기업연계프로젝트

부동산 뉴스 데이터를 활용한
부동산 정보 큐레이션 서비스

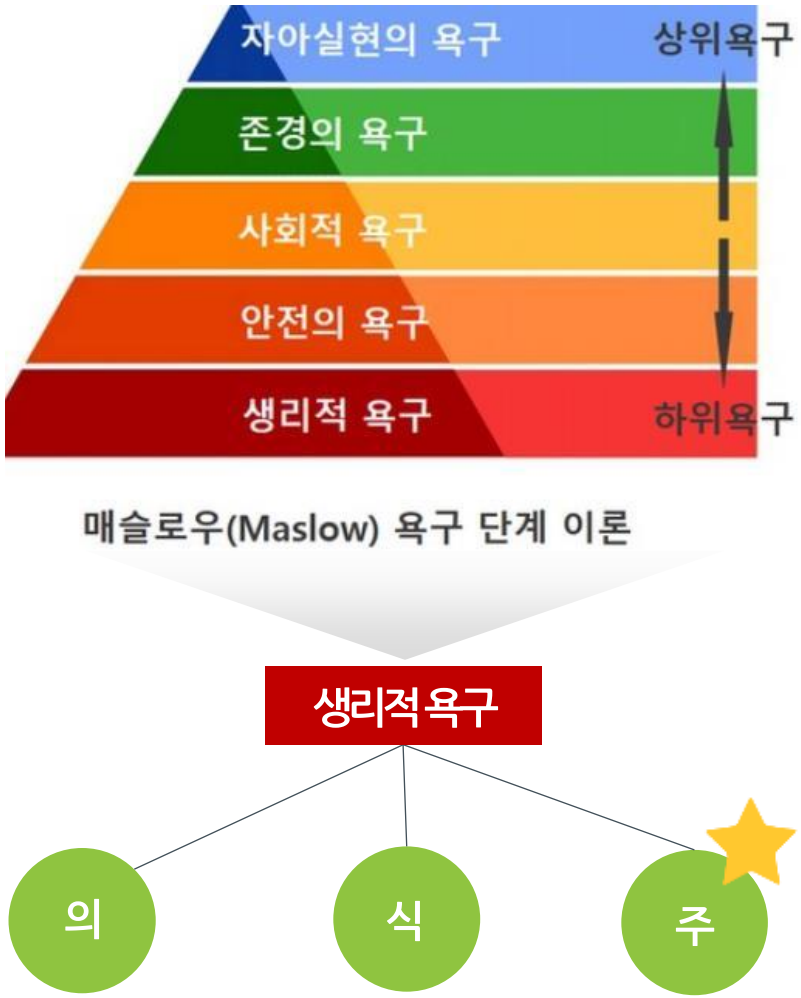
삼성전자 x YOSIGO
멘토 : 강민구, 멘티 : 김영진, 한승주

CONTENTS

Part 1. 프로젝트 개요

1. 프로젝트 배경 및 목표
2. 팀 구성 및 역할

프로젝트 개요



프로젝트 배경 및 목표

- **배경 :**
  ① 생리적 욕구는 매슬로우 기본욕구 5단계 중 **가장 하위 욕구**이며, 의식주를 충족해야 기초적인 생활이 가능함
  ② 특히, 집은 최근 주택가격의 급격한 상승, 부동산 정책 변화, 지역 간 가격 편차가 커져서 **사회문제가 여전히 심각한 상황**임
  ③ 정형 데이터 외 **비정형 데이터**로 수요자들에게 부동산 정보를 제공하는데 목적으로 함

- **목표 :** 부동산 뉴스를 분류 · 정제하여 아파트매매가격과의 연관성 분석 및 가격 예측, 토픽별 주요 뉴스 및 키워드 등 소비자에게 도움될 수 있는 비정형 부동산 정보를 제공한다.

프로젝트 개요

팀 구성 및 역할

강민구 멘토

프로젝트 총괄 업무(PM)

김영진 멘티

한승주 멘티

공동 작업

프로젝트 기획, 데이터 수집, 데이터 탐색,
데이터 전처리, 데이터 모델링, 결과보고서 작성

Part 2. 분석계획 수립

1. 데이터구성도
2. 분석환경 및 사용기술
3. Timeline

분석계획 수립

데이터 구성도

구분	활용 데이터	데이터명	중요도	shape	데이터 소스
정형	아파트 매매 실거래가	apt_deal_2017_to_2021_08_FINAL.csv	선택	(2,777,486 , 12)	- 공공데이터 오픈 API로 수집
정형	아파트 매매 가격지수	아파트매매가격지수.xlsx	필수	(679, 19)	- KB부동산에서 아파트 가격지수 시계열 데이터로 제공
비정형	부동산 뉴스 데이터	NewsResult_2010to202109_FIN.csv	필수	(310,019 , 19)	- 빅카인즈에서 제공하는 뉴스들을 '부동산 AND 아파트' 키워드가 있는 뉴스를 수집
비정형	부동산 도메인 감성라벨링	news_title_top1000_pos_neg.csv	필수	(1,000, 4)	- 단어 빈도 상위 1000개를 뽑아 각 단어별로 아파트 매매 상승 단어 1, 하락 단어 -1, 중립 0으로 라벨링
정형	군산대 감성사전	SentiWord_info.json	필수	(14,854, 1)	- 표준국어대사전을 구성하는 각 단어의 뜻풀이를 분석하여 긍부정어를 추출
정형	행정구역 GeoJSON 파일	TL_SCCO_CTPRVN.json	필수		- 지역별 시각화를 위해 수집

분석계획 수립

분석환경 및 사용기술

Analysis Library

NumPy 1.19.5

Pandas 1.2.4

matplotlib 3.3.4

Folium 0.12.1

RegEx 2.2.1

PYTHON pickle

KoNLPy 0.11.1

seaborn 1.2.4

GENSIM topic modelling for humans 3.6.0

scikit learn Machine Learning with Scikit-Learn 0.23.2

JSON 2.0.9

Version Control

GitHub

git

Development Tools

python 3.8.8

jupyter 6.3.0

분석계획 수립

TIMELINE

8월 19일 ~
프로젝트 기획안 작성
- 프로젝트 아이디어 구상
- 프로젝트 기획서 작성

8월 25일 ~
데이터 수집
- 네이버 뉴스 크롤링
- 빅카인즈 오픈 API 활용
- 아파트매매 실거래가 오픈 API 활용

9월 6일 ~
데이터 형태 파악
- 부동산 뉴스데이터 구조
- 아파트매매 실거래가 데이터 구조

9월 13일 ~
데이터 전처리
- 이상치, 결측치 확인
- 파생변수 생성

9월 20일 ~
텍스트 전처리
- Text Clearing
  (regex, stop words 활용)
- Tokenizing
- Vectorization
  - Bag of Words
  - TF-IDF

10월 18일 ~
데이터 분석
- 문서분류(Clustering)
- 감성분석
- 아파트 가격 예측 모델링

10월 25일 ~
모델링 검증 및 최종 결과물
- 회귀분석 및 이진분류 모델 검증
- 텍스트 전처리 재작업
- 최종 결과물 시각화
- 결과보고서 작성

Part 3. 데이터분석 및 결과

1.데이터수집
2.데이터전처리
3.데이터분석

데이터 분석 및 결과

데이터 수집 | 데이터 전처리 | 데이터 분석

1 데이터 확보

뉴스 기사 DATA

- 빅카인즈에서 제공하는 Open API 활용
- 빅카인즈 홈페이지를 통해 검색한 뉴스의 메타데이터를 다운로드
- 발행일, 제목, 카테고리, 키워드, 특성(top50), 본문(최대 200자) 등 확보
- [2010.01.01 ~ 2021.09.30] **총 310,019건**의 데이터 확보

아파트 매매 DATA

- 국토교통부에서 제공하는 아파트 실거래가 API를 통해 수집
- 아파트명, 건축년도, 전용면적, 층 , 거래금액, 주소, 법정동, 거래일자 등 수집
- [2017.01.01 ~ 2021.08.31] **총 2,777,486건**의 데이터 확보
- KB부동산, 주별 아파트매매가격지수 데이터 수집, **총 679건**의 데이터 확보

2 뉴스 데이터 형태

published_at	title	keyword	content
2010-01-01	꽁꽁 얼었던 서울 집값 이번주 소폭 상승	상승,서울,집값,소폭,마지막,마지막,서울,지역,부동산,시장,거래,매매가,활기,마무리,아파트,연속,집값,서울,지역,9월,상승,부동산114,부동산,정보,업체,부	지난해 마지막 주 서울지역 부동산시장은 거래와 매매가가 조금 활기를 찾은 상태에서 마무리됐다. 재건축 아파트 값이 3주 연속 올랐으며 서울 지역 집값은
2010-01-01	[부동산 칼럼]부동산시장 차별화 가능성 크다	가능성,부동산,시장,차별,가능,새해,상반기,부동산,시장,상승요인,하락요인,충돌,양상,전개,주택산업연구원,통계,필자,인천지역,전문가들,동의,인천지	새해 상반기의 부동산 시장은 상승요인과 하락요인이 서로 충돌하는 양상으로 전개될 것으로 보인다. 최근 주택산업연구원도 이와 비슷한 통계를 발표했는데, 필자를 비롯한 인천
2010-01-01	경인년 부동산 핫 이슈 '강남재건축 위례신도시 분양'	강남재건축,경인년,부동산,이슈,강남재,건축,위례,도시,분양,부동산,시장,이슈,전문가들,2008년,글로벌,금융,위기,부동산,가격,상승세,전망,상반기,금리,	올해 부동산 시장은 어떤 이슈로 움직일까. 전문가들은 2008년과 같은 글로벌 금융위기가 다시 불어닥치지 않는다면 올해 부동산 가격은 대체로 상승세를 띨 것으로 전망한다. 다만
2010-01-01	[창간 50주년 도약! 2010 경제]'꽃피는 4월' 불황뒤 '땅'이 굳어진다	도약,창간,주년,2010,꽃피,4월,불황뒤,경기,부동산,시장,부동산,1분기,침체기,보금자리주택사업,시작,4월,침체,1년,호황기,예상,금리인상,정부정책,속도,	[경인일보=최규원기자] 2010년 부동산 경기 '맑음'. 2010년 부동산 시장은 1분기 다소 침체기를 겪은 뒤 보금자리주택사업이 본격 시작되는 4
2010-01-01	인천 아파트 시가총액 5년새 2배	인천,아파트,시가,총액,2배,시가총액,인천,지역,아파트,시가,총액,5년,증가,정보업체,부동산,정보,업체,닥터아파트,628만,기준,전국,아파트,201가구,시	[경인일보=임승재기자] 인천지역 아파트 시가총액이 지난 5년동안 두배 가까이 증가한 것으로 나타났다. 30일 부동산 정보업체인 닥터아파트에 따르
2010-01-02	지역우선 공급비율 손질 서울 거주자 청약기회 줄어든다	지역,공급,비율,손질,서울,거주자,청약,기회,부동산시장,정부,정책,제도,변경,좌우,전망,부동산,향방,가름,변수들,시장,내부,변수,수요자,투자,자들,관심,부	올해 부동산시장은 정부 정책과 제도 변경에 크게 좌우될 전망이다. 부동산 경기의 향방을 가름할 시장 내부 변수들이 눈에 띄는 게 없는 데다 수요자나 투자자들의 관심 또한 부동

- **뉴스 제목(title)** : 감성분석 실시
- **키워드(keyword)** : 벡터화, 필터링, 불용어 처리 등

데이터 분석 및 결과

데이터 수집 → 데이터 전처리 → 데이터 분석

1 결측치 처리

news_id	0
published_at	0
provider	0
byline	51274
title	0
category_1	0
category_2	55529
category_3	113543
category_incident_1	265365
category_incident_2	294826
category_incident_3	297734
PS	164251
LC	8114
OG	3559
keyword	0
feature_top50	2
content	0
url	60384
except	273023

dtype: int64

feature_top50 2 → 결측치 제거

2 뉴스 필터링

① **부동산** and **아파트**만 포함된 뉴스

② category_1 = **경제>부동산** 해당 뉴스

③ 뉴스 제목에서 홍보성 기사, 아파트 단지의 금액(실거래가) 상승/하락 기사 등 무의미한 뉴스문서는 제외

④ **중복** 기사 제거

310,019개 뉴스 → 115,519개 뉴스

3 불용어 처리(stop words)

① 정규표현식(Regex) : 단어 정제

② 단어 길이가 **1 이하** 단어

③ 단어 빈도수 **3 이하** 등장한 단어

④ **영단어** 제외

⑤ 최종 단어 필터링 :
no_below(**50**), no_above(**0.7**)

264,512개 단어 → 15,297개 단어

데이터 분석 및 결과

4 파생변수 생성(감성지수_비지도학습)

1차	부동산도메인감성사전
내용	빈도수 높은 상위 1,000개 단어에 긍부정 라벨링 작업 실시 -상승 단어 +1, 하락 단어 -1, 중립 단어 0
예시	

	0	1	sj_긍부정	yj_긍부정	통일	score
0	거래	16134	0	0	0	0
1	아파트	14114	0	0	0	0
2	부동산	13368	0	0	0	0
3	전용	13144	0	0	0	0
4	서울특별시	9921	0	0	0	0

+

2차	군산대감성사전
내용	표준국어대사전을 구성하는 각 단어의 뜻풀이를 분석하여 긍부정어를 추출(가중치 0.25)
예시	

	word	word_root	polarity
0	(-:	(	1
1	(;_;)	(;_;)	-1
2	(^.^)	(^.^)	1
3	(^-^)	(^-^)	1
4	(^.^~	(	1

※ 뉴스제목 ➔ 형태소 분석(Okt) ➔ 1, 2차 감성사전 ➔ 감성지수 산출

감성지수 높은 뉴스 제목 상위 10개

```
[('낙찰가율 70%벽 돌파 토지 경매시장 고공행진 재개발 재건축도 꿈틀', 5),
 ("'재건축 연한 단축'호재에 1980년대 아파트 인기 낙찰가율 급등", 5),
 ("서울 도심 재개발 재건축 아파트에 관심 몰려 '북한산 더샵' 분양 인기 상승", 5),
 ('대형 개발호재 쏟아지는 평택, 집값상승에 새 아파트로 수요자들 몰려', 5),
 ('도시개발사업 기대 급증, 개발호재 기대해도 괜찮을까', 5),
 ('文 "집값 안정화" 주장했지만 서울 집값 상승폭 확대...전셋값은 9년만 최대상승', 5),
 ('文 공급확대 주문에 커지는 재개발 재건축 규제 완화론 "신도시 만으론 부족"', 4.75),
 ('2분기 재건축-재개발 분양대전, 흥행성적 재건축>재개발 전망', 4.5),
 ('서울 청약경쟁률 상위 10개 중 7개가 재건축 재개발 '반포 래미안 아이파크'도 인기 기대', 4.5),
 ('신혼부부희망타운 인기, 장단점은? 당첨되면 로또? 신혼희망타운 인기 쏠림 현상', 4.5)]
```

감성지수 낮은 뉴스 제목 하위 10개

```
[("[정부 부동산 규제 가시화] 부산 부동산, 금리 인상 대출 규제 투기 지구 '3각 파도'에 기 꺾일까", -5.5),
 ('금리인상 악재 현실화된 부동산시장 하락세 심화 우려', -5.5),
 ('탈세 편법대출 난무 9억 이상 고가주택 불법거래 811건', -5.5),
 ("'풍선효과' 우려없나...고가 다주택자 대출규제 없어 한계도", -5.25),
 ('대출규제로 거래심리 '꽁꽁' 신도시 분양권 웃돈 반토막', -5),
 ("[2018 부동산] 대출 규제, 입주 폭탄 '악재' 보유세 인상 '변수'", -5),
 ("대출규제 엎친 데 금리인상 덮쳐 '거래절벽' 심화", -5),
 ("[금리인상 부동산 영향] 예고된 '악재' 서울 지방 간 양극화 심화 우려(종합)", -5),
 ('수도권 비규제 지역 서울 9억 이하 아파트 '이상 과열' 규제 비켜간 곳 '풍선효과'', -5),
 ('전세대출 규제 시행 혼란 없었지만 '풍선효과' 우려', -5)]
```

데이터 분석 및 결과

데이터 수집 > 데이터 전처리 > 데이터 분석

부동산 뉴스 데이터로 아파트매매가격 예측하기

- Work Flow(모델링)

Data Transform

- **Text Vectorization**
  TFIDF 방식
- **클러스터링**
  문서분류(kmeans)
- **감성분석 (비지도 학습)**
  (부동산 도메인 + 감성사전)
- **차원축소**
  코사인 유사도 거리 계산
  (15,297차원 -> 200차원)
- **시간(주, 월) 단위로 합치기**

>>

Cross Validation

- **전체 data set :**
  2010 ~ 2021년 09월
- **Train/Valid set :**
  2010 ~ 2020년
  Hold out 방식
- **Test set :**
  2021.01~2021.09

>>

Modeling

- **회귀분석 모델 :**
  Linear, Lasso, Ridge,
  XGBoost, Random forest

>>

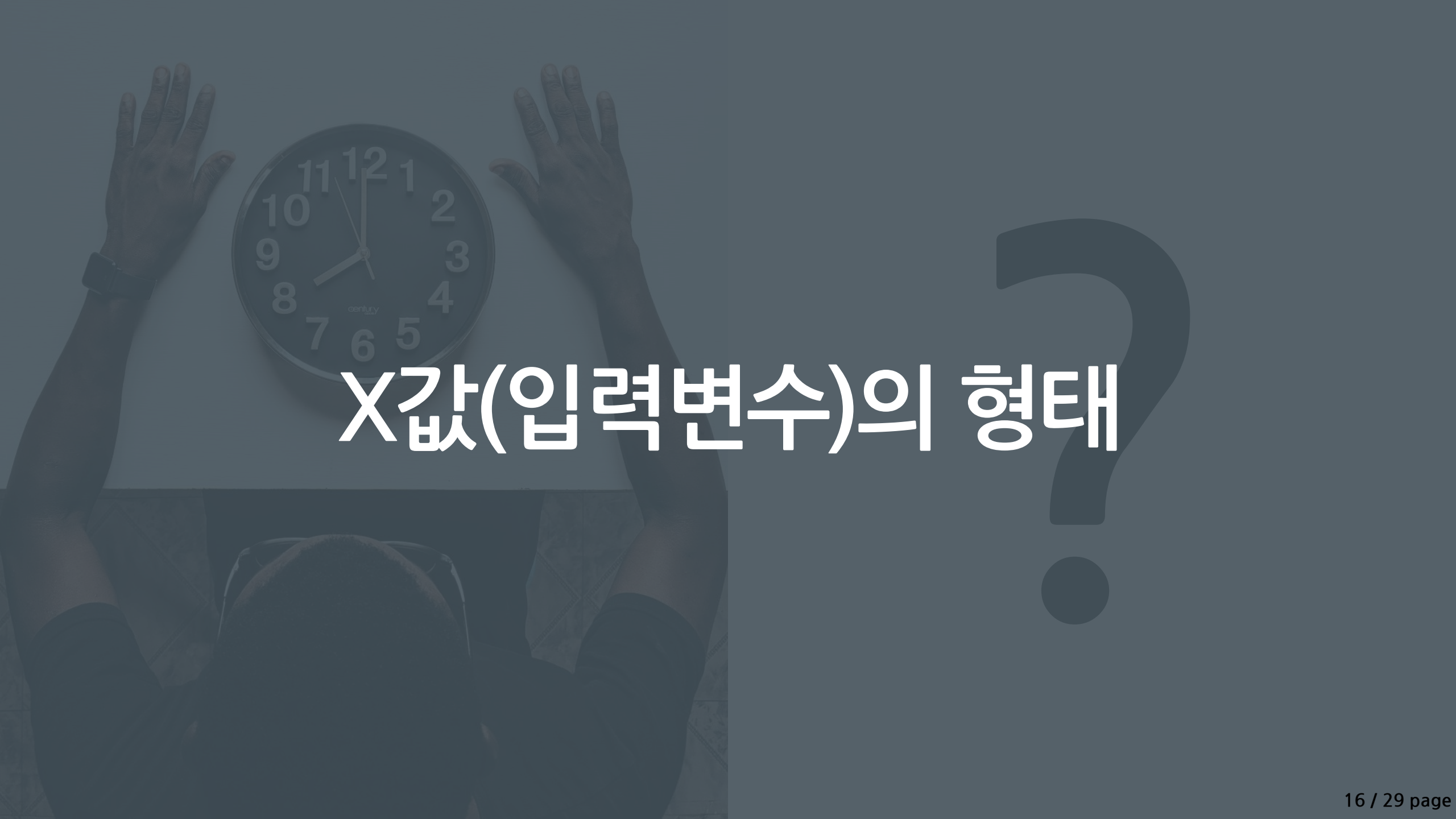
Evaluation

- **평가지표 비교**
  RMSE, R^2 등
- **최종 모델 선택**
- **예측값 시각화**
  Folium 활용

부동산 뉴스 데이터로 아파트매매가격 예측하기

- Data Transform

?



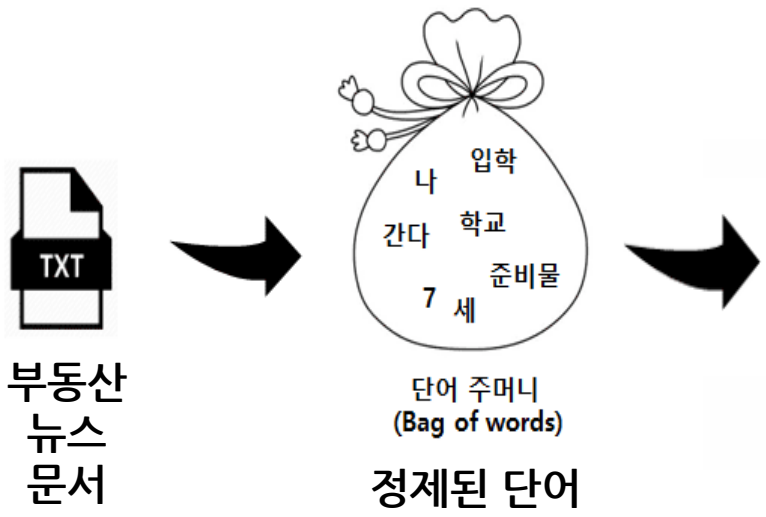
X값(입력변수)의 형태

데이터 분석 및 결과

부동산 뉴스 데이터로 아파트매매가격 예측하기

- 입력변수 생성

STEP 1 Text Vectorization



부동산 뉴스 문서

단어 주머니 (Bag of words)

정제된 단어

TF-IDF

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

$tf_{i,j}$ = number of occurrences of i in j
df_i = number of documents containing i
N = total number of documents

words
[상승, 서울, 집값, 소폭, 마지막, 마지막, 서울, 지역, 부동산, 시장, 거래...
[가능성, 부동산, 시장, 차별, 가능, 새해, 상반기, 부동산, 시장, 상승요인,...
[강남재건축, 경인년, 부동산, 이슈, 강남재, 건축, 위례, 도시, 분양, 부동산...
[도약, 창간, 주년, 꽃피, 불황뒤, 경기, 부동산, 시장, 부동산, 침체기, 보...
[인천, 아파트, 시가, 총액, 시가총액, 인천, 지역, 아파트, 시가, 총액, 증...
[지역, 공급, 비율, 손질, 서울, 거주자, 청약, 기회, 부동산시장, 정부, 정...
[서울, 강남, 연속, 상승, 강남, 건축, 아파트, 상승, 연속, 서울, 아파트시...
[실투자, 확정, 보장, 원대, 원룸텔, 초역, 세권, 수익, 삼성홈플러스, 뉴코아...
[지속, 서울아파트, 전셋값, 학군, 수요, 상승세, 겨울, 방학, 학군, 수요, ...
[강남, 반등, 서울, 소폭, 거래시장, 서울, 아파트, 거래, 시장, 하락세, 소...

TF-IDF.shape

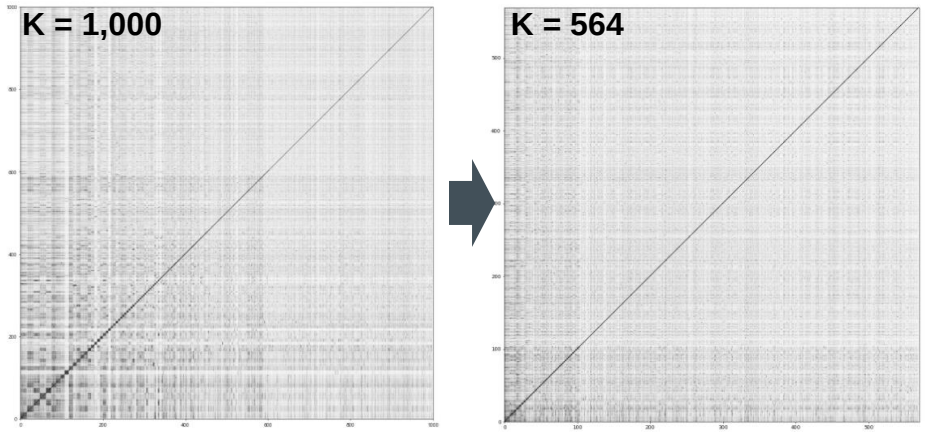
(115519, 15297)

데이터 분석 및 결과

부동산 뉴스 데이터로 아파트매매가격 예측하기

- 입력변수 생성

STEP 2 Clustering (Spherical Kmeans)



- K-means 방식이 토픽모델링보다 좀더 효과적임
- **15,297개 단어**를 **1,000개 단어**로 1차 Clustering 함
- 각 군집의 centroid로 계산하여 유사한 군집끼리 Merge하여 **564개**로 **2차 군집**을 진행함

최종 군집 개수 : 564개 ➔ 토픽 사이즈(문서개수) 상위 200개

서울 accuracy(R^2)					경기 accuracy(R^2)				
차원	Y = 매매지수		Y = 편차*		차원	Y = 매매지수		Y = 편차*	
	Linear	XGB	Linear	XGB		Linear	XGB	Linear	XGB
50	0.937	0.960	0.739	0.820	50	0.927	0.942	0.789	0.850
100	0.966	0.963	0.816	0.755	100	0.955	0.940	0.874	0.871
200	**0.985**	**0.967**	**0.832**	**0.815**	**200**	**0.979**	**0.946**	**0.840**	**0.869**
300	0.989	0.961	0.650	0.833	300	0.989	0.943	0.742	0.866

- 상위 200차원으로 결정한 이유
  - ▶ 토픽 사이즈(문서 개수)별로 정렬하여 **빈도수가 극히 낮은 군집은 이상치로 판단**
  - ▶ **차원이 커질수록** 매매지수 Linear 모델의 **정확도가 높아지는 경향을 보임**
  - ▶ 종속변수를 편차로 살펴보면, **300차원은 정확도가 오히려 낮아짐**
  - ▶ **지역별, 모델별, 편차** 등 종합적으로 고려 시, 최적의 군집 개수를 상위 200개로 결정함

*※ 편차(변동률) : (금주 매매지수 - 전주 매매지수) / 전주 매매지수 * 100*

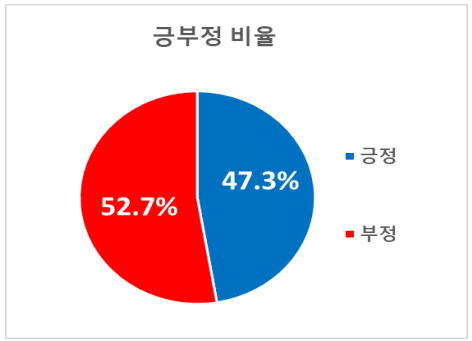
데이터 분석 및 결과

부동산 뉴스 데이터로 아파트매매가격 예측하기

- 입력변수 생성

STEP 3 Sentiment Analysis

뉴스 제목을 형태소 분석기(OKT)로 감성분석 실시
(부동산 도메인 리벨링 + 군산대 감성사전)



부동산 도메인에 대한 감성분류의 가중치는 **1**,
군산대 감성사전에 대한 가중치는 **0.25**를 부여함

STEP 4 201차원 생성

1. 200개 군집의 Centroid와 TF-IDF 벡터 간 코사인 유사도 거리 계산
2. 감성지수 = 문서별 SUM(긍정 + 부정)
3. 201차원 = 문서별 200개 군집의 코사인 유사도 거리 + 감성지수

{'2009-53': array(
[1.79433329, 1.52463016, 2.15365873, 2.26776357, 1.06855394, 1.42162746, 1.0527757 , 0.67464791, 2.8837213 , 1.22211754, 1.87284005, 1.22918716, 0.86516892, 2.26816271, 0.98840285,
1.76459505, 1.12885645, 0.98077146, 0.76336828, 0.90352487, 1.04288834, 0.95239239, 1.30034425, 1.21978633, 1.21566134, 1.69732418, 1.62700259, 0.71355259, 1.30899804, 1.08189694,
1.27087769, 1.49114934, 0.99557656, 1.37081964, 1.32140399, 1.84364521, 1.08136899, 0.79937258, 1.5091041 , 1.59999188, 1.19296292, 1.40748985, 1.2025894 , 1.58983019, 0.81789092,
1.59611698, 0.96968548, 0.99733083, 1.50170921, 0.51945656, 1.71918022, 1.04792621, 0.81413448, 1.5340278 , 1.85450875, 0.75934472, 1.16946354, 2.03437253, 0.91214028, 1.43279149,
2.21836629, 1.26076433, 0.93343968, 2.79834628, 0.52373788, 1.55749399, 1.91535476, 1.39834761, 1.34619031, 2.02209776, 1.49642448, 1.72319123, 0.64447382, 1.4848489 , 0.98721403,
1.30460094, 1.46825679, 1.16412573, 1.48662144, 1.48788787, 0.85035142, 1.59969898, 1.23980494, 0.51783364, 1.74958208, 1.00228952, 1.48286987, 1.47272842, 0.72867957, 1.71124769,
1.83711833, 0.94876016, 1.39244742, 1.38738524, 2.12709101, 1.12029449, 1.21460766, 1.72978795, 2.4079185 , 0.98165522, 1.08143358, 1.41232894, 1.11274245, 1.00666624, 1.14737749,
1.2301941 , 1.56375279, 1.24837055, 0.88136517, 1.20037563, 0.20136851, 1.80931793, 1.72315487, 1.21567121, 1.34752312, 1.28058225, 1.54654548, 1.82432985, 0.8834132 , 1.07110056,
1.84603085, 1.26789115, 1.7237678 , 1.06119464, 2.34176936, 1.43355677, 1.34451707, 1.01900327, 1.16317868, 2.33105061, 1.72246617, 1.08540006, 0.69638179, 1.2158045 , 1.32847256,
2.56289048, 1.28912189, 0.68140268, 0.51365696, 3.10776961, 0.93745979, 0.78881352, 1.09738267, 0.6898406 , 1.34686876, 0.8293329 , 1.09472044, 1.6695367 , 1.05419276, 0.69838969,
1.57889739, 0.86443658, 1.28746759, 0.86950395, 1.85243525, 1.42429128, 1.36589567, 0.71762522, 1.50493515, 1.10260497, 2.26671719, 1.31184148, 1.38541846, 0.5956388 , 1.9660705 ,
0.52270208, 0.90938087, 1.19366075, 1.62133771, 2.11357909, 1.46169706, 0.88676187, 0.68789251, 3.3413818 , 2.13957162, 0.77076955, 1.31603034, 1.43734707, 1.19234637, 1.45543886,
1.10497302, 1.38924869, 1.71099313, 0.64293408, 2.07082126, 1.09829855, 1.74916222, 1.19838646, 1.53321898, 1.37593343, 1.15280707, 1.24642629, 1.41473069, 0.95392912, 1.00957156,
1.04135872, 1.29330191, 1.64420949, 1.04844311, 1.26140743, 11.25]),

데이터 분석 및 결과

데이터 수집 → 데이터 전처리 → 데이터 분석

부동산 뉴스 데이터로 아파트매매가격 예측하기

- 입력변수 생성

STEP 5 최종 X값 (입력변수)

Y= 아파트 매매 가격지수

	구분	서울	부산	대구	인천	광주
0	2008-04-07	81.083635	58.646524	62.623283	92.962950	60.227006
1	2008-04-14	81.285271	58.719607	62.600014	93.271035	60.305642
2	2008-04-21	81.503276	58.803415	62.572266	93.648677	60.361562
3	2008-04-28	81.667358	58.861788	62.550464	94.043377	60.405527
4	2008-05-05	81.823950	58.906905	62.527563	94.335416	60.472733

- 종속변수 : 주 단위로 구분 ➔ 입력변수도 **주 단위로 통일화**
- 입력변수는 종속변수의 **1개월 전** 시점의 값으로 데이터 변환
  (논문 참고 *➔ 뉴스가 아파트가격에 영향을 미치는 시점은 1개월 전이 가장 유의함)

데이터 변환

- **4주 합계** : 4주 전부터 금주까지 문서별 벡터 합계
- **4주 평균** : 4주치 문서별 벡터값의 합계 / 4주치 문서 수
- **1주 평균** : ∑1주치 문서별 벡터값의 평균값(1~4주 평균을 SUM함)

※ 참고 논문 : 아파트 매매가격과 부동산 온라인 뉴스의 교차상관관계와 인과관계 분석(박재수, 이재수)_대한국토·도시계획학회)

데이터 분석 및 결과

부동산 뉴스 데이터로 아파트매매가격 예측하기

- 종속변수 생성

STEP 6 종속변수

Y = raw data(매매지수)

	구분	서울	부산	대구	인천	광주
0	2008-04-07	81.083635	58.646524	62.623283	92.962950	60.227006
1	2008-04-14	81.285271	58.719607	62.600014	93.271035	60.305642
2	2008-04-21	81.503276	58.803415	62.572266	93.648677	60.361562
3	2008-04-28	81.667358	58.861788	62.550464	94.043377	60.405527
4	2008-05-05	81.823950	58.906905	62.527563	94.335416	60.472733

데이터 변환

Y = 편차

기준 주차

	2009-53	2010-01	2010-02	2010-03	2010-04	2010-05	2009-53	2009-53
서울	0	0.02336	0.053445	0.028653	0.090269	0.057536	0.063221	0.028095
부산	0	0.131157	0.16916	0.18886	0.275412	0.297428	0.397713	0.127223
대구	0	0.001991	0.075715	0.036254	0.043611	0.036162	0.038563	0.009358
인천	0	-0.01353	0.013262	-0.04002	-0.01086	0.029286	-0.01723	-0.01833
광주	0	-0.01006	0.082659	-3.84E-05	0.007374	0.06435	0.096786	0.009733
대전	0	0.131026	0.188868	0.273727	0.365487	0.247518	0.316619	0.13685
울산	0	0.043005	0.19703	0.093427	0.035727	0.169179	0.045586	0.050226
경기	0	-0.01039	-0.02784	0.004548	-0.00633	0.033078	-0.00517	-0.00213
강원	0	0.011143	-0.04235	0.186121	0.065153	0.087252	0.041864	0.009887
충북	0	0.116313	0.074126	0.093895	0.072052	0.091897	0.063679	0.053305
충남	0	-0.00771	0.017517	0.042573	0.018495	0.013831	0.104168	0.022478

*※ 편차(변동률) : (금주 매매지수 - 전주 매매지수) / 전주 매매지수 * 100*

데이터 분석 및 결과

부동산 뉴스 데이터로 아파트매매가격 예측하기

서울 Linear 모델 예측 시각화

회귀모형 결과

- **최종 모델 : Linear Regression model**

서울	X 형태	Y = 매매지수		Y = 편차	
		Linear	XGB	Linear	XGB
accuracy	4주 합계	0.789	0.592	0.220	0.468
	4주 평균	**0.984**	**0.952**	**0.766**	**0.827**
	1주 평균	0.99	0.987	0.763	0.816
RMSE	4주 합계	39.221	28.56	2.401	0.378
	4주 평균	**26.373**	**25.034**	**0.124**	**0.085**
	1주 평균	23.554	33.833	0.145	0.695

① 정확도와 RMSE를 확인한 결과,
▶ **4주 평균**이 정확도와 평균 제곱근 오차(RMSE)가 **가장 우수한 결과**
▶ **Linear 모델 (Y = 편차)** 를 사용한 결과값이 **예측력이 가장 좋음**
② 4주 평균과 1주 평균의 예측 값은 실제 값과 **유사한 경향성**을 보임
▶ 매수심리나 매도심리 같은 **하나의 척도**로 활용 가능 (트렌드를 보여줄 수 있음)
③ 예측값을 시도별로 folium을 활용하여 **지도로 시각화**함
▶ 지역별로 매매 변동률을 시각적으로 확인 가능
HTML 실행

데이터 분석 및 결과

부동산 뉴스 데이터로 아파트매매가격 예측하기

- 문서 분류(Clustering)

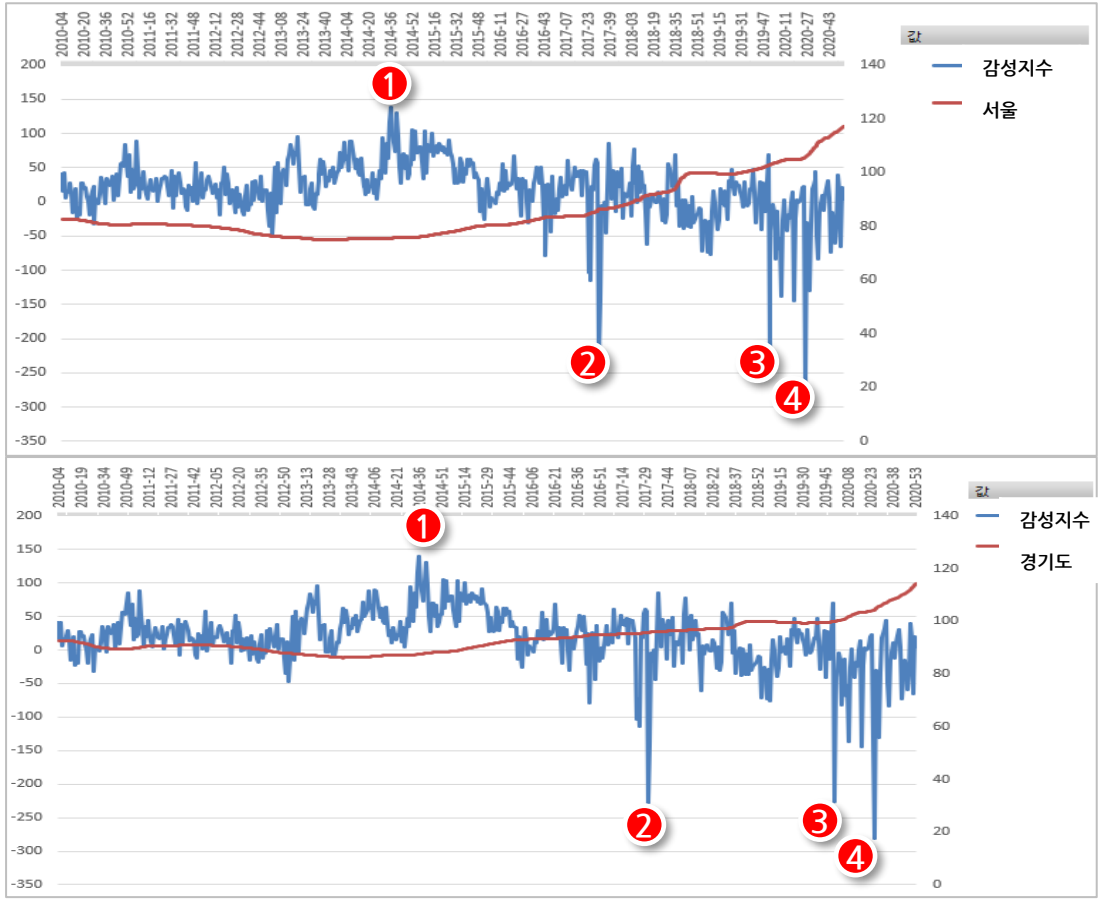
Topic Top5

No.	35번 군집	16번 군집	26번 군집	36번 군집	42번 군집
주제	부동산 시장 현황	정책, 규제	투자, 수익률	서울 집값	분양

➔ *문서 분류 시각화 확인하러 가기 : spherical_kmeans_vis.html#topic=36&lambda=1&term=*

데이터 분석 및 결과

감성지수와 아파트매매가격지수와의 상관관계



서울특별시, 경기도 지역
➔음의관계

001 >> **2014년 40주차 = *+130.25점***

2014년 9월 1일 부동산 정책 발표

※ **주요 내용** - 재건축 완화, 택지개발촉진법 폐지, 주택청약 간소화 등

002 >> **2017년 31주차 = *-227.75점***

2017년 8월 2일 부동산 정책 발표

※ **주요 내용** - 투기지역, 투기과열지구 지정, 대출 및 주택청약규제 등

003 >> **2019년 51주차 = *-226.25점***

2019년 12월 16일 부동산 정책 발표

※ **주요 내용** - 고강도 세금, 대출 대출 규제 등

004 >> **2020년 25주차 = *-281.5점***

2020년 6월 17일 부동산 정책 발표

※ **주요 내용** - 규제지역 확대, 토지거래허가구역 , 실거주 요건 강화 등

데이터 분석 및 결과

주요 뉴스 & keywords 추출

주요뉴스 제목

trained TextRank, n Documents = 22

#11 (1.64) : "저렴한 전세 어디 없나요?" 무주택자들의 설움 '반전세 월세' 더 늘듯
#13 (1.5) : [추석 이후 주택시장 전망] "공급 전세물량 동반 감소"...가을 이사철이 불안하다
#14 (1.35) : "바로 옆집인데 10억 비싸네" '뻥튀기 호가' 난리난 이곳
#7 (1.3) : “세 올려 稅내자”...초고가 월세시대
#15 (1.26) : 서울 전세 평당 1억3000만원 역대 최고
#5 (1.16) : 불장 이어가는 인천 부동산, ‘청라한양수자인 레이크블루’ 호가 얼마?
#1 (1.15) : ‘구해줘! 홈즈’ 나왔던 ‘9억원’ 광진구 아파트의 1년 후 근황[이슈픽]
#10 (1.13) : 지난해보다 더 오른 아파트값, 추석 후엔 잡힐까
#4 (1.09) : <포럼>부동산 재앙 재확인한 국책硏 보고서
#8 (1.04) : 세금發 '아파트 거래절벽'
#3 (0.93) : "30평대 아파트 10억 돌파 속출" '공급 감소' 대전 집값 상승률 광역시 1위
#2 (0.902) : 올해 외지인 아파트 매입 비중 28.1% 2006년 이래 최고
#16 (0.884) : '전세 평당 1억' 잇따라..."전세값 상승률 법시행 직전 1년 3배"
#0 (0.873) : [영상] 공세권 갖춘 이천자이 더 파크, ‘강남 접근성’ 내세웠지만 ‘공원 빼면 글쎄’
#12 (0.842) : DL이앤씨, 집코노미 박람회에 서울 'e편한세상 강일 어반브릿지' 출품
#6 (0.807) : 홍남기 "전 월세 가격 안정 방안, 올해 안에 찾겠다"
#9 (0.797) : 부산에서 가장 비싼 아파트 브랜드는?
#21 (0.751) : 김만배 누나와 거래, 부친 다운계약 의혹 尹, 정면돌파 나섰다
#19 (0.696) : 중국인, 3년 동안 국내 아파트 3조 매수 미국인 2조원 넘어
#20 (0.67) : 천화동인 이사, 2019년 윤석열 부친 자택 매입 尹측 "시세보다 싸게 팔아”
#18 (0.651) : 박영수 특검 딸, 화천대유 아파트 분양받았다 "집값 2배 뛰어"
#17 (0.598) : 전남도, 부동산실거래 정밀조사 불법증여 의심 등 9건 적발

Keywords Top 30

8.122 아파트
5.695 월세
4.485 가격
4.396 부동산
4.388 서울
4.223 전세
3.463 상승
3.197 시장
2.877 거래
2.830 보증금
2.521 매물
2.446 계약
2.271 기준
2.203 단지
2.008 공급
1.988 집값
1.813 주택
1.748 시행
1.707 매매
1.706 1억
1.697 임대차
1.676 전셋값
1.622 지역
1.593 물량
1.583 수도
1.501 전용

2021년 9월 서울 관련 뉴스

① 2021년 9월 한달 뉴스를 대상으로 서울특별시 아파트매매가격지수와 관련 있는 뉴스 기사 제목과 Keywords를 **TextRank**로 확인함

② 주요 뉴스의 제목을 살펴보면 전월세, 공급량 감소와 가격 상승에 대한 기사 내용들이 대부분임

③ 핵심 Keywords는 아파트(8.1)가 가장 높고, 그 다음은 월세(5.7), 가격(4.5), 부동산 및 서울(4.4), 전세(4.2) 등의 순서로 나타남

➔ **주요 뉴스와 핵심 키워드를 살펴보면, 2021년 9월은 아파트 가격 상승에 대한 불안감 때문에 전월세를 찾는 수요자들이 많으나 공급량이 부족한 상황임을 짐작할 수 있음**

Part 4,

결론 및 한계점

1. 결론
2. 한계점

결론 및 한계점

결론

1. 비정형 데이터를 벡터화시켜 예측 모델링 구현
2. 토픽별 아파트매매가격지수와의 연관성 파악
3. 감성지수와 아파트매매가격지수의 반비례 관계

결론 및 한계점

한계점

① **정제, 혼재, 세분류**에 따른 세부 작업 필요성
② **아파트매매가격에 영향을 주는 변수들 多**
➔ **정형 데이터와의 결합 시, 주요 피쳐 중 하나로 활용 가능**

정제

텍스트 정제

- 아파트 가격 상승/하락에 영향을 주지 않는 무의미한 키워드들에 대해 **추가로 정제하거나 필터링**이 이루어진다면 예측 모델링의 성능이 개선될 것

혼재

상승, 하락 뉴스 기사의 혼재

- 부동산 뉴스는 아파트가격이 상승하고 있어도 하락에 대한 우려 기사가 존재하고 반대로 하락할 때에는 상승할 것이라는 뉴스기사가 존재하여, **아파트 가격을 예측하기 위한 입력변수만으로 사용하기엔 한계점이 존재**

세분류

지역별 뉴스 Filtering

- 입력변수로 “부동산&아파트” 키워드로 검색되는 모든 뉴스에 대한 문서를 사용하였는데, **지역별로 뉴스를 filtering** 할 수 있는 작업이 이루어진다면 **지역별 가격예측 고도화 가능**함

Q&A