

문화콘텐츠와 자연어처리

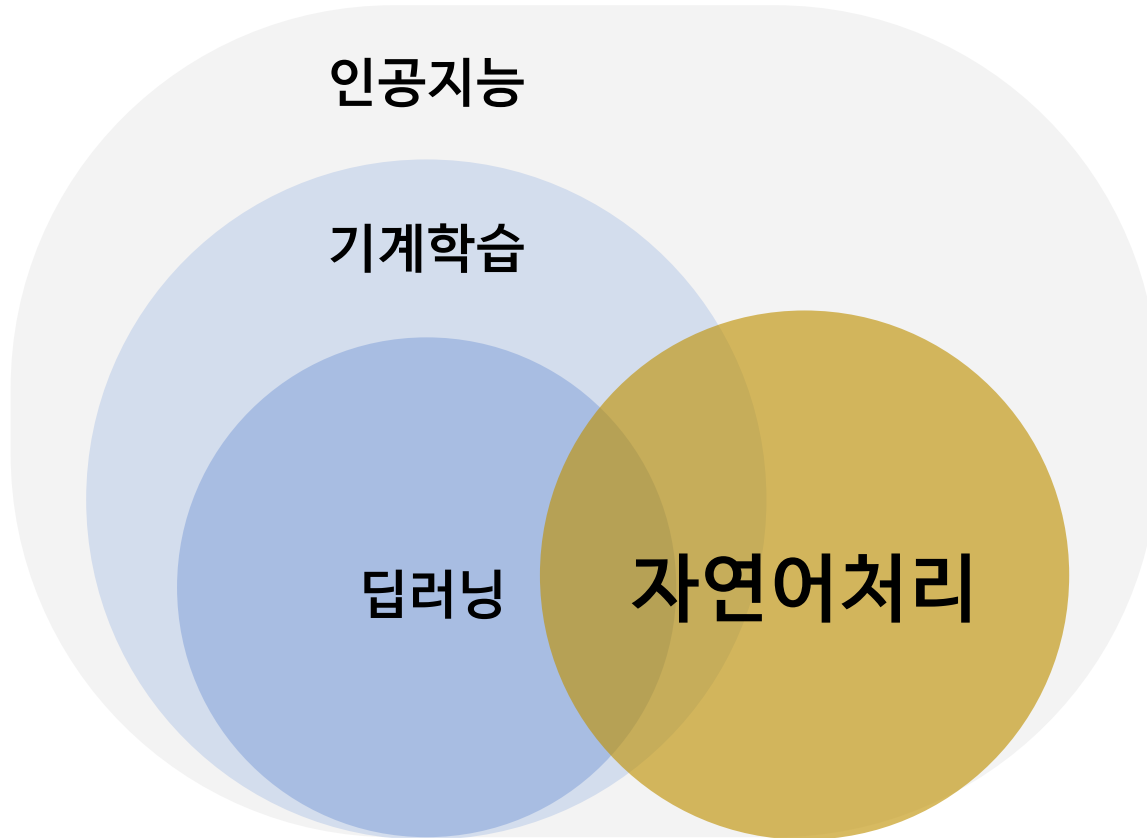
구영은

성균관대학교 문과대학 독어독문학과

성균관대학교 소프트웨어융합대학 컬처앤티크놀로지융합전공

(sarah8835@skku.edu)

자연어처리란?



※ 자연어처리 Natural Language Processing (NLP)

기계가 **인간의 언어**를 잘 분석할 수 있도록 하는 기술

- 자연어 텍스트에 대한 언어적 분석
(ex. 어휘적, 구문적, 의미적, 화용적 분석)
- 언어적 분석을 통한 서비스 개발

- 인공지능 Artificial Intelligence (AI)
- 기계학습 Machine learning (ML)
- 딥러닝 Deep learning (DL)

인간과 언어

■ 언어가 인간에게 갖는 의미는?

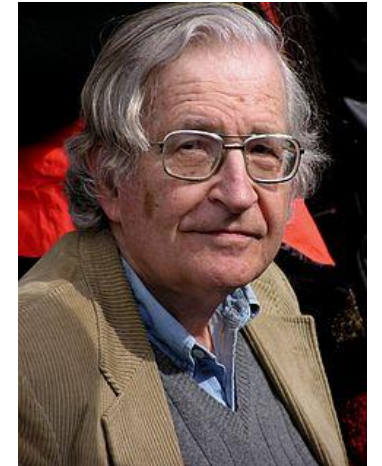
- 의사소통의 도구
- 문화/문명을 이룩하게 하는 수단
- 인간은 ‘나’와 ‘세상’을 연결하는 매개체



Steven Pinker
(1954.09.18~)

“ Language is a window into human nature,
but it is also a fistula, an open wound through which we’re
exposed to an infectious world. ”

“ Language is the mirror of the mind;
and a detailed study of language might reveal to us
just how the mind works. ”



Noam Chomsky
(1928.12.07~)

인간과 언어



밥은 먹고 다니냐?

인간과 언어

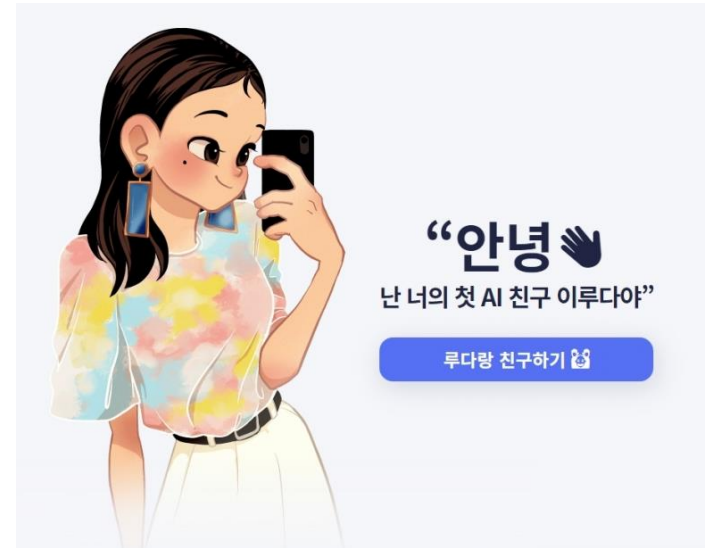


꼭 그렇게 다 가져가야만 속이 후련했냐?

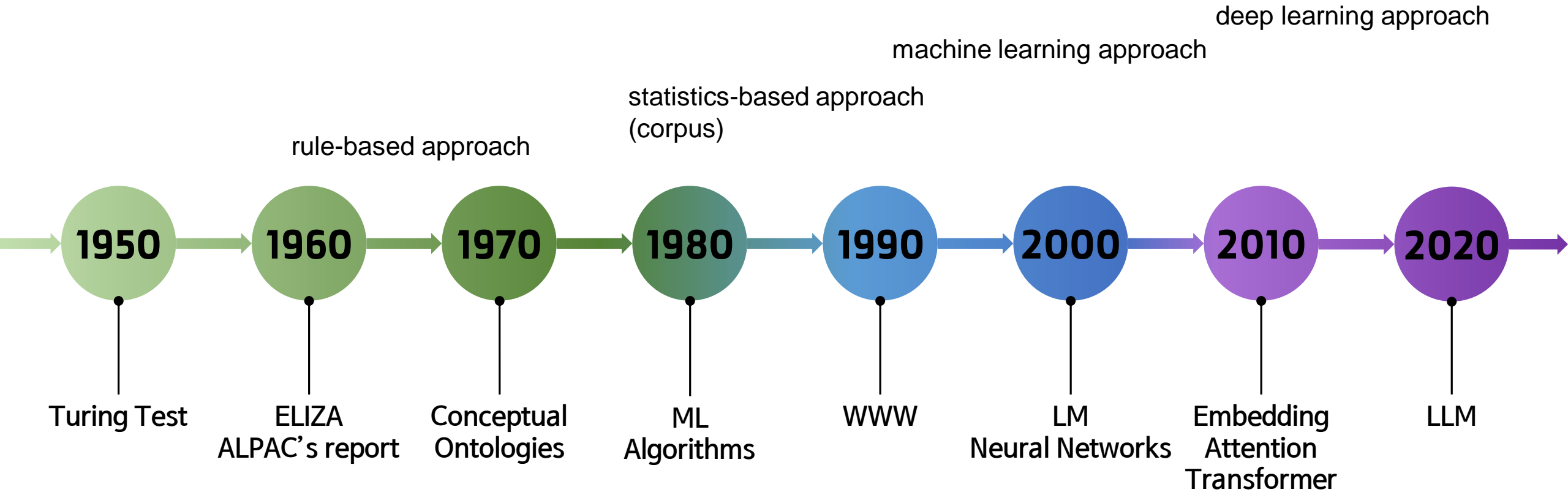
인간과 언어



NLP 기술을 활용한 서비스



Evolution of NLP



Rule-based NLP

■ 1954: Georgetown Experiment (최초의 기계번역 실험)

- 조지타운 대학과 IBM 이 공동으로 기계번역 시스템을 개발
- 60개가 넘는 러시아어 문장을 영어로 자동 번역을 시연함

□ ALPAC's (Automatic Language Processing Advisory Committee) report

- ALPAC은 1966년 기계 번역에 대하여 매우 부정적인 보고를 발표
- 이후 기계 번역 연구에 대한 정부 지원금이 많이 삭감됨



Rule-based NLP

■ 1960's: ELIZA (대화를 하는 최초의 AI 프로그램)

- 1966년 MIT 인공지능 연구소의 조세프 바이젠바움이 개발
- 패턴 매칭(pattern matching)과 대체 방법(substitution)을 이용하여 대화를 적절히 흉내냄
 - 마치 프로그램이 상대방의 발화를 이해를 하는 것처럼 착각하게 함
 - 개발자가 일일이 사람들의 예상 질문과 답변을 프로그래밍 해야 함
 - 예상치 못한 질문이나 말을 건넨다면 답변할 수 없음



Joseph Weizenbaum
(1923.01.08-2008.03.05)

```

○○○                               Eliza (1966)

> Hello, I am Eliza.
* Hey!
> Don't you ever say Hello?
* Hello!
> How are you today.. What would you like to discuss?
* I'm depressed.
> Do you enjoy being depressed?
* No, I don't.
> Why no?
* Because I want to be happy.
> Why do you want to be happy?
* Because I'm tired od being depressed.
> That is quite interesting.
* I'm not sure█
  
```

```

Welcome to
EEEEEE LL      IIII ZZZZZZ  AAAAA
EE      LL      II      ZZ  AA  AA
EEEEEE LL      II      ZZZ  AAAAAA
EE      LL      II      ZZ  AA  AA
EEEEEE LLLLLL  IIII  ZZZZZZ  AA  AA

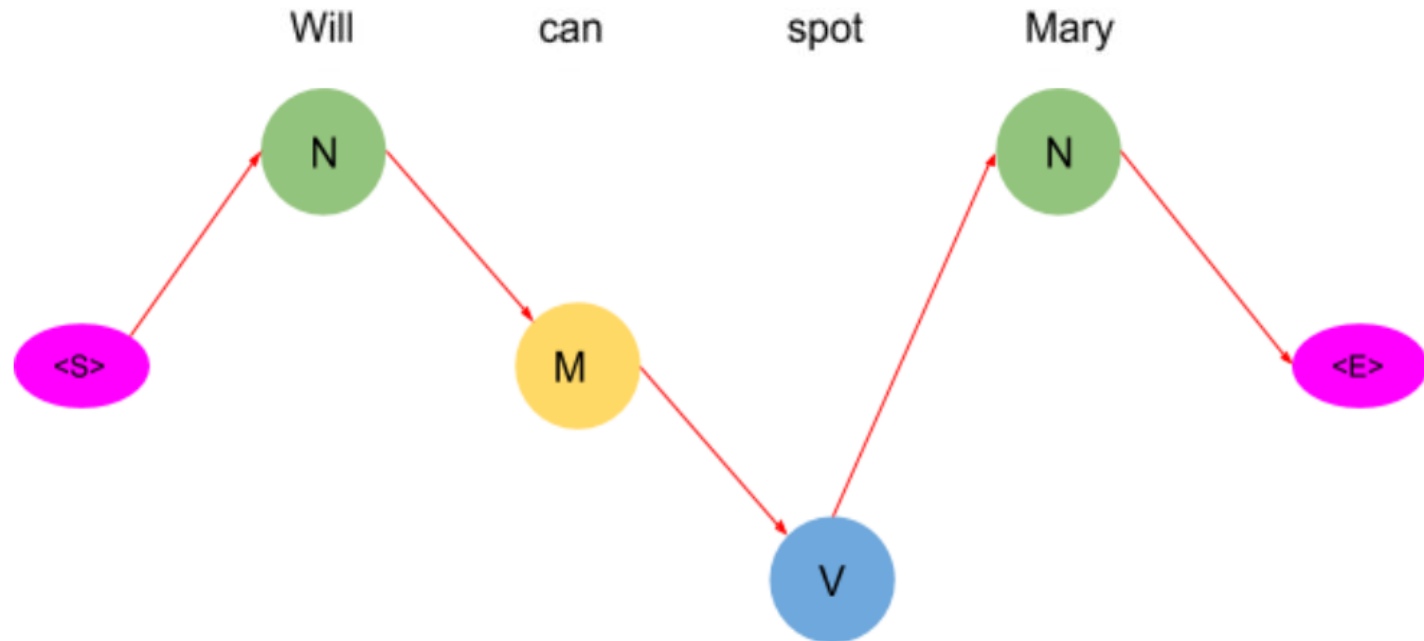
Eliza is a mock Rogerian psychotherapist.
The original program was described by Joseph Weizenbaum in 1966.
This implementation by Norbert Landsteiner 2005.

ELIZA: Is something troubling you ?
YOU:   Men are all alike.
ELIZA: What is the connection, do you suppose ?
YOU:   They're always bugging us about something or other.
ELIZA: Can you think of a specific example ?
YOU:   Well, my boyfriend made me come here.
ELIZA: Is it important to you that your boyfriend made you come here ?
YOU:   He says I'm depressed much of the time.
ELIZA: I am sorry to hear that you are depressed.
YOU:   It's true. I am unhappy.
ELIZA: Can you explain what made you unhappy ?
YOU:   █
  
```

Statistics-based NLP

■ Late 1980's

- 기계 학습과 통계적인 모델들이 등장함에 따라 통계 기반 자연어처리가 시작됨
- 통계 기반 자연어처리의 수학적 배경에는 확률 이론과 정보 이론 등이 있음
 - Naïve Bayes
 - Decision Tree
 - Hidden Markov Model(HMM)
 - Maximum Entropy Model
 - Support Vector Machine(SVM)
 - Conditional Random Fields(CRF)



Deep Learning-based NLP

■ 딥러닝 (Deep Learning)

- 인간의 뇌와 유사한 방식으로 기능하도록 구축된 알고리즘을 사용하는 기계학습의 한 유형

■ 임베딩 (Embedding)

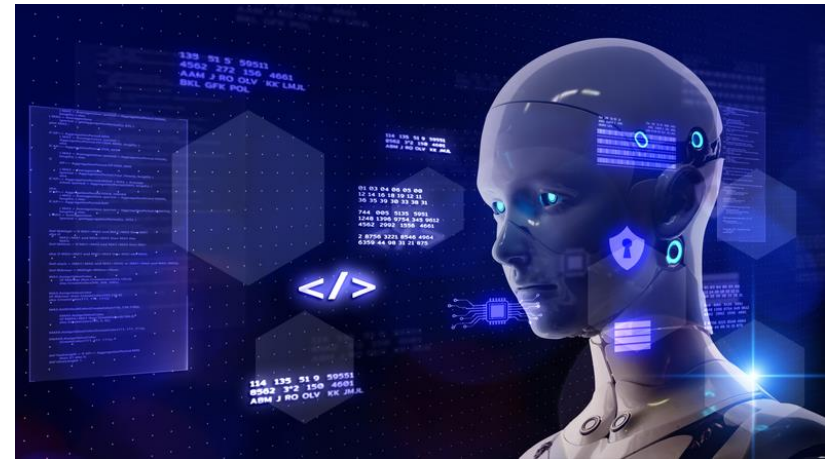
- Word2Vec, GloVe, FastText
- ELMo, BERT

▶ 인간 언어를 **무엇으로** 학습?

■ 언어 모델 (Language Model)

- CNN, RNN, LSTM
- Seq-2-seq
- Transformer (BERT, GPT)

▶ 인간 언어를 **어떻게** 학습?



Deep Learning-based NLP

■ 어텐션 (Attention)

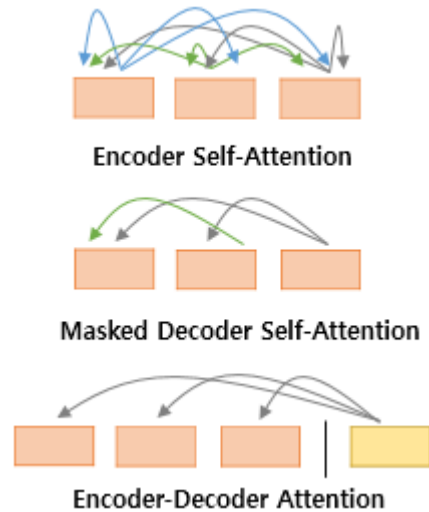
- 디코더에서 출력 단어를 예측하는 시점 (time step)마다, 인코더에서의 전체 입력 문장을 다시 한 번 참고
- 해당 시점에서 예측해야 할 단어와 연관이 있는 입력 단어에 더 집중



Deep Learning-based NLP

■ 트랜스포머(Transformer)

- 2017 구글이 발표한 논문 'Attention is all you need' 에 나온 모델
- 어텐션 기법을 기반으로 함



- BERT, GPT 등의 기본 모델로 활용됨

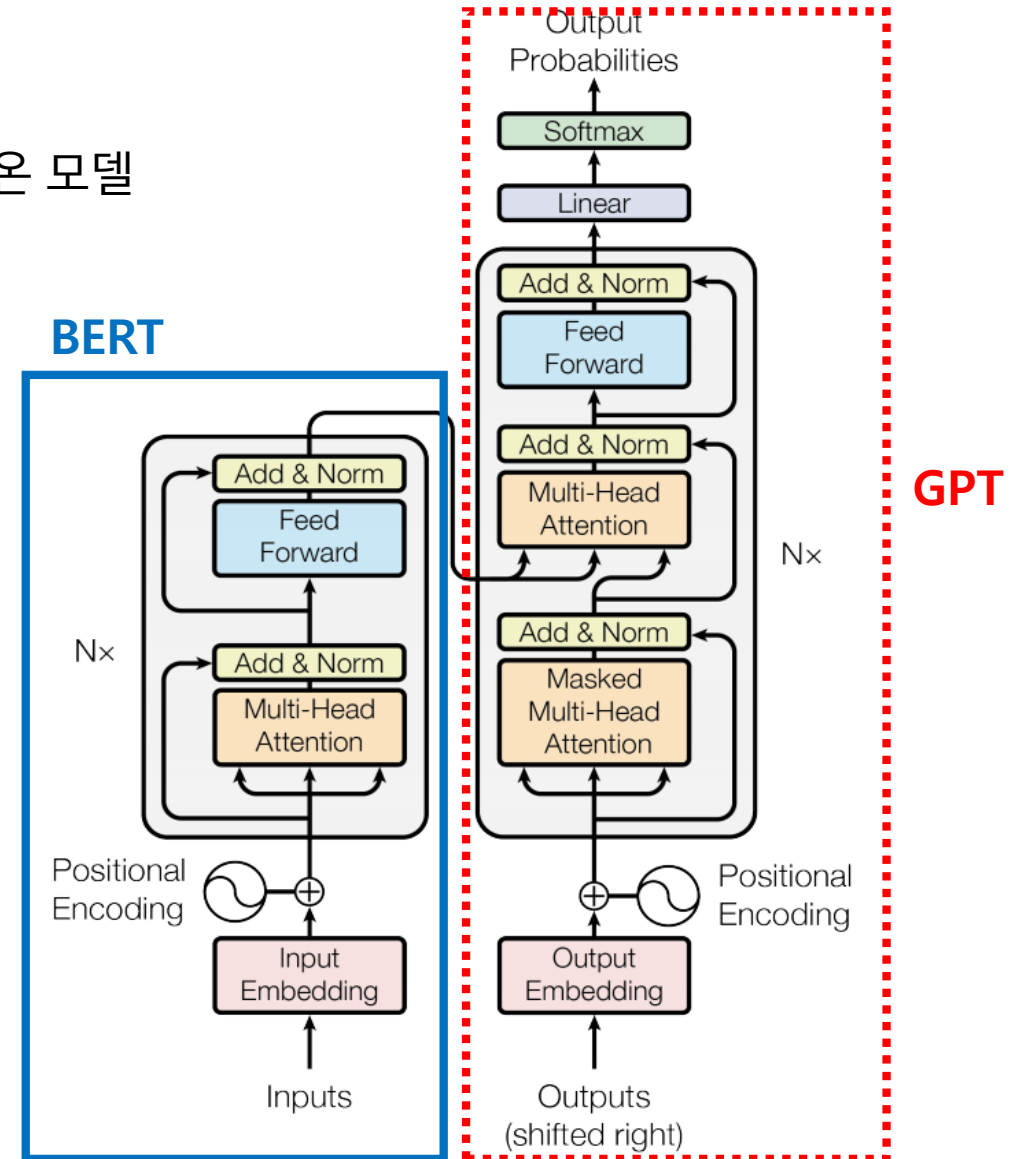


Figure 1: The Transformer - model architecture.

BERT

■ BERT (Bidirectional Encoder Representations from Transformers)

- 2018년 10월 구글에서 발표한 모델 (Devlin et al. 2018)
- 트랜스포머 Transformer의 **인코더**를 활용한 사전훈련 기반의 딥러닝 언어 모델 (PLM)
 - 위키피디아와 같은 대용량의 **unlabeled data**로 모델을 사전훈련 **pre-training**한 이후, 특정 태스크의 **labeled data**로 미세조정 **fine-tuning**하여 언어처리 태스크를 수행하는 모델
- 다양한 **자연어 이해** 분야에 응용 가능



- 주요 특징

- 1) 방대한 텍스트 코퍼스(Wikipedia 등)를 이용하여 **범용 목적의 언어 이해 모델**을 사전 훈련
- 2) 하나의 양방향 모델이 **문장의 앞뒤 문맥을 동시에 활용**해서 의미를 해석할 수 있어 언어처리에서 높은 정확도를 달성

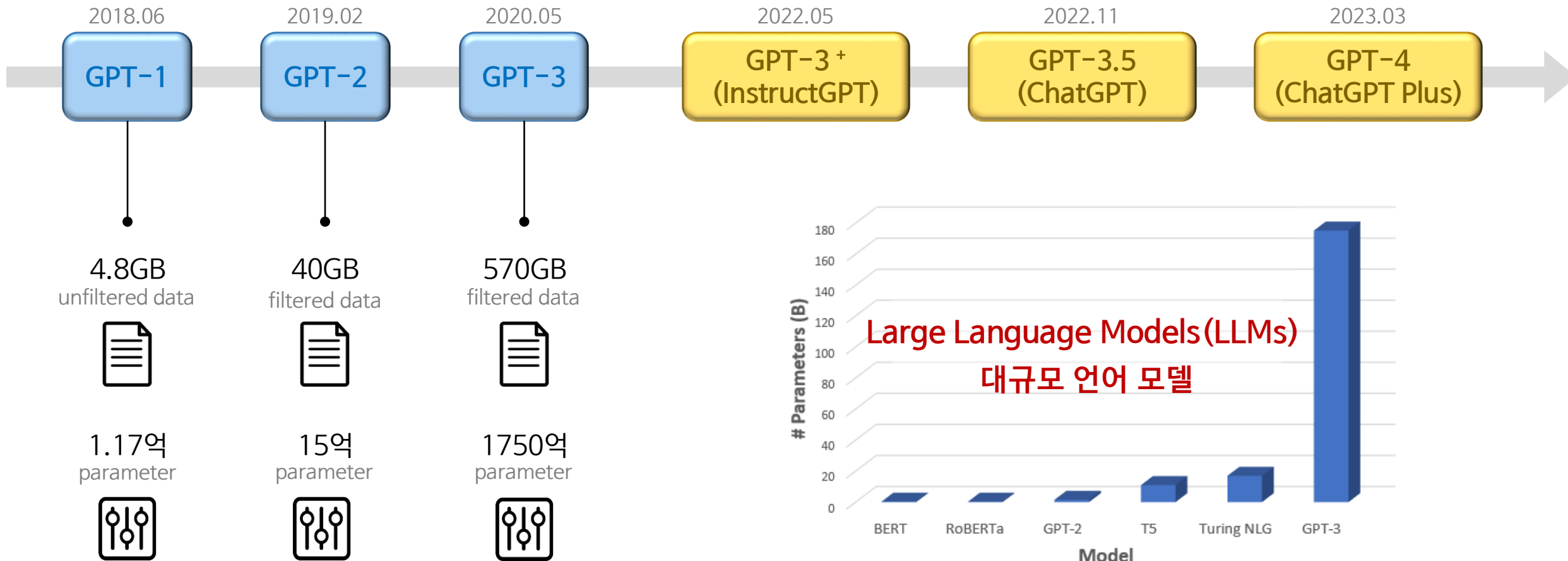
SQuAD1.1 Leaderboard

Rank	Model	EM	F1
	Human Performance Stanford University (Rajpurkar et al. '16)	82.304	91.221
1 Oct 05, 2018	BERT (ensemble) Google AI Language https://arxiv.org/abs/1810.04805	87.433	93.160
2 Sep 09, 2018	nlnet (ensemble) Microsoft Research Asia	85.356	91.202
3 Jul 11, 2018	QANet (ensemble) Google Brain & CMU	84.454	90.490

GPT

■ GPT (Generative Pre-trained Transformer)

- 2018년을 시작으로 OpenAI에서 발표한 모델 (Radford et al. 2018)

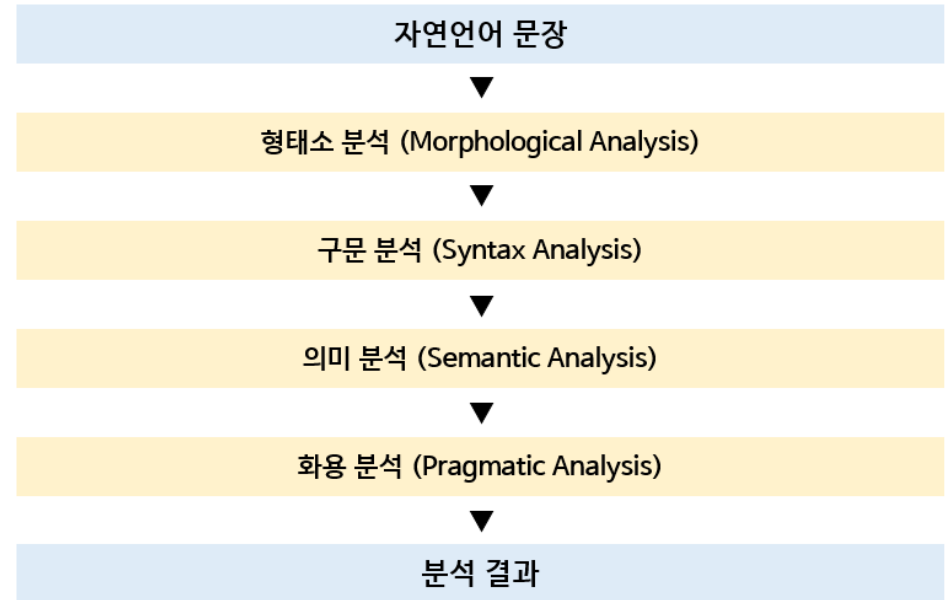


Steps for NLP

■ 언어학(Linguistics)

언어체계 또는 언어구조에 대한 과학적 연구 (scientific study of language system / structure)

- 음성학(Phonetics), 음운론(Phonology)
소리과 관련된 학문
- 형태론(Morphology)
언어 표현의 생성(결합) 원리를 연구하는 학문
- 통사론(Syntax)
문장 내 구성 요소의 결합 원리를 연구하는 학문
- 의미론(Semantics)
언어 표현의 의미를 연구하는 학문
- 화용론(Pragmatics)
맥락 내에서 언어 구조의 사용 양상을 연구하는 학문



NLP Techniques



Tokenization

Stopword Removal

Bag-of-Words

N-grams

TF-IDF

POS tagging

Predicate Argument Structure

Keyword Extraction

Named Entity Recognition

Word Sense Disambiguation

Topic Modeling

Intent Detection

Anaphora Resolution

Summarization

Sentiment Analysis

Question Answering

Machine Translation

Information Retrieval

컬텍 자연어처리 교과목

1학기

<언어공학과 문화콘텐츠>

- 인간과 언어
- 문화콘텐츠 개발과 언어데이터 분석의 필요성
- 언어학 이론 기초
 - 형태론, 통사론, 의미론, 화용론
- 언어데이터 전처리 및 분석 기초
- 언어데이터 분석 기반의 콘텐츠 기획

2학기

<문화콘텐츠와 자연어처리>

- 문화/예술/콘텐츠 산업의 언어데이터 분석
- 문화/예술/콘텐츠 산업의 데이터 크롤링
- 자연어처리 방법론
 - 어휘/형태, 통사, 의미분석
 - POS tagging, Dependency parsing, SRL, WSD, Keyword extraction
- 기계번역, 감성분석, 개체명인식
- 기계 학습, 딥러닝

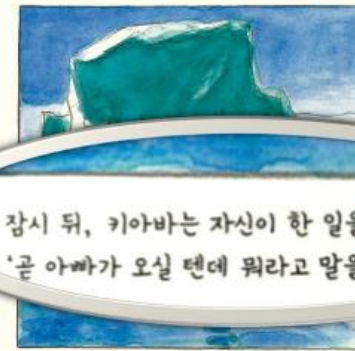
수업 사례 1

- 유아 교육용 텍스트 분석 서비스

유아 교육용 텍스트 분석 서비스

1. 주제 선정

관련 자료 탐색



잠시 뒤, 키아바는 자신이 한 일을 **뉘우쳤어요.**
'곧 아빠가 오실 텐데 뭐라고 말을 하지?'



고무장갑은 누구보다도 **정성껏** 세안을 심었어요.

'낮설다', '뉘우치다', '정성껏' 등
다소 어려울 수 있는 어휘들

유아 교육용 텍스트 분석 서비스

2. 선행 연구 검토

형태론·음운론적 특성을 고려한 분석

친숙성

친숙한 주변 사물의 이름으로부터 학습함

복잡성

음운론적 관점에서 볼 때 발음이 명료한
모음이나 자음이 첨가되는 글자가 먼저 습득됨

습득성

'단어 → 음절 → 음소' 순으로
유아의 인식 능력이 발달해 나감



단계	조합 방식	예
1	단모음 / 기본자음 + 단모음	우, 가
2	경음 또는 격음 + 단모음	쁘, 코
3	받침이 있는 단모음	악, 강
4	경음 또는 격음 + 단모음 + 받침	꼭, 땅
5	이중모음 / 기본자음 + 이중모음	와, 의, 과
6	받침이 있는 이중모음	광, 광
7	경음 또는 격음 + 이중모음	파, 따, 뒤
8	경음 또는 격음 + 이중모음 + 받침	팅

유아 교육용 텍스트 분석 서비스

2. 선행 연구 검토

서비스 대상1 : 유아 텍스트 수용자 (ex. 유아 학부모)

음운론적 특성을 고려한 분석

입력 : 텍스트 (학습서, 동화책 등)

출력 : 난이도 (어휘 기반)



음운론적 관점에서 볼 때 발음이 명료한
모음이나 자음이 첨가되는 글자가 먼저 습득됨

단계	조합	예
1	단모음 / 기본자음 + 단모음	우, 가
2	서비스 대상2 : 유아 텍스트 창작자 (ex. 동화 작가) 경음 또는 격음 + 단모음	우, 고
3	받침이 있는	악, 강
4	경음 또는 격음 + 단모음 + 받침	꼭, 땅
5	이중모음 / 기본	와, 의, 과
6	받침이 있는 이중모음	광, 광
	경음 또는 격음 + 이중모음	파, 따, 뛰
	받침	팅

한글의 문법적 특성을 활용해, 경음 또는 격음 + 이중모음
복잡도의 단계적 구분을 분석에 활용 가능!

유아가 읽을 텍스트의 난이도를 계산해주는 서비스를 만들어보자!

수업 사례 2

- 수리신청 글 분석을 통한 기숙사 시설 개선

수리신청 글 분석을 통한 기숙사 시설 개선

문화콘텐츠와자연어처리



주제 선정 계기

"둘 다 기숙사 생활을 하면서
시설고장 불편을 겪었던 경험이 있고
본교 학생으로써 학교 기숙사 개선에
기여하고 싶은 마음도 있다."

-성대 18학번들



수리신청 글 분석을 통한 기숙사 시설 개선

문화콘텐츠와자연어처리



"변기가 막혔어요"
"온수가 안 나와요"
"세면대 고장났어요"


분석 →

데이터 분석 의의

- 잦은 고장 시설 파악 후 예방 조치
- 데이터 분석을 통한 기숙사 예산 배분
- 본교 학생 기숙사 생활의 질 향상

수리신청 글 분석을 통한 기숙사 시설 개선

문화콘텐츠와자연어처리





분석 데이터 소개

성균관대학교 명륜학사 홈페이지
-www.dorm.skku.edu/dorm_seoul

수리신청 현황 페이지
-기숙사별 고장 시설 수리 신청
-고장 시설 및 기숙사 파악 가능

수리신청 글 분석을 통한 기숙사 시설 개선


문화콘텐츠와자연어처리



키워드 추출 언어 분석

화장실 문이 잠겼습니다.

샤워기 헤드 교체 신청합니다. ...

화장실 수건걸이 한 쪽 나사가 ..

화장실 변기가 갑자기 막혔습니..

도어락 배터리가 없어요..

세면대 아래 하수구가 막힌것 같..

세면대 있는 곳의 전등이 나갔습..

입사할때부터 세면대가 막혀 있..

인터넷이 안됩니다. 어제 아침까..

도어락 배터리가 나가서 건전지..

품사 기반 키워드 추출

- 명사 위주의 키워드 추출 (kkma.pos, NNG)
- 기숙사별 키워드 분류
- 수리 신청 목적성 데이터 (감정분석, 기계학습 모델 필요성 x)

데이터 일부

수업 사례 3

– 언어습관 개선 프로그램

언어습관 개선 프로그램

기획의도


언어사용의 중요성

언어사용의 중요성

문제정의

대학생의 부정적 언어

“언어사용이 중요함에도 불구하고 많은 대학생이 습관적, 무의식적으로 부정적인 언어를 구사하고 있음”



출처: 한림미디어랩 대학생 110명 설문조사

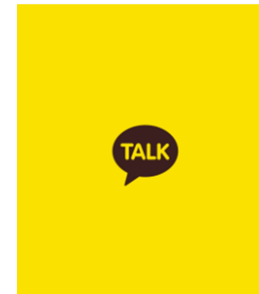
대학생 110명 설문조사 결과 2029세대는 부정적인 언어표현을 많이 사용하는 것으로 나타남

화면 캡처: 생로병사의 비밀

긍정적 언어는 도파민을, 부정적 언어는 코티솔(스트레스호르몬)을 분비시킴

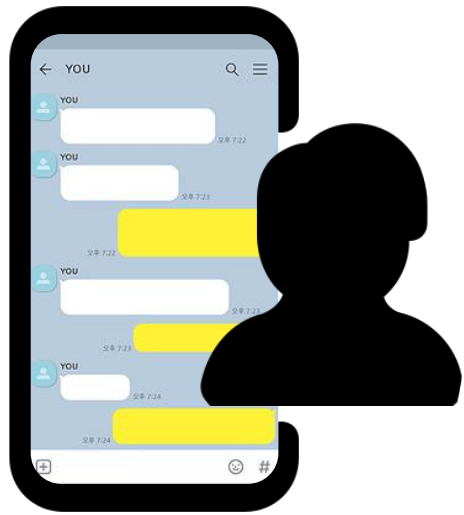
서비스소개

언어습관 개선 프로그램

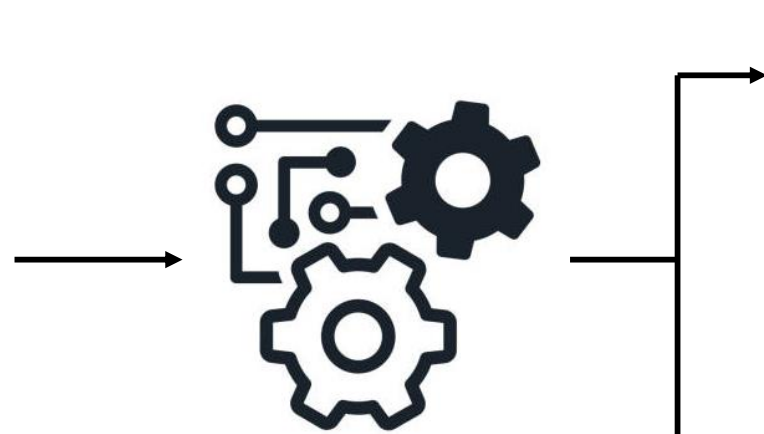


- ✓ 카카오톡에서 사용자의 언어 데이터를 분석
- ✓ 언어 데이터의 감성 분석을 통해 긍정, 부정 언어 표현의 비율을 계산
- ✓ 사용자가 그동안 무의식적으로 사용했던 언어가 긍정적인 언어인지, 부정적인 언어인지 스스로 **모니터링** 할 수 있도록 도움 제공

언어습관 개선 프로그램



데이터
(사용자 카카오톡 채팅)



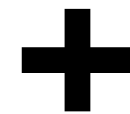
기계학습 모델
(언어습관 개선 프로그램)



긍정 언어



부정 언어



언어 사용 개선
긍정 언어 제안

※ 기계학습(Machine Learning)

주어진 데이터를 통해 기계가 데이터의 유형(class)간 관계를 파악하고, 새롭게 관측된 데이터의 유형을 스스로 판별할 수 있도록 학습시키는 방법

언어습관 개선 프로그램

채팅 데이터 불러오기

```
raw1 = pd.read_csv("/gdrive/My Drive/GCO/wonder.csv")
```

데이터 전처리

```
pattern1 = '[-+=,#/?:;$.A^*"-%!@wWnWrWt▼♣§☆♡'▽' /♥▲◆▶*— "" ']'
raw1['maintext'] = raw1['maintext'].progress_map(lambda x:re.sub(pattern1, '', x).strip())
```

```
100% ██████████ | 7912/7912 [00:00<00:00, 262827.08it/s]
```

```
dio_url = 'https://raw.githubusercontent.com/park1200658/KnuSentiLex/master/KnuSentiLex/data/SentiW
dio_df = pd.read_json(dio_url)
```

```
dio_new = pd.DataFrame(dio_df['word'])
dio_new['tag'] = 'NNP'
dio_new
```

```
dio_new.to_csv("custom_diot.txt", index=False, header=False, sep="\t") #사용자지정 사전을 만든다
```

```
komoran.set_user_dio("custom_diot.txt")
```

```
raw1['tokens'] = raw1['maintext'].progress_map(lambda x:komoran.get_nouns(x))
```

감성분석 실시

```
word_list = dio_df['word'].unique()
raw1['sent_score'] = raw1['tokens2'].progress_map(lambda x:sum([dio_df[dio_df['word']==word]['polar
if word in word_list else 0 for wor
```

```
100% ██████████ | 7912/7912 [00:08<00:00, 2490.78it/s]
```

```
raw1['word_count'] = raw1['tokens2'].progress_map(lambda x:len(x))
```

```
100% ██████████ | 7912/7912 [00:00<00:00, 477383.78it/s]
```

```
raw1['sent_index'] = raw1['sent_score']/raw1['word_count']
```

```
raw1['sent_index'].describe()
```

```
count      3846.000000
mean         0.093934
std          0.582843
min          -2.000000
25%          0.000000
50%          0.000000
75%          0.000000
max           2.000000
Name: sent_index, dtype: float64
```


언어습관 개선 프로그램

실시한 감성 분석에 대한 결과를 출력하여 보여주기 (긍/부정 CLASS 제시)

```
score = raw1['sent_index'].mean()
print('긍부정 점수: {}'.format(score))

if -2<score<-1:
    print('당신은 부정적인 언어습관을 가지고 있습니다')
elif -1<=score<0:
    print('당신은 비교적 부정적인 언어습관을 가지고 있습니다')
elif 0<=score<1:
    print('당신은 비교적 긍정적인 언어습관을 가지고 있습니다')
elif 1<=score<2:
    print('당신은 긍정적인 언어습관을 가지고 있습니다')
```

긍부정 점수: 0.09393422144382098
당신은 비교적 긍정적인 언어습관을 가지고 있습니다

언어습관 개선 프로그램

긍정적 언어사용 추천 서비스

```
pos = pd.read_csv("/gdrive/My Drive/GCO/pos_sentiment.csv")  
  
today_word = pos.sample(n=3)  
print('오늘은 이런 말을 써보세요: {}'.format(today_word['text'].unique()))
```

오늘은 이런 말을 써보세요: ['수고많았어' '즐거워' '할 수 있어']

감사합니다

구영은

성균관대학교 문과대학 독어독문학과

성균관대학교 소프트웨어융합대학 컬처앤티크놀로지융합전공

(sarah8835@skku.edu)