1. 분석 단계별 절차

→텍스트 데이터 전처리

csv 파일에 인덱스를 생성하고, document number를 생성하였다. 문서를 개별 문장 분리하여 형태소로 토큰화 하였다. 불용어 처리는 따로 하지 않았고, 형태소 분석에 필요한 어간을 추출하였다.

→특징변수 추출 및 표현

반도기반 TF_IDF를 적용하였고, 형태소 분석기는 Komoran 을 사용하였다.

분석에 포함한 형태소는 “일반명사NNG, 고유명사NNP, 형용사VA, 일반부사MAG, 접속부사MAJ” 이다.

→특징변수 시각화

트리맵과 워드 클라우드를 사용하여 특징 변수를 시각화 하였다.

→토픽분석(LDA) 기법 선택

TF-IDF 기준으로 Term-document matrix 생성 → LDA 모델 생성→ LDA 모델 결과 출력

1. 분석 단계별 절차

→텍스트 네트워크 분석 기법 선택

Term-document matrix 생성→ co-occurence matrix 생성→ word similirity matrix 생성→ 어휘 공기(단어 동시 등장) 네트워크 생성→ 최소 신장 트리 기반의 어휘 공기(단어 동시 등장) 네트워크 생성

→시스템 결과 테스트 및 평가

분석 결과 해석과 한계점 평가

2. 특징변수 표현 및 추출방법 설명

→특징변수 표현 방법: TF-IDF 반도기반

텍스트 분석에 사용되는 머신러닝 알고리즘은 입력 변수로 숫자로 표현된 값만 인식할 수 있으므로 주어진 문서의 의미 단어를 숫자로 변환하여야 한다. 즉 자연어를 숫자로 변환해야 하는데 이를 특징 변수 표현이라고 한다. 텍스트 분석에 사용되는 알고리즘은 특징 변수를 기반으로 출력 변수를 생성하므로 특징 변수가 분석 알고리즘의 정확성, 효율성에 미치는 영향이 매우 크다. Bag of Word는 자연어 처리에서 널리 쓰이는 가장 기본적인 특징 표현을 위한 방법으로 bow 단계를 거친 후 TF 나 TF_IDF 빈도 기반 특징 변수 추출이 가능하다. 문서 집합에서 문서- 단어 행렬로 표현할 때 행렬의 값을 가중치라고 한다. BoW는 문서에 등장한 단어의 빈도를 가중치로 사용한다. 본 분석에 사용된 빈도 기반 특징 변수 추출 방법은 TF_IDF 이다. TF_IDF 는 의미없이 흔히 나오는 단어를 제외 하는 방법으로 단어 빈도-역문서 빈도를 의미한다. TF_IDF 는 문서 별로 자주 등장하는 단어에 낮은 가중치를 주고 드물게 나오는 단어에는 높은 가중치를 주는 방법으로 단어 빈도와 문서 빈도 역수를 사용해 가중치를 생성한다.

→형태소 분석기: Komoran 형태소 분석기 사용

형태소 분석이란 의미의 초소 단위인 형태소를 사용해 단어가 어떻게 형성되었는 지에 대한 문법적 분석이다. 하나의 단어로부터 의미를 갖는 최소 단위인 형태소로 분절하는 과정이다. 문서 별 단어를 분리하여 단어 별 품사를 태깅하여 추후 분석에 사용한다. 접미사가 발달한 교착어인 한국어는 후보 생성 단계에서 형태소를 분리하고, 변이 형태소의 경우 원형으로 복원한다. 다음으로 후보들 중에서 사전 검색을 하고 결합 제약 규칙 등을 적용하는 후보 선택 단계를 거친다. 형태소 분석기를 통한 입력 문장은 형태소 별로 나눠진 목록으로 출력된다. 품사 태깅 과정에서는 출력문을 입력 받아 각 형태소에 품사를 태깅하므로 형태소 분석이 완료된다. 본 분석에서는 자바로 구현된 Komoran 형태소 분석기를 사용하였다. Komoran 형태소 분석기는 분석 속도는 느리지만 체언과 용언 모두에서 일정 수준 이상의 분석 정확도를 보이며 지속적으로 개발하고 있다 . KoNLPy 라이브러리를 설치하여 파이썬에서 Komoran 형태소 분석기를 사용할 수 있었다. 접두사, 접미사, 어미 등을 제외한 의미를 지닌 어근 만으로 이루어진 어휘 형태소를 사용하였다. 명사, 동사, 형용사, 부사 네 가지를 포함하려 하였으나 Komoran 에서 동사 추출 시 어근만으로 단어를 알아보기가 어려워 분석에는 "**일반명사NNG, 고유명사NNP, 형용사VA, 일반부사MAG, 접속부사MAJ** "다섯 가지 품사만 포함하였다.

2. 특징변수 표현 및 추출방법 설명

→시각화 방법 : 워드 클라우드와 트리맵 사용

워드 클라우드는 텍스트에서 빈번히 사용된 키워드를 시각적으로 표시하는 텍스트 마이닝 방법으로, 단어의 사용빈도가 높을수록 그 단어를 강조하기 위해 크게 표현한다. 사용빈도가 높은 단어일수록 글씨 크기를 크게 표현함으로써 문서에서 강조하고자 하는 말을 한눈에 볼 수 있도록 하는 시각화 법이다. 가중치 값을 워드 클라우드의 크기 조정의 기반이 되는 값으로 매개함으로써 분석 대상에 중요하고 의미 있는 단어들을 크게 표현 한다. 본 분석에서는 TF-IDF 빈도기반을 바탕으로 시각화하였다.

트리맵 또한 문서 내 등장 키워드를 시각화 하여 정보를 전달하는 유용한 방법이다. 트리맵을 사용하면 중첩 직사각형을 사용하여 데이터를 계층형식으로 사용할 수 있다. 사각형이 클수록 가중치가 크다는 것을 의미하고 문서에서 중요한 단어이다. 직사각형의 상대적 크기를 통해서 단어의 중요도 차이를 시각적으로 분명히 파악할 수 있다는 장점이 있다 .

3. 특징변수 표현 및 추출 결과

<빈도기반(TF-IDF)으로 추출된 워드클라우드(좌) 와 트리맵(우)>

3. 특징변수 표현 및 추출 결과

→결과

빈도기반(TF-IDF)으로 추출된 워드 클라우드와 트리맵(전 슬라이드 그림)을 살펴보면 소득이 가장 큰 텍스트로 도출되었고 이어서 가계, 소비 , 증가, 대출, 분기 순으로 컸다. 가계부채, 정부, 주택, 크다, 상승, 정책, 지난해, 지출 , 금융, 경제 , 올해 ,더, 하락 등 경제 상황에 관련한 단어의 빈도가 높은 것을 알 수 있다. 분기 별 가계의 소득, 소비, 대출 , 부채 등 증가 등의 단어도 상위 빈도 100개 안으로 도출되었다.

트리맵과 워드 클라우드는 특징 변수 표현을 사용하여 단어 추출하여 시각화 한 결과를 한 눈에 직관적으로 파악할 수 있도록 도와준다. 단어 간 연결 부분은 해석자의 몫이겠지만 문서의 내용이 어떠한 주제를 다루는 지 추측할 수 있게 해준다.

본 데이터의 트리맵과 워드 클라우드는 소득과 지출 등 금융에 관한 내용일 것으로 예상할 수 있지만 그러나 증가와 감소, 소비와 대출 등 다양한 단어들도 상위 빈도 100개의 단어로서 함께 추출되었다는 점을 미뤄보면 소비가 증가 하였는 지 혹은 대출이 증가 하였는 지 세세한 내용까지 정확히 파악하기 힘들다.

경제와 관련된 자료는 증가, 감소와 같이 수치를 나타내는 단어에 민감하여 이에 따라 내용이 매우 달라질 수 있기 때문에 단어의 인과 관계를 파악하려 한다면 특징 변수 표현을 통해 추출한 시각화 자료만으로 결과 해석에 어려움이 있다.그러나 문서의 내용을 직관적으로 한눈에 큰 그림으로 파악할 수 있다는 점에서 매우 유용함을 알 수 있다.

4. 분석 기법 설명

→토픽 모델:

라벨에 의존하지 않으며 데이터 내 숨겨진 패턴을 찾는 비지도 학습법이다.토픽 모델은 문서에 포함된 이슈를 추출하기 위한 목적으로 사용된다. 문서들을 군집화 할 시 배타적으로 집단을 나누는 텍스트 클러스터링과 달리 문서들과 몇 개의 토픽들을 연계시킨다. 텍스트 단어들을 수학 및 통계적인 방법을 사용하여 대량의 문서 컬렉션 안에 내재된 주제(토픽)들을 빠르게 추출하는 텍스트 분석 기법이다. 문서 집합을 분석하여 각 문서의 토픽의 단어 분포를 산출하고, 토픽 모델을 통해 문서 집합 내에 감추어진 화제 혹은 정리된 개념들을 도출할 수 있고 문서의 주제도 추론 가능하다. 토픽 분석은 처음 잠재 의미 분석으로 시작하였고 최근에는 잠재 디리클레 할당 모델 기법이 대표적으로 사용된다.

잠재 디리클레 할당 모델이란 문서의 토픽 분포와 토픽 단어 분포가 디리클레 분포를 따른다는 가정하에 문서들의 토픽을 도출하는 방법이다. 실제 관찰 가능한 문서의 단어를 바탕으로 분석자가 알고 싶은 잠재 변수인 토픽의 단어 분포, 문서의 토픽 분포를 베이지안 모델에 의해서 추정한다. 잠재 디리클레 할당 모델은 문서의 토픽 분포에 의해 토픽이 결정되고, 다시 해당 토픽의 단어 분포에 의해 단어가 확률적으로 선택된 결과로 문서가 생성되었다고 가정한다. 여러 문서가 있을 때 문서들 간의 공통적인 토픽을 파악할 때 유용하다. 즉 잠재 디리클레 할당 모델의 핵심 프로세스는 문서가 핵심 토픽의 단어 분포로부터 생성되는 과정을 확률 모형으로 모델링한 결과를 나타낸다.

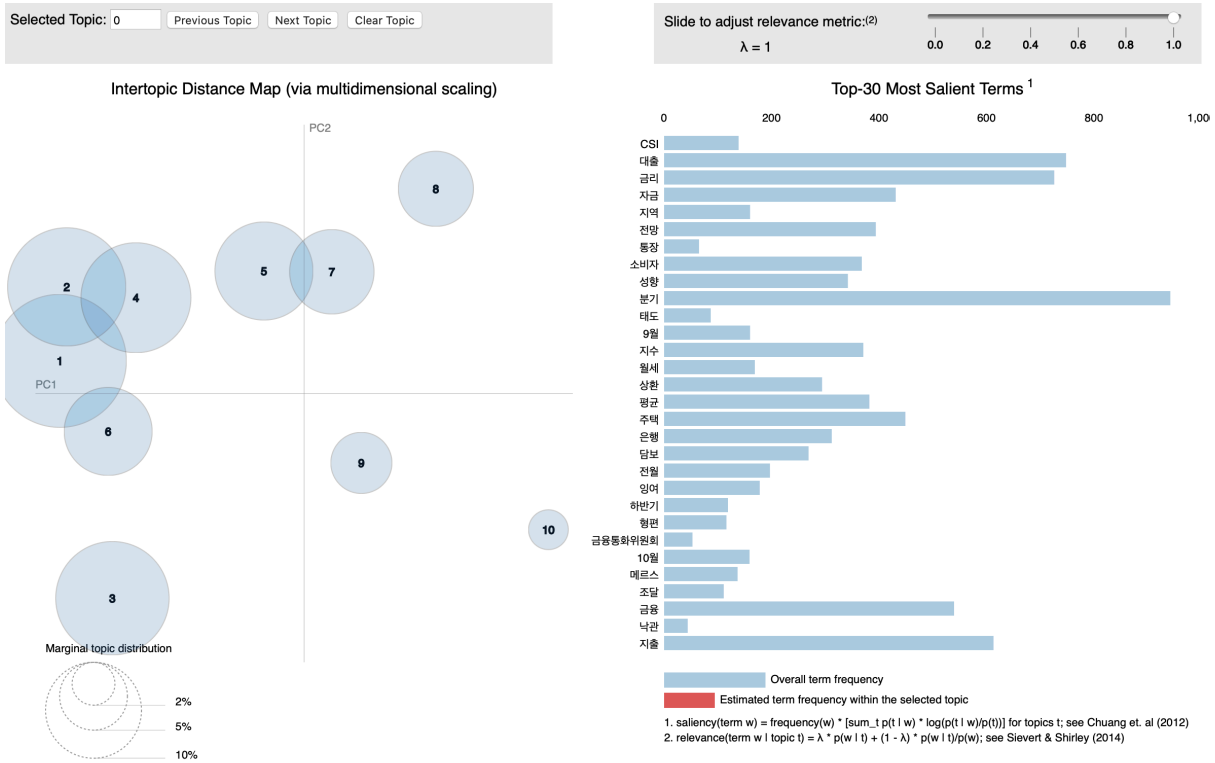
→네트워크 분석:

텍스트에서 나온 언어들 사이의 연결과 개념들의 연결망을 추출하는 것을 가능하도록한 방법론이다. 텍스트를 구성하는 언어와 이러한 언어 간의 관계를 분석한다는 측면에서 언어네트워크라고도 한다. 개체 자체의 속성에 중점을 두기 보다 수학의 그래프이론을 이용하여 객체 간의 관계를 분석한다. 단순히 특정 개념이 문서 집합에 얼마나 많이 등장 하였는 지에 그치지 않고, 그 개념이 다른 개념들과 어떤 관계를 가지는 지 또한 단어들이 특정한 패턴을 이루며 배열되어 있는 지에 대한 구조에 대한 분석도 가능하다. 텍스트 네트워크 분석은 언어 의미 네트워크 분석과 달리 텍스트에 직접적으로 드러나 있는 관계만을 대상으로 분석한다.

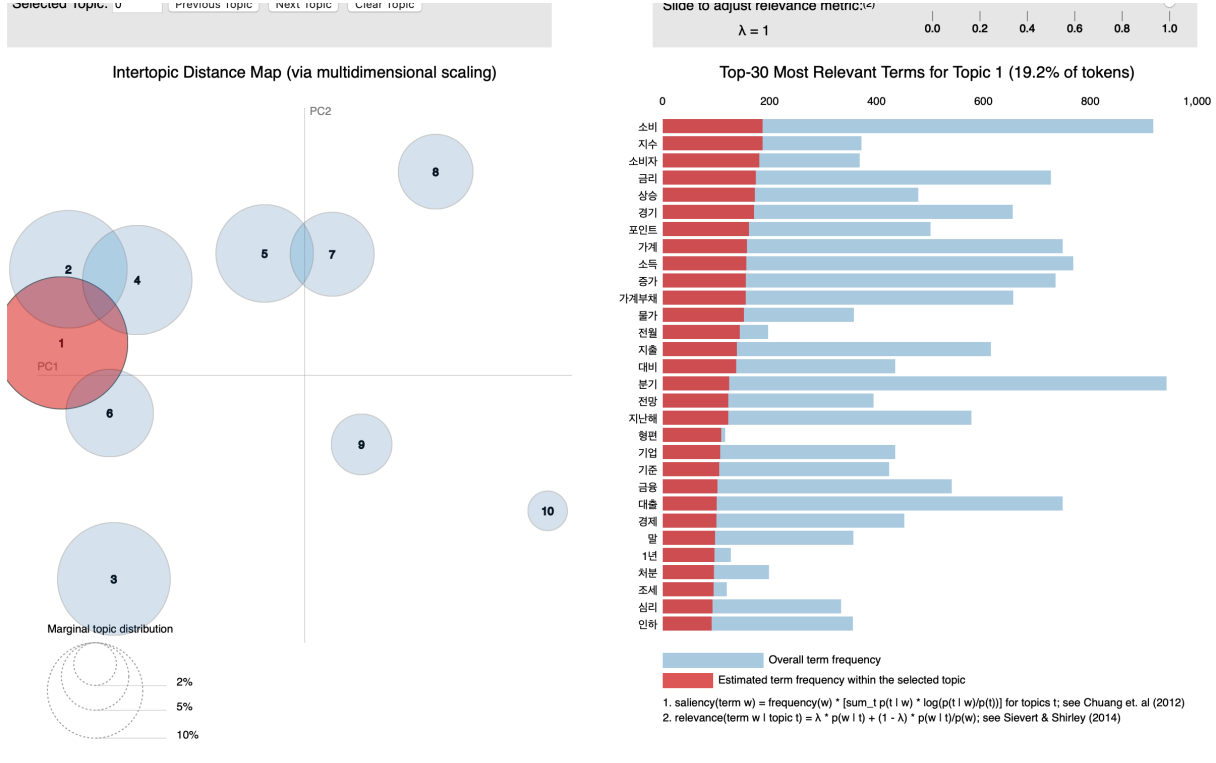
네트워크 분석의 그래프의 종류는 방향 그래프, 무방향 그래프, 가중 그래프 등이 있다. 그래프의 속성을 나타내는 지표로는 연결성(도달 가능성, 거리, 경로 개수), 중심성(중개 중심성, 근접 중심성, 고유백터중심성)등이 있다. 텍스트 분석 네트워크는 단어 별 네트워크 중앙성을 측정하여 핵심 단어와 연관어와의 관계를 시각화하고 단어 간 관계, 맥락을 통해 의미를 도출한다.

5. 분석 결과 및 한계점

토픽모델 분석 결과



<토픽이 선택되지 않았을 때>



<토픽1>

5. 분석 결과 및 한계점

→토픽 모델 분석 결과 해석:

연구자가 임의로 토픽 10개를 설정하였으므로 10개의 원이 생성되었고 , 각 원은 토픽을 나타내었다. 각 원의 넓이는 코퍼스 내에서 N개의 전체 토큰들에 대한 비율을 나타내는 것으로 토픽1이 그 넓이가 가장 컸고 토픽 10으로 갈 수록 그 넓이가 작아졌다. Marginal topic distribution이 10%정도에서 2%정도로 작아졌다.

각각 토픽의 단어들의 확률분포를 살펴보았을 때 각각의 토픽의 주제 및 이슈를 추정하여 보았다.

토픽이 선택되지 않았을 때 코퍼스내에서 중요한 단어는 분기, 대출, 금리, 지출, 금융, 주택, 자금, 전망, 상환 등의 순 이였다. 코퍼스는 경제 관련된 내용으로 분기별 금융시장의 예측과 전망 그리고 대응에 관한 것으로 추정할 수 있다. 토픽 모델은 전 슬라이드의 그림과 같이 그래프로 분석결과를 나타내어 토픽 별 단어의 분포와 빈도를 쉽게 파악할 수 있어 매우 편리하다.

토픽 1을 선택하였을 때 주어진 토픽에 의해 생성된 용어의 출현 횟수는 빨간색 막대그래프로 표시되며 어느 단어가 많이 출현 하였는 지 나타낸다. 소비, 지수, 소비자, 금리, 상승, 경기, 포인트, 가계, 소득, 증가, 가계부채, 물가, 전월, 지출, 대비, 분기, 전망, 지난해, 형편, 기업, 기준, 금융, 대출, 경제, 말, 1년, 처분, 조세, 심리, 인하 의 순 이였다.

이 단어들을 추정해보 토픽1은 경제와 관련된 주제로 금리 상승에 따른 소비 지수와, 소비자 심리, 가계 지출 및 소득, 가계 부채, 물가, 지출 등 분기 별로 결산하여 그 결과를 공유하고 작년과 비교하여 기업과 가계 등 경제 형편이 나아졌는 가를 평가 한다고 할 수 있다.

5. 분석 결과 및 한계점

위와 같은 방법으로 각각의 10개 토픽의 주제를 유추해 보면 아래와 같다.

토픽1: 금리상승에 따른 가계 형편

토픽2: 분기별 가계 소득과 부채

토픽3: 소비 지출 증가

토픽4: 금리와 대출

토픽5: 가계부채로 인한 정부의 정책

토픽6: 메르스로 인한 소비자심리지수(CSI) 변동

토픽7: 경기 활성화 및 부동산 규제

토픽8: 대출과 상환에 대한 정부 정책

토픽9: 경기부양 정책의 효과

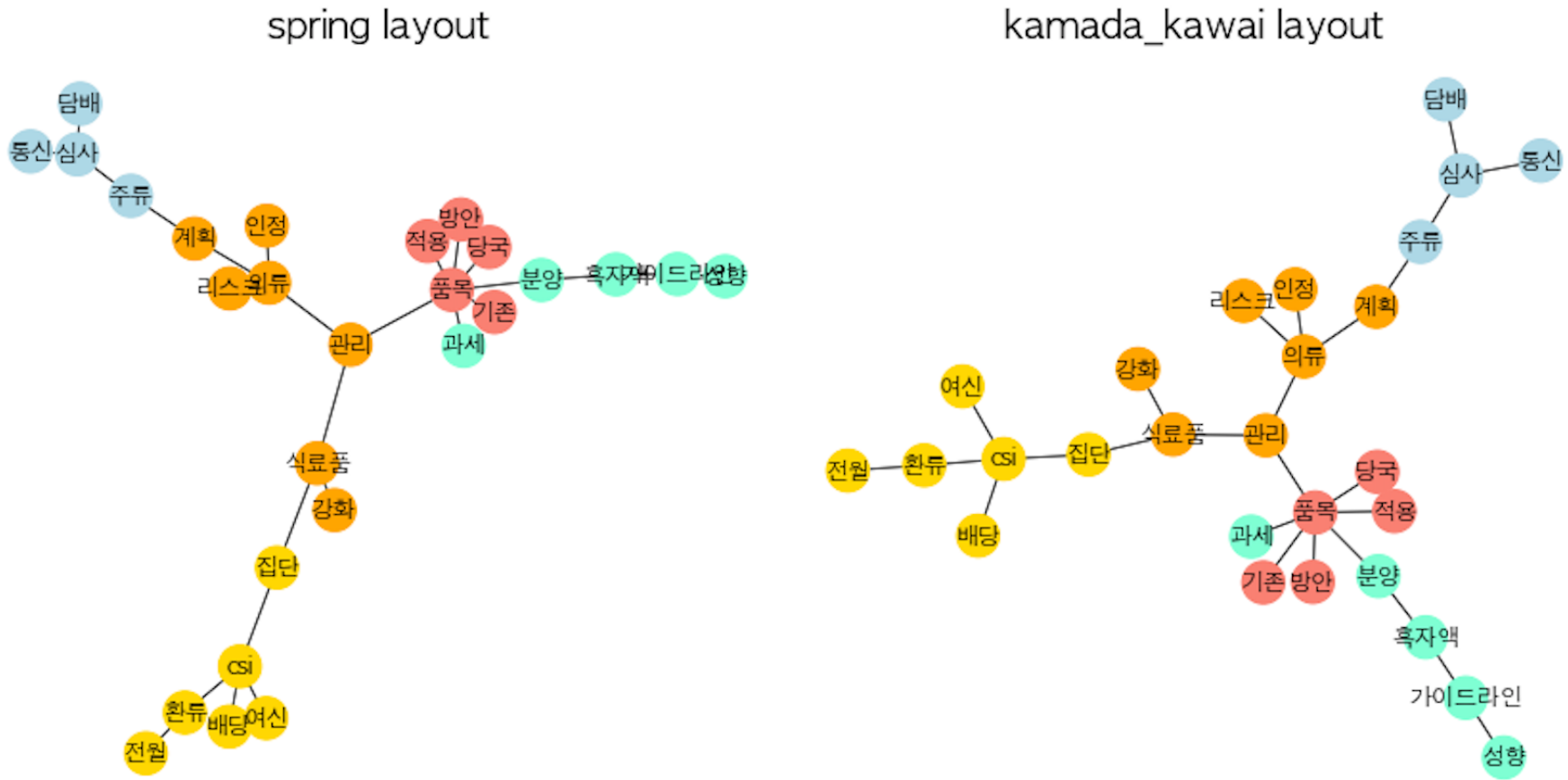
토픽10: 이주열 장관의 발표와 관련한 경제 상황

→토픽 분석 한계점:

위와 같이 토픽 별 단어의 확률 분포에 따라 각각의 주제를 어림잡아 유추해 볼 수 있었지만 조사와 어미를 제외한 단어들만으로는 인과관계 파악이 힘들어 세세한 주제를 정확하게 파악하기는 어려웠다. 그리고 도메인 지식이 없다면 단어들로만으로 의미 있는 분석을 내놓기에는 한계가 있었다. 경제에 관한 토픽이라는 것 정도는 알지만 정확히 해석하기는 힘들었다. 그러나 어떠한 내용에 관한 것인지 큰 그림은 그릴 수 있을 만큼 충분한 정보들이 제공되었다., 뉴스기사가 발행된 시점 등의 정보가 추가로 제공된다면 좀 더 정확한 해석이 가능할 것이다. 이로보아 토픽 모델 분석을 시도할 때는 도메인 지식이 무엇보다도 중요하다.

5. 분석 결과 및 한계점

네트워크 분석 결과



5. 분석 결과 및 한계점

→네트워크 분석 결과 해석:

텍스트 네트워크 분석 결과 전 슬라이드의 그래프를 얻을 수 있었다. 최소 신장 트리 기반 네트워크 모델로 추출된 spring layout와 Kadama_kawai layout을 살펴보면 관리를 중심으로 네트워크가 형성되어있고 동시 등장 단어는 같은 색깔로 표시되어 있다. 분석 결과 아래와 같이 분류된다.

주황:관리, 식료품, 강화, 리스크, 의류, 계획, 안정

빨강:적용, 방안, 당국, 품목, 기존

파랑: 담배, 통신, 심사, 주류

그린:과세, 분양, 흑자, 가이드라인, 성향

노랑:집단, CSI, 환류, 배당, 여신, 전월

단어 간의 관계는 선을 통해서 유추할 수 있다. 본 분석 결과에서는 무방향의 그래프로 표현되어 방향성은 나타나 있지 않은 연결 네트워크를 생성하였다. 그래프를 해석해 보면 다음과 같다. 당국에서는 식료품, 의류 등 의 품목에 대한 기존과 달리 관리 강화 계획을 발표한 것으로 보여지고 이에 따라 과세가 달라지고 이에 대한 가이드라인이 제공된 것으로 추정할 수 있다. 담배, 통신, 주류 등 심사에도 적용되었고, 소비자심리지수(CSI)는 전월과 어떠한 변동사항이 있을 것으로 예상되며 여신심사에도 영향이 있을 것으로 보인다.

5. 분석 결과 및 한계점

→토픽 모델과 텍스트 네트워크 분석 모델 비교 및 한계점:

처음 출발은 두 모델 모두 똑같은 조건의 형태소 분석으로 시작하였지만 그 분석 결과는 조금 달랐다. 토픽 모델은 문서의 주제를 기반으로 네트워크 분석은 단어의 관계를 기반으로 분석을 하였기 때문에 큰 그림으로는 경제에 관한 내용이라는 것에서 두 분석에서 같았지만 세부적으로 볼 때는 다른 분석의 결과가 도출 되었다. 토픽 모델의 경우에는 부동산, 물가 가계부채, 소비자 체감과 같은 금융과 관계된 내용이 도출되었지만 네트워크 분석 모델의 경우에는 정부의 규제 정책에 따른 경제적 파장이 주된 내용으로 판단되었다.

당초 생각은 두가지 기법이 조금 달라도 도출된 내용은 세부적인 부분까지 같은 것이라고 예상하였으나 의외로 둘은 조금 성격이 다른 단어들을 각각 도출하여 (중복적인 단어도 있었지만) 해석의 방향이 조금 달라졌다. 그 이유는 토픽 모델과 텍스트 네트워크 분석 모델은 각각 주제 중심, 관계 중심으로 그 목적이 달랐기 때문에 그 결과가 달라진 것으로 볼 수 있고 그 파장은 생각보다 컸음을 확인할 수 있다.

텍스트 네트워크 분석 모델 또한 토픽 모델과 마찬가지로 도메인 지식이 없다면 해석의 정확도가 떨어지지만 네트워크 연결 선을 통해서 단어 간 관계를 파악할 수 있고 인과관계에 대해 어느정도 유추 가능하였다. 그러나 확률 기반으로 단어의 중요도를 도출해주는 토픽 모델과 달리 주제를 파악하는 데는 유용하지 않았다.

위 내용을 간추려보면 연구자가 도메인 지식이 풍부해야 텍스트 분석 결과 해석이 용이하며 연구자가 각각의 텍스트 분석 기법의 특징을 잘 이해하여야만 분석 목적에 맞는 기법을 선택 할 수 있음을 본 프로젝트를 통해 알 수 있었다.