자연어 처리

최용복 9기
김남호 9기

목차

NLP?

1 자연어

2 자연어 처리

3 KoNLPy

NLP!

Natural Language
자연어

자연어란?

자연어

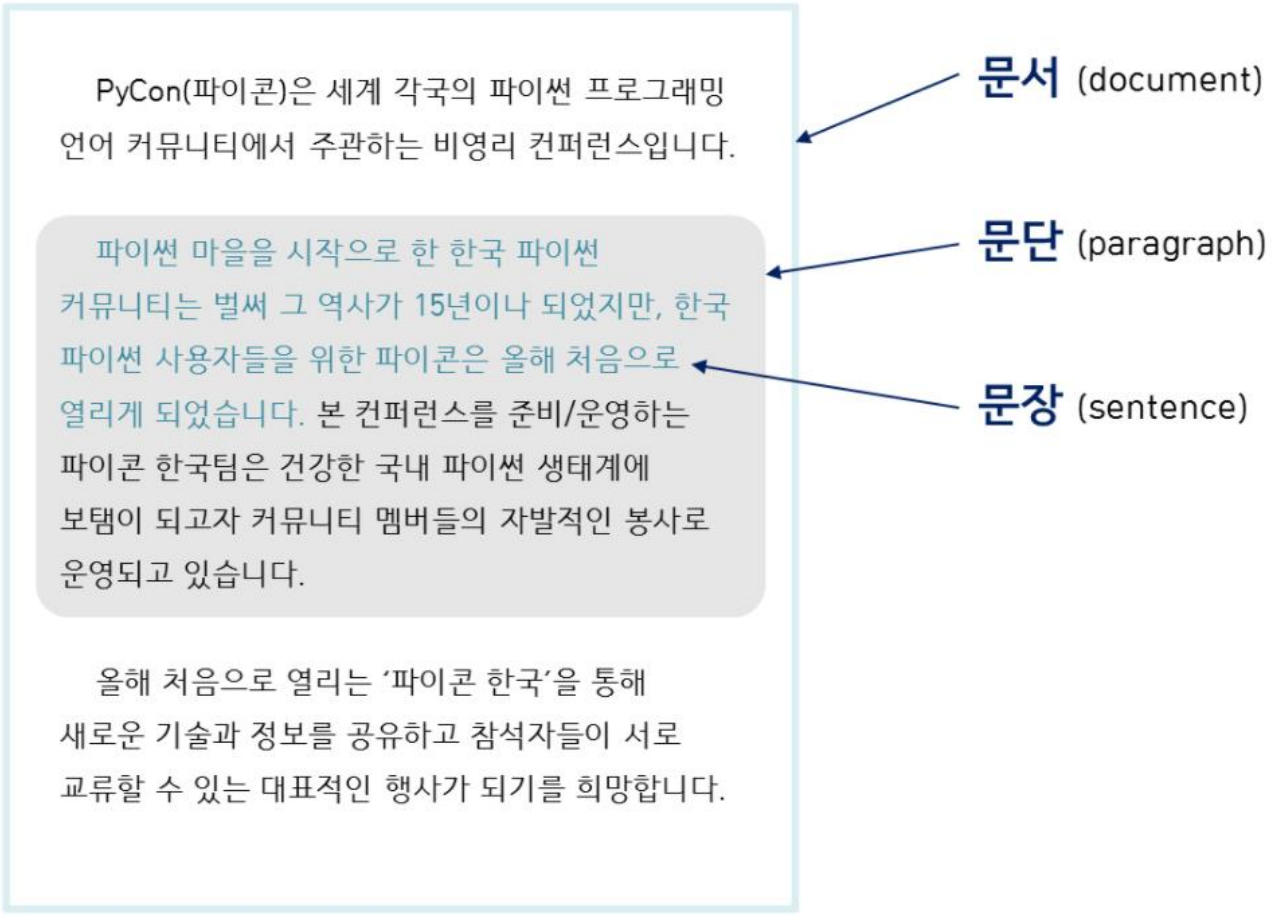
사람들이 일상적으로 쓰는 언어를
인공적으로 만들어진 언어인 인공어와 구분하여 부르는 개념

자연어의 구성

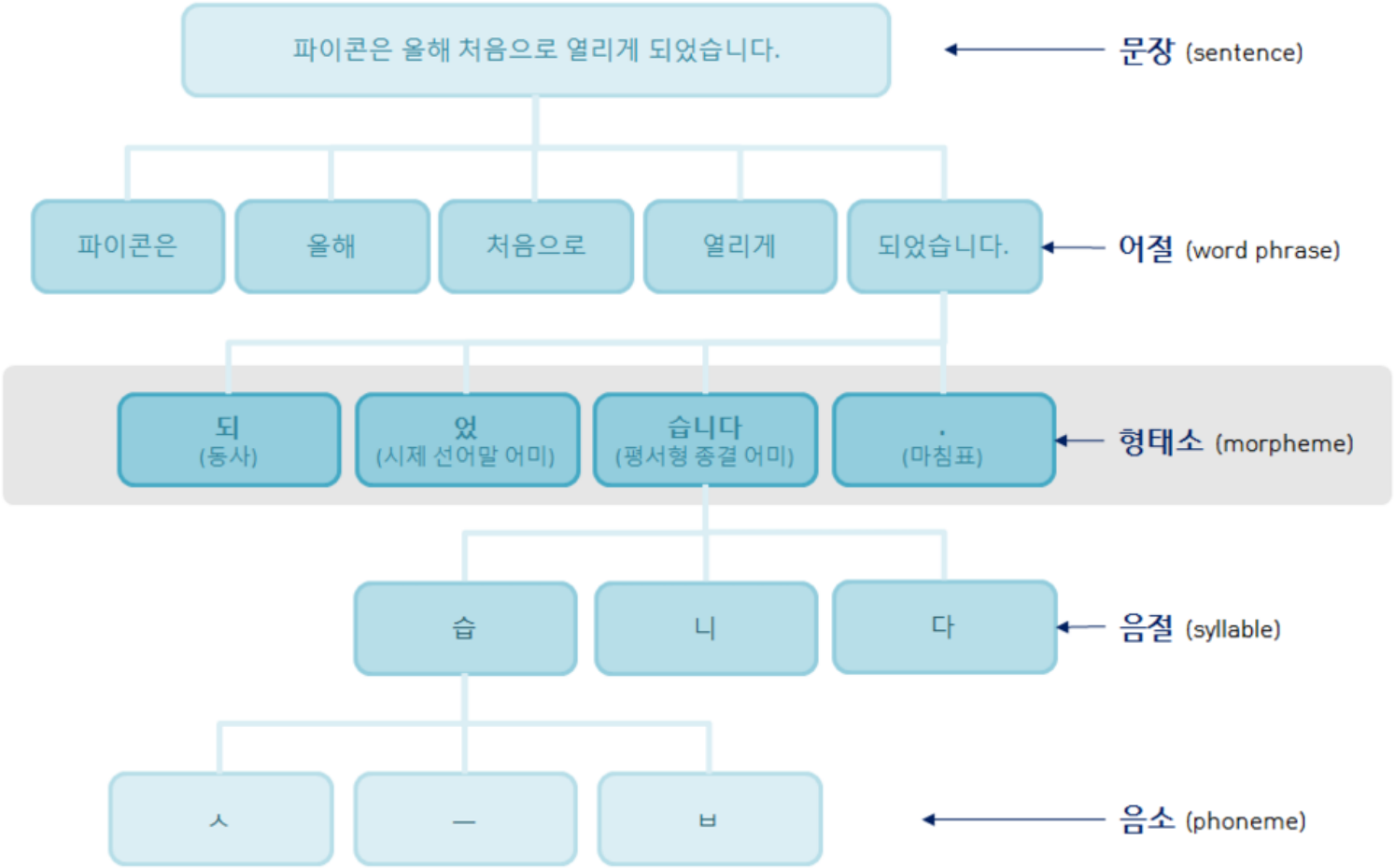
말

글

자연어의 구성


PyCon(파이콘)은 세계 각국의 파이썬 프로그래밍 언어 커뮤니티에서 주관하는 비영리 컨퍼런스입니다.
파이썬 마을을 시작으로 한 한국 파이썬 커뮤니티는 벌써 그 역사가 15년이나 되었지만, 한국 파이썬 사용자들을 위한 파이콘은 올해 처음으로 열리게 되었습니다. 본 컨퍼런스를 준비/운영하는 파이콘 한국팀은 건강한 국내 파이썬 생태계에 보탬이 되고자 커뮤니티 멤버들의 자발적인 봉사로 운영되고 있습니다.
올해 처음으로 열리는 '파이콘 한국'을 통해 새로운 기술과 정보를 공유하고 참석자들이 서로 교류할 수 있는 대표적인 행사가 되기를 희망합니다.
문서 (document)
문단 (paragraph)
문장 (sentence)

자연어의 구성

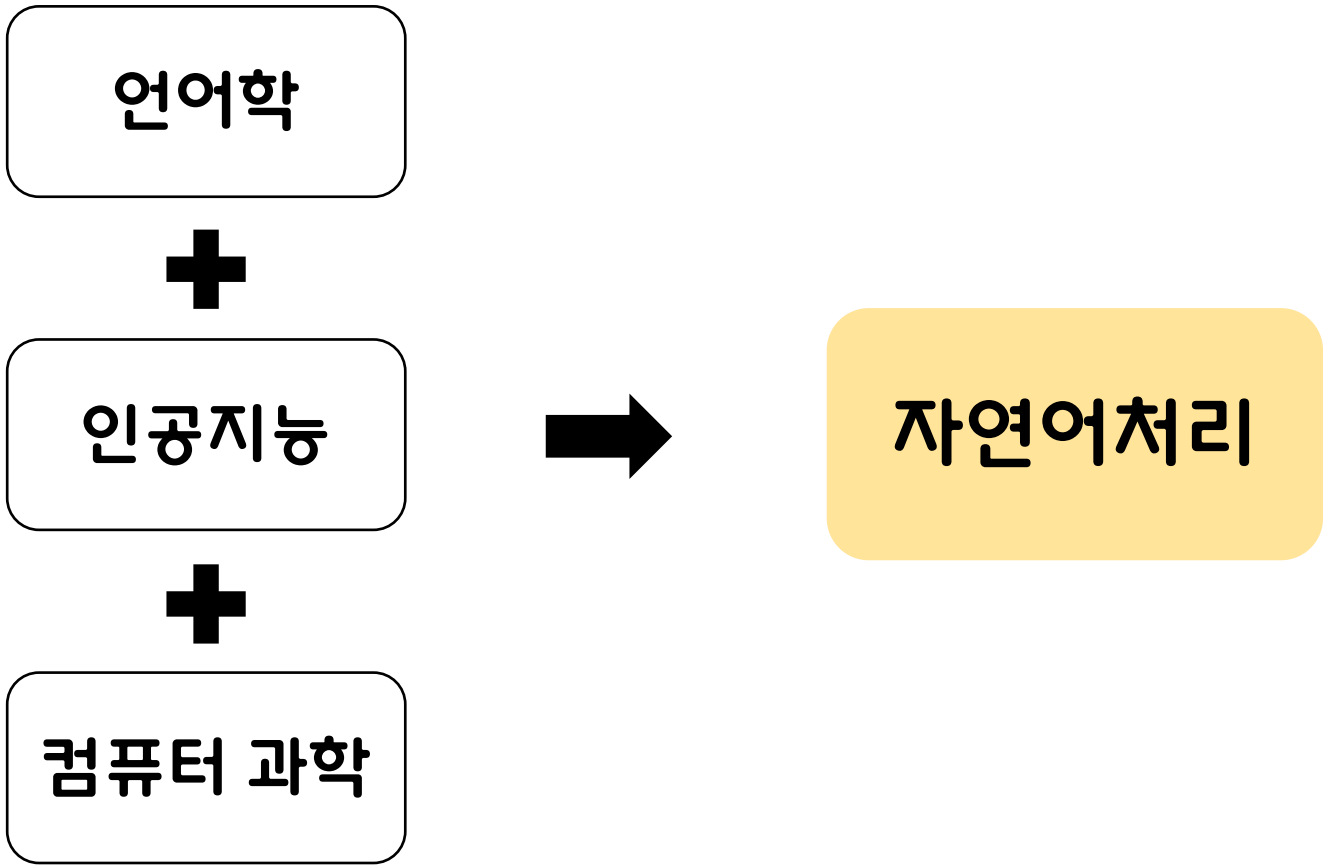

파이콘은 올해 처음으로 열리게 되었습니다.
문장 (sentence)
파이콘은
올해
처음으로
열리게
되었습니다.
어절 (word phrase)
되
(동사)
었
(시제 선어말 어미)
습니다
(평서형 종결 어미)
.
(마침표)
형태소 (morpheme)
습
니
다
음절 (syllable)
ㅅ
ㅡ
ㅂ
음소 (phoneme)

Natural Language Process
자연어 처리

자연어 처리

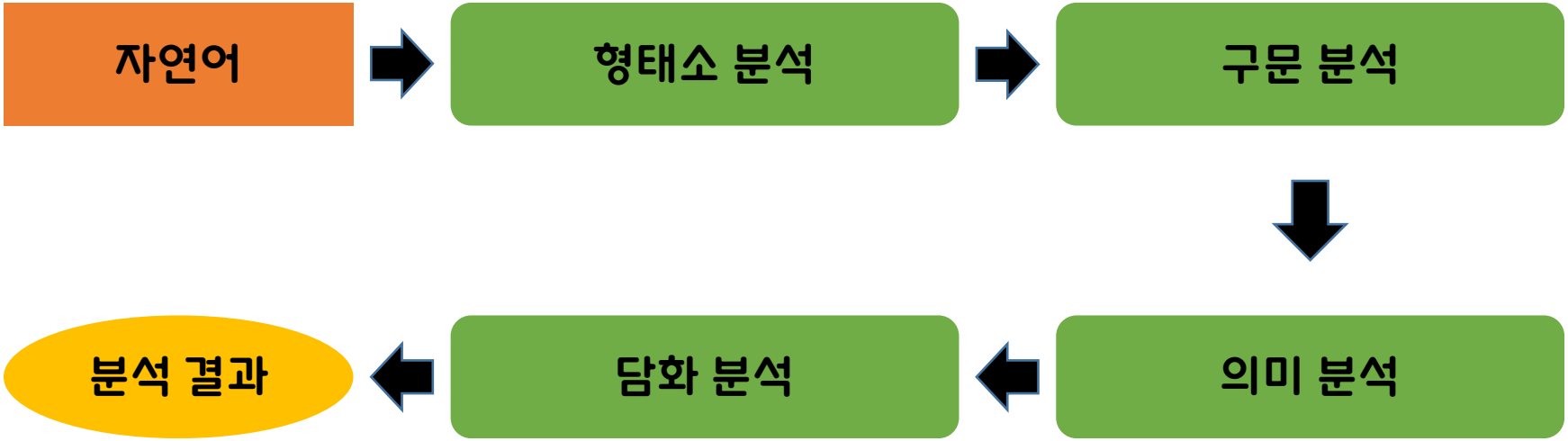
자연어를 분석하여 컴퓨터가 이해할 수 있는 형태로 만들거나
그러한 형태를 다시 인간이 이해할 수 있는 언어로 표현하는 제반 기술

자연어 처리



분석 단계

자연어 분석 단계



형태소 분석

- 형태소 단위 분석
- 사전 정보와 형태소 결합 정보 이용
- 언어마다 다른 난이도
- 검색 엔진

구문 분석

- 형태소들의 역할 분석
- 문장 성분 구별
- 문장의 구조를 파악

의미 분석

- 문장의 '의미'를 파악
- 중의성 해소
- 대용어 처리

담화 분석

- 문장 간의 연관관계 파악
- 전/후 문맥 정보 이용
- 심층적 의미 파악

자연어 분석

형태소 분석

구문 분석

의미 분석

담화 분석

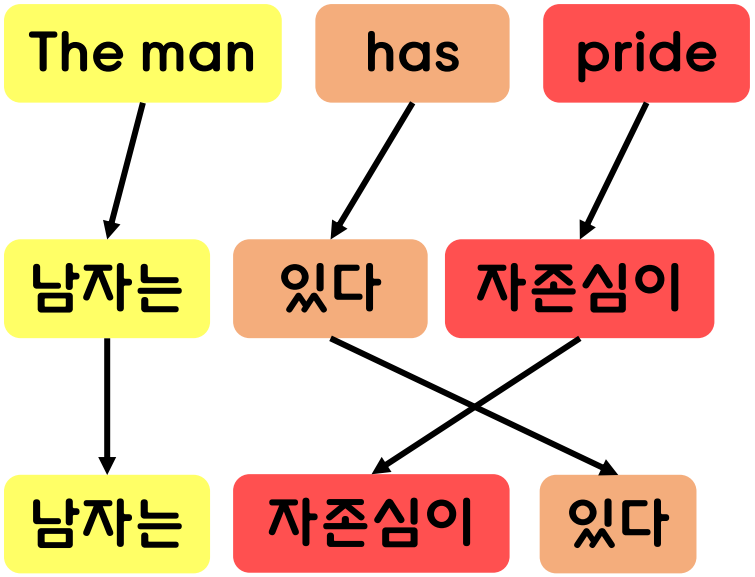
응용 기술

응용 기술

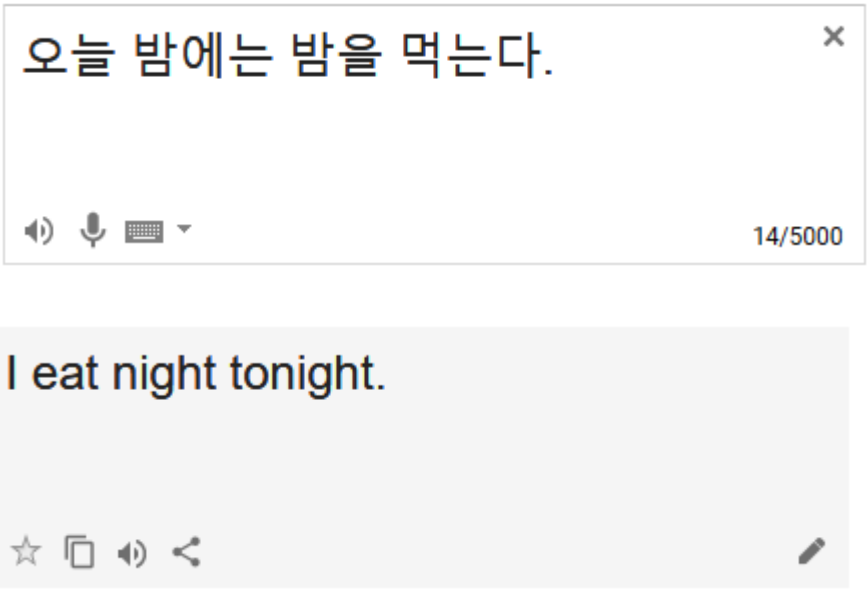
쉬움	중간	어려움
스펠링 체크	서류 형태 해석	기계번역
키워드 검사	구문해석	감정분석
유사어 감지		질의응답시스템

기계번역

문구기반 번역의 예



단어의 모호성



감정분석

문장으로부터 감정을 판단

긍정적 - 매우 재미있다. 아무리 놀아도 실증이 나지 않는다.

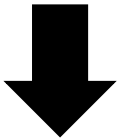
-> 0.86

부정적 - 설치하지마. 데이터만 낭비한다.

-> -0.68

자연어 처리의 어려움

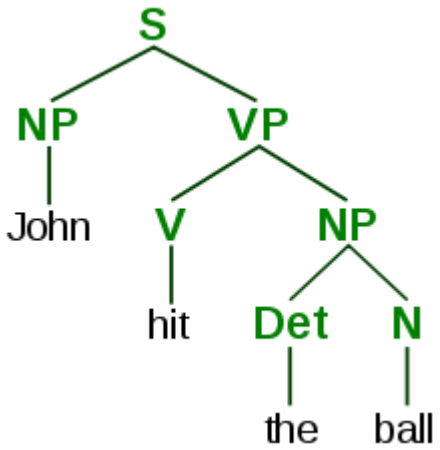
언어, 상황, 환경, 지각 지식의 학습 및 표현의 복잡함

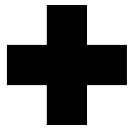


자연어 처리 연구는 오늘날에도 현재진행형

KoNLPy

형태소 분석기



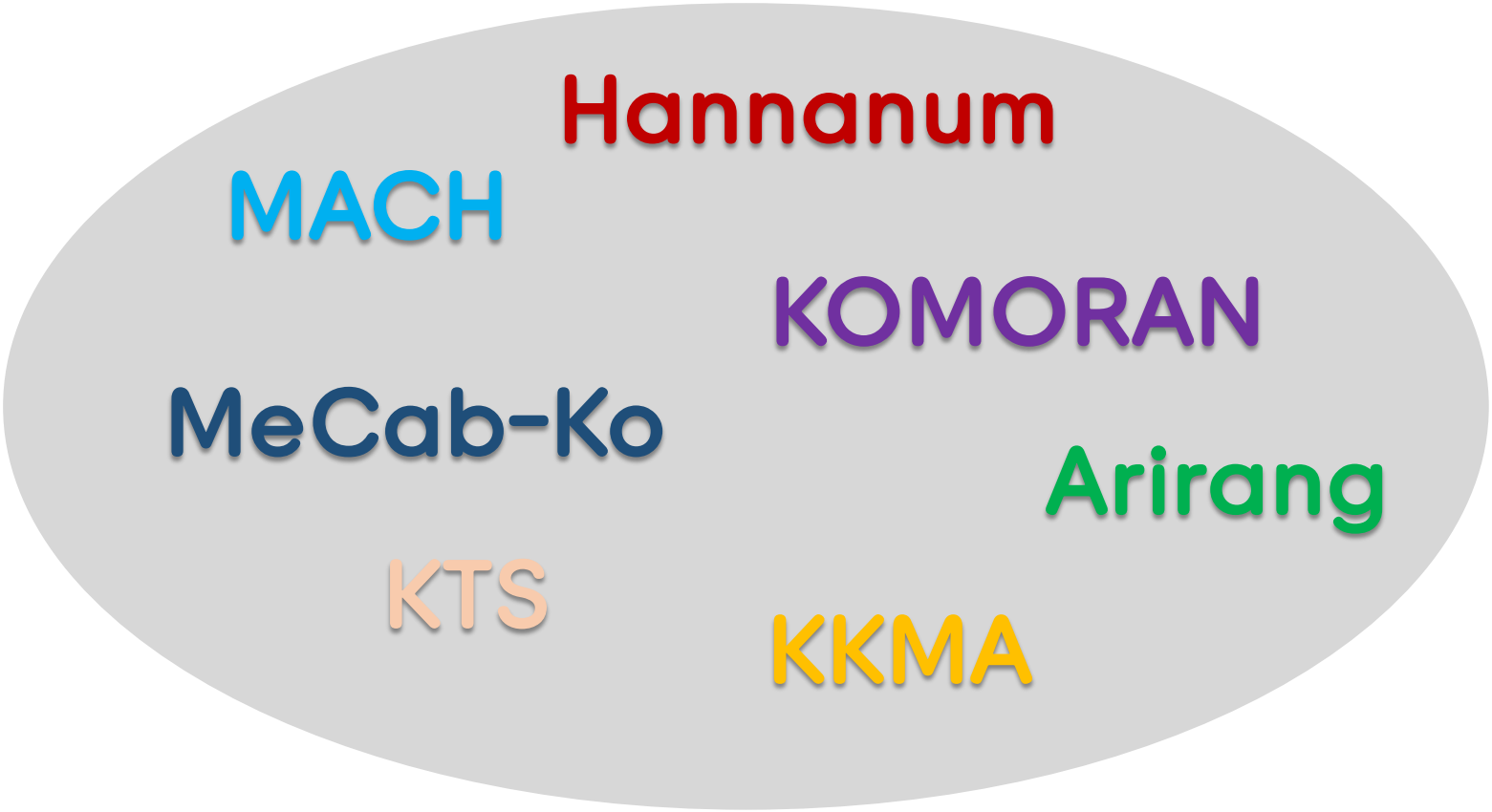


알고리즘

코퍼스

구조를 이룬 텍스트 집합
영역 규정

형태소 분석기



NLP package

KoNLP

Korean Natural Language Processing

NLTK

Natural Language Toolkit

KoNLP

But!

Open-source
Hannanum 형태소 분석기
세종계획(한국어 코퍼스, 정부사업)
NLP를 위한 각종 함수
Documentation

뭔가 아쉽...

NLTK

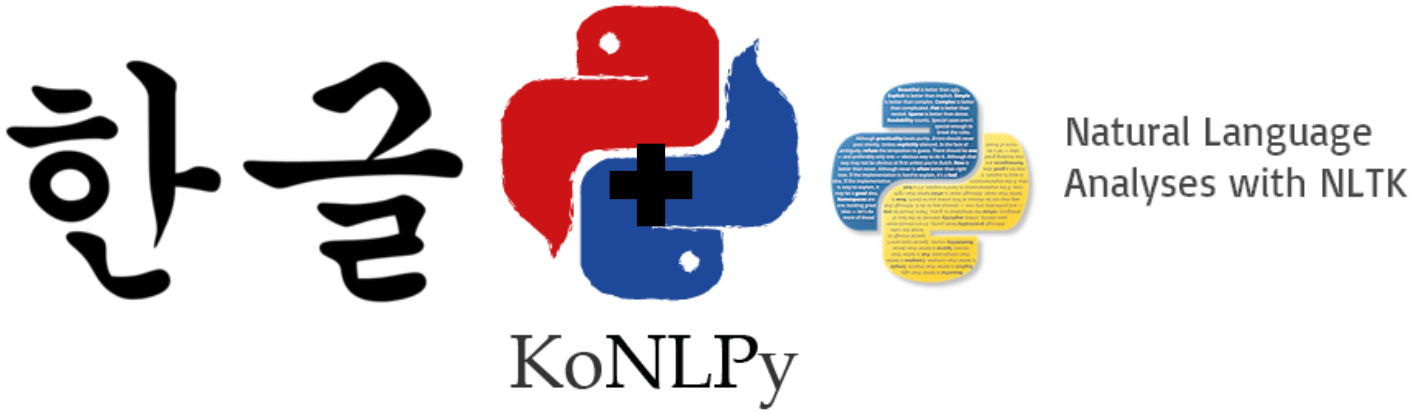
Open-source
Python
방대한 코퍼스(주로영어)
NLP를 위한 방대한 함수
language-free

But!

빈약한 한국어 지원

Natural Language
Analyses with NLTK

KoNLPy



Open-source
Python
많은 한글 코퍼스
많은 형태소 분석기
편리한 함수(ex. pprint)

KoNLPy

형태소 분석 모듈

Kkma, Hannanum,
Twitter, Komoran

```
From konlpy.tag
Import Kkma
```

기존 모듈

크롤링
웹프로그래밍
데이터 분석

```
Import urllib
```

Synergy

형태소 분석

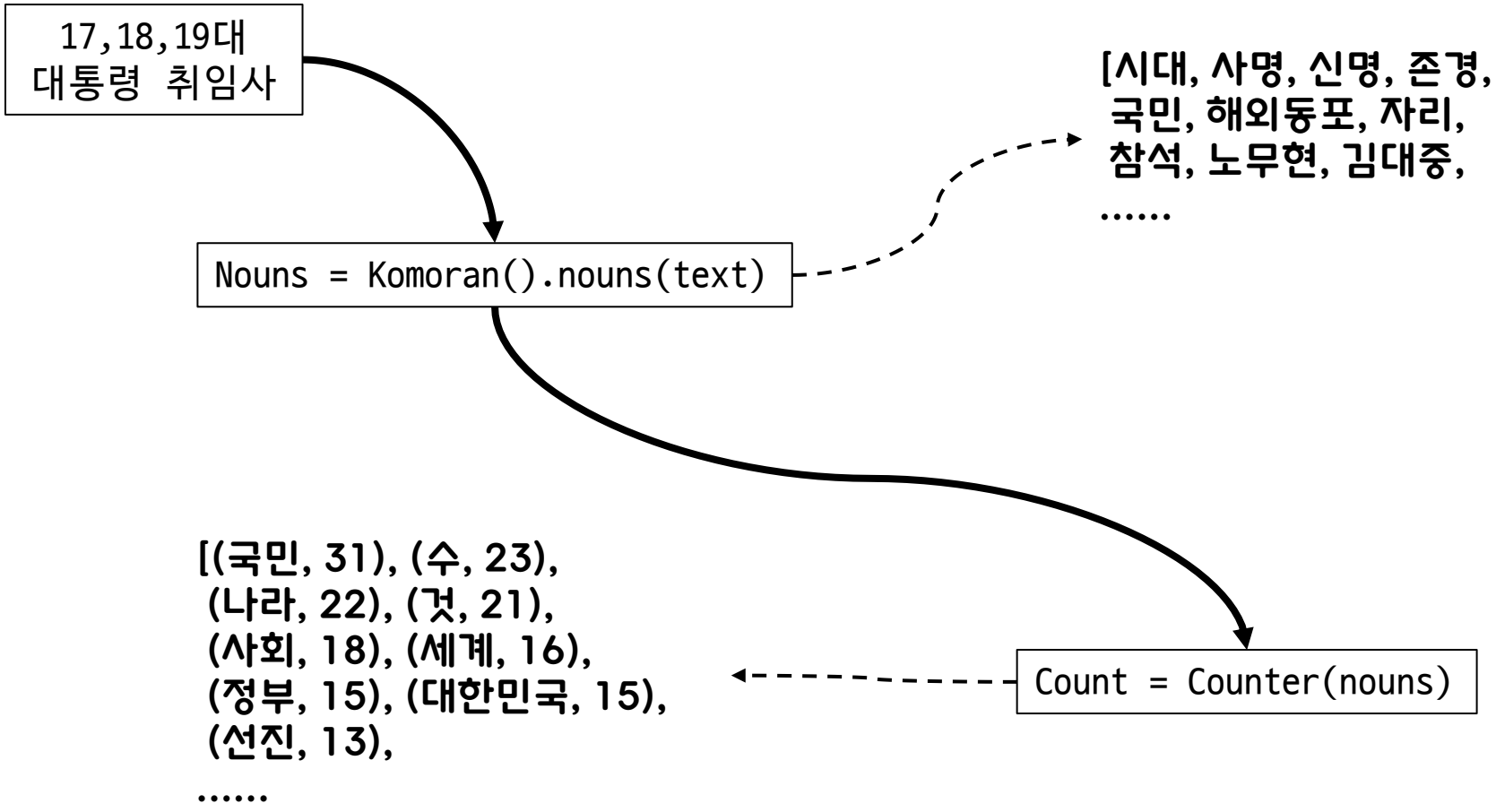
Kkma().pos('text')
이제 저는 여러분과 함께
자신감을 가지고
미래로 가는 길을
찾아 열어가고자 합니다.

분석결과

[(이제, MAG),
(저, NP),
(는, JX),
(여러분, NP),
(과, JKM),
(함께, MAG),
(자신감, NNG),
(을, JKO),
(가지, VV),
(고, ECE),
(미래, NNG),
(로, JKM),
(가, VV),
(는, ETD),
(길, NNG),
(을, JKO),
(찾, VV),
(아, ECD),
(열, VV),
(어, ECD),
(가, VV),
(고자, ECD),
(하, VV),
(ㅂ니다, EFN),
(., SF)]

Kkma (ntags=56)	
Tag	Description
NNG	보통명사
NNP	고유명사
NNB	일반 의존 명사
NNM	단위 의존 명사
NR	수사
NP	대명사
VV	동사
VA	형용사
VXV	보조 동사
VXA	보조 형용사
VCP	긍정 지정사, 서술격 조사 '이다'
VCN	부정 지정사, 형용사 '아니다'
MDN	수 관형사
MDT	일반 관형사
MAG	일반 부사
MAC	접속 부사
...	

주제 추출(WordCloud)



주제 추출(WordCloud)

Count와 pytagclound 패키지를 이용하여 워드 클라우드 생성

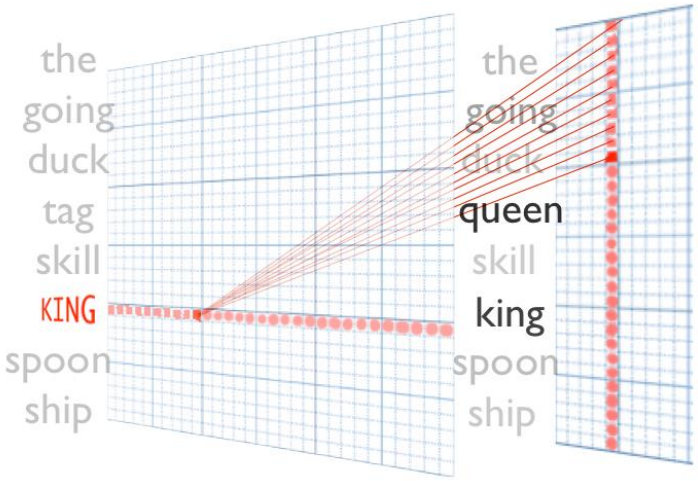
연관어 찾기

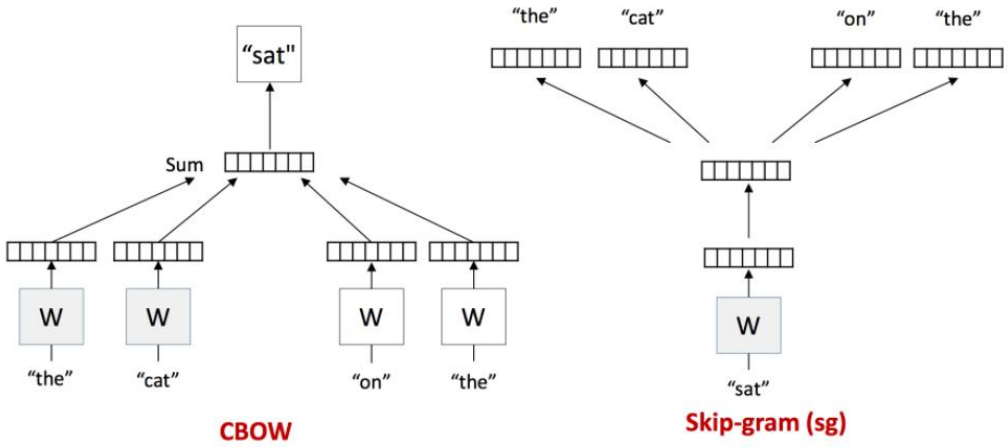
빈출 명사 != 글의 주제

언어의 의미적인 접근이 필요

연관된 단어 찾기(Word2Vec)

언어의 벡터화
근처에 있는 단어 = 연관된 단어





단어(=벡터)는 인공 신경망을 통한 학습가능

연관된 단어 찾기(Word2Vec)

gensim

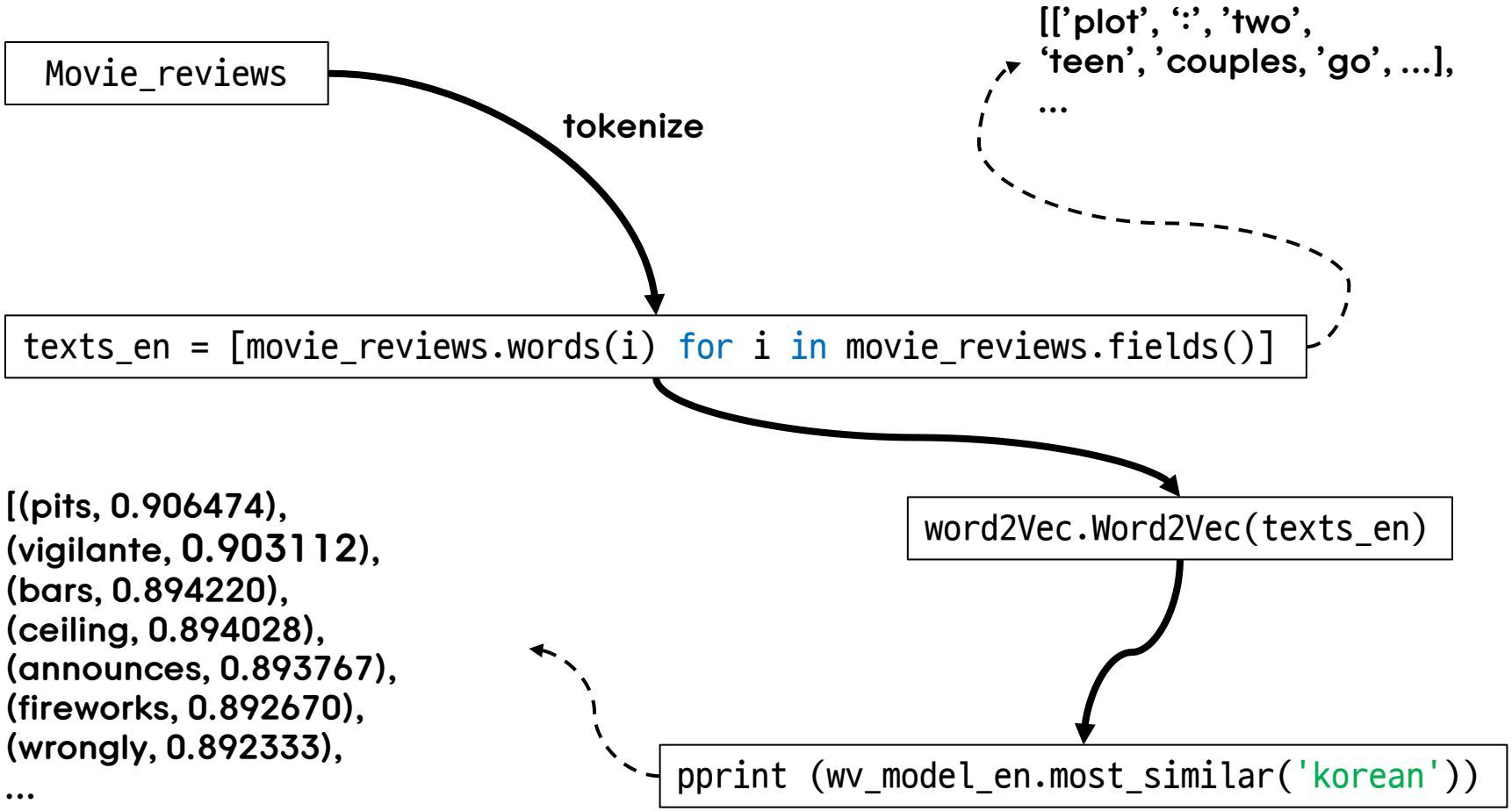
Word2Vec
(rather than numPy)

+

Natural Language
Analyses with NLTK

Corpus
(영어로 진행)

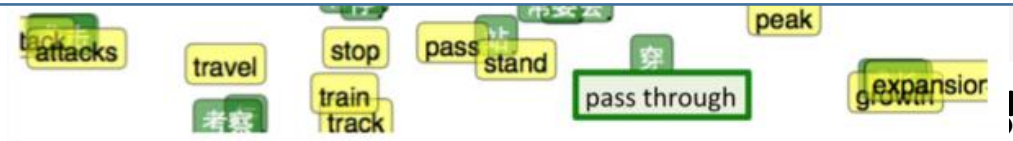
연관된 단어 찾기(Word2Vec)



NLP using Word2Vec

Word2Vec => Sentence2vec, Paragraph2Vec, Doc2Vec

정확도가 비약적으로 향상될 예정



기계 번역

태의 질의응답

QnA

Thank you