KoNLPy 형태소 분석기

국민대학교 소프트웨어학부
강 승 식

KoNLPy 형태소 분석기

- 한나눔(Hannanum)
- 꼬꼬마(Kkma)
- 코모란(Komoran)
- Okt(Open Korea Text) → 구) Twitter 형태소 분석기
- 메캅(Mecab) → 윈도에서 지원 안함! 별도 설치 필요
- soynlp
- 카이(khaiii)

KoNLPy 설치 방법

https://konlpy-ko.readthedocs.io/ko/v0.4.3/install

윈도우

1. Java 1.7+이 설치되어 있나요?
2. JAVA_HOME 설정하기
3. JPype1 (>=0.5.7)을 다운로드 받고 설치. 다운 받은 *.whl* 파일을 설치하기 위해서는 *pip* 을 업그레이드 해야할 수 있습니다.

```
> pip install --upgrade pip
> pip install JPype1-0.5.7-cp27-none-win_amd64.whl
```

4. 명령 프롬프트로 KoNLPy 설치하기

```
> pip install konlpy
```

→ pip3 install konlpy

경고:

- KoNLPy의 `Mecab()` 클래스는 윈도우에서 지원되지 않습니다.

[1] 64비트 윈도우에는 *win-amd64*, 32비트 윈도우에는 *win32* 라고 쓰여진 파일을 사용합니다.

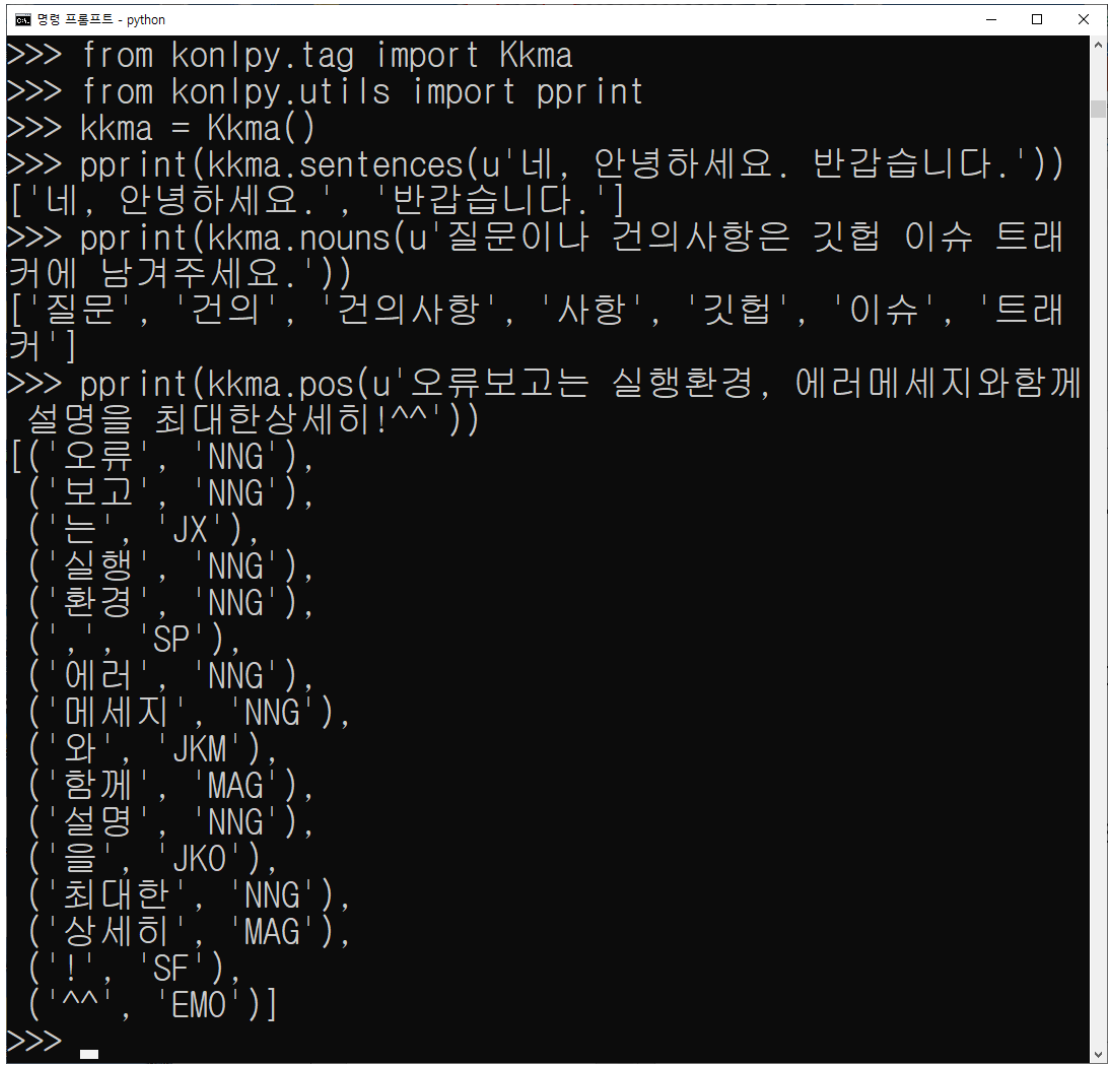
[2] MinGW/MSYS 또는 Cygwin을 설치한 후 압축을 푸실 수 있습니다. Git을 같이 사용하시는 경우 Git BASH 가 좋은 옵션일 수 있습니다. 또는 7zip 을 이용해 *tar* 파일의 압축을 푸실 수 있습니다.

실행 예제: Kkma

```
from konlpy.tag import Kkma
from konlpy.utils import pprint

kkma = Kkma()
pprint(kkma.sentences(u'네, 안녕하세요. 반갑습니다.'))
pprint(kkma.nouns(u'질문이나 건의사항은 깃헙 이슈 트래커에 남겨주세요.'))
pprint(kkma.pos(u'오류보고는 실행환경, 에러메세지와함께 설명을 최대한상세히!^^'))
```

API 문서 -- https://konlpy.org/ko/v0.5.2/#api



실행 예제: Komoran

```
from konlpy.tag import Komoran
komoran = Komoran(userdic='/tmp/dic.txt')

print(komoran.morphs(u'우왕 코모란도 오픈소스가 되었어요'))
print(komoran.nouns(u'오픈소스에 관심 많은 멋진 개발자님들!'))
print(komoran.pos(u'혹시 바람과 함께 사라지다 봤어?'))
```

실행 예제: Hannanum

```
from konlpy.tag import Hannanum
hannanum = Hannanum()

print(hannanum.analyze(u'롯데마트의 흑마늘 양념 치킨이 논란이 되고 있다.'))
print(hannanum.morphs(u'롯데마트의 흑마늘 양념 치킨이 논란이 되고 있다.'))
print(hannanum.nouns(u'다람쥐 헌 쳇바퀴에 타고파'))
print(hannanum.pos(u'웃으면 더 행복합니다!'))
```

실행 예제: Okt

```
from konlpy.tag import Okt
okt = Okt()

print(okt.morphs(u'단독입찰보다 복수입찰의 경우'))
print(okt.nouns(u'유일하게 항공기 체계 종합개발 경험을 갖고 있는 KAI는'))
print(okt.phrases(u'날카로운 분석과 신뢰감 있는 진행으로'))
print(okt.pos(u'이것도 되나욬ㅋㅋ'))
```

실행 예제: Mecab

```
from konlpy.tag import Mecab
mecab = Mecab()

print(mecab.morphs(u'영등포구청역에 있는 맛집 좀 알려주세요.'))
print(mecab.nouns(u'우리나라에는 무릎 치료를 잘하는 정형외과가 없는가!'))
print(mecab.pos(u'자연주의 쇼핑몰은 어떤 곳인가?'))
```

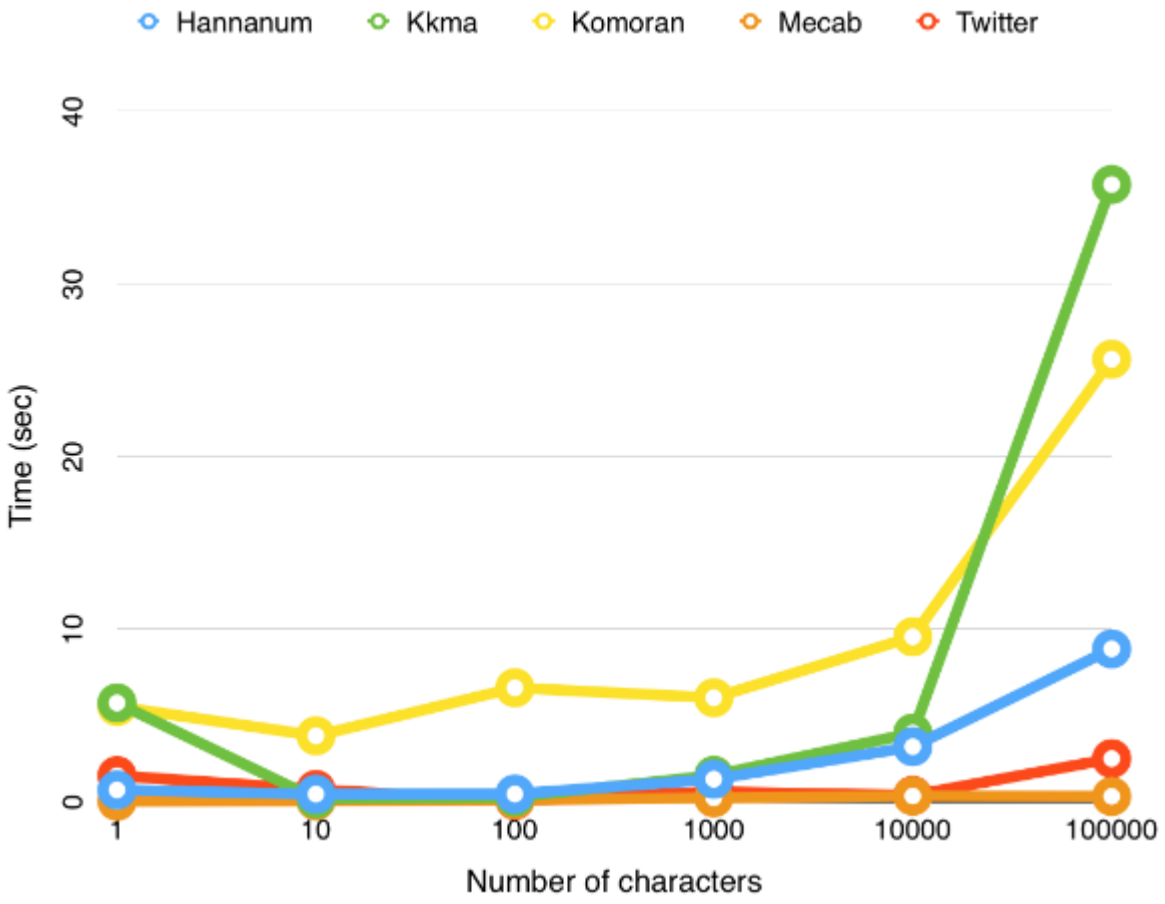
형태소 분석기: 실행시간 비교

1. 로딩 시간: 사전 로딩을 포함하여 클래스 로딩 시간

- Kkma: 5.6988 secs
- Komoran: 5.4866 secs
- Hannanum: 0.6591 secs
- Twitter: 1.4870 secs
- Mecab: 0.0007 secs

2. 실행시간: 10만 문자 문서의 pos 메소드 실행 시간

- Kkma: 35.7163 secs
- Komoran: 25.6008 secs
- Hannanum: 8.8251 secs
- Twitter: 2.4714 secs
- Mecab: 0.2838 secs



성능비교(파이썬 2.7버전) -- https://github.com/konlpy/konlpy/blob/master/docs/morph.py

한국어 품사 태그 비교표

https://docs.google.com/spreadsheets/d/1OGAjUvalBuX-oZvZ_-9tEfYD2gQe7hTGsgUpiiBSXI8/edit#gid=0

Sejong project (ntags=42)		Sim Gwangsub project (ntags=26)		Twitter Korean Text (ntags=19)		Komoran (ntags=42)		Mecab-ko (ntags=43)		Kkma (ntags=10)	Kkma (ntags=30)	Kkma (ntags=56)		Hannanum (ntags=9)		Hannanum (ntags=22)		Hannanum (ntags=26)		Hannanum (ntags=69)		Example
Tag	Description	Tag	Description	Tag	Description	Tag	Description	Tag	Description	Tag	Tag	Tag	Description	Tag	Description	Tag	Description	Tag	Description	Tag	Description	
																				NCPA	동작성 명사	
																		NCP	서술성명사	NCPS	상태성 명사	
																				NCN	비서술성 명사	어린이집
NNG	일반 명사					NNG	일반 명사	NNG	일반 명사			NNG	보통명사			NC	보통명사	NCN	비서술성명사	NCR	비서술성-직위 명사	
																				NQPA	성	손
																				NQPB	이름	예진
																				NQPC	성+이름	손예진
NNP	고유 명사	NN	명사			NNP	고유 명사	NNP	고유 명사			NNP	고유명사			NQ	고유명사	NQ	고유명사	NQQ	기타-일반	대한민국
																				NBN	비단위성 의존명사	때문
		NX	의존 명사					NNB	의존 명사			NNB	일반 의존 명사							NBS	비단위성 의존명사 --하다 붙는 것	
NNB	의존 명사	UM	단위 명사			NNB	의존 명사	NNBC	단위를 나타내는 명사		NN	NNM	단위 의존 명사			NB	의존명사	NB	의존명사	NBU	단위성 의존명사	퍼센트
																				NNC	양수사	백만
NR	수사	NU	수사			NR	수사	NR	수사		NR	NR	수사			NN	수사	NN	수사	NNO	서수사	첫째
																				NPP	인칭대명사	자기
NP	대명사	NP	대명사	Noun	명사 (Nouns, Pronouns, Compan	NP	대명사	NP	대명사	N	NP	NP	대명사	N	체언	NP	대명사	NP	대명사	NPD	지시대명사	
																				PVD	지시 동사	
VV	동사	VV	동사	Verb	동사	VV	동사	VV	동사		VV	VV	동사			PV	동사	PV	동사	PVG	일반 동사	
																				PAD	지시 형용사	
VA	형용사	VA	형용사	Adjective	형용사	VA	형용사	VA	형용사		VA	VA	형용사			PA	형용사	PA	형용사	PAA	성상 형용사	
		VX	보조 동사									VXV	보조 동사									아니하
VX	보조 용언	AX	보조 형용사			VX	보조 용언	VX	보조 용언		VX	VXA	보조 형용사			PX	보조 용언	PX	보조 용언	PX	보조 용언	
VCP	긍정 지정사	CP	서술격 조사 '이다'			VCP	긍정 지정사	VCP	긍정 지정사			VCP	긍정 지정사, 서술격 조사 '이									
VCN	부정 지정사					VCN	부정 지정사	VCN	부정 지정사	V	VC	VCN	부정 지정사, 형용사 '아니다	P	용언							아니
		DN	수 관형사									MDN	수 관형사									
																				MMD	지시 관형사	다른 (분들은)
MM	관형사	DT	일반 관형사	Determiner	관형사 (ex: 새, 헌, 참, 첫, 이, 그, 7	MM	관형사	MM	관형사		MD	MDT	일반 관형사			MM	관형사	MM	관형사	MMA	성상 관형사	
MAG	일반 부사			Adverb	부사 (ex: 잘, 매우, 빨리, 반드시, 과	MAG	일반 부사	MAG	일반 부사			MAG	일반 부사							MAG	일반 부사	전부, 직접, 미리
MAJ	접속 부사			Conjunction	접속사	MAJ	접속 부사	MAJ	접속 부사		MA	MAC	접속 부사							MAJ	접속 부사	또는, 다만
		AD	부사							M				M	수식언	MA	부사	MA	부사	MAD	지시 부사	
IC	감탄사	EX	감탄사	Exclamation	감탄사 (ex: 헐, 어머나, 얼씨구)	IC	감탄사	IC	감탄사	I	IC	IC	감탄사	I	독립언	II	감탄사	II	감탄사	II	감탄사	어머나
JKS	주격 조사					JKS	주격 조사	JKS	주격 조사			JKS	주격 조사							JCS	주격 조사	
JKC	보격 조사					JKC	보격 조사	JKC	보격 조사			JKC	보격 조사							JCC	보격 조사	
JKG	관형격 조사					JKG	관형격 조사	JKG	관형격 조사			JKG	관형격 조사							JCM	관형격 조사	
JKO	목적격 조사					JKO	목적격 조사	JKO	목적격 조사			JKO	목적격 조사							JCO	목적격 조사	
JKB	부사격 조사					JKB	부사격 조사	JKB	부사격 조사			JKM	부사격 조사							JCA	부사격 조사	으로, 에서, 에게
JKV	호격 조사					JKV	호격 조사	JKV	호격 조사			JKI	호격 조사							JCV	호격 조사	
JKQ	인용격 조사					JKQ	인용격 조사	JKQ	인용격 조사		JK	JKQ	인용격 조사							JCR	인용격 조사	
JC	접속 조사					JC	접속 조사	JC	접속 조사		JC	JC	접속 조사							JCJ	접속격 조사	이나
																JC	격조사	JC	격조사	JGT	공동격 조사	
																				JXC	통용 보조사	부터, 까지
JX	보조사					JX	보조사	JX	보조사		JX	JX	보조사			JX	보조사	JX	보조사	JXF	종결 보조사	(종습니다) 그려.
		JO	조사	Josa	조사 (ex: 의, 에, 에서)					J				J	관계언	JP	서술격 조사	JP	서술격 조사	JP	서술격 조사	
												EPH	존칭 선어말 어미									

말뭉치

- https://konlpy.org/ko/v0.5.2/data/

1. kolaw: 한국 법률 말뭉치: constitution.txt
2. kobill: 대한민국 국회 의안 말뭉치: 1809890.txt, 1809899.txt

```
from konlpy.corpus import kolaw
c = kolaw.open('constitution.txt').read()
print(c[:10])

from konlpy.corpus import kobill
d = kobill.open('1809890.txt').read()
print(d[:15])
```

형태소 분석 사전

Hannanum 시스템 사전

KAIST 말뭉치를 이용해 생성된 사전. (4.7MB)

`./konlpy/java/data/kE/dic_system.txt` 에 위치해있으며, 아래에서 파일의 일부를 보실 수 있습니다.:

```
...
나라경제        ncn
나라기획        nqq
나라기획회장    ncn
나라꽃  ncn
나라님  ncn
나라도둑        ncn
나라따르        pvg
나라링링프로덕션        ncn
나라말  ncn
나라망신        ncn
나라박물관      ncn
나라발전        ncpa
나라별  ncn
나라부동산      nqq
나라사랑        ncn
나라살림        ncpa
나라시  nqq
나라시마        ncn
...
```

사용자 사전에 새로운 항목을 추가하기 위해서
는 `./konlpy/java/data/kE/dic_user.txt` 를 수정하시면 됩니다.

Kkma 시스템 사전

세종 말뭉치를 이용해 생성된 사전. (32MB)

꼬꼬마 형태소 분석기의 `.jar` 파일 안에 위치해 있습니다. 사전 파일을 직접 보기 위해서는 꼬꼬마 미러 를 확인해보시기 바랍니다. `kcc.dic` 는 다음과 같은 형태를 가집니다.:

```
아니/IC
후우/IC
그래서/MAC
그러나/MAC
그러니까/MAC
그러면/MAC
그러므로/MAC
그런데/MAC
그리고/MAC
따라서/MAC
하지만/MAC
...
```

Mecab 시스템 사전

세종 말뭉치로 만들어진 CSV 형태의 사전. (346MB)

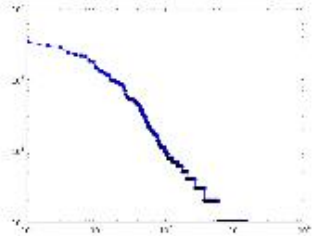
컴파일 된 사전은 `/usr/local/lib/mecab/dic/mecab-ko-dic` (또는 MeCab 설치시 지정한 경로)에 있으며, 원본 사전은 소스코드 에서 확인하실 수 있습니다.
`CoinedWord.csv` 파일의 일부를 아래에서 보실 수 있습니다.:

```
가오티,0,0,0,NNG,*,F,가오티,*,*,*,*,*
갑툭튀,0,0,0,NNG,*,F,갑툭튀,*,*,*,*,*
강퇴,0,0,0,NNG,*,F,강퇴,*,*,*,*,*
개드립,0,0,0,NNG,*,T,개드립,*,*,*,*,*
갠소,0,0,0,NNG,*,F,갠소,*,*,*,*,*
고퀄,0,0,0,NNG,*,T,고퀄,*,*,*,*,*
광삭,0,0,0,NNG,*,T,광삭,*,*,*,*,*
광탈,0,0,0,NNG,*,T,광탈,*,*,*,*,*
굉천,0,0,0,NNG,*,T,굉천,*,*,*,*,*
국을,0,0,0,NNG,*,T,국을,*,*,*,*,*
귀요미,0,0,0,NNG,*,F,귀요미,*,*,*,*,*
...
```

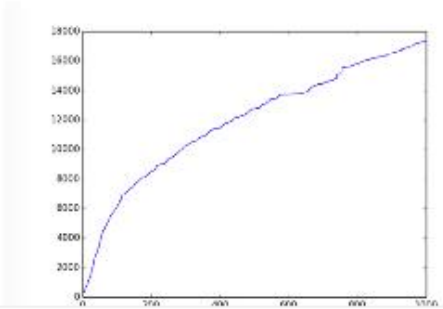
사용 예시

https://konlpy.org/ko/v0.5.2/examples/

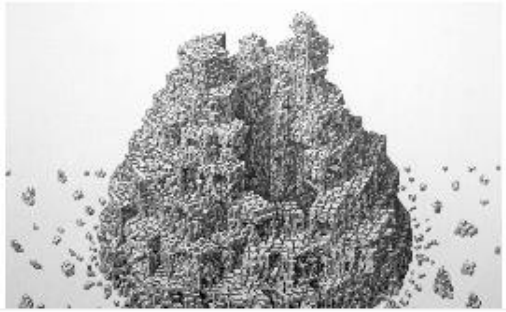
문서 탐색하기



말뭉치 탐색하기

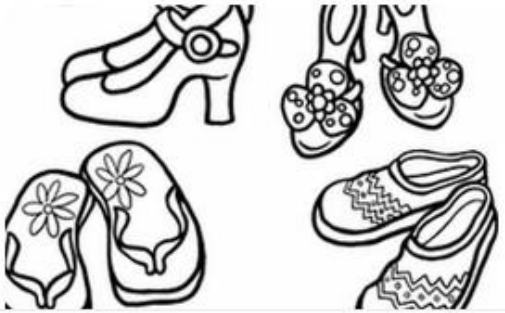


랜덤 텍스트 생성하기



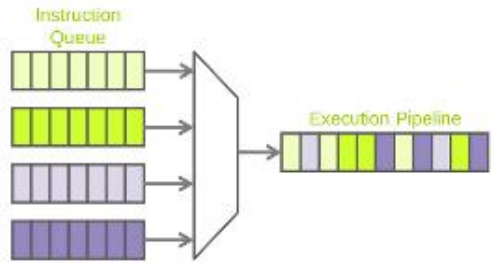
워드클라우드 그리기

연어(collocation) 찾기



구문 분석

KoNLPy를 이용한 멀티쓰레딩



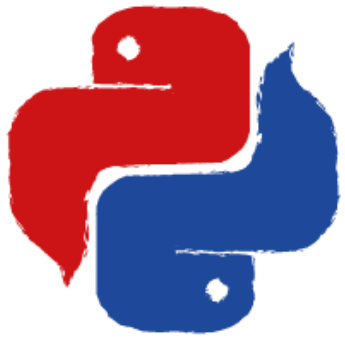
- 문서 탐색하기 -- https://konlpy.org/ko/v0.5.2/examples/explore/
- 연어(collocation) 찾기 -- https://konlpy.org/ko/v0.5.2/examples/collocations/
- 구문 분석 -- https://konlpy.org/ko/v0.5.2/examples/chunking/#chunking
- 랜덤 텍스트 생성하기 -- https://konlpy.org/ko/v0.5.2/examples/generate/

- 워드클라우드 그리기 -- https://konlpy.org/ko/v0.5.2/examples/wordcloud/

- KoNLPy를 이용한 멀티쓰레딩 -- https://konlpy.org/ko/v0.5.2/examples/multithreading/
- 말뭉치 탐색하기 -- https://konlpy.org/ko/v0.5.2/examples/corpus/

참고문헌

https://konlpy.org/ko/v0.5.2/references/



KoNLPy

Table of Contents

참고문헌

주석:

Please modify this document if anything is erroneous or not included. Last updated at 2019년 12월 03일.

한국어 형태소 분석기

한국어 텍스트를 분석할 때 가장 기본적으로 행해야하는 것은 형태소 분석입니다. 이를 위해 다양한 프로그래밍 언어로 된 여러 라이브러리가 있습니다:

C/C++

- MeCab-ko (2013) - By Yong-woon Lee and Youngho Yoo GPL LGPL BSD
- UTagger (2012) - By Joon-Choul Shin, Cheol-Young Ock* (Ulsan University) GPL custom
  - 신준철, 옥철영, 기분석 부분 어절 사전을 활용한 한국어 형태소 분석기 (A Korean Morphological Analyzer using a Pre-analyzed Partial Word-phrase Dictionary), 정보과학회논문지: 소프트웨어 및 응용, 제39권 제5호, 2012.

KoNLTK(또는 konlp) 형태소 분석기

- https://konltk.github.io/

C> pip install konlp
C> python

```
from konlp.kma.klt2000 import klt2000
klt = klt2000()
simple_txt = "자연어처리와 인공지능 수업에서는 한국어 텍스트 처리 기법을 배웁니다."
klt.morphs(simple_txt)
klt.nouns(simple_txt)
klt.pos(simple_txt)
klt.sent_tokenize(simple_txt)
```