[Mini_Project]

고소영

- **주제 선정 과정: 1)** 카카오톡 대화 분석 2) 자연어처리 취업 스펙 우선순위 정하기 3) 지금.. 최종주제..

- **Title:** 무슨 논문을 쓰면 좋을까?: 자연어를 키워드로 한 학술 트렌드 엿보기
- **Source:** www.riss.kr
- Technique: Web_crawling_technique
- Knowledge_related: Text_mining
- **Data_Crawling_scheme**

예시)

구분	연도	제목	페이지 링크	국문 초록
학위논문	2019	자연어처리 모델~	http://www.riss.kr/search/detail/DetailView.do?~~	자연어 처리는 ~

**구분유형: 1) 학위논문 2) 국내학술지논문

☑ 알아야 공부하고, 알아야 쓰고, 알아야 관심을 가질 수 있다.

[데이터 분석 실행 계획]

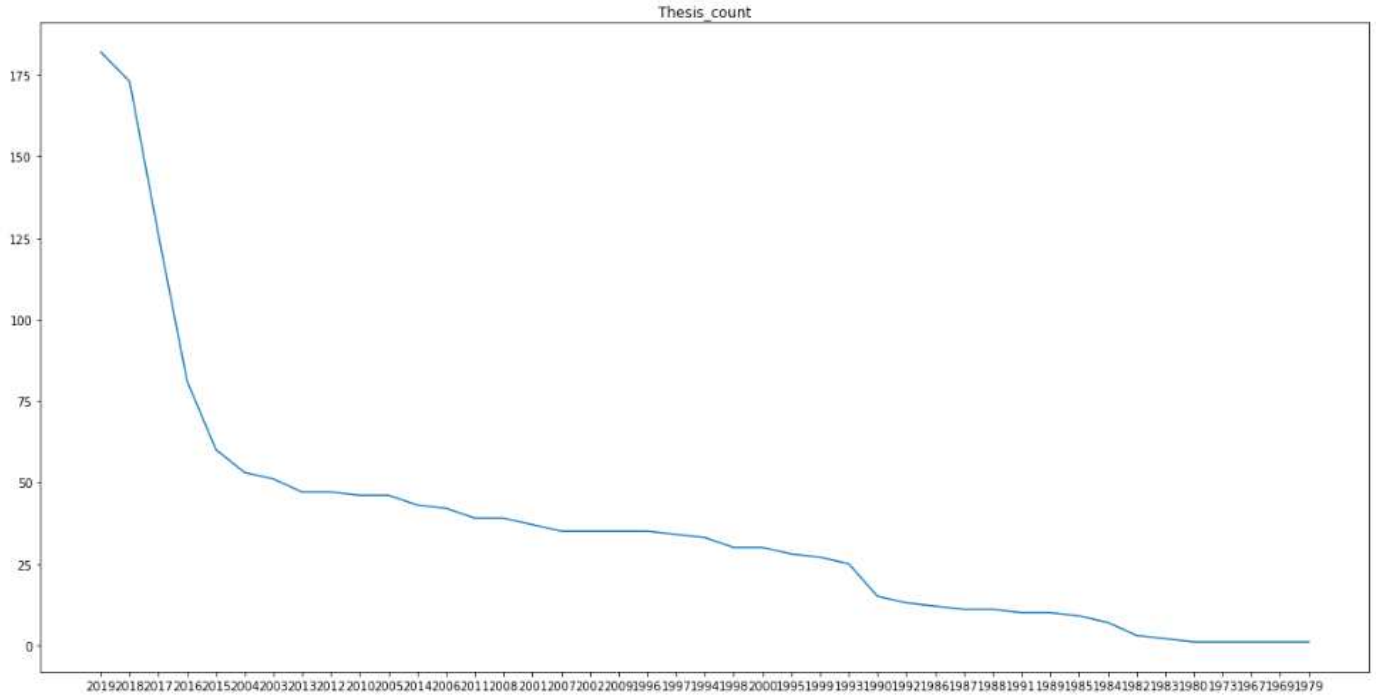
1) 연도별 키워드(빈출단어) Top10 (제목 / 초록기준)
2) 자연어를 중심어로 한 주변 단어 추출하기 (window - 2정도로 할 계획)
~~3) 학위논문 / 학술지 논문 비교~~
4) 문서 유사도 분석 - 논문간
5) Word_Cloud 만들기 (~~연도별~~, 전체)
6) Word2Vec -> 유사단어검사
7) 딥러닝으로 단어 예측하기

결과 1. 데이터 크롤링 결과

데이터 규모 : 학위논문 660건, 국내 학술지 논문 907건 => **총 1567건 수집**

	Gubun	thesis_year	thesis_title	subpage_link	abstract
0	국내학술지논문	1992	자연어 활용(2) : 지능형 지리교육 시스템을 위한 자연어 인터페이스에 관한 연구...	http://www.riss.kr/search/detail/DetailView.do...	1. 서론 2. 지능형 지리교육 시스템의 구조 3. IGTS의 지식베이스 구성 4....
1	국내학술지논문	2011	자연어 인터페이스를 위한 관계에 대한 자연어 표현 자동 수집 방법 = Autom...	http://www.riss.kr/search/detail/DetailView.do...	관계에 대한 다양한 자연어 표현을 다루는 것은 구조 정보에 대한 자연어 질의 인터페...
2	국내학술지논문	2011	자연어 인터페이스를 위한 관계에 대한 자연어 표현 자동 수집 방법 = Autom...	http://www.riss.kr/search/detail/DetailView.do...	관계에 대한 다양한 자연어 표현을 다루는 것은 구조 정보에 대한 자연어 질의 인터페...
3	국내학술지논문	2016	뉴스 기사의 자연어처리 = Natural Language Processing o...	http://www.riss.kr/search/detail/DetailView.do...	뉴스 기사가 빅데이터화함에 따라 뉴스 분석에서 컴퓨터 보조 질적 자료분석 소프트웨어...
4	국내학술지논문	2012	자연어 처리의 현황과 전망	http://www.riss.kr/search/detail/DetailView.do...	자연어 처리는 IT 분야의 기술적 산업적 측면뿐 아니라, 인간의\n 언어 지식을 객...

[연도별 논문 수 그래프]



** 좌측) 2019 - 우측) 1979

결과 2. 데이터 분석

1) 워드클라우드

[자연어 관련 논문 초록 기준 빈출 단어 워드 클라우드]

2) 연도별 키워드 Top10 (분석 - 분석 - 정보 - 데이터)

```
analyze_2019=myokt.nouns(abs_all)

stopwords=['연구', '방법', '서론', '본론', '결론', '등재', '요약']
nouns_extract2019=[word for word in analyze_2019 if len(word)>1]
nouns_final2019=[word for word in nouns_extract2019 if word not in stopwords]

keywords_2019=pd.Series(nouns_final2019)

keywords_2019.value_counts()
```

```
분석      425
데이터     386
모델      286
시스템     273
기반      257
       ...
귀납        1
종성        1
보전        1
이선희       1
몽호        1
Length: 3078, dtype: int64
```

```
analyze_2018=myokt.nouns(abs_all2018)

stopwords=['연구', '방법', '서론', '본론', '결론', '등재', '요약']
nouns_extract2018=[word for word in analyze_2018 if len(word)>1]
nouns_final2018=[word for word in nouns_extract2018 if word not in

keywords_2018=pd.Series(nouns_final2018)

keywords_2018.value_counts()
```

```
분석      417
정보      292
모델      282
데이터     240
한국      227
       ...
도급        1
암시        1
```

```
analyze_2017=myokt.nouns(abs_all2017)

stopwords=['연구', '방법', '서론', '본론', '결론', '등재', '요약']
nouns_extract2017=[word for word in analyze_2017 if len(word)>1]
nouns_final2017=[word for word in nouns_extract2017 if word not in

keywords_2017=pd.Series(nouns_final2017)

keywords_2017.value_counts()
```

```
정보      329
분석      319
활용      214
데이터     210
기반      192
       ...
실내        1
소우        1
양질        1
이민정       1
```

[2014년]

```
thesis_2014=df.groupby('thesis_year').get_group('2014')
ab_2014=thesis_2014['abstract']
abs_all2014=''
for ab in ab_2014.values:
    abs_all2014=abs_all2014+ab

analyze_2014=myokt.nouns(abs_all2014)

stopwords=['연구', '방법', '서론', '본론', '결론', '등재
nouns_extract2014=[word for word in analyze_2014 if len
nouns_final2014=[word for word in nouns_extract2014 if

keywords_2014=pd.Series(nouns_final2014)

keywords_2014.value_counts()
```

```
분석      107
담론       94
시스템      92
모델       83
기반       73
```

** 최신 논문 10선

```
df.sort_values(by='thesis_year').tail(10)['thesis_title']
```

```
1424    소셜 미디어 참여에 관한 연구 동향과 쟁점의 변화 : 네트워크 분석과 클러스터링 기...
1103    A Study on the Extraction of Knowledge Graph f...
838     다문화 영역 이중언어 교육 연구 동향 분석  = Research Trends of...
1327                    어텐션 메커니즘을 활용한 특허 문서의 다중 레이블 분류
1108    계층구조 주목 메카니즘 기반 순환 신경망을 통한 발화 의도 분류 = Intent C...
1431    Context-Aware Cross-Sentence Relation Extracti...
1324    개체-인식 주의집중 매커니즘 기반 양방향 LSTM 네트워크를 통한 의미적 관계 분류...
790      '파우스트의 탄식'과 피구라 - 매체사적 관점에서 바라본 디지털 시대 문예학의 가능...
1098    SNS 정보 기반의 인공지능 챗봇 플랫폼 설계 및 구현 = Designing and...
1263    사회관계망 코퍼스 비정형 토큰의 전처리 및 후처리 부분 패턴문법 연구 = A Stu...
Name: thesis_title, dtype: object
```

Task2. 문서 유사도 분석 - 유사 논문 리스트 출력

```
def get_recommendations(title, cosine_sim=cosine_sim):

    idx = indices[title]
    sim_scores = list(enumerate(cosine_sim[idx]))
    sim_scores = sorted(sim_scores, key=lambda x: x[1], reverse=True)
    sim_scores = sim_scores[1:11]
    thesis_indices = [i[0] for i in sim_scores]
    return df['thesis_title'].iloc[thesis_indices]

get_recommendations('Word2vec와 Label Propagation을 이용한 감성사전 구축 방법에 대한 연구')
```

```
1432                              오피니언 분류의 감성사전 활용효과에 대한 연구
765                             오피니언 분류의 감성사전 활용 효과에 대한 연구
622     오피니언 분류의 감성사전 활용효과에 대한 연구   = A Study on the E...
1257                  오피니언 마이닝 기반의 장르별 감정폭을 적용한 영화 추천 시스템
1520                              단어 임베딩을 활용한 텍스트 임베딩 모델 연구
1063                                 R을 이용한 텍스트 마이닝에 대한 연구
1190    빅데이터 관리를 위한 오피니언 감성사전 모델 설계 = Design of Opinio...
1369                  감성사전 기반 Word2vec 자질을 이용한 감성 분류 시스템
907                             자연어 저장소를 이용한 자연어 질의처리에 관한 연구
1474    소셜 미디어의 감성 분석에 기반한 컨텐츠 추천 방법에 관한 연구 = Study on...
Name: thesis_title, dtype: object
```

** 코사인 유사도 기법 사용, sklearn / TFIDF 사용

Task3. '자연어' 논문 Word2Vec

Okt 사용, 단어 토큰화 이후 이에 해당하는 단어만 추출하여 gensim Word2Vec에 적용

```
from gensim.models import Word2Vec

model=Word2Vec(result, size=100, window=5, min_count=5, workers=4, sg=0)
```

```
sim_words=model.wv.most_similar('자연어')
```

```
sim_words
```

```
[('자연언어', 0.8951966762542725),
 ('음성', 0.7996607422828674),
 ('음성인식', 0.7858203649520874),
 ('기계', 0.7753868103027344),
 ('이해', 0.766119122505188),
 ('한글', 0.739138662815094),
 ('주로', 0.7356473207473755),
 ('모듈', 0.7356259822845459),
 ('처리', 0.7348905801773071),
 ('번역', 0.7316057682037354)]
```

```
sim_2=model.wv.most_similar('모델')
```

```
sim_2
```

```
[('분류기', 0.847148597240448),
 ('비지', 0.8452481031417847),
 ('준지', 0.8413642644882202),
 ('방식', 0.8391686677932739),
 ('학습', 0.8386950492858887),
 ('능가', 0.8345831632614136),
 ('최적', 0.8331055641174316),
 ('를', 0.8084697723388672),
 ('랭킹', 0.799949586391449),
 ('거울', 0.7939714789390564)]
```

[유사 단어 검색 결과 (자연어, 모델)]