

Deep Learning for NLP

Deep Learning Theory and Software

NLP and Information Retrieval

Practices

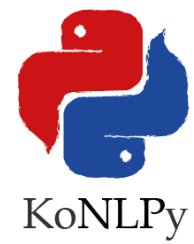
Practices

Prof. Joongheon Kim

School of Computer Science and Engineering, Chung-Ang University, Seoul, Korea

[https://sites.google.com/site/joongheonkim/
joongheon@gmail.com](https://sites.google.com/site/joongheonkim/joongheon@gmail.com)

- KoNLPy
 - <https://konlpy-ko.readthedocs.io/>
 - Python package that is for processing Korean language information



KoNLPy is a Python package for Korean natural language processing.

목차

- KoNLPy: 파이썬 한국어 NLP
 - 거인의 어깨 위에 서기
 - 라이선스
 - 참여하기
- 시작하기
- 사용하기
- API
- 인덱스와 표

Translations

English
한국어

Donate

Note:
You are not using the most up to date version of the library. [v0.5.1](#) is the newest version.

KoNLPy: 파이썬 한국어 NLP

build passing docs passing

KoNLPy("코엔엘파이"라고 읽습니다)는 한국어 정보처리를 위한 파이썬 패키지입니다. 설치 방법은 [이 곳을](#) 참고해주세요.

NLP를 처음 시작하시는 분들은 [시작하기](#)에서 가볍게 기본 지식을 습득할 수 있으며, KoNLPy의 사용법 가이드는 [사용하기](#), 각 모듈의 상세사항은 [API](#) 문서에서 보실 수 있습니다.

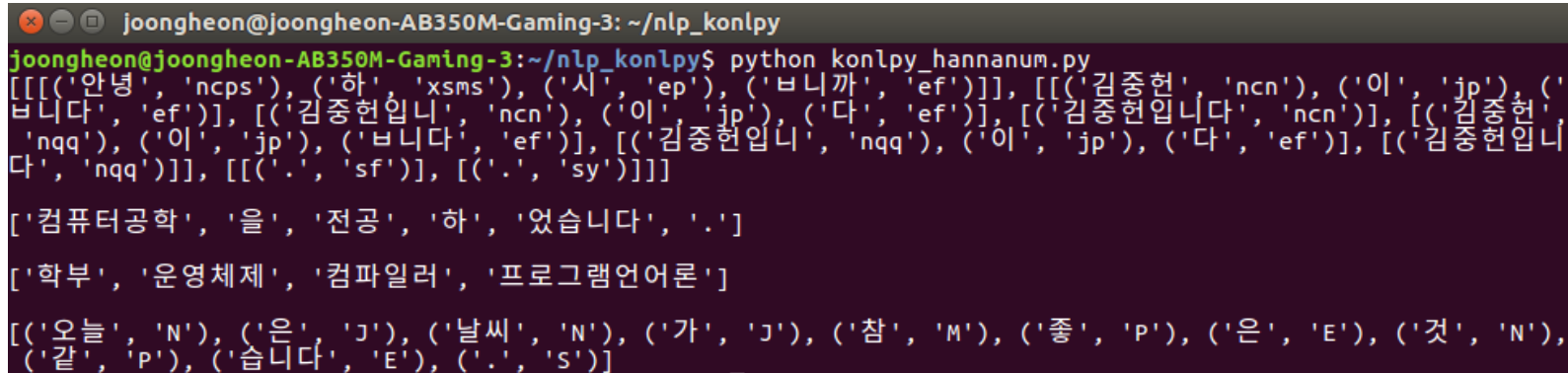
```
>>> from konlpy.tag import Kkma
>>> from konlpy.utils import pprint
>>> kkma = Kkma()
>>> pprint(kkma.sentences(u'네, 안녕하세요. 반갑습니다.'))
[네, 안녕하세요...,
반갑습니다.]
>>> pprint(kkma.nouns(u'질문이나 건의사항은 깃헙 이슈 트랙커에 남겨주세요.'))
[질문,
건의,
건의사항,
사항,
깃헙,
이슈,
트래커]
>>> pprint(kkma.pos(u'오류보고는 실행환경, 메러메세지와함께 설명을 최대한상세히!^^'))
[(오류, NNG),
(보고, NNG),
(는, JX),
(실행, NNG),
(환경, NNG),
```

- Installation Commands
 - **pip install konlpy**
 - **sudo apt-get install openjdk-8-jdk**

```

1  # -*- coding: utf-8 -*-
2  from konlpy.tag import Hannanum
3  hannanum = Hannanum()
4  print(hannanum.analyze('안녕하십니까 김중헌입니다.'))
5  print("")
6  print(hannanum.morphs('컴퓨터공학을 전공했습니다.'))
7  print("")
8  print(hannanum.nouns('학부에서 운영체제와 컴파일러 그리고 프로그래머언어를 가르치고 있습니다.'))
9  print("")
10 print(hannanum.pos('오늘은 날씨가 참 좋은 것 같습니다.'))

```



```

joongheon@joongheon-AB350M-Gaming-3: ~/nlp_konlpy
joongheon@joongheon-AB350M-Gaming-3:~/nlp_konlpy$ python konlpy_hannanum.py
[[[('안녕', 'ncps'), ('하', 'xsms'), ('시', 'ep'), ('버니까', 'ef')], [(('김중헌', 'ncn'), ('이', 'jp'), ('버니다', 'ef'))], [(('김중헌입니', 'ncn'), ('이', 'jp'), ('다', 'ef'))], [(('김중헌입니다', 'ncn'))], [(('김중헌', 'nqq'), ('이', 'jp'), ('버니다', 'ef'))], [(('김중헌입니', 'nqq'), ('이', 'jp'), ('다', 'ef'))], [(('김중헌입니다', 'nqq'))], [(('.', 'sf'))], [(('.', 'sy'))]]

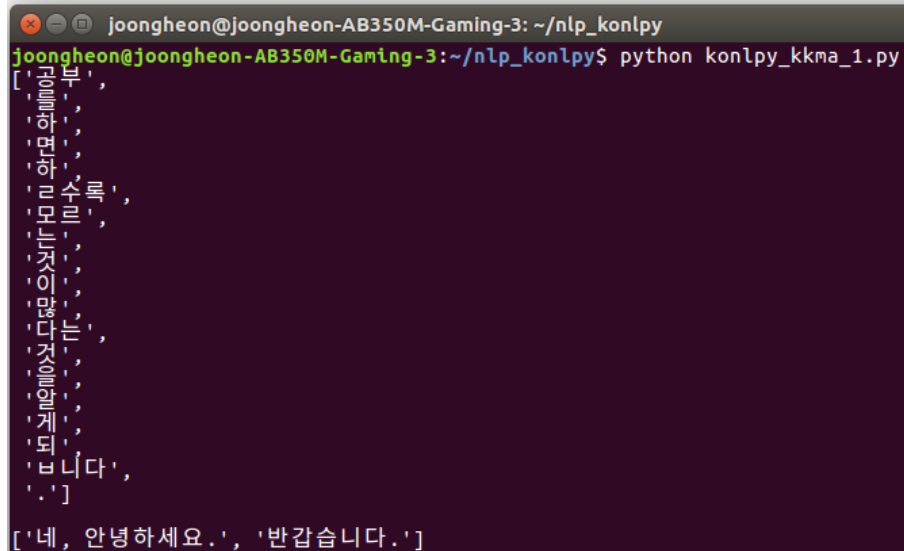
['컴퓨터공학', '을', '전공', '하', '었습니다', '.']

['학부', '운영체제', '컴파일러', '프로그래머언어']

[(('오늘', 'N'), ('은', 'J'), ('날씨', 'N'), ('가', 'J'), ('참', 'M'), ('좋은', 'P'), ('은', 'E'), ('것', 'N'), ('같', 'P'), ('습니다', 'E'), (('.', 'S'))]

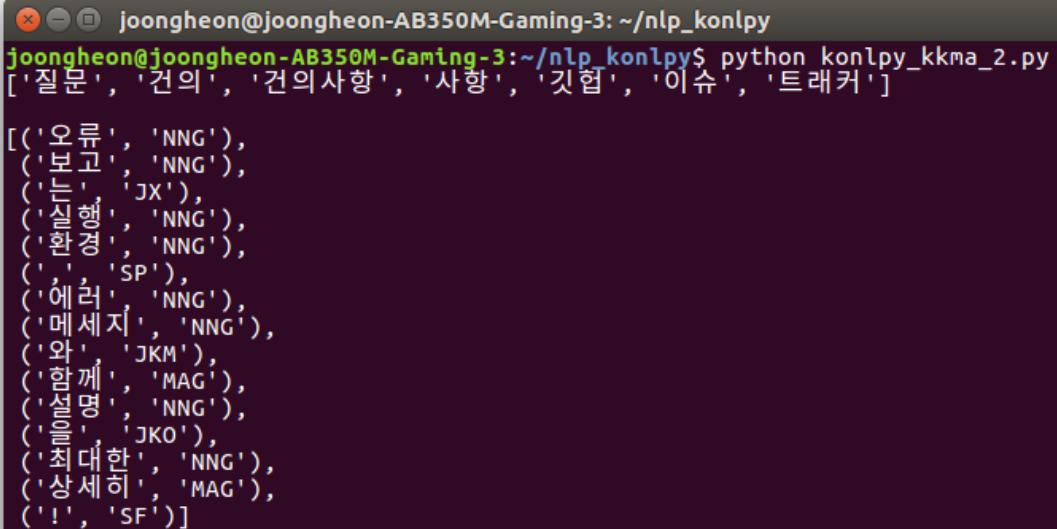
```

```
1 # -*- coding: utf-8 -*-
2 from konlpy.tag import Kkma
3 from konlpy.utils import pprint
4 kkma = Kkma()
5 pprint(kkma.morphs('공부를 하면할수록 모르는게 많다는 것을 알게 됩니다.'))
6 print("")
7 pprint(kkma.sentences('네, 안녕하세요. 반갑습니다.'))
8 print("")
```



```
joongheon@joongheon-AB350M-Gaming-3: ~/nlp_konlpy
joongheon@joongheon-AB350M-Gaming-3:~/nlp_konlpy$ python konlpy_kkma_1.py
['공', '부', '하', '면', '하', '수', '로', '크', '모', '르', '는', '것', '이', '많', '다', '는', '것', '을', '알', '게', '되', '니', '다', '.']
['네, 안녕하세요.', '반갑습니다.']
```

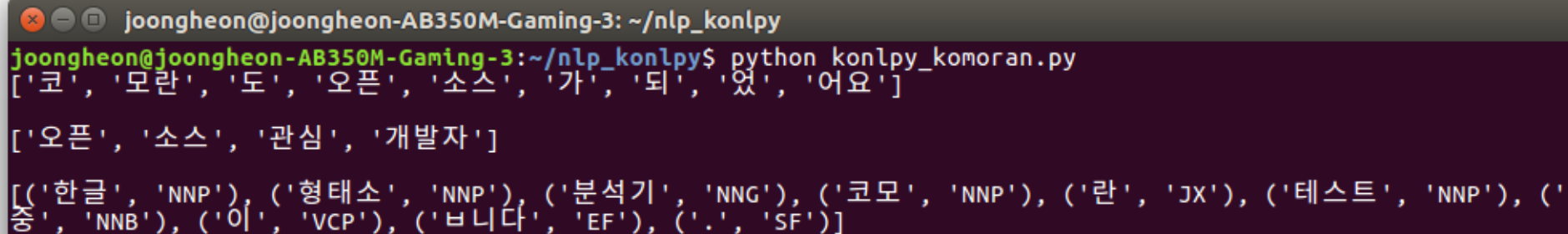
```
1 # -*- coding: utf-8 -*-
2 from konlpy.tag import Kkma
3 from konlpy.utils import pprint
4 kkma = Kkma()
5 pprint(kkma.nouns('질문이나 건의사항은 깃헙 이슈 트래커에 남겨주세요.'))
6 print("")
7 pprint(kkma.pos('오류보고는 실행환경, 에러메세지와함께 설명을 최대한상세히!'))
8 print("")
```



```
joongheon@joongheon-AB350M-Gaming-3: ~/nlp_konlpy
joongheon@joongheon-AB350M-Gaming-3:~/nlp_konlpy$ python konlpy_kkma_2.py
['질문', '건의', '건의사항', '사항', '깃헙', '이슈', '트래커']

[('오류', 'NNG'),
 ('보고', 'NNG'),
 ('는', 'JX'),
 ('실행', 'NNG'),
 ('환경', 'NNG'),
 ('', 'SP'),
 ('에러', 'NNG'),
 ('메세지', 'NNG'),
 ('와', 'JKM'),
 ('함께', 'MAG'),
 ('설명', 'NNG'),
 ('을', 'JKO'),
 ('최대한', 'NNG'),
 ('상세히', 'MAG'),
 ('!', 'SF')]
```

```
1 from konlpy.tag import Komoran
2 komoran = Komoran()
3 print(komoran.morphs('코모란도 오픈소스가 되었어요'))
4 print("")
5 print(komoran.nouns('오픈소스에 관심 많은 멋진 개발자님들!'))
6 print("")
7 print(komoran.pos('한글형태소분석기 코모란 테스트 중 입니다.'))
8 print("")
```



```
joongheon@joongheon-AB350M-Gaming-3: ~/nlp_konlpy
joongheon@joongheon-AB350M-Gaming-3:~/nlp_konlpy$ python konlpy_komoran.py
['코', '모란', '도', '오픈', '소스', '가', '되', '었', '어요']

['오픈', '소스', '관심', '개발자']

[('한글', 'NNP'), ('형태소', 'NNP'), ('분석기', 'NNG'), ('코모', 'NNP'), ('란', 'JX'), ('테스트', 'NNP'), ('중', 'NNB'), ('이', 'VCP'), ('버니다', 'EF'), ('.', 'SF')]
```

```

1 from konlpy.tag import Komoran
2 komoran = Komoran(userdic='lib/dic.txt')
3 print(komoran.morphs('코모란도 오픈소스가 되었어요'))
4 print("")
5 print(komoran.nouns('오픈소스에 관심 많은 멋진 개발자님들!'))
6 print("")
7 print(komoran.pos('혹시 바람과 함께 사라지다 봤어?'))
8 print("")

```

lib/dic.txt

1	코모란	NNP	
2	김중헌	NNP	
3	오픈소스	NNG	
4	바람과 함께 사라지다	NNP	

```

joongheon@joongheon-AB350M-Gaming-3: ~/nlp_konlpy
joongheon@joongheon-AB350M-Gaming-3:~/nlp_konlpy$ python konlpy_komoran2.py
['코모란', '도', '오픈소스', '가', '되', '었', '어요']

['오픈소스', '관심', '개발자']

[('혹시', 'MAG'), ('바람과 함께 사라지다', 'NNP'), ('보', 'VV'), ('았', 'EP'), ('어', 'EF'), ('?', 'SF')]

```



```
1 import sys
2 from bs4 import BeautifulSoup
3 from urllib.request import urlopen
4
5 original_stdout = sys.stdout
6 file = open('file_171023.txt', 'w', encoding='UTF-8')
7 sys.stdout = file
8
9 url = urlopen('https://www.boannews.com/media/view.asp?idx=57628&skind=0')
10 soup = BeautifulSoup(url, "lxml")
11
12 title = soup.find_all("div", {"id": "news_title02"})
13 contents = soup.find_all("div", {"id": "news_content"})
14
15 print(title)
16 print(contents)
17
18 sys.stdout = original_stdout
19 file.close()
```

```
파일(F) 편집(E) 보기(V) 검색(S) 도구(T) 문서(D) 도움말(H)
열기(O) [F] 저장(S)
1 [<div id="news_title02">[긴급] 제주항공 도메인 통해 ‘마이랜섬’ 랜섬웨어 유포</div>]
2 [<div id="news_content">
3 <b>제주항공 도메인 비롯한 국내 다수 웹사이트, 광고 배너 통해 마이랜섬 유포</b><br/><br/>[보안뉴스 김경애 기자] 국내 저가항공사인 제주항공 도메인(www.jejuair.co.kr)를 비롯해 국내 여러 웹사이트에서 한국을 타깃으로 제작된 ‘
마이랜섬(My Ransom)’ 랜섬웨어가 지속적으로 유포되고 있어 웹사이트 이용자들의 각별한 주의가 요구된다. <br/><br/><div class="news_image"><span class="img"></span><p class="txt">▲제주항공 도메인에서 마이랜섬 랜섬웨어가 유포된 정황 화면[이미지=보안뉴스 입수]</p></div><br/>본지가 취재한 바에 따르면 23일 오전 10시 8분경 제주항공 도메인에서
마이랜섬 랜섬웨어가 유포된 정황이 포착됐으며, 현재(오후 2시 44분)까지도 계속 유포되고 있는 것으로 분석됐다. 특히, 웹사이트 접속만으로 마이랜섬 랜섬웨어에 감염될 수 있는 드라이브 바이 다운로드 방식으로 유포되고 있어 추가 피
해가 예상되고 있다. <br/><br/>현재 유포되고 있는 마이랜섬 랜섬웨어는 광고 서버를 통해 유포되고 있는데, IP로 인증하기 때문에 다른 IP에서 접속할 경우 또 다시 감염되는 것으로 분석됐다. <br/><br/>위협정보대응 전문 서비스를 제
공하는 제로써트(ZeroCERT)에 따르면 홈페이지 광고프로그램을 통해 유포하는 멀버타이징 방식으로 마이랜섬 랜섬웨어가 유포되고 있으며, 광고 서버나 도메인은 바뀔 수 있기 때문에 제주항공 웹사이트의 광고배너가 유포에 악용된 것으로
분석했다.<br/><br/>현재 제주항공의 경우 www.jejuair.net을 대표 홈페이지로 사용하고 있지만, 본지 확인 결과 현재 호스팅 업체에 등록된 도메인 이름은 www.jejuair.co.kr이다. 등록인과 책임자 모두 제주항공으로 되어 있는 상태
다.<br/><br/>더군다나 호스팅 업체에서 확인되는 도메인 네임서버는 도메인 관리자 홈페이지에서 변경이 가능해 서버 해킹과 관리자 계정 탈취 가능성도 제기되고 있다. <br/><br/>이와 관련 제로써트 측은 “서버 해킹을 통해 랜섬웨어 유
포에 악용됐거나 도메인 관리자 계정 탈취로 연결될 수 있어 서버 보안뿐만 아니라 관리자 계정 관리도 중요하다”며, “이에 IP를 중간해서 가로챘는지 여부와 서버 해킹 및 관리자 계정 탈취 여부 등에 대한 철저한 조사가 필요하다”고 강조
했다. <br/><br/>본지는 제주항공 측의 입장을 듣기 위해 계속 통화를 시도하고 있지만 아직 연결되지 않고 있다. 이렇듯 제주항공 도메인 주소를 통해 랜섬웨어가 계속 유포되고 있는 만큼 웹사이트 이용자들은 백신과 소프트웨어를 최신
버전으로 업데이트해 유지하고, 악성코드가 제거되기 전까지 가급적 제주항공 도메인(www.jejuair.co.kr)에 접속하지 않는 것이 바람직하다. <br/><br/>한편, 마이랜섬 랜섬웨어는 케르베르(Cerber) 랜섬웨어의 변형으로 매그니튜드 익스
플로잇 킷을 통해 유포되는 것으로 알려졌는데, 이로 인해 파이어아이 등 해외 보안업체나 언론을 중심으로 매그니베르 또는 매그니버 랜섬웨어로 불리기도 한다. 그러나 본지에서는 국내에서 최초 발견한 순천향대 SCH사이버보안연구센터가
작성한 마이랜섬 랜섬웨어로 통일해 사용한다. <br/>[김경애 기자(<a href="mailto:boan3@boannews.com">boan3@boannews.com</a>)]<p align="center">www.boannews.com</p> 무단전재-재배포금지<br/></div>]
```

```
1 from newspaper import Article # pip install newspaper3k
2 url1 = 'http://v.media.daum.net/v/20170604205121164' #크롤링할 url 주소 입력
3 url2 = 'http://news.nate.com/view/20180810n20003?mid=n0412'
4 url3 = 'https://www.nytimes.com/2018/09/03/opinion/ukraine-corruption-h
5 a = Article(url2, language='ko') #언어가 한국어이므로 language='ko'로 설정
6 b = Article(url3, language='en') #언어가 한국어이므로 language='ko'로 설정
7 a.download()
8 a.parse()
9 print(a.title) #기사 제목 가져오기
10 print("")
11 print(a.text[:150]) #기사 내용 가져오기 (150자)
12 print("")
13 b.download()
14 b.parse()
15 print(b.title) #기사 제목 가져오기
16 print("")
17 print(b.text[:150]) #기사 내용 가져오기 (150자)
18 print("")
19 #print(a.text[:]) #기사 내용 가져오기
```

```
joongheon@joongheon-AB350M-Gaming-3: ~/nlp_konlpy
```

```
joongheon@joongheon-AB350M-Gaming-3:~/nlp_konlpy$ python konlpy_newspaper.py
```

```
[취재파일] 폭염, 실내에서 생활하는 젊고 건강한 사람은 괜찮을까?
```

확대 사진 보기

올여름 기록적인 폭염으로 발생한 온열질환자는 지난해(2017년) 여름 전체 환자 1천 574명보다 이미 2.2배나 많고 질병관리본부가 온열질환 응급실감시체계 운영을 시작한 2011년 이래 역대 최고치를 기록 중이다.

온열질환자를 연령별로 보면

Opinion | How Ukraine Is Fighting Corruption One Heart Stent at a Time

The procurement program saved the Ukrainian government \$3 million on stents alone – and \$222 million overall . And the savings, according to Health Min