경제 텍스트 데이터를 활용한 키워드 분석방안

서울시립대학교 전자전기컴퓨터공학부
김한준 교수
khj@uos.ackr

목차

연구배경

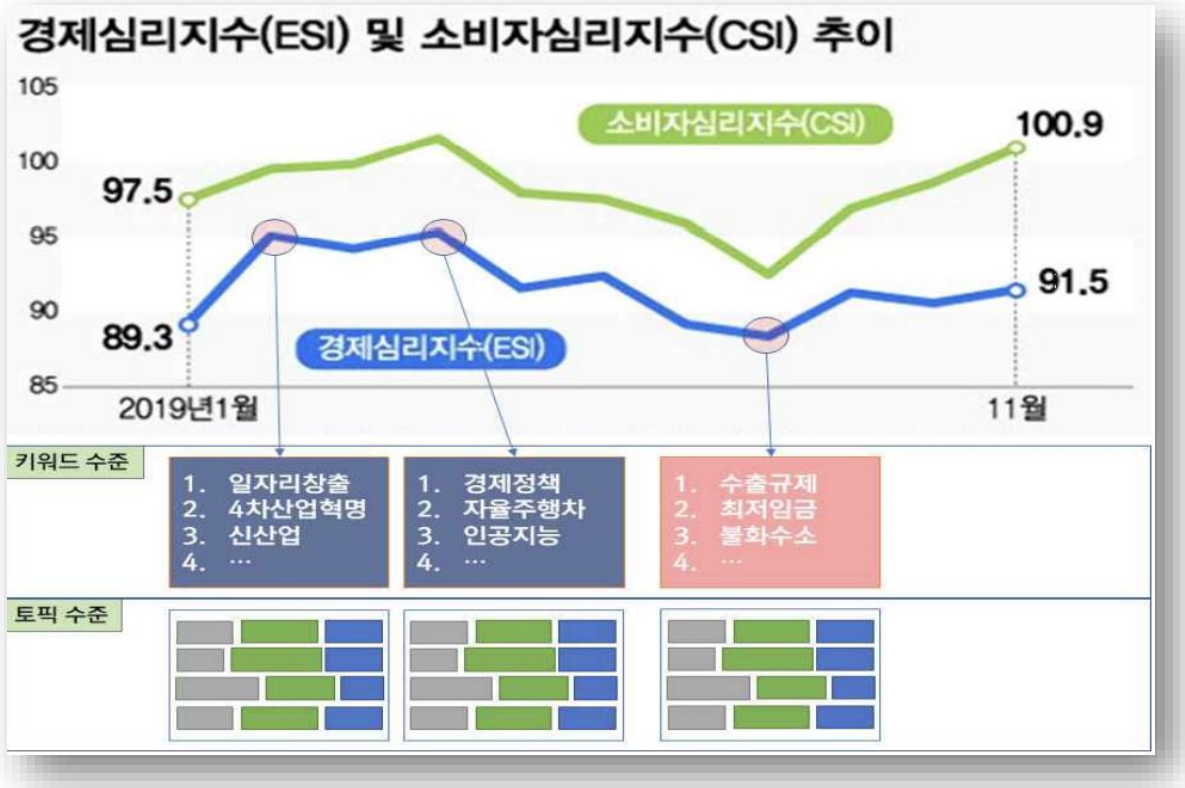
- **경제 텍스트마이닝**
  - 국내외 경제의 빠른 변화에 따라 신속한 경제동향 분석과 경제정책 수립·점검을 위해서, 경제 관련 뉴스 기사에 대한 텍스트마이닝 기술의 활용 증가
  - text classification/clustering, keyword extraction, information extraction, sentiment analysis 등

- **경제심리 관련 키워드 분석**
  - 경제심리에 영향을 미치는 보다 정확한 키워드 및 토픽 정보를 추출하여, 경제심리지수(ESI)와의 관련성을 추적

연구목표

연구목표

- 특정 시기의 경제심리지수가 임계점 이상(또는 이하)로 증가(또는 감소) 상태인 경우에, 해당 기간의 뉴스 기사로부터 관련 요인을 추론할 수 있는 키워드 및 토픽의 추출



기존 topic modeling 기술의 문제점

1. 토픽 내 키워드들이 분절된 단일 단어가 많아서 해당 토픽의 유의미성이 낮음
2. 하나의 토픽은 연관정보가 없는 키워드들의 집합으로 표현-> 요인분석 어려움

Topic Modeling (LDA) 분석

- Topic Modeling 결과: 하나의 topic은 키워드들의 집합체로서, 키워드 간 연관성을 고려하지 않기 때문에 (스토리텔링 수준) 내재된 의미 파악이 어려움

토픽모델링

전체	Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic9	Topic10
미국	피해자	미국	인상	금리	사고	주가	중국	민주당	게임	결혼
중국	이유	북한	경영	취업	헝가리	상승	미국	청와대	협력	배우
북한	판단	트럼프	취업	유지	유람선	외국인	화웨이	자유	모델	마약
국민	범행	장관	부담	수출	수색	하락	홍콩	국민	글로벌	여성
사고	여성	스웨덴	고용	개선	인양	점수	삼성전자	여사	브랜드	무대
트럼프	변호사	양국	현대중공업	인하	실종자	수익률	반도체	문재인	선정	고백
헝가리	단체	멕시코	개선	일자리	부다페스트	거래	무역전쟁	비판	스타트업	상대
장관	피해	방문	노조	경기	다뉴브강	상장	장비	장관	소비자	등장

연구개발내용 (1/2)

- 클러스터링 기반 경제 관련 뉴스 데이터 수집
  - 최근 경제 상황 은 정치/사회적 상황, IT기술, 세계정세 등과 긴밀하게 연관
  - 토픽 내 키워드들은 ‘경제’와 직/간접적으로 연관된 용어 포괄 필요
  - 뉴스 기사 데이터에 대한 클러스터링 및 엔트로피 분석

- 토픽모델에 사용되는 양질의 키워드 추출
  - 4단계 불용어 제거 작업 수행
  - 품사 태깅 결과를 활용하여 유의미한 복합단어 추출

연구개발내용 (2/2)

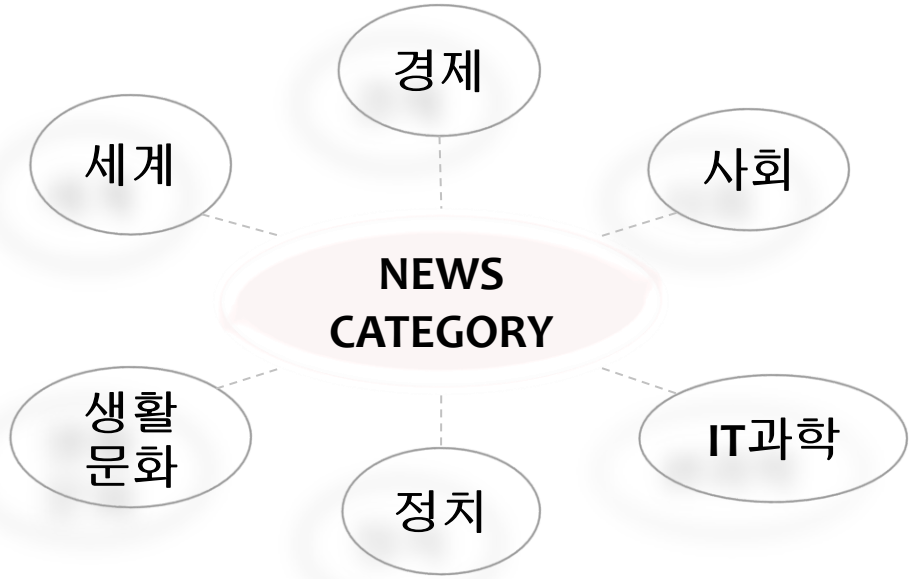
- 토픽 내 키워드들간 의미적 연관성을 표현한 토픽그래프의 생성
  - 각 토픽에 포함된 단어들간의 의미적 연관분석(Association Analysis)을 수행하여, 임계값 이상의 연관도를 가지는 단어들간의 관계성을 그래프(또는 네트워크) 형태(topic graph)로 표현
  - 토픽 내 연결된 키워드 중의 일부 는 이전 단계에서 인식하지 못한 복합어가 추출
  - 키워드간 의미적 연관도 산정을 위해 뉴스 데이터에 대한 Word2Vec 학습모델 및 Node2Vec 비지도 임베딩 학습모델의 생성
- 완전그래프 인식을 통한 토픽 내 소단위 개념의 추출
  - 토픽그래프 내 단어 연관망을 구성한 후, 완전그래프 단위를 인식함으로써 토픽 내 소단위 개념을 추출

뉴스 데이터

분석 활용 데이터

- 네이버 뉴스 기사 2005.01.01 ~ 2020.03.22
- 날짜 별 CSV 파일 형태 (2,000 ~ 20,000 개의 레코드로 구성)

PUBLISH_DATE	CATEGORY1	CATEGORY2	AUTHOR	TITLE	CONTENT
출간 일자	분류 코드	카테고리	신문사	제목	내용



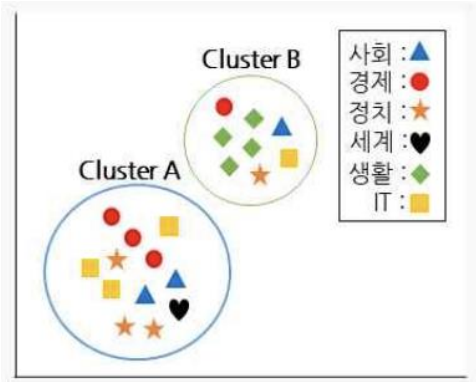
<뉴스 카테고리>

부분 발췌

- 편의상, 2019.05.19 ~ 2019.06.15 기간 동안 각 날짜 별로 500개의 기사를 랜덤 추출하여 총 14,000개의 데이터 수집
- 데이터의 CONTENT 컬럼 추출

클러스터링 기반 경제 뉴스 데이터 선정

- **뉴스기사 데이터에 대한 클러스터링 수행**
  - 클러스터 = 유사한 텍스트의 집합체
  - K-means 알고리즘 활용



- **클러스터 엔트로피 산정**

$$Entropy_A = -\left(\frac{2}{12}log\frac{2}{12} + \frac{3}{12}log\frac{3}{12} + \frac{3}{12}log\frac{3}{12} + \frac{1}{12}log\frac{1}{12} + \frac{3}{12}log\frac{3}{12}\right) = 1.545$$

$$Entropy_B = -\left(\frac{1}{8}log\frac{1}{8} + \frac{1}{8}log\frac{1}{8} + \frac{1}{8}log\frac{1}{8} + \frac{4}{8}log\frac{4}{8} + \frac{1}{8}log\frac{1}{8}\right) = 1.386$$

$$Entropy_W = (Entropy_A * \frac{12}{20}) + (Entropy_B * \frac{8}{20}) = 1.481$$

Number of cluster	Category in cluster	Weighted entropy
K=1	{사회, 경제, 정치, 생활, 세계, IT}	2.076
K=2	{경제} | {사회, 정치, 생활, 세계, ,IT}	0.556
	{경제, IT} | {사회, 정치, 생활, 세계}	0.609
	{사회, 경제, 정치, IT} | {생활, 세계}	0.449
	…	
	{사회, 경제, 정치, 세계, IT} | {세계}	0.345
	{사회, 경제, 정치, 세계, IT} | {생활}	0.217
K=6	{사회} / {경제} / {정치} / {생활} / {세계} / {IT}	1.5328

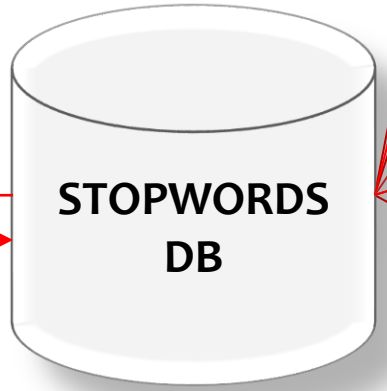
불용어(STOPWORD) 처리

- 뉴스기사 내 유의미하지 않은 (Topic Modeling에 도움이 되지 않는) 단어들을 제거하여 Topic Modeling 결과의 품질을 높이는 것이 목적
- MySQL을 사용하여 불용어(stopwords) DB 구축 및 연동

뉴스기사

LOAD

INSERT



<불용어 Database 구축>

[영상] "헝가리 다뉴브강 침몰 유람선에 한국인 33명 탑승"

입력 2019-05-30 08:24 | 수정 2019-05-30 08:37

앵커

정리를 해보자면 현지시간 29일 저녁 9시에 헝가리 부다페스트 다뉴브강에서 유람선이 침몰을 했고 요.

인솔자까지 한국인 33명이 탑승을 하고 있었습니다.

지금까지 모두 7명이 사망한 것으로 외교부에서 확인이 되고 있습니다.

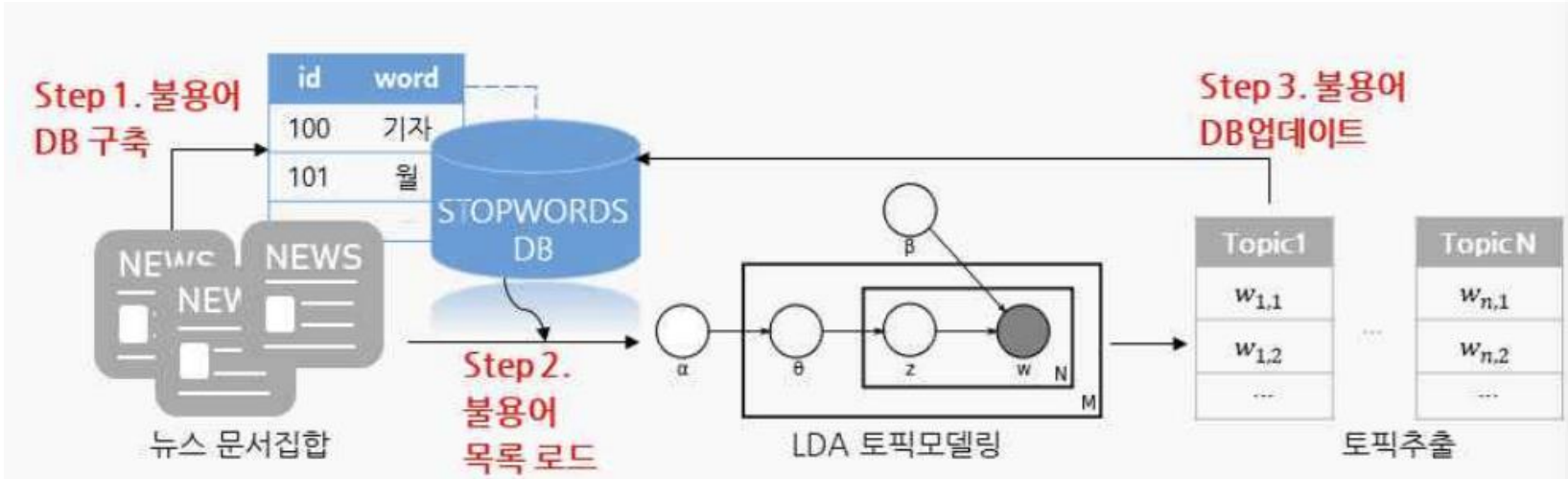
그렇다면 자세한 내용 국제부 연결해서 조금 더 알아보도록 하겠습니다.

윤효정 기자 전해주시죠.

기자

헝가리 부다페스트 다뉴브강에서 현지 시간 29일 밤 9시쯤 우리 시간으로는 오늘 새벽 4시쯤 한국인 단체관광객을 태운 유람선이 다른 크루즈선과 충돌한 뒤 침몰했습니다.

불용어 처리



stopword_no	stopword	stopword_no	stopword
635	그	481	개발
620	그곳	622	개요
179	그래픽	875	개월
607	그러나	308	개인
570	그런데	491	개최
554	그리고	281	갤러리
488	그림	467	거의
578	금융	409	검찰
307	금지	454	검토
523	기간	633	것
181	기관	368	결과
473	기념	147	결정
35	기능	346	하고
107	기록	26	하기
529	기반	566	하나
203	기사	411	하는
27	기술	453	하루
562	기억	456	하지만
372	기업	528	학교
535	기온	550	학생
54	기자	250	한국
559	기존	304	한국경제
76	기준	67	한국일보
574	기획재	459	한다는

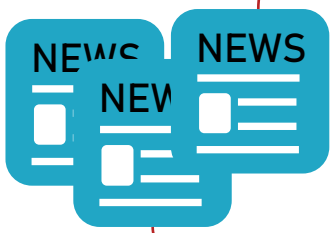
- 기본 한글 불용어 등록: 접속사, 부사 등
- 문서출현빈도 DF 값이 큰 단어를 불용어로 등록: '신문', '기자' 등
- 토픽모델링을 통해 추출된 토픽의 결과를 확인하여, 불용어 수준 단어 등록

형태소 분석: 품사 태깅

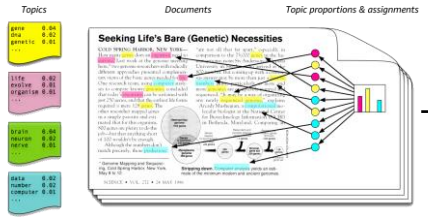
- 한글형태소분석: kkma, Komoran, Okt, Mecab 등

단어	Okt	Mecab
허위자료	허위/Noun + 자료/Noun	허/IC + 위자료/NNG
병역필	병역/Noun + 필/Noun	병/NNG + 역필/NNG
실업난	실업난/Noun	실업/NNG + 난/NP+JX
취준생	취준생/Noun	취/NNP + 준/NNP + 생/NNP
평창동계올림픽	평창동계올림픽/Noun	평창동/NNP + 계/XSN + 올림픽/NNP
호캉스	호캉스/Noun	호/NNG + 캉/JKB + 스/IC
최대주주	최대/Noun + 주주/Noun	최/XPN + 대주주/NNG
세계인권선언	세계인권선언/Noun	세계인/NNG + 권/XSN + 선언/NNG
불법조업	불법/Noun + 조업/Noun	불/XPN + 법조/NNG + 업/NNG
...		

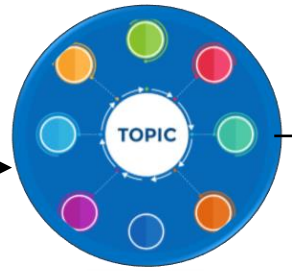
토픽그래프 생성



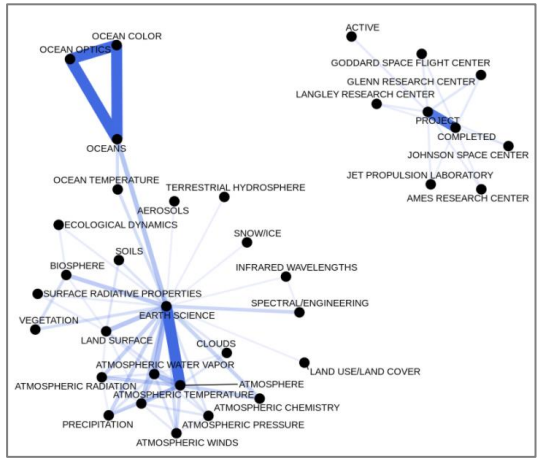
뉴스기사



LDA

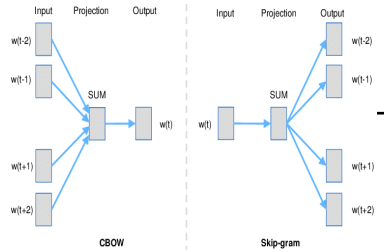


토픽모델

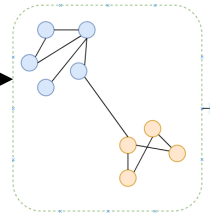


토픽그래프

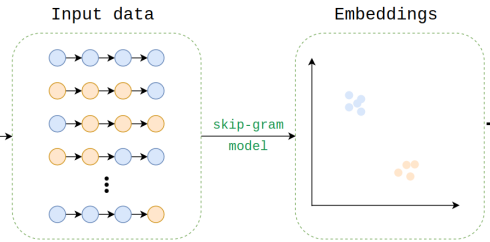
Topic 내 키워드 간
시맨틱 유사도 계산



Word2Vec 모델



가중그래프



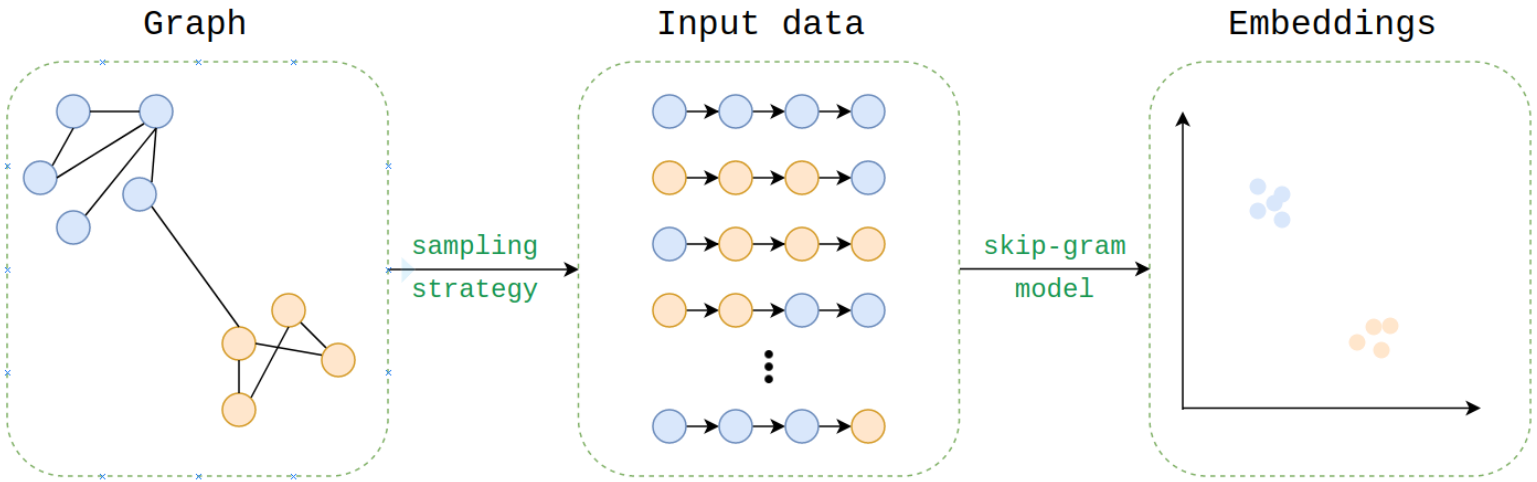
Node2Vec 모델

Step 2. 시맨틱 유사도 기반 토픽그래프 생성

Step 1. 이중 임베딩 모델 (Word2Vec & Node2Vec) 구축

Node2Vec Embedding

- Node2Vec 이란?
  - Graph Embedding 기법 중 하나로서, graph 내 각 node를 vector로 표현하는 representational learning 기법
  - 기존 임베딩 기법은 feature 자체를 학습하는데, Node2Vec을 이용하면 graph 내 연결 패턴을 학습하여 보다 정교한 embedding이 가능함



토픽그래프 생성

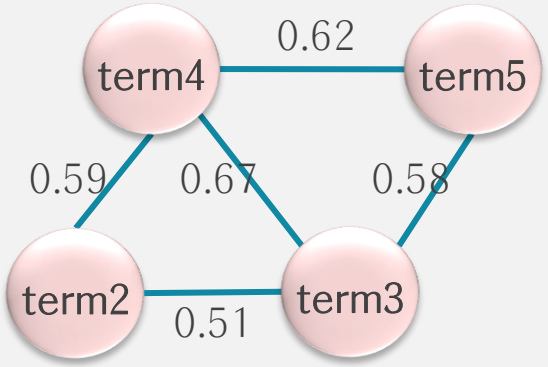
	term1	term2	term3	term4	term5
term1	1.00	0.48	0.38	0.44	0.25
term2	0.48	1.00	0.51	0.59	0.22
term3	0.38	0.51	1.00	0.67	0.58
term4	0.44	0.59	0.67	1.00	0.62
term5	0.25	0.22	0.58	0.62	1.00

단어-단어 유사도 행렬

상위 5개

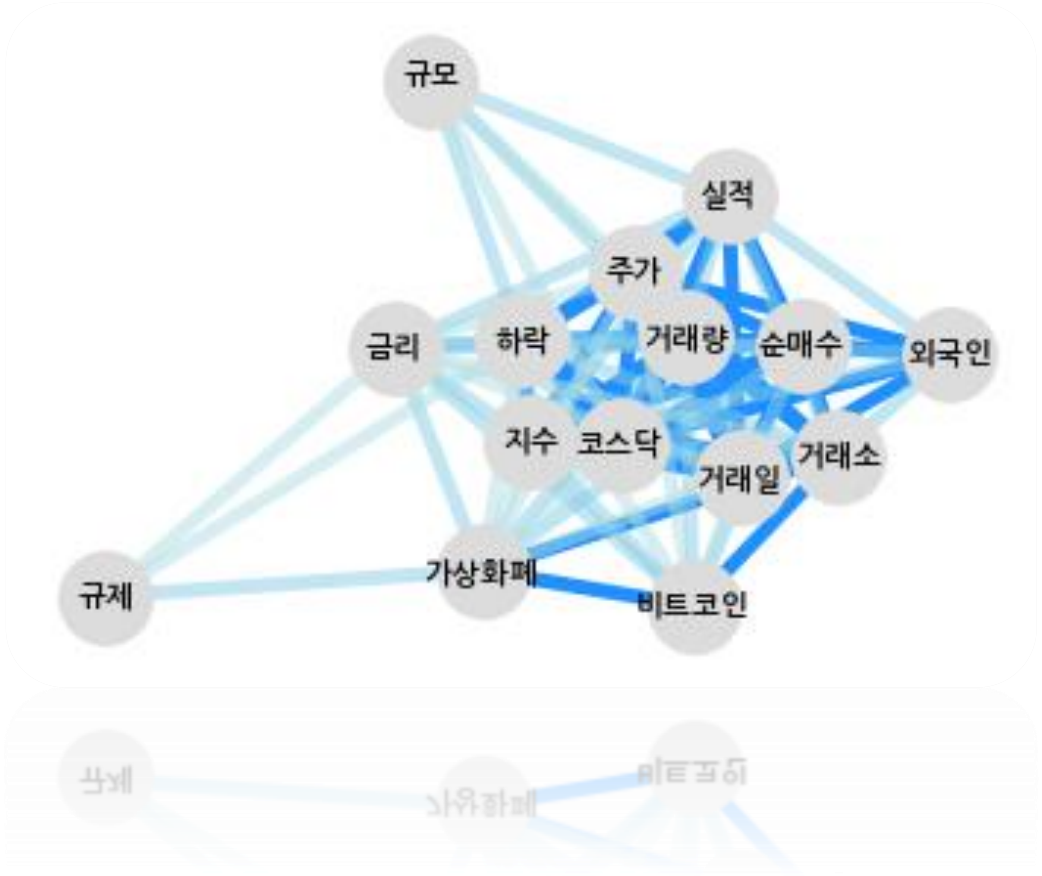
※가능한 전체 간선 수 : 10개

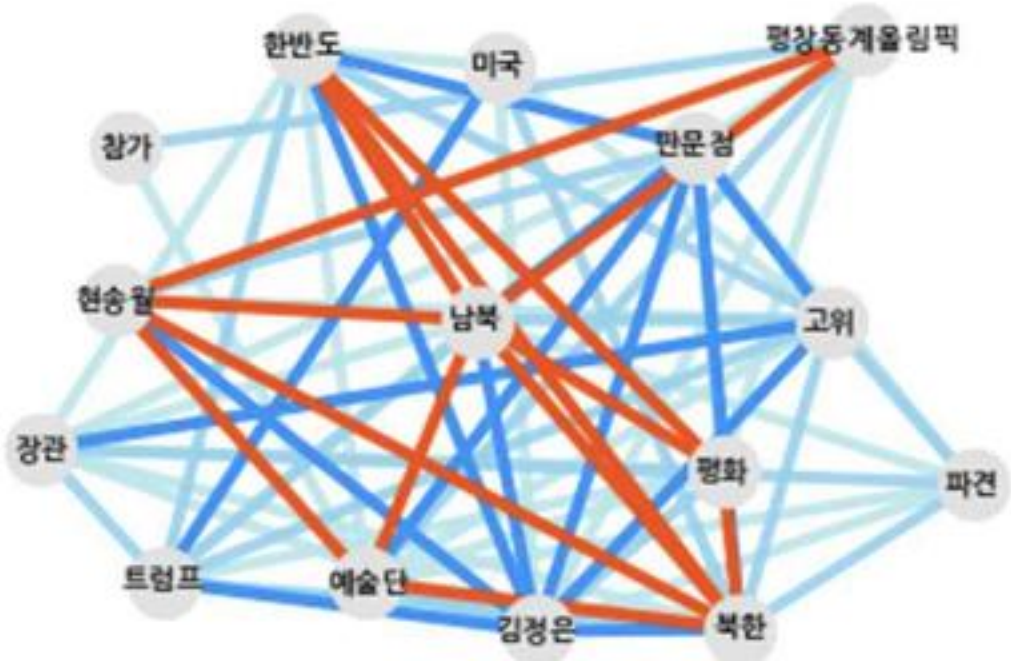
상위값 기반 토픽그래프



토픽그래프

토픽그래프 생성

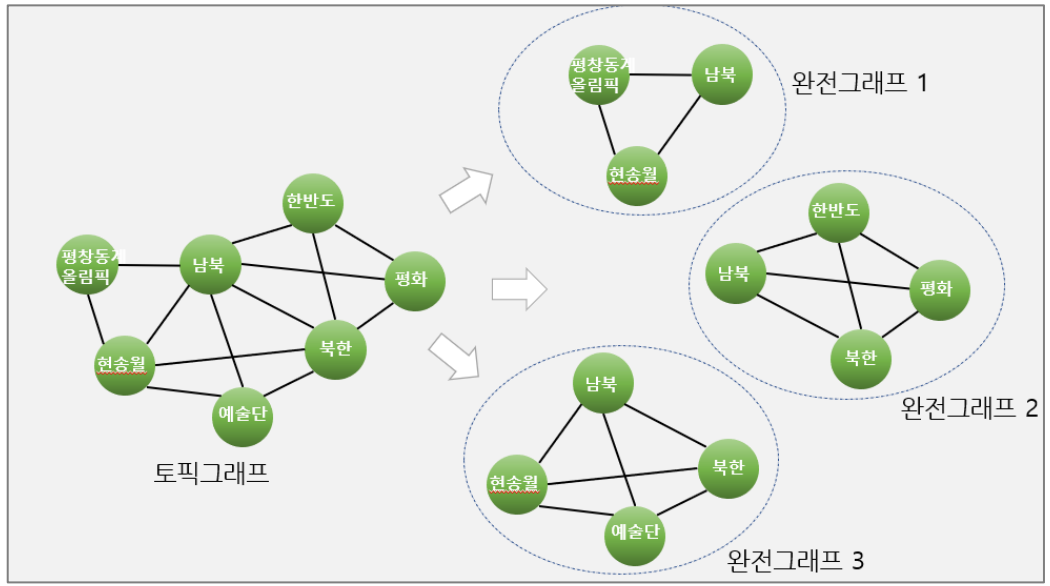

규모
실적
주가
금리
하락
거래량
순매수
외국인
지수
코스닥
거래일
거래소
규제
가상화폐
비트코인


한반도
미국
평창동계올림픽
참가
판문점
현송월
남북
고위
장관
평화
파견
트럼프
예술단
김정은
북한

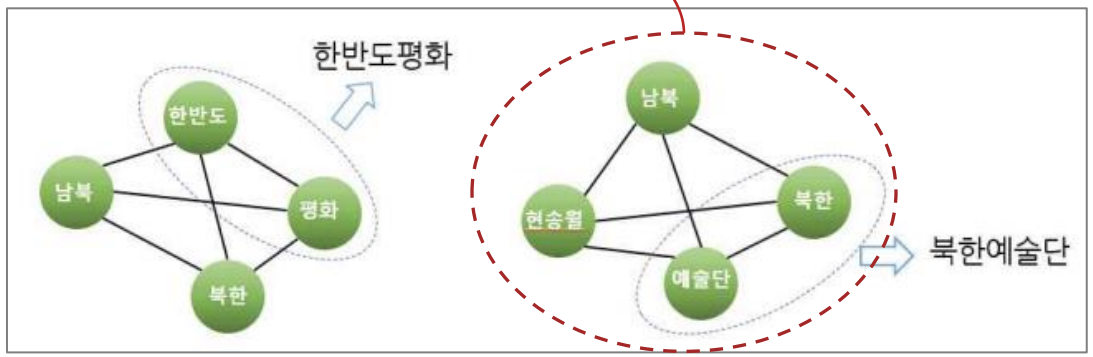
토픽그래프 생성

완전그래프 생성

- 완전그래프(complete graph): 모든 node간 edge존재
- 소개념 또는 복합어 추출 가능



[남북] 회담에 참여한 [북한예술단]의 [현송월]



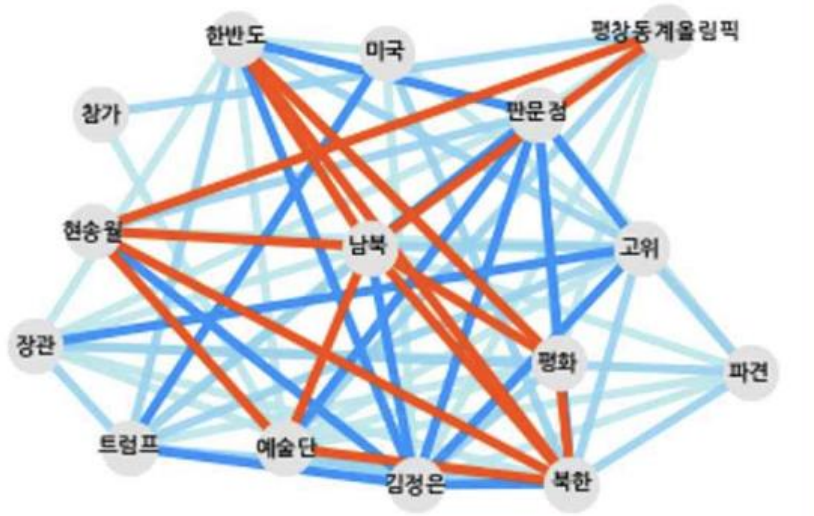
토픽그래프 분석

2018년 1월 뉴스 기사에 대한 토픽그래프 분석

- 2018년 1월 경제심리지수(ESI)는 97.3이며, 전월 대비 2.6p 하락

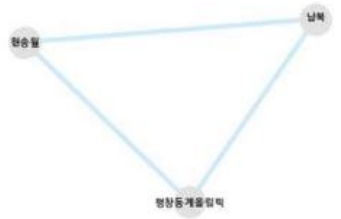
토픽2	
북한	0.044
남북	0.024
평창	0.017
올림픽	0.013
통일부	0.012
파견	0.011
미국	0.011
트럼프	0.011
대표단	0.010
참가	0.010
장관	0.009
평창동계 올림픽	0.009
예술단	0.008
판문점	0.008
한반도	0.007

○ 토픽그래프

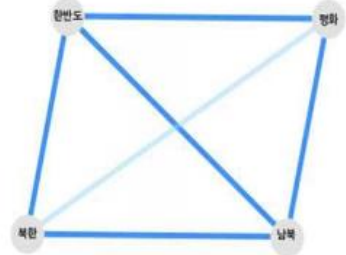


※ 완전그래프는 빨간색으로 표시

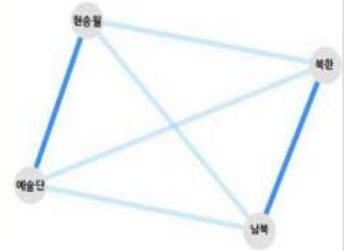
○ 완전그래프 1



○ 완전그래프 2



○ 완전그래프 3



▪ 토픽그래프의 해석

완전그래프 1	• [평창동계올림픽]은 [남북] 관계 개선에 도움을 준 가운데, 화제가 되고 있는 [현송월] • [남북] 회담에 데뷔한 [현송월]이 참여한 [평창동계올림픽]
완전그래프 2	• [한반도] [평화]를 위해 [북한]과 진행된 [남북] 회담 • [북한]과의 [남북] 회담을 통해 다가오는 [한반도][평화]
완전그래프 3	• [남북] 회담에 데뷔한 [북한] [예술단]의 [현송월]
토픽그래프	• 평창동계올림픽에 현송월이 일원인 북한의 예술단이 참가하고, 판문점에서 한반도 평화 협정이 오고가는 등 남북 간 교류가 이루어짐

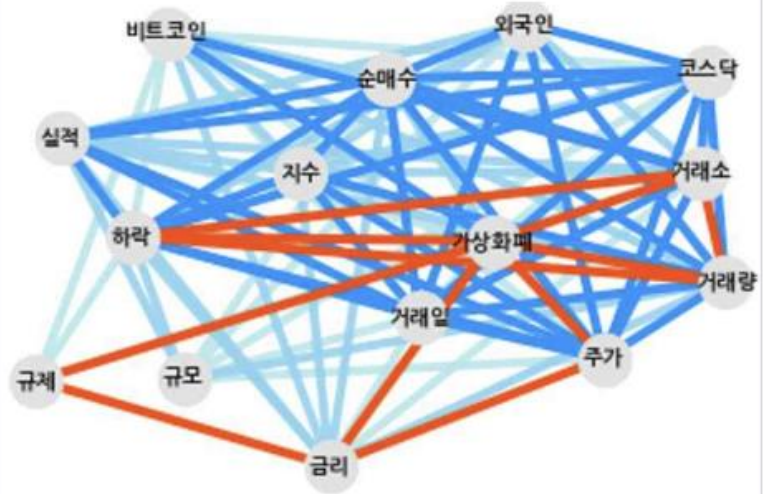
토픽그래프 분석

2018년 1월 뉴스 기사에 대한 토픽그래프 분석

- 2018년 1월 경제심리지수(ESI)는 97.3이며, 전월 대비 2.6p 하락

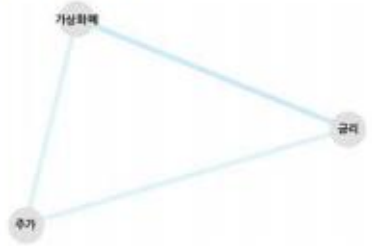
Topic5	
가상화폐	0.021
상승	0.016
거래	0.015
거래소	0.010
하락	0.009
외국인	0.009
코스닥	0.007
비트코인	0.007
거래량	0.007
투자자	0.007
주가	0.006
실적	0.005
지수	0.005
연구원	0.004
규모	0.004

○ 토픽그래프



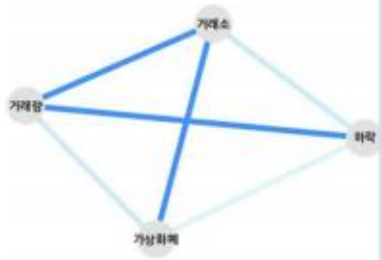
※ 완전그래프는 빨간색으로 표시

○ 완전그래프 1



○ 완전그래프 2

○ 완전그래프 3



▪ 토픽그래프의 해석

완전그래프 1	• [가상화폐]는 [금리]와 [주가]에 영향을 끼침
완전그래프 2	• [가상화폐]의 [규제]가 발생함
완전그래프 3	• [가상화폐] [거래소]에서 [거래량]이 [하락]함
토픽그래프	• 가상화폐인 비트코인 규제 정책이 시행됨으로, 관련 거래량 및 주가 등이 하락함

결론

요약

- 국내 경제 관련 뉴스기사를 통합·연계하여 경제지수에 영향을 미치는 키워드 및 토픽 정보 추출
- 경제심리지수(ESI)와의 관련성 추적을 위한 토픽그래프 생성 알고리즘 개발
  1. 클러스터링 기반 포괄적 경제 관련 카테고리의 선정
  2. W2V+N2V 병합 임베딩 기반 정확한 연관도 산정
  3. 토픽모델 내 키워드에 대한 토픽그래프