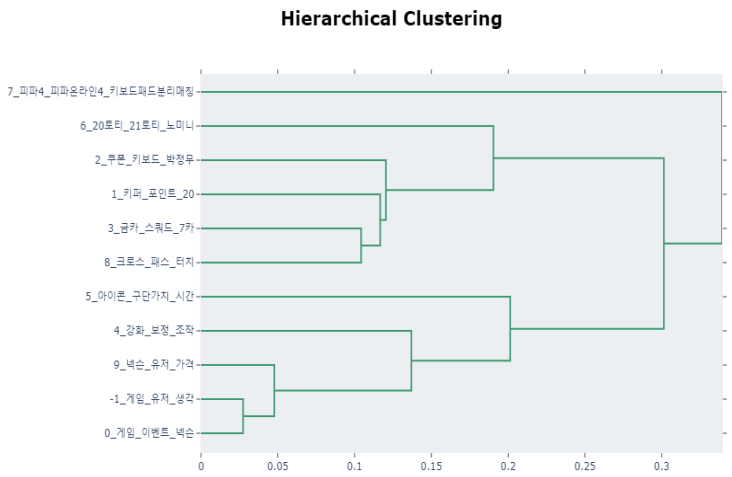
Uninterpretable	시즌 & 체감	쿠폰 & 이벤트	게임 내 이슈	플레이 개선점	구단가치	Uninterpretable	Uninterpretable	시즌카드	Uninterpretable
게임	선수	쿠폰	게임	박정무	100억	손흥민	패드	챔스	넥슨
피파	게임	첼시	망겜	포메이션	20억	선수	접속	뮌헨	넥슨아
넥슨	유저	넥슨	인게임	팀평	50억	이적시장	내팀	스쿼드	넥슨님
축구	시즌	섭종	쿠폰	스쿼드	30억	경기	키보드	토츠	유저
패드립	넥슨	스쿼드	피파온라인4	키보드자동수비상향	200억	점검	스텟	19챔스	게임
금카	체감	게임	버그	키보드패드분리매칭	40억	단일	컴퓨터	첼시	넥슨이
버그	확률	시즌	피파4불매운동	레벨업	첼시	시세	원천	챔스시즌	현질
피파4	축구	이벤트	오류	90실축비주류포지션	70억	단일팀	무한대	호나우두	점검
스쿼드	출시	유저	엔진	크로스중거리헤딩상향	90억	서버	현혹	네이마르	체감
인게임	플레이	2차	피파4	단일	5억	추천	해킹	음바페	박정무

3. 토픽 모델링

3) FIFA-BERTopic

기존의 모델인 KcBERT에 크롤링을 통해 수집한 FIFA 도메인 단어들을 추가 사전 학습하여 의미를 학습한 모델, 기존의 30,000건의 단어 말뭉치에 수집한 20,826건의 도메인 명사를 추가하여 학습을 수행함

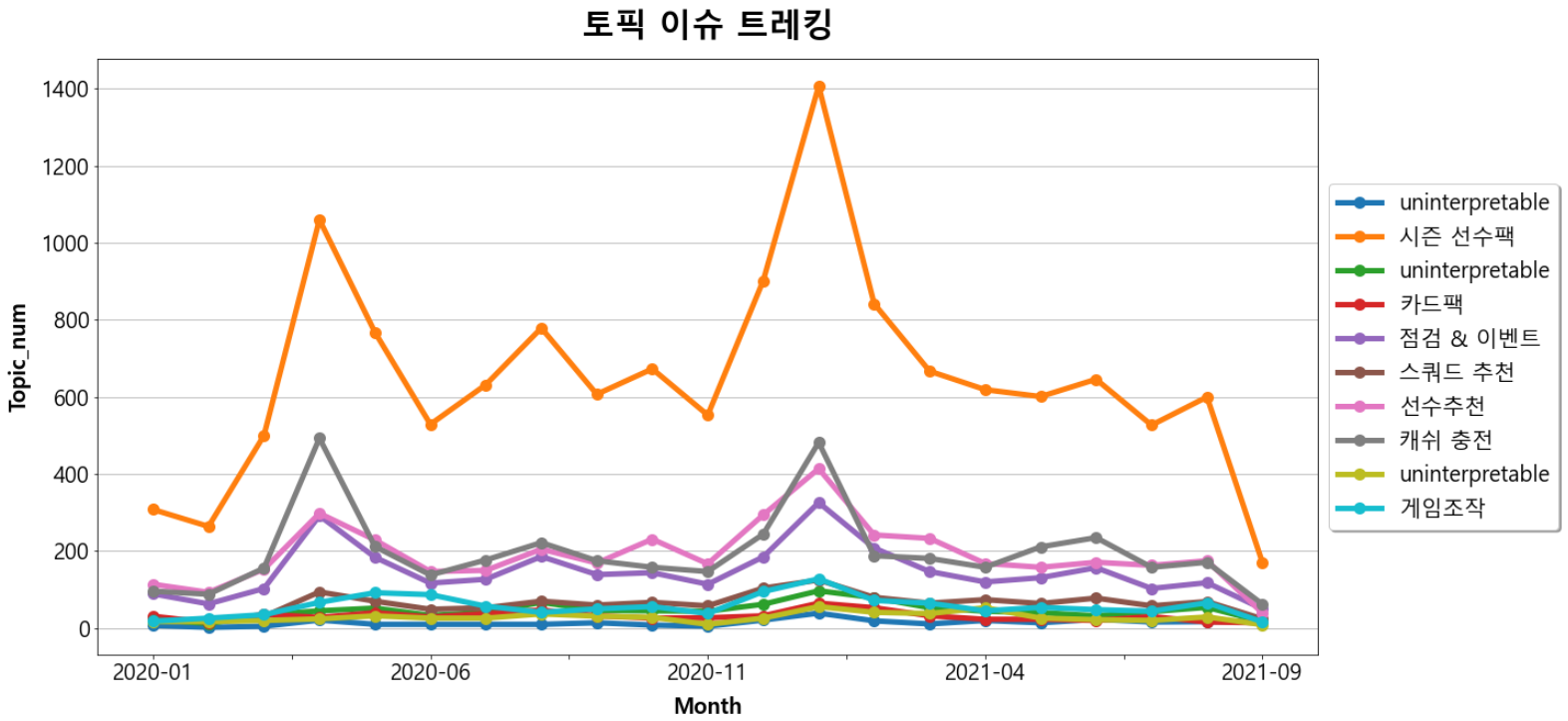




개선 & 소통	Uninterpretable	쿠폰 & 이벤트	스쿼드 평가	강화확률 이슈	구단가치 보존	시즌 선수팩	점검 & 플레이 개선	플레이 메타	게임 개선점
게임	키퍼	쿠폰	금카	강화	아이콘	20토티	피파4	크로스	넥슨
이벤트	포인트	키보드	스쿼드	보정	구단가치	21토티	피파온라인4	패스	유저
넥슨	20	박정무	7카	조작	시간	노미니	키보드패드분리매칭	터치	가격
문제	레벨	2차쿠폰	레알	몰수패	넥슨	가격	키보드자동수비상향	골대	불매운동
피파	레벨업	패스	밀란	패드립	넥슨님	10억	개인기대폭축소	순비피	문제
개선	골키퍼	1차쿠폰	가요	넥슨	게임	5억	크로스중거리헤딩상향	크로스충	피파
축구	보상	2차	주세요	오류	이게	100억	실축존재	헤딩	박정무
때문	최근	1차	레알마드리드	보정이	가격	1억	90실축비주류포지션	박제	개돼지
시간	스트레스	피파4불매운동	추천	풀게이지	피파	아이폰	연장점검	대한민국	기업
플레이	사용	수비수	마드리드	이게	시대	30억	피파온라인	포메이션	때문

4. 이슈 트래킹

1) LDA

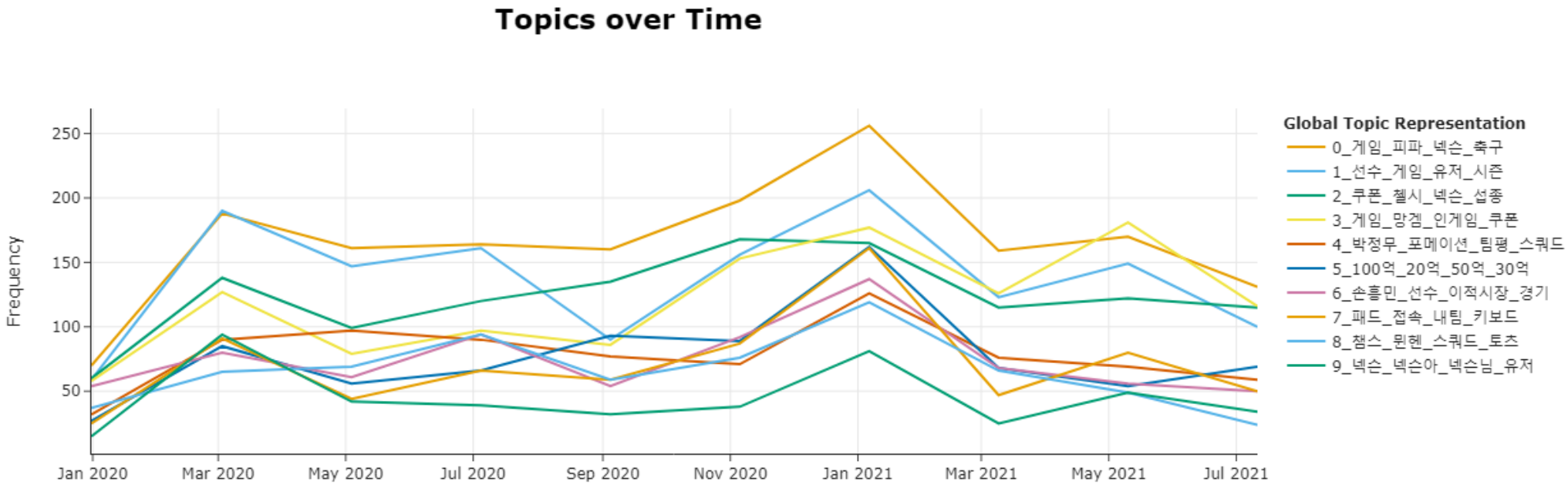


LDA 토픽 모델링의 결과를 통해 2020.01.01~ 2021.09.11 동안의 이슈에 대한 트래킹을 실시

LDA 토픽 모델링은 모든 문서에 대하여 정해진 토픽을 할당해야한다. 따라서, 전혀 의미를 반영하지 못하는 문서로 인해 토픽에 노이즈가 형성되는 한계가 존재한다.

4. 이슈 트래킹

2) BERTopic

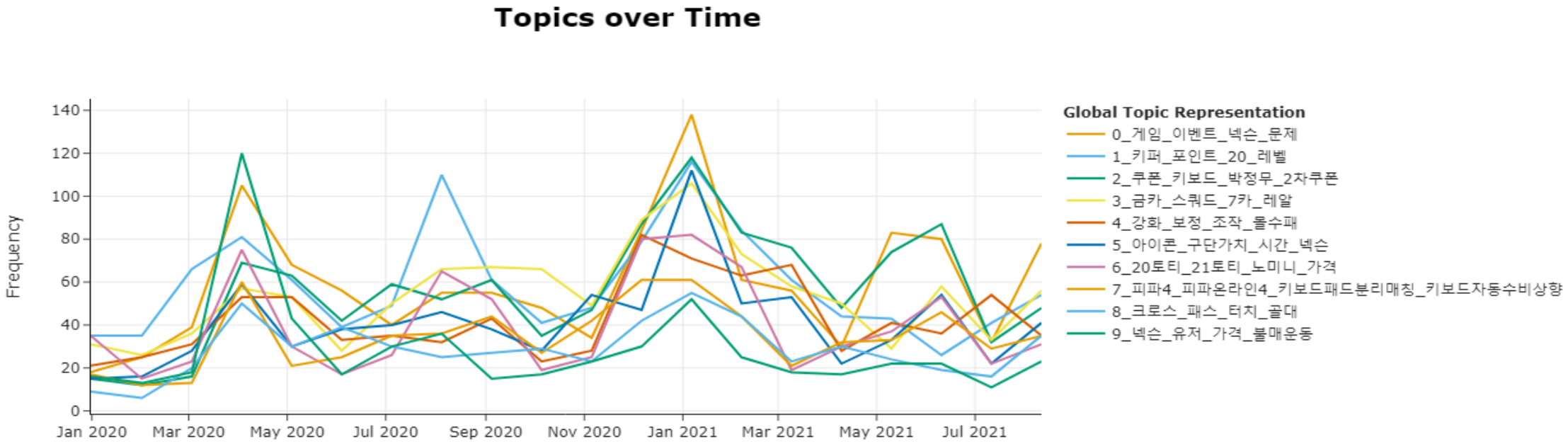


KcBERT 토픽 모델링의 결과를 통해 2020.01.01~ 2021.09.11 동안의 이슈에 대한 트래킹을 실시

KcBERT의 경우, 피파 온라인 내의 단어를 학습하지 못한 상태로 고품질의 문장 임베딩을 추출하기 어렵다는 한계가 존재한다.

4. 이슈 트래킹

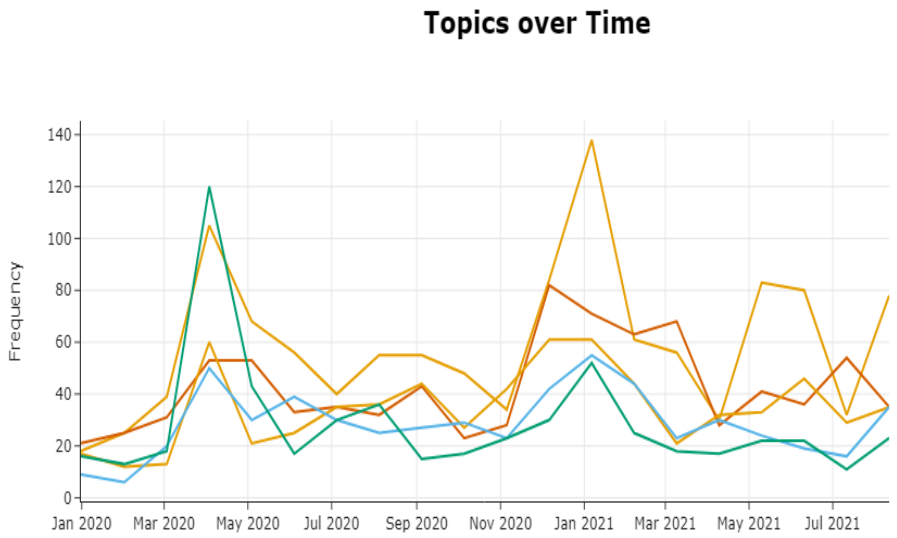
3) FIFA-BERTopic

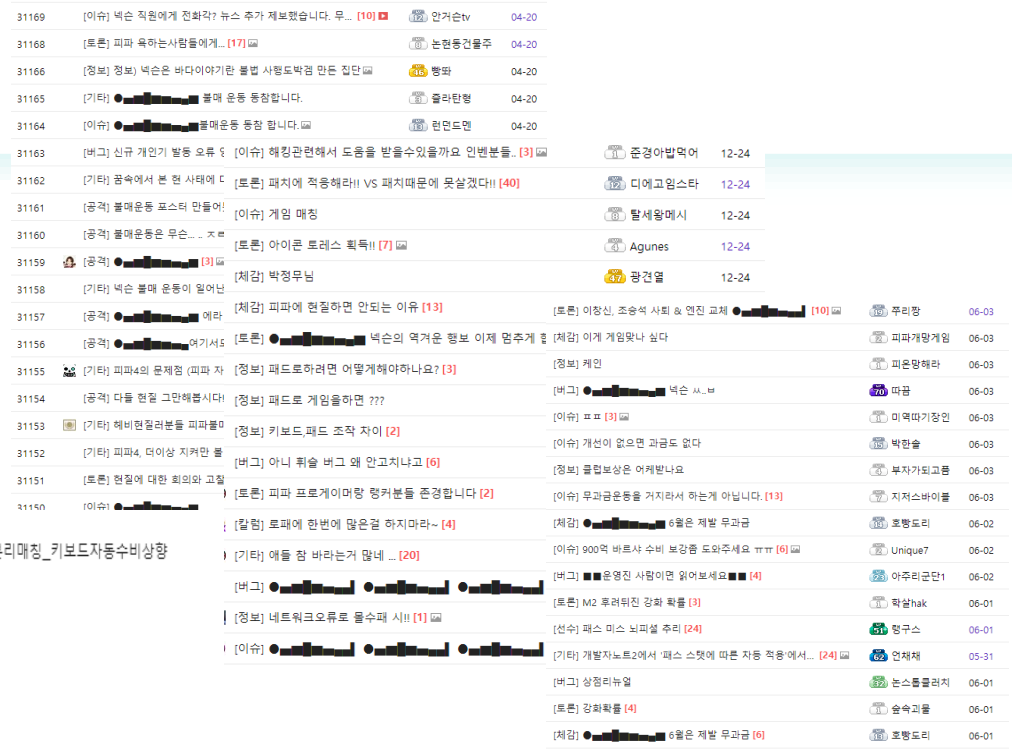


FIFA-BERT 토픽 모델링의 결과를 통해 2020.01.01~ 2021.09.11 동안의 이슈에 대한 트래킹을 실시

피파온라인 내의 도메인 단어를 학습함에 따라, 고품질의 문장 임베딩을 추출할 수 있으며 좀 더 다이나믹한 이슈 트래킹 결과를 구축

5. 결과 및 기대효과





1) 여러 토픽 결과들 중, “개선 & 소통”, “강화확률 이슈”, “플레이 개선”, “플레이 메타”, “게임 개선점 " 토픽은 피파온라인4 운영에 있어서 끊임없이 관리해야 하는 이슈이다. 실제로 21.04, 20.12~21.02, 21.06에 “개선 & 소통”, “게임 개선점 " 토픽이 상승하는 것을 볼 수 있으며 해당 시기에 많은 유저들이 게시판을 통해 불만, 개선점을 남긴 것을 확인할 수 있다.

2) 추후, 대규모의 게임 텍스트 말뭉치를 구축하여 FIFA-BERT모델을 구축한다면 더욱 고품질의 임베딩을 추출 가능하며, 트래킹 기간을 매주 업데이트로 설정하여 업데이트 이후 유저들의 의견 및 개선점을 파악할 수 있을 것으로 기대한다.

THANK YOU