**비즈니스 프로그래밍 22-2 기말프로젝트**

**경영학과 201921487 김서준**

**-목차-**

1. 서론

1.1 연구 주제 및 선정 배경

1.2 연구 절차 설명

2. 본론

2.1 데이터 수집 및 저장

2.2 형태소 분석

2.3 데이터 특징 파악

2.4 데이터 시각화

3. 결론

3.1 결과 요약

3.2 한계점

3.3 연구의의

**1.  서론**

**1.1 연구 주제 및 선정 배경**

연구주제 :

 비즈니스 프로그래밍의 기말과제의 주제로 정한 것은 **‘신도시 관련 뉴스제목 분석을 통한 해당 지역 부동산 성공 및 실패 요인 도출 및 예상’**이다. 그렇다면 이러한 주제를 정하게 된 배경은 무엇일까?

선정배경 :

 먼저 부동산 시장은 대한민국에서 규제와 세금이 가장 강력한 시장 중의 하나이다.

부동산은 정치.경제 측면에 미치는 영향이 다른 종류의 재산에 비해서 상대적으로 강력하다. 그렇기에 부동산은 대한민국에서 가장 정치적.경제적으로 항상 논란의 대상이 된다.

 부동산은 특히 서민경제와 관계가 깊다. 일반적인 중산층의 부채와 자산에서 부동산의 비중이 상당하기 때문이다. 부동산을 소유했거나 부동산에 투자를 한 사람의 입장에서 부동산 가격의 하락과 상승으로 부터 오는 손해와 수익은 일반적으로 주식, 소비, 경기 등으로 부터 오는 손해와 수익과 비교하여 그 경제의 단위 자체가 큰 차이를 보인다. 그렇기 때문에 부동산 분야에서는 정부의 적절한 대응이 필요하며 부동산 정책은 정부의 경제 능력을 평가하는 중요한 잣대이다.

 분석 대상을 2기 신도시와 관련된 뉴스제목으로 정한 이유는 다음과 같다.

1. 신도시 건설 정책은 아마 국민들에게 가장 잘 알려진 정부의 부동산 정책 일 것이다.
2. 신도시 건설 사업은 기수를 나누어 수도권 안 여러 지역에서 진행했고 또 진행중이다.
3. 지역별로 사업의 성공과 실패가 나뉘는 정책이다.
4. 2기 신도시의 경우 이미 지역별로 성공(판교신도시).실패(운정신도시)의 귀추가 나누어졌다.
5. 뉴스제목은 뉴스 기사 중 가장 중요한 부분을 담고 있다.
6. 해당 신도시 들은 입주 시작일이 특정되어 있다.

**-연구 가설-**

1. 제목에서 추출되는 텍스트들이 성공한 2기 신도시인 판교신도시는 긍정적일 것이고 실패한 2기 신도시인 운정신도시는 부정적일 것이다.
2. 신도시 사업의 성공.실패와 연관되는 여러가지 키워드들이 도출될 것이다.
3. 2기 신도시 관련 뉴스기사를 분석한 결과로 추후 n기 신도시의 결과를 미리 예측해 볼 수 있을 것이다.

**1.2 연구 절차 진행**

**1.2.1 데이터 수집 및 저장 :** 네이버뉴스 검색을 통하여 판교신도시와 운정신도시 관련 뉴스 제목을 수집하여 jsonl 파일로 저장한다. 크롤링하는 뉴스제목들이 있는 뉴스가 작성된 기간은 각각의 신도시 입주일 이전 1년간으로 설정한다.(ex. 운정신도시의 입주일은 2009.06.09 이므로 운정신도시 관련 기간 설정 2008.06.09 ~ 2009.06.09, 판교신도시 입주일이 2008.12.31 이므로 판교신도시 관련 뉴스 기간 설정 2007.12.31 ~ 2008.12.31)

**1.2.2 데이터 전처리(형태소 분석) :** 판교신도시,운정신도시 별로 나뉘어 저장된 파일들을 konlpy 패키지를 활용하여 형태소 분석을 진행한다. 이때 분석 후 저장할 형태소는 "NNG"(일반명사), "NNP"(고유명사),"VA"(형용사), "XR"(어근) 으로 한다.

**1.2.3 데이터 분석 :** 형태소 분석을 통해 얻은 데이터를 이용하여 단어 빈도 분석, 불용어 처리 등을 진행한다. 또한 데이터들을 워드클라우드, squarify 등을 통해 시각화한다. 이러한 분석들로 어떠한 키워드를 통해 신도시 사업의 성공.실패를 예측해 볼 수 있는지를 궁극적으로 도출한다.

2. 본론

2.1 데이터 수집 및 저장

#<파일 형식관계로 같이 제출한 ipynb 파일에 코드와 주석으로 작성함>

2.2 형태소 분석

#<파일 형식관계로 같이 제출한 ipynb 파일에 코드와 주석으로 작성함>

2.3 데이터 특징 파악

운정 신도시와 판교신도시 각각 300개의 뉴스제목을 크롤링하고 그 안에서 형태소 분석을 통해 중복 포함 1000개의 형태소를 추출한 결과는 다음과 같다

|  |
| --- |
| 운정신도시 |
| {'신도시': 50, '분양': 47, '가구': 21, '전매': 19, '미분양': 18, '단지': 18, '수도권': 17, '부동산': 16, '공급': 15, '제한': 15, '아파트': 14, '완화': 12, '임대': 11, '시장': 11, '택지': 10, '주택': 10, '공공': 9, '복합': 9, '마련': 9, '개발': 8, '수혜': 8, '청약': 8, '분양권': 8, '내년': 8, '경기': 8, '공사': 7, '건설': 7, '용지': 6, '전철': 6, '관심': 6, '개통': 5, '서울': 5, '한양': 5, '수자': 5, '양수': 5, '자인': 5, '보상': 5, '상가': 5, '상업': 5, '주공': 5} |
| 판교신도시 |
| {'신도시': 84, '분양': 48, '입주': 40, '가구': 29, '내년': 28, '상가': 24, '분당': 16, '시장': 15, '전매': 15, '단지': 15, '아파트': 14, '부동산': 13, '개통': 12, '도로': 11, '청약': 10, '대형': 9, '공급': 9, '공사': 9, '새해': 8, '완화': 8, '제한': 8, '수도권': 7, '알짜': 7, '연기': 7, '로또': 7, '주공': 7, '주택': 6, '집값': 6, '임대': 5, '규제': 5, '마지막': 5, '경기': 5, '시설': 5, '유찰': 5, '지하': 5, '차도': 5, '부영': 4, '내달': 4, '마련': 4, '하락': 4} |

위의 결과를 통해 비교해본다면 먼저

긍정적인 형태소들은 실패한 신도시 건설 사업으로 여기어지는 ‘운정신도시’보다 성공한 것으로 여기어지는 ‘판교신도시’ 가 긍정적인 느낌을 전달하는 형태소들인 ‘신도시’, ‘입주’, ‘내년’, ‘분당’, ‘상가’, ‘개통’, ‘도로’ 등이 보다 높은 빈도수를 차지하는 것으로 볼 수 있다.

부정적인 형태소들에 있어서는 ‘운정신도시’가 ‘판교신도시’보다 ‘미분양’, ‘공급’, ‘제한’ 등의 형태소들이 더 많은 빈도수를 차지하고 있는 것으로 나타난다.

**2.4 데이터 시각화**

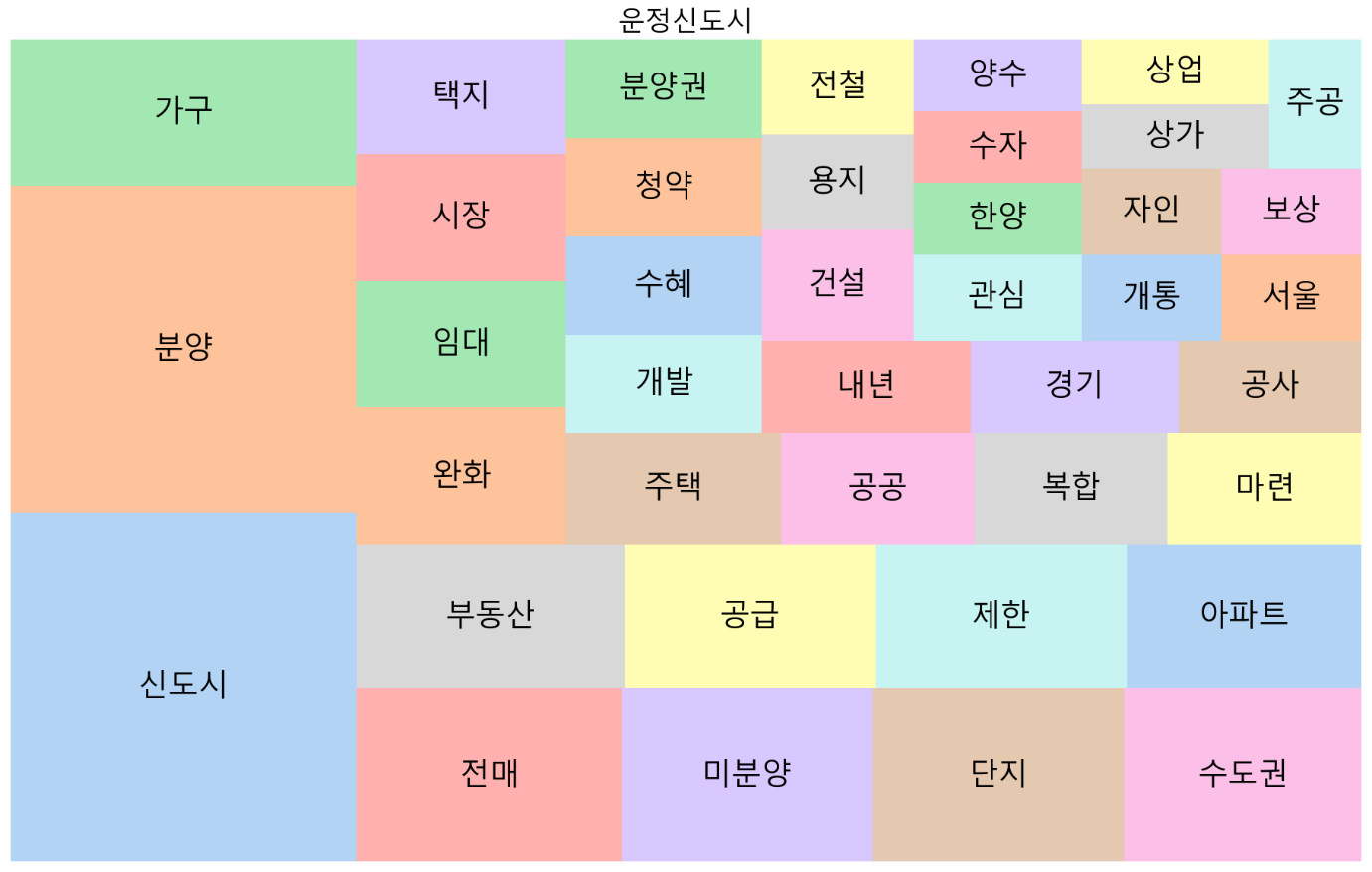
#<과정은 코드와 주석으로 같이 제출한 ipynb 파일에 작성함>

**2.4.1워드클라우드**





**2.4.2 Squarify**





**3. 결론**

**3.1 결과 요약 :**

이번 비즈니스 프로그래밍 기말 프로젝트를 통한 형태소 분석과 빈도분석을 통해 얻을 수 있는 결론으로는

첫째: ‘판교신도시’는 ‘운정신도시’에선 없거나 매우 낮은 빈도수를 보이는 미래를 나타내는 형태소인 ‘내년’과 전통적인 부촌이자 가장 성공한 1기 신도시인 ‘분당’등이 상대적 그리고 절대적으로도 높은 빈도수를 보이고 있으며 ‘신도시’ 라는 형태소의 빈도 또한 상대적으로 훨씬 앞서고 있다.

둘째: 좋은 교통 상황을 대변해주는 형태소인 ‘도로’, ‘개통’등에 있어 ‘운정신도시’의 형태소 빈도수 분석결과보다 2배가량 많은 빈도수를 보이고 있다는 점 을 뽑을 수 있다.

셋째: 마지막으로 부정적인 느낌을 주는 형태소인 ‘미분양’의 경우 ‘운정신도시 에서는 5위의 높은 빈도수를 보여주고 있지만 판교신도시의 경우 아예 형태소 빈도분석에 존재하지 않는 다는 것을 보여주며 ‘제한’의 경우 ‘운정신도시’가 빈도에 있어 약 2배가량 더 높다.

따라서 만약 추후 n기 신도시에 대한 입주 시작일이 공표된다면 해당 신도시 또한 입주시작일 이전 1년간의 뉴스제목을 위의 과정과 같은 크롤링과 분석을 통해 해당 신도시 건설 사업의 성공과 실패를 예측해 볼 수 있을 것으로 여기어진다.

만약 미래에 특정 n기 신도시 관련하여 텍스트 마이닝을 진행한 결과에서 ‘신도시’라는 형태소와 미래를 특정하는 형태소, 부촌을 특정하는 형태소 그리고 좋은 교통상황을 나타내는 형태소의 빈도가 상대적, 절대적으로 높게 나오고 또한 ‘미분양’ ‘제한’ 등의 형태소의 빈도는 상대적, 절대적으로 적게 나온다면 우리는 해당 신도시 건설 사업이 성공할 가능성이 높다고 예측해 볼 수 있다.

**3.2 한계점 :**

이 프로젝트의 결론이 주는 한계점으로는

1. TD-IDF 분석을 진행하지 않았다는 점 따라서 [텍스트 마이닝](https://ko.wikipedia.org/w/index.php?title=%ED%85%8D%EC%8A%A4%ED%8A%B8_%EB%A7%88%EC%9D%B4%EB%8B%9D&action=edit&redlink=1)에서 이용하는 어떤 단어가 특정 문서 내에서 얼마나 중요한 것인지를 나타내는 [통계](https://ko.wikipedia.org/wiki/%ED%86%B5%EA%B3%84)적 수치를 고려하여 분석되지 않은 값이라는 점.

2. 불용어를 주관적 기준으로 선정했다는 점

3. 기사 내용이 포함되진 않은 뉴스제목만을 분석했다는 점

4. 해당 지역의 지리적, 경제적, 사회적 특성은 고려하지 않았다는 점

5. 주관적인 기준으로 2기 신도시의 특정 신도시 건설사업을 성공, 실패로 나누었다는 점

이렇게 5가지가 있다.

**3.3 연구의의**

**프로젝트의 연구의의:**

이 프로젝트는 위에 열거한 것처럼 5가지의 명확한 한계점을 가지지만 위의 한계점들을 가지는 것까지 고려했을 때도 설정했던 가설인

1. 제목에서 추출되는 텍스트들이 성공한 2기 신도시인 ‘판교신도시’는 긍정적일 것이고 실패한 2기 신도시인 ‘운정신도시’는 부정적일 것이다.

2. 신도시 사업의 성공, 실패와 연관성이 있는 여러가지 키워드들이 도출될 것이다.

이 두가지를 일정 수준에서는 만족하기 때문에 의의를 가진다고 생각한다.