웹과텍스트마이닝개론
프로젝트 중간보고서

통계학과 2129045 박경숙

목차

진행 상황

I 서론

새로운 교통수단으로 제시되는 드론은 현재 본격적인 상용화 준비에 있다. 올해 2월에 발행된 세계 화물용 드론 시장 보고서에 의하면, 화물용 드론 시장 규모는 최근 몇 년동안 급격히 확대되어 2023년 13억 1,000만 달러에서 2024년에는 17억 6,000만 달러에 달하고, 연평균 복합 성장률(CAGR) 34.7%로 성장할 것으로 전망하였다.

이러한 화물용 드론의 산업 동향과 사용 현황을 파악하고, 관련 뉴스의 내용에 따른 주가 변동의 상관관계를 통계적으로 시계열 분석해보고자 한다.

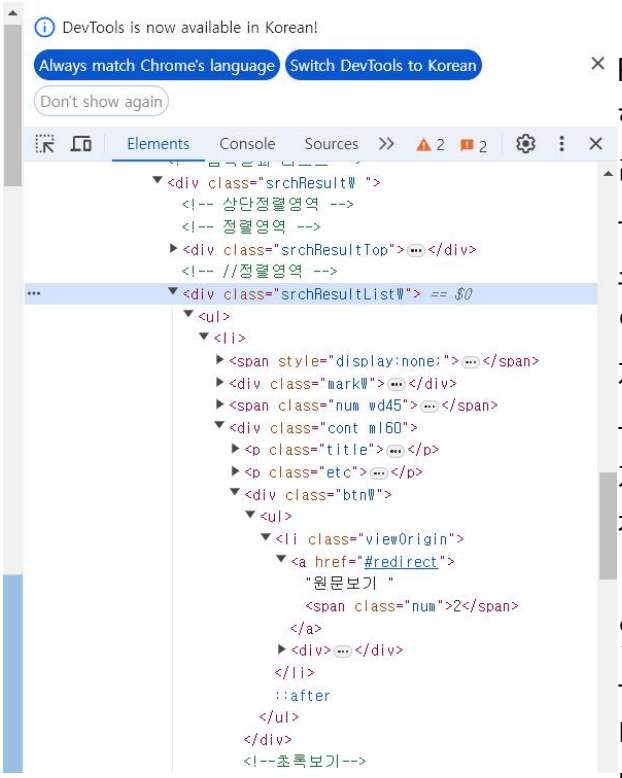
II-1 데이터 설명

RISS에서 화물용 드론에 관련한 학술논문을 검색하였을 때, 다음과 같은 문제점이 있었다. 키워드를 "화물용 드론" 또는 "화물 드론"으로 설정한 경우 모두 결과가 1건만 출력되었다. 또한 논문의 초록이 한국어가 아닌 영어로 기재된 경우가 드물게 존재하였다. 따라서 검색 키워드를 "드론"으로 설정하여 범위를 넓게 하고, 이후의 형태소 분석을 위해 국내 학술논문만을 검색하도록 하여 언어를 한국어로 제한하였다. 그 결과 총 1113건의 관련 논문의 제목, 저자, 출간연도 그리고 초록을 수집하였다.

NAVER News에서 뉴스 본문을 가져오는 과정에서는 두 가지의 문제가 발생하였다. 먼저, 뉴스 링크의 형식이 다른 경우가 존재하여 일관된 방법으로 html parsing 할 수 없어 제외해야만 했다. 또한 NAVER API에서 한번에 출력가능한 뉴스의 개수를 1000개로 제한되었다. 만약 키워드를 "드론"으로 하여 검색한다면 최근 3달간의 뉴스가 출력되고 그보다 더 이전의 뉴스에 접근할 수가 없다. 따라서 키워드를 "화물 드론"으로 설정하여 범위를 좁히고, sort_type을 'sim'으로 지정하여 관련도순으로 정렬하였다. 총 436개 관련 기사의 제목, 링크, 보도일 그리고 본문을 수집하였다.

II-2 데이터 수집

beautifulsoup을 사용하여 드론' 키워드를 갖는 RISS 국내 학술논문의 제목, 저자, 출간연도 그리고 초록을 데이터 프레임으로 저장하였다.



RISS 사이트에서 조건에 맞게 검색한 결과 홈페이지이다. 제시된 논문 리스트는 "srchResultList₩" 클래스를 가진다. 논문 제목은 하위 클래스인 "p.title"에 있고, 저자, 학회, 연도는 "p.etc"에 띄어쓰기를 구분자로 하여 함께 존재한다. 따라서 구분자 '₩n'을 기준으로 리스트로 저장한 다음, 인덱싱하여 각각을 저장하였다.

한편, 초록은 "p.preAbstract"에 있다. 일부 초록이 존재하지 않는 논문의 경우 예외 처리하였다. 논문 데이터프레임(df_RISS)은 다음과 같다.

	제목	초록	저자	학회	출간연도
0	영상 드론의 운동성과 보기 양식에 관한 소고	드론은 3차원 입체 공간에서 활동하는 증강현실 기계다. 이 글은 문화기술의 재매개적...	임종수(Jongsoo Lim)	한국언론학회	2017
1	드론영상의 보기 양식과 하이퍼 리얼리티 연구	None	임종수(Jongsoo Lim),이소현(Sohyun Lee)	KBS 공영미디어 연구소	2018
2	드론 활용 목표물 추적 응용에서의 인공지능 작업 실행 효율 비교 분석	None	손경환(Kyunghwan Son),David Hostallero,김대우(Daewoo...	한국통신학회	2018
3	메쉬 네트워크에서 드론의 유용성 검증	This paper investigates the usefulness of dron...	정예림(Yerim Chung),박태준(Taejoon Park),민윤홍(Yunhong...	한국경영과학회	2016
4	집회 및 시위에 관한 법률상 경찰의 드론에 의한 채증 규정 연구	먼저 현 집시법 개정안 제19조 제3항에서 경찰의 집회 장소에서 드론의 사용을 일체...	이희훈	한국입법학회	2022
...	...	...	...	...	...
1108	핀테크 혁명과 보험업의 미래	본 연구에서는 디지털화와 인공지능 관련하여 보험업의 미래모습과 법적인 과제에 대해 ...	이성남(Sungnam Lee)	한국보험법학회	2022
1109	자유학년제 연구학교의 4차 산업혁명 관련 교육프로그램에 참여한 특수학교 교사의 경험	Based on the experience of teachers who operat...	김혜진,임경원	한국통합교육학회	2022
1110	전문대학 성인학습자의 비교과교육 요구분석에 기반한비교과교육 개선방안 연구	A Study on Improvement Plans for Extracurricul...	정복임	인문사회 21	2022
1111	인공지능 강사에 의한 ESG경영 교육이 학습자의 ESG에 대한 자기효능감에 미치는 영향	[Purpose]There are few studies on whether AI-b...	유가애,배수진,권오병	한국경영교육학회	2022
1112	Cyber상의 위협과 재난으로부터 자유와 안전을 보장할 제어국가의 법적 규제 - 헌...	The 4th industrial revolution is changing the ...	성봉근	한국국가법학회	2022

1113 rows × 5 columns

NAVER News의 경우 사전에 NAVER API를 발급받은 다음, 작업 환경에 "NAVER_Client_ID.txt", "NAVER_Client_Secret.txt"을 저장하여 접속하였다. 그 다음 조건에 맞게 검색한 결과 한 번에 100개씩 총 1000개 뉴스의 정보를 얻을 수 있었다. 100개의 뉴스 정보들은 items 객체 안에서 리스트로 존재하였고, 리스트의 각 요소는 딕셔너리 형태로 title, originallink, link, description의 key, pubDate를 가지고 있다. originallink는 원문의 링크이므로 제외하고, description은 본문 전체를 나타내지 않으므로 제외하였다. 또한

pubDate는 'Wed, 23 Feb 2022 17:15:00 +0900'의 형태를 띄고 있어 시각화를 위해 '2022-2-23'의 형태로 수정하였다. link에 접속한 다음 "article.go_trans_article_content"에 접근하여 존재하는 텍스트를 병합하여 수집하였다. 뉴스 데이터프레임(df_NAVER)은 다음과 같다.

	뉴스 제목	뉴스 링크	뉴스 보도	뉴스 본문
0	중동으로 눈 돌린 중국 드론?...이항, 첫 자율 유인 비행	https://n.news.naver.com/mnews/article/374/000...	2024-5-8	₩n₩n₩n₩n₩n₩n₩n₩n₩n[시험 비행하는 EH216-S (홍콩 사우스차이...
1	포천시, 드론·UAM 활용 기회발전특구 유치에 '바짝'	https://n.news.naver.com/mnews/article/018/000...	2024-5-8	₩n₩nKDI·軍·민간기업 참여 '군용드론 시험평가지원센터 구축 용역' 개시₩n드론...
2	항공기 외관 검사 '4시간→50분' 단축 가능... 드론 개발 각축전 치열	https://daily.hankooki.com/news/articleView.ht...	2024-5-10	None
3	[산업워치] 대한항공 '노사상생 협약식'·한화오션 '초대형 컨테이너선 인도'...	http://www.newswatch.kr/news/articleView.html?...	2024-5-10	None
4	中 드론 제조업체 이항, 중동서 첫 자율 유인 비행 완료	https://n.news.naver.com/mnews/article/001/001...	2024-5-8	₩n₩nUAE 아부다비서 시험 비행..."중동서 드론 상업비행 곧 시작"₩n₩n₩n...
...	...	...	...	...
995	프랑스 드론 시장 동향	http://news.kotra.or.kr/user/extra/kotranews/b...	2021-4-2	None
996	소형 드론들간 협력으로 무거운 짐 배송하는 기술 개발	http://www.irobotnews.com/news/articleView.htm...	2021-3-29	None
997	영국 항모에서 이착함에 성공한 모하비 드론 [최현호의 무기인사이드]	https://n.news.naver.com/mnews/article/081/000...	2023-11-25	₩n₩t₩t₩t[서울신문 나우뉴스]₩n₩n₩n₩n₩n₩n₩n₩nHMS 프린스 오...
998	로봇에 드론까지...배민·도미노·KT '무인 배달' 속속 나서	http://www.opinionnews.co.kr/news/articleView....	2021-3-25	None
999	"우크라, 4천km 떨어진 러 시베리아 화물열차 공격 "	https://www.jeonmae.co.kr/news/articleView.htm...	2023-12-1	None

1000 rows × 4 columns

III-1 네트워크 분석을 통한 드론의 국내 연구 동향

[데이터 전처리]

앞서 수집한 두 데이터 프레임에서 텍스트에 해당하는 부분인 df_RISS['초록']과 df_NAVER['뉴스 본문'] 각각 array type에서 list type으로 변환한 다음 병합하여 길이가 1446인 리스트(data)를 만들었다. 전체 텍스트에 대하여 키워드 후보가 되는 명사를 추출하기 위해 형태소 분석을 진행하였다. 한국어에 대한 형태소 분석능력이 비교적 뛰어나다고 알려진 Kiwi를 사용하여 data를 하나의 텍스트(text)로 병합하여 태그가 'NN'으로 시작하거나 'SL'로 시작하는 경우 반환하는 명사 추출함수 "noun_extractor_kiwi"를 설정하였다.

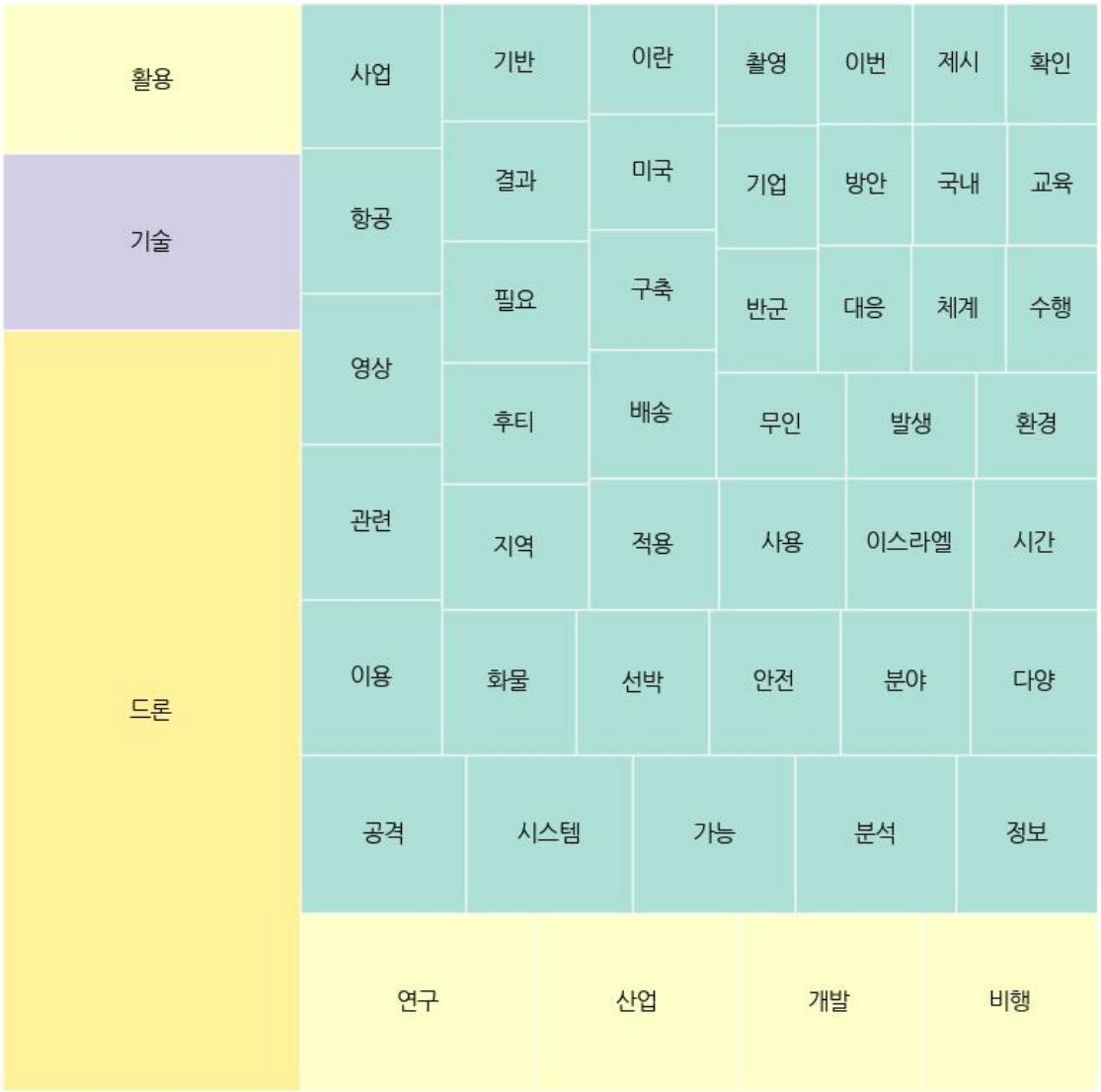
[시각화 결과]

아래는 앞서 진행한 형태소 분석 결과에 대한 wordcloud 이미지이다. 왼쪽은 태그를 모든 명사로 설정한 경우이고, 오른쪽은 태그를 고유명사인 'NNP'로만 제한한 경우이다.

(1)모든 명사태그를 적용

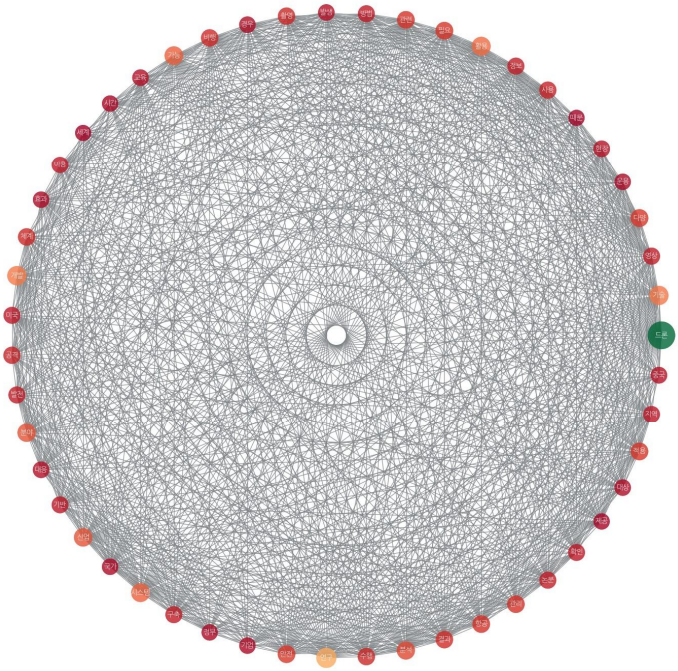
(2)고유명사만 적용

아래는 Treemap의 결과이다.

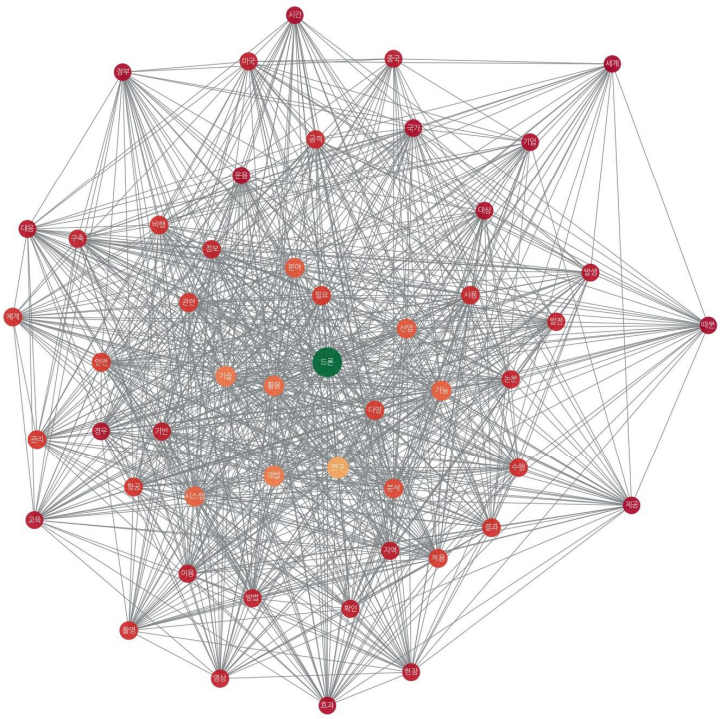


biagram을 통해 두 단어간 토큰을 만들고 networkx에서 시각화를 위해 노드 수를 9489개에서 51개로, edge 수를 86338개에서 960개로 조정하였다.

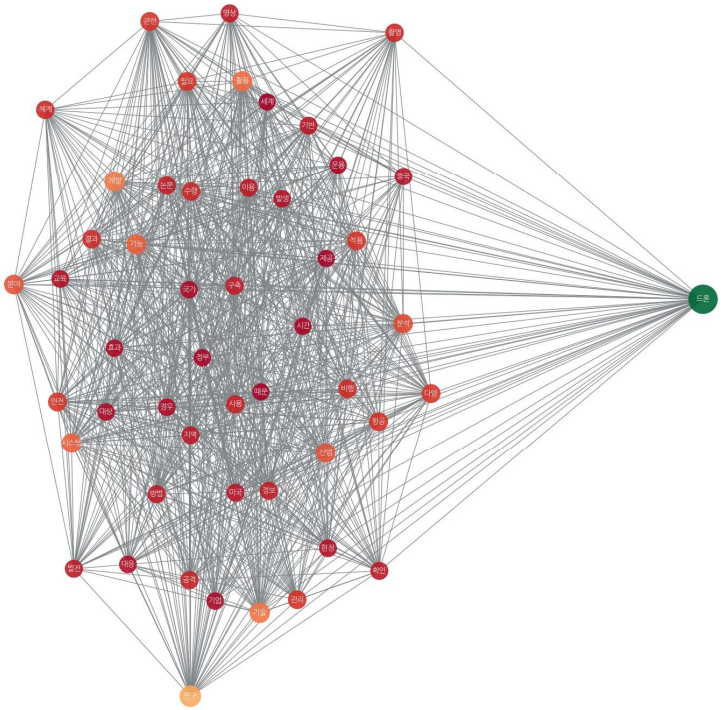
Circular Layout about Drone



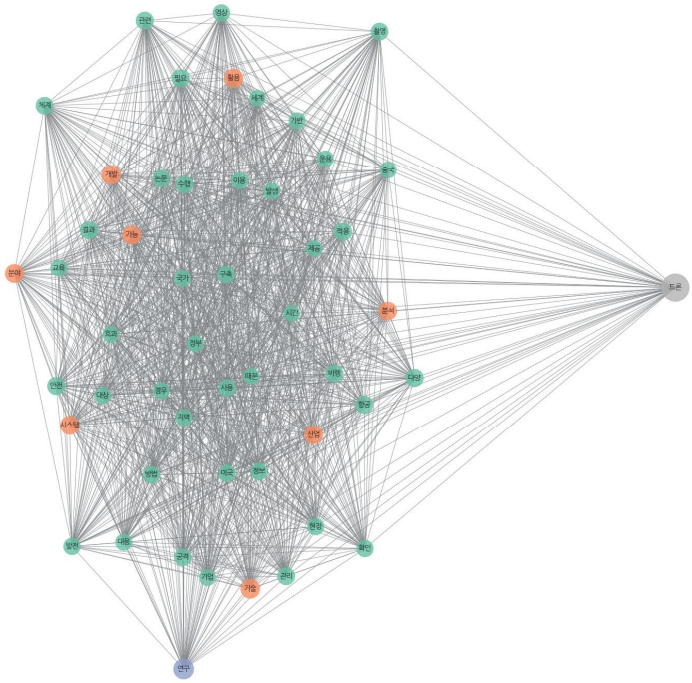
Spring Layout about Drone

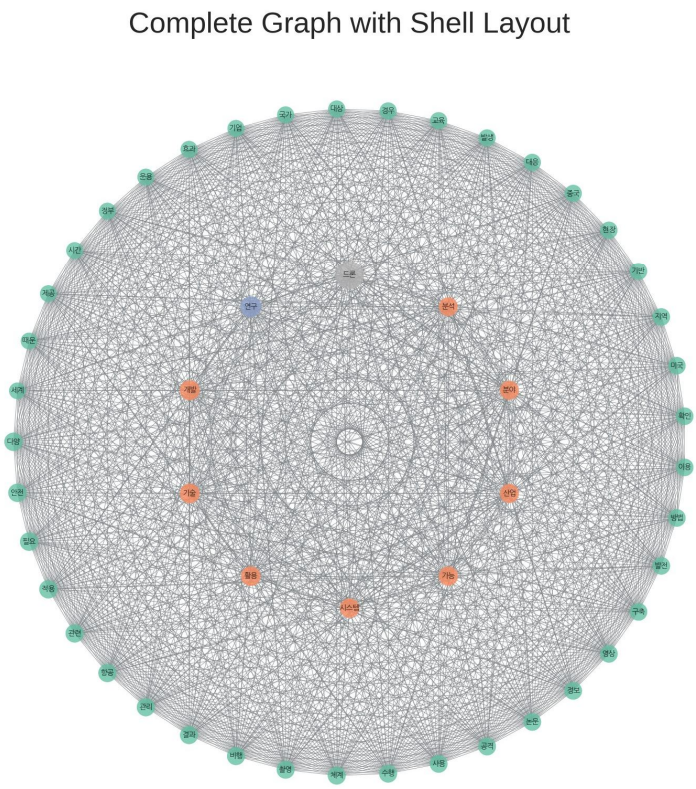


Kamada Kawai Layout about Drone



Complete Graph with Kamada Kawai Layout





향후 활동 계획

1. 관련 뉴스의 내용에 따른 주가 변동의 상관분석
2. 가능하다면 Topic Modeling을 사용하여 잠재 의미 분석 진행