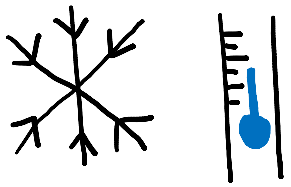
파이썬 자연어 처리

이성주
seongjoo@codebasic.io



Day 2

```
def func(...)
    for x in [    ]:
        if/elif/else
    return (a, b)
```

Plain Text
<XML>

말뭉치
Corpus

$ jupyter notebook

In[ ]:

Kernel
S
python

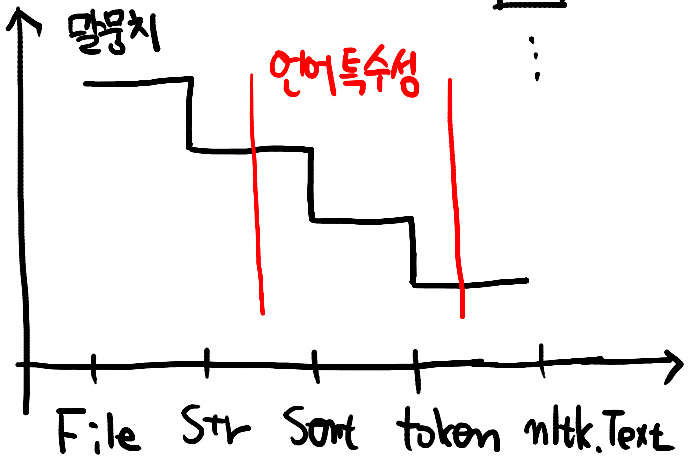
Pandas

I	V
D	A
X	L
	S

	Columns
I	
D	VALS
X	

Series DataFrame

- .value_counts()
- .sort_values()
- .drop_duplicates()



NLTK.org

.similar()
.collocation()

jupyter Day 2

Logout

Trusted | Python 3

File Edit View Insert Cell Kernel Widgets Help

Code

In [1]:
```
from pandas import Series, DataFrame
import pandas as pd

%matplotlib inline
```

nltk_data
└ corpora ~ 말뭉치들
  └ PlaintextCorpusReader └ gutenberg

In [2]:
```
from nltk.corpus import gutenberg
```

.fileids() { └ austen-emma.txt ⋮ }

In [3]:
```
gutenberg
```

Out[3]:
```
<PlaintextCorpusReader in 'C:\Users\student\AppData\Roaming\nltk_data\corpora\gutenberg'>
```

In [4]:
```
gutenberg.fileids()
```

Out[4]:
```
['austen-emma.txt',
 'austen-persuasion.txt',
 'austen-sense.txt',
 'bible-kjv.txt',
 'blake-poems.txt',
 'bryant-stories.txt',
```

jupyter Day 2

Logout

Trusted | Python 3

File Edit View Insert Cell Kernel Widgets Help

Code

CorpusReader

In [5]:
```
raw_text = gutenberg.raw('austen-emma.txt')
```

In [6]:
```
raw_text[:100]
```

"106자" 윗줄 open( ).read( ) → str

Out[6]: str
```
'[Emma by Jane Austen 1816]₩n₩nVOLUME I₩n₩nCHAPTER I₩n₩n₩nEmma Wood
house, handsome, clever, and rich, with a'
```

CorpusReader

In [7]:
```
tokens = gutenberg.words('austen-emma.txt')
```

sents → words

In [9]:
```
tokens[:10]
```

문장개념

Out[9]:
```
['[', 'Emma', 'by', 'Jane', 'Austen', '1816', ']', 'VOLUME', 'I',
'CHAPTER']
```

In [ ]:

Trusted | Python 3

File Edit View Insert Cell Kernel Widgets Help

STD LIB

In [10]:
```
import re
```
Regular Expression "정규표현식" ~ 문자열 패턴

In [18]:
```
re.split('\W+', raw_text)[:10]
```
\W+ "영문자, 숫자 한 글자 이상"

Out[18]:
```
['', 'Emma', 'by', 'Jane', 'Austen', '1816', 'VOLUME', 'I', 'CHAPT
ER', 'I']
```

CorpusReader.words(fileids)

In [13]:
```
word_tokens = nltk.tokenize.word_tokenize(raw_text)
```
← 언어 특수성

영어

In [14]:
```
word_tokens[:10]
```

Out[14]:
```
['[', 'Emma', 'by', 'Jane', 'Austen', '1816', ']', 'VOLUME', 'I',
'CHAPTER']
```

CorpusReader.sents()

In [15]:
```
sent_tokens = nltk.tokenize.sent_tokenize(raw_text)
```
영어 ~ 한국어

In [17]:
```
sent_tokens[:3]
```

Out[17]:
```
['[Emma by Jane Austen 1816]₩n₩nVOLUME I₩n₩nCHAPTER I₩n₩n₩nEmma Wo
odhouse, handsome, clever, and rich, with a comfortable home₩nand
happy disposition, seemed to unite some of the best blessings₩nof
existence; and had lived nearly twenty-one years in the world₩nwit
h very little to distress or vex her.',
 "She was the youngest of the two daughters of a most affectionat
```

Trusted | Python 3

File Edit View Insert Cell Kernel Widgets Help

In [20]:
```
def 어휘통계산출(raw_text):
    word_tokens = nltk.tokenize.word_tokenize(raw_text)
    sent_tokens = nltk.tokenize.sent_tokenize(raw_text)
    어휘_토큰 = Series(
        t for t in word_tokens).str.lower().drop_duplicates()
    글자수 = len(raw_text)
    단어수 = len(word_tokens)
    문장수 = len(sent_tokens)
    어휘수 = len(어휘_토큰)
    return Series({
        '글자': 글자수,
        '단어': 단어수,
        '문장': 문장수,
        '어휘': 어휘수
    })
```

In [21]:
```
어휘통계산출(raw_text)
```

Out[21]:
```
글자    887071
단어    191673
문장      7493
어휘      8000
dtype: int64
```

In [ ]:

Trusted | Python 3

File Edit View Insert Cell Kernel Widgets Help

In [26]:
```
집계결과 = []
for fid in gutenberg.fileids():
    통계 = 어휘통계산출(gutenberg.raw(fid))
    통계.name = fid
    집계결과.append(통계)
```

In [27]:
```
pd.concat(집계결과, axis=1).T
```

Out[27]:

fid	글자	단어	문장	어휘
austen-emma.txt	887071	191673	7493	8000
austen-persuasion.txt	466292	97888	3654	5967
austen-sense.txt	673022	141367	4833	6681
bible-kjv.txt	4332554	946812	29812	17188
blake-poems.txt	38153	8239	355	1534
bryant-stories.txt	249439	55621	2715	4011
burgess-busterbrown.txt	84663	18542	1001	1553
carroll-alice.txt	144395	33310	1625	2838
chesterton-ball.txt	457450	97203	4624	8485

Trusted | Python 3

File Edit View Insert Cell Kernel Widgets Help

In [28]:
```
말뭉치_통계 = pd.concat(집계결과, axis=1).T
```

오름차순 = "아님"

In [31]:
```
말뭉치_통계.sort_values(by='어휘', ascending=False)
```

Out[31]:

	글자	단어	문장	어휘
melville-moby_dick.txt	1242990	254989	9852	18717
bible-kjv.txt	4332554	946812	29812	17188
whitman-leaves.txt	711215	149198	3827	13739
milton-paradise.txt	468220	95709	1835	9282
edgeworth-parents.txt	935158	209090	10096	8818
chesterton-ball.txt	457450	97203	4624	8485
chesterton-brown.txt	406629	85412	3712	8058
austen-emma.txt	887071	191673	7493	8000
austen-sense.txt	673022	141367	4833	6681
chesterton-thursday.txt	320525	69408	3588	6491
austen-persuasion.txt	466292	97888	3654	5967
shakespeare-hamlet.txt	162881	36326	2355	4812

✓

내림차순

Trusted | Python 3

File Edit View Insert Cell Kernel Widgets Help

In [40]:
```
어휘다양성.sort_values().plot(kind='barh')
```

NLTK

CorpusReader

XMLCorpusReader

NPSChatCorpusReader

```
...
```

In [41]:
```
from nltk.corpus import nps_chat
```

In [42]:
```
nps_chat
```

Out[42]:
```
<NPSChatCorpusReader in 'C:₩₩Users₩₩student₩₩AppData₩₩Roaming₩₩nltk_data₩₩corpora₩₩nps_chat'>
```

<xml>

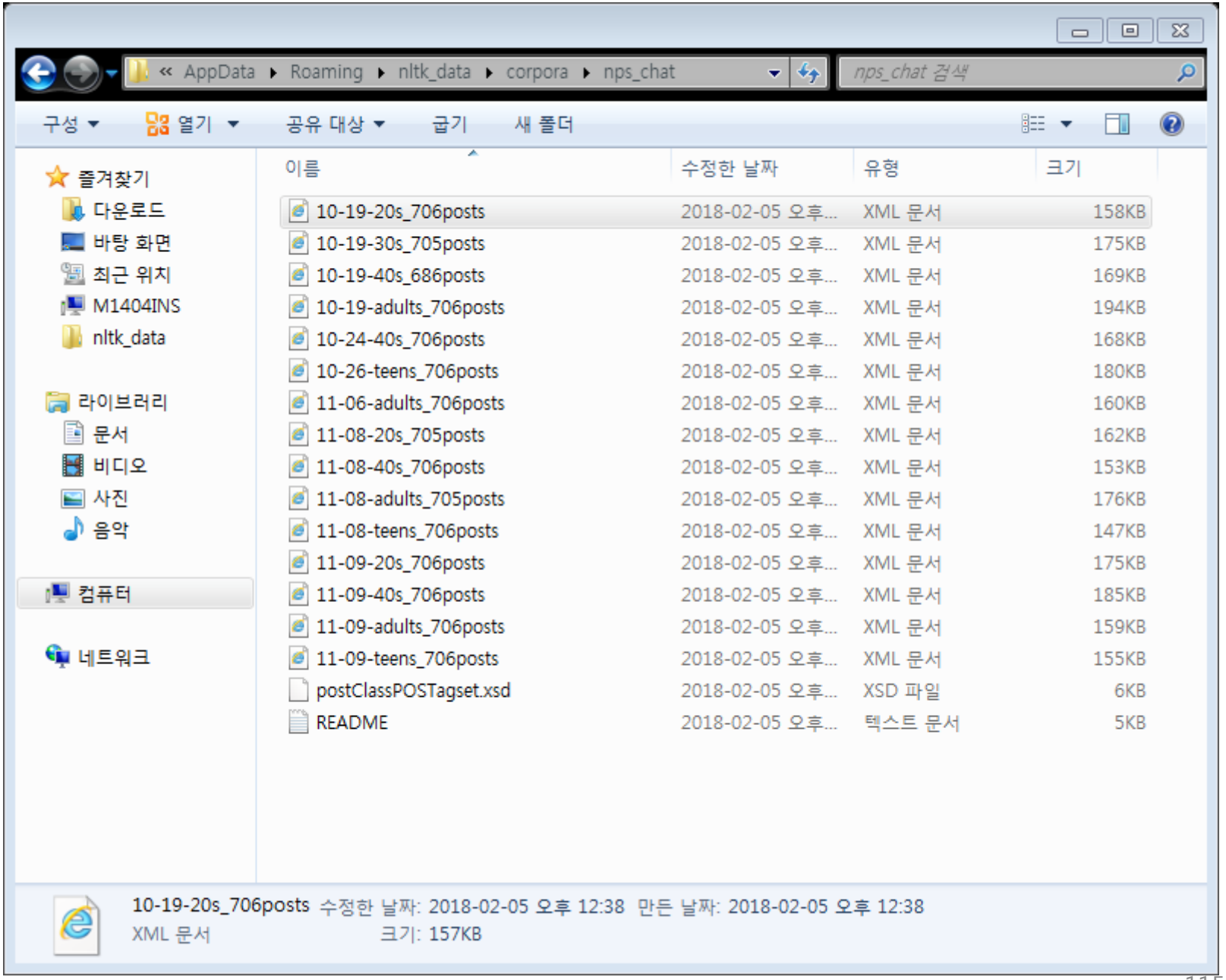
In [44]:
```
tokens = nps_chat.posts('10-19-20s_706posts.xml')
```

In [45]:
```
tokens
```

Out[45]:
```
[['now', 'im', 'left', 'with', 'this', 'gay', 'name'], [':P'],
...]
```

In [ ]:
```

```


« AppData ▸ Roaming ▸ nltk_data ▸ corpora ▸ nps_chat
nps_chat 검색
구성
열기
공유 대상
굽기
새 폴더
즐겨찾기
다운로드
바탕 화면
최근 위치
M1404INS
nltk_data
라이브러리
문서
비디오
사진
음악
컴퓨터
네트워크
이름
수정한 날짜
유형
크기
10-19-20s_706posts 2018-02-05 오후... XML 문서 158KB
10-19-30s_705posts 2018-02-05 오후... XML 문서 175KB
10-19-40s_686posts 2018-02-05 오후... XML 문서 169KB
10-19-adults_706posts 2018-02-05 오후... XML 문서 194KB
10-24-40s_706posts 2018-02-05 오후... XML 문서 168KB
10-26-teens_706posts 2018-02-05 오후... XML 문서 180KB
11-06-adults_706posts 2018-02-05 오후... XML 문서 160KB
11-08-20s_705posts 2018-02-05 오후... XML 문서 162KB
11-08-40s_706posts 2018-02-05 오후... XML 문서 153KB
11-08-adults_705posts 2018-02-05 오후... XML 문서 176KB
11-08-teens_706posts 2018-02-05 오후... XML 문서 147KB
11-09-20s_706posts 2018-02-05 오후... XML 문서 175KB
11-09-40s_706posts 2018-02-05 오후... XML 문서 185KB
11-09-adults_706posts 2018-02-05 오후... XML 문서 159KB
11-09-teens_706posts 2018-02-05 오후... XML 문서 155KB
postClassPOSTagset.xsd 2018-02-05 오후... XSD 파일 6KB
README 2018-02-05 오후... 텍스트 문서 5KB
10-19-20s_706posts 수정한 날짜: 2018-02-05 오후 12:38 만든 날짜: 2018-02-05 오후 12:38
XML 문서 크기: 157KB

```
<!-- edited with XMLSpy v2007 sp1 (http://www.altova.com) by Eric
Forsyth (Naval Postgraduate School) -->
- <Session xsi:noNamespaceSchemaLocation="postClassPOSTagset.xsd"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
  - <Posts>
      - <Post user="10-19-20sUser7" class="Statement">
          now im left with this gay name
        - <terminals>
            <t word="now" pos="RB"/>
            <t word="im" pos="PRP"/>
            <t word="left" pos="VBD"/>
            <t word="with" pos="IN"/>
            <t word="this" pos="DT"/>
            <t word="gay" pos="JJ"/>
            <t word="name" pos="NN"/>
          </terminals>
      </Post>
      - <Post user="10-19-20sUser7" class="Emotion">
          :P
        - <terminals>
            <t word=":P" pos="UH"/>
```

Trusted | Python 3

File Edit View Insert Cell Kernel Widgets Help

In [40]:
```
어휘다양성.sort_values().plot(kind='barh')
```

...

In [41]:
```
from nltk.corpus import nps_chat
```

In [42]:
```
nps_chat
```

Out[42]:
```
<NPSChatCorpusReader in 'C:₩₩Users₩₩student₩₩AppData₩₩Roaming₩₩nltk_data₩₩corpora₩₩nps_chat'>
```

In [44]:
```
tokens = nps_chat.posts('10-19-20s_706posts.xml')
```

CorpusReader

Statement

In [45]:
```
tokens
```

Out[45]:
```
[['now', 'im', 'left', 'with', 'this', 'gay', 'name'], [':P'],
...]
```

In [ ]:

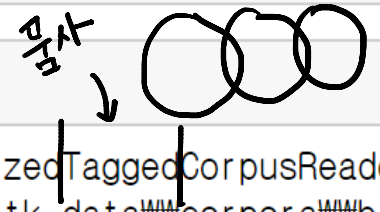
Trusted | Python 3

File Edit View Insert Cell Kernel Widgets Help

```
...]
```

In [46]:
```
from nltk.corpus import brown
```

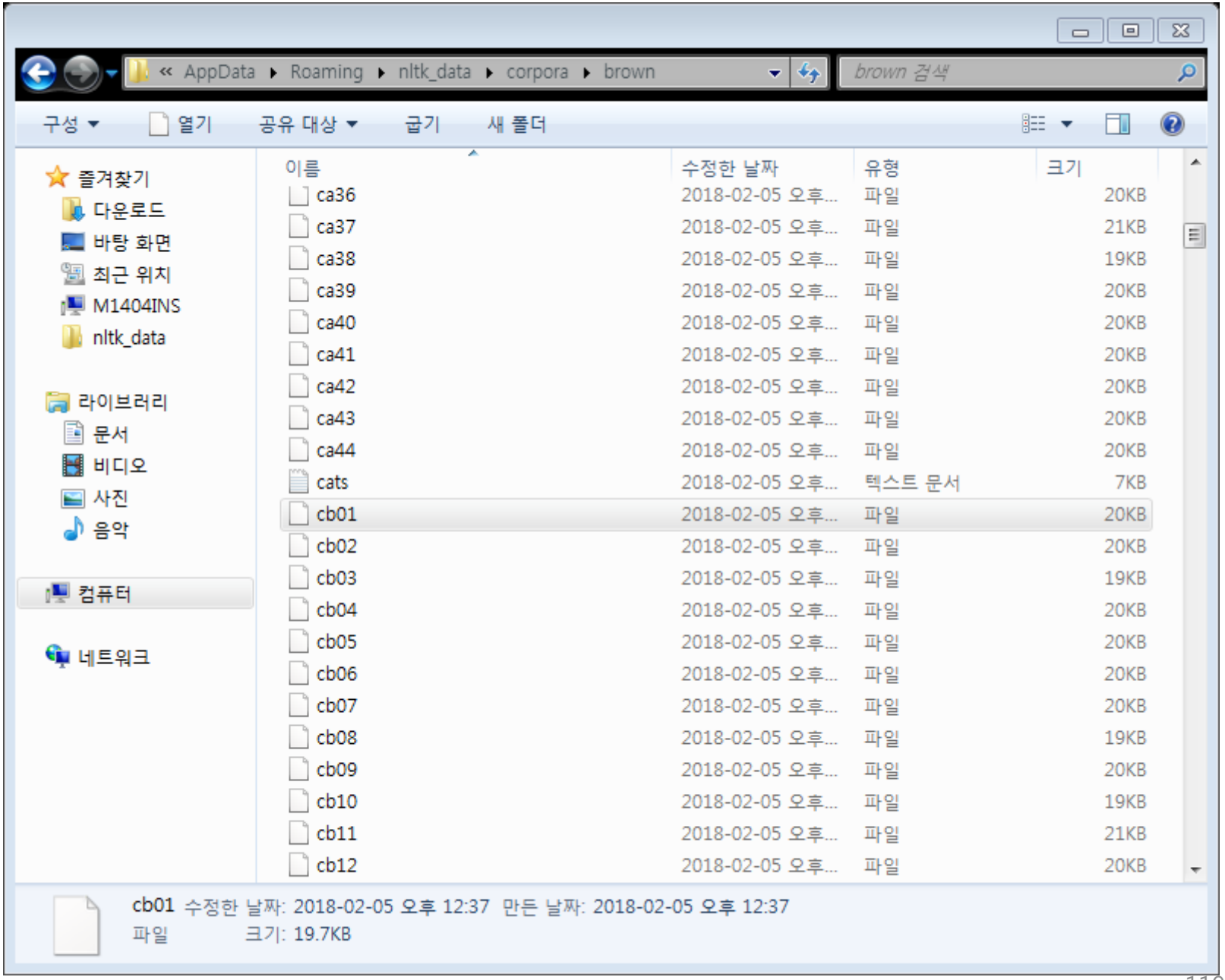
In [48]:
```
brown
```



Out[48]:
```
<CategorizedTaggedCorpusReader in 'C:₩₩Users₩₩student₩₩AppData₩₩Roaming₩₩nltk_data₩₩corpora₩₩brown'>
```

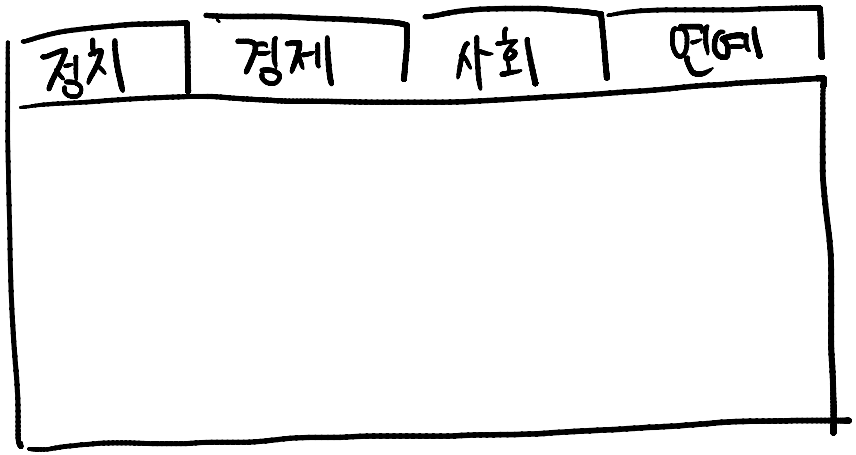
In [47]:
```
brown.categories()
```

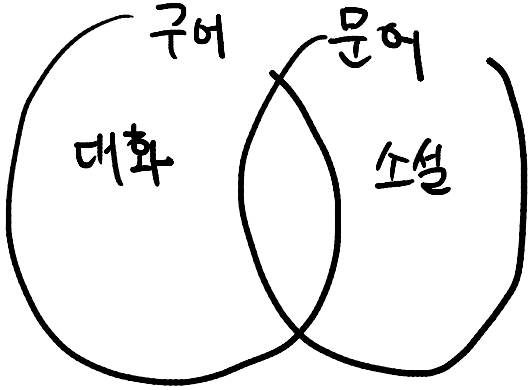
Out[47]:
```
['adventure',
 'belles_lettres',
 'editorial',
 'fiction',
 'government',
 'hobbies',
 'humor',
 'learned',
 'lore',
 'mystery',
 'news',
 'religion',
 'reviews',
 'romance',
 'science_fiction']
```


« AppData ▸ Roaming ▸ nltk_data ▸ corpora ▸ brown
brown 검색
구성
열기
공유 대상
굽기
새 폴더
즐겨찾기
다운로드
바탕 화면
최근 위치
M1404INS
nltk_data
라이브러리
문서
비디오
사진
음악
컴퓨터
네트워크
이름
수정한 날짜
유형
크기
ca36 2018-02-05 오후... 파일 20KB
ca37 2018-02-05 오후... 파일 21KB
ca38 2018-02-05 오후... 파일 19KB
ca39 2018-02-05 오후... 파일 20KB
ca40 2018-02-05 오후... 파일 20KB
ca41 2018-02-05 오후... 파일 20KB
ca42 2018-02-05 오후... 파일 20KB
ca43 2018-02-05 오후... 파일 20KB
ca44 2018-02-05 오후... 파일 20KB
cats 2018-02-05 오후... 텍스트 문서 7KB
cb01 2018-02-05 오후... 파일 20KB
cb02 2018-02-05 오후... 파일 20KB
cb03 2018-02-05 오후... 파일 19KB
cb04 2018-02-05 오후... 파일 20KB
cb05 2018-02-05 오후... 파일 20KB
cb06 2018-02-05 오후... 파일 20KB
cb07 2018-02-05 오후... 파일 20KB
cb08 2018-02-05 오후... 파일 19KB
cb09 2018-02-05 오후... 파일 20KB
cb10 2018-02-05 오후... 파일 19KB
cb11 2018-02-05 오후... 파일 21KB
cb12 2018-02-05 오후... 파일 20KB
cb01 수정한 날짜: 2018-02-05 오후 12:37 만든 날짜: 2018-02-05 오후 12:37
파일 크기: 19.7KB

뉴스/신문

Categorized CorpusReader





Trusted | Python 3

File Edit View Insert Cell Kernel Widgets Help

In [49]:
```
news_tokens = brown.words(categories='news')
```

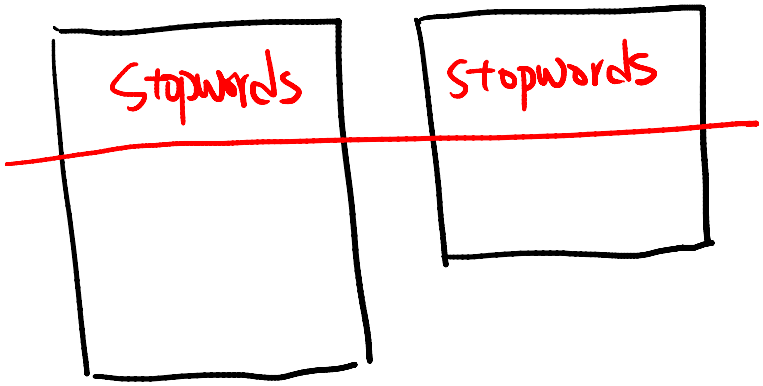
In [50]:
```
news_tokens
```

Out[50]:
```
['The', 'Fulton', 'County', 'Grand', 'Jury', 'said', ...]
```

In [53]:
```
Series(t for t in news_tokens).str.lower().value_counts()
```

Out[53]:

the	6386
,	5188
.	4030
of	2861
and	2186
to	2144
a	2130
in	2020
for	969
that	829
is	733
``	732
was	717
''	702
on	691
he	642
at	636



연습

brown 말뭉치의 모든 카테고리에 대해 관심 토큰의 도수를 집계해 비교해 봅시다.

In [56]:
```
집계 = []
for 분류 in brown.categories():
    tokens = brown.words(categories=분류)
    토큰도수 = Series(t for t in tokens).str.lower().value_counts()
    토큰도수.name = 분류
    집계.append(토큰도수[관심단어])
```

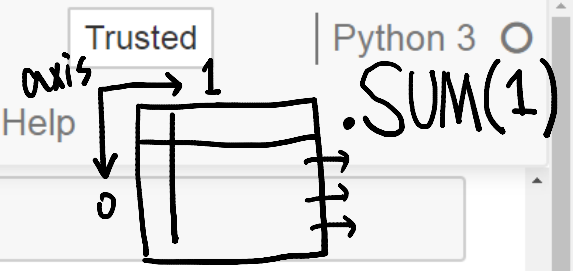
In [58]:
```
pd.concat(집계, axis=1).T
```

Out[58]:

	can	could	may	might	must	will
adventure	48	154	7	59	27	51
belles_lettres	249	216	221	113	171	246
editorial	124	57	79	39	55	235
fiction	39	168	10	44	55	56
government	119	38	179	13	102	244
hobbies	276	59	143	22	84	269

Trusted | Python 3

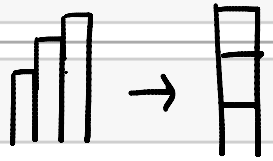
File Edit View Insert Cell Kernel Widgets Help



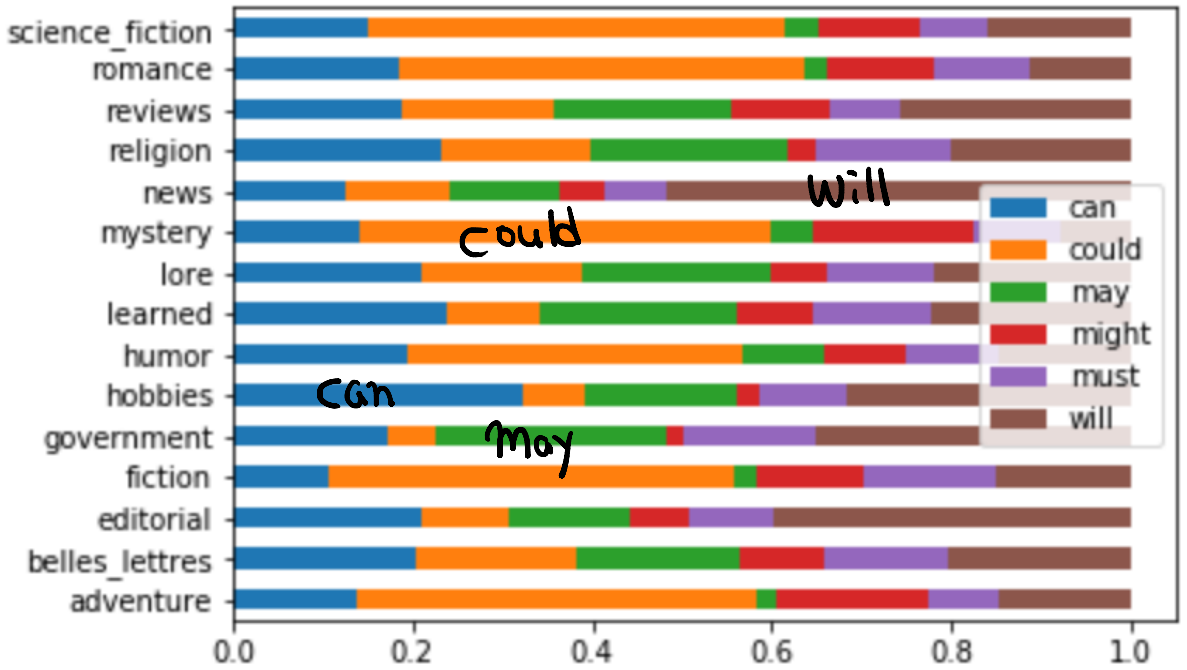
In [59]:
```
분류별_통계 = pd.concat(집계, axis=1).T
```

In [62]:
```
분류별_비율 = 분류별_통계.div(분류별_통계.sum(1), axis=0)
```

In [65]:
```
분류별_비율.plot(kind='barh', stacked=True)
```



Out[65]: <matplotlib.axes._subplots.AxesSubplot at 0xf08b160>



In [ ]:

Trusted | Python 3

File Edit View Insert Cell Kernel Widgets Help

...

In [66]:
```
from nltk.corpus import udhr as 인권선언
```

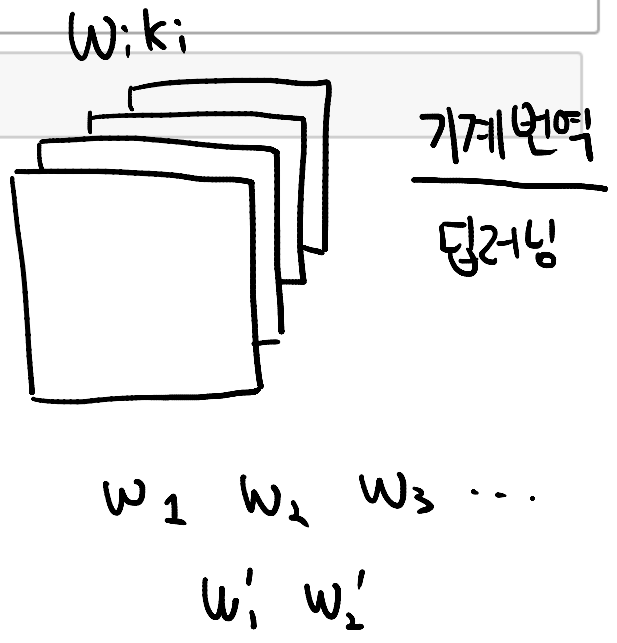
In [70]:
```
인권선언
```

Out[70]:
```
<UdhrCorpusReader in 'C:₩₩Users₩₩student₩₩AppData₩₩Roaming₩₩nltk_data₩₩corpora₩₩udhr'>
```

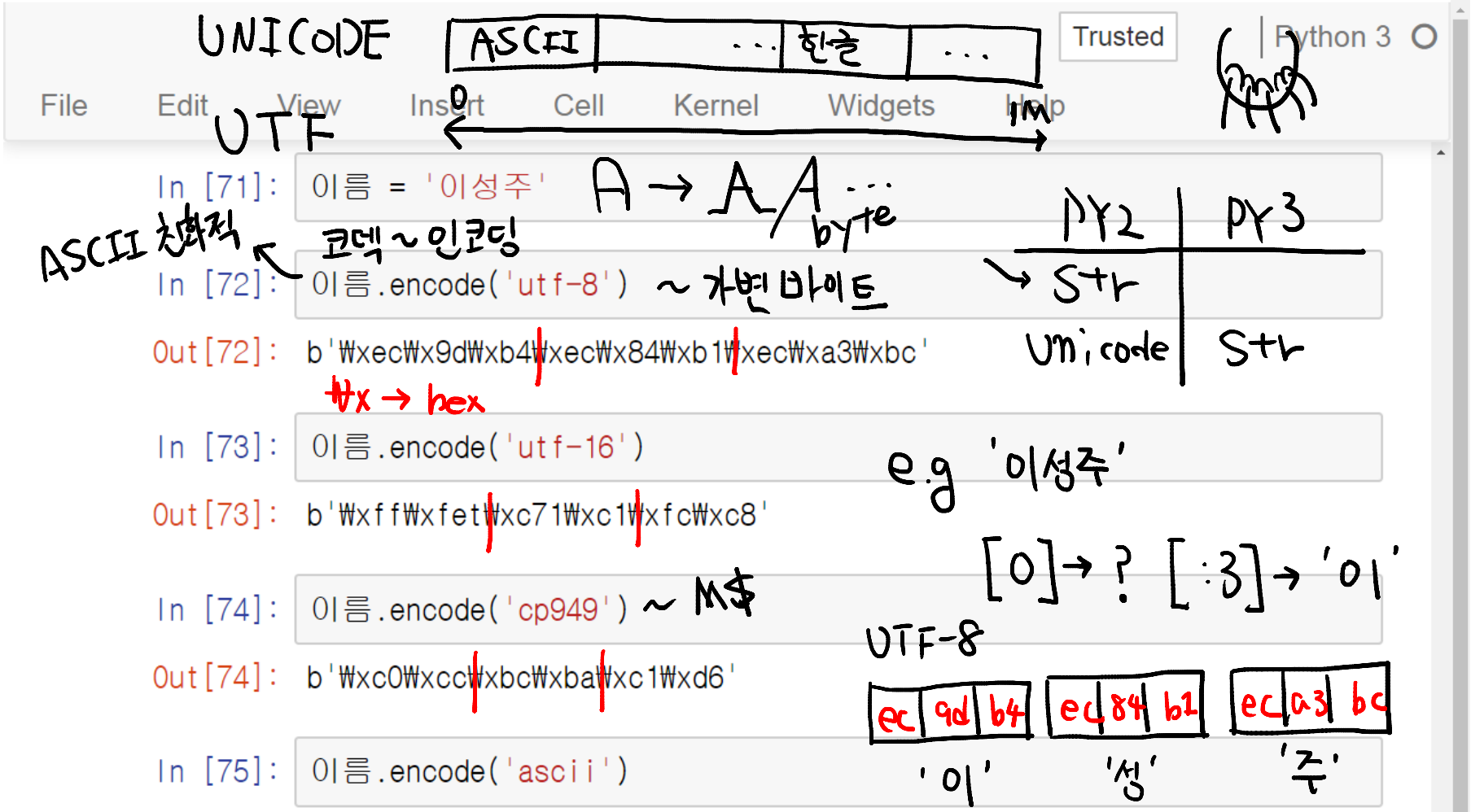
In [69]:
```
인권선언.fileids()[140:150]
```

Out[69]:
```
['Kazakh-UTF8',
 'Kiche_Quiche-Latin1',
 'Kicongo-Latin1',
 'Kimbundu_Mbundu-Latin1',
 'Kinyamwezi_Nyamwezi-Latin1',
 'Kinyarwanda-Latin1',
 'Kituba-Latin1',
 'Korean_Hankuko-UTF8',
 'Kpelewo-UTF8',
 'Krio-UTF8']
```

↖ 인코딩



In [ ]:



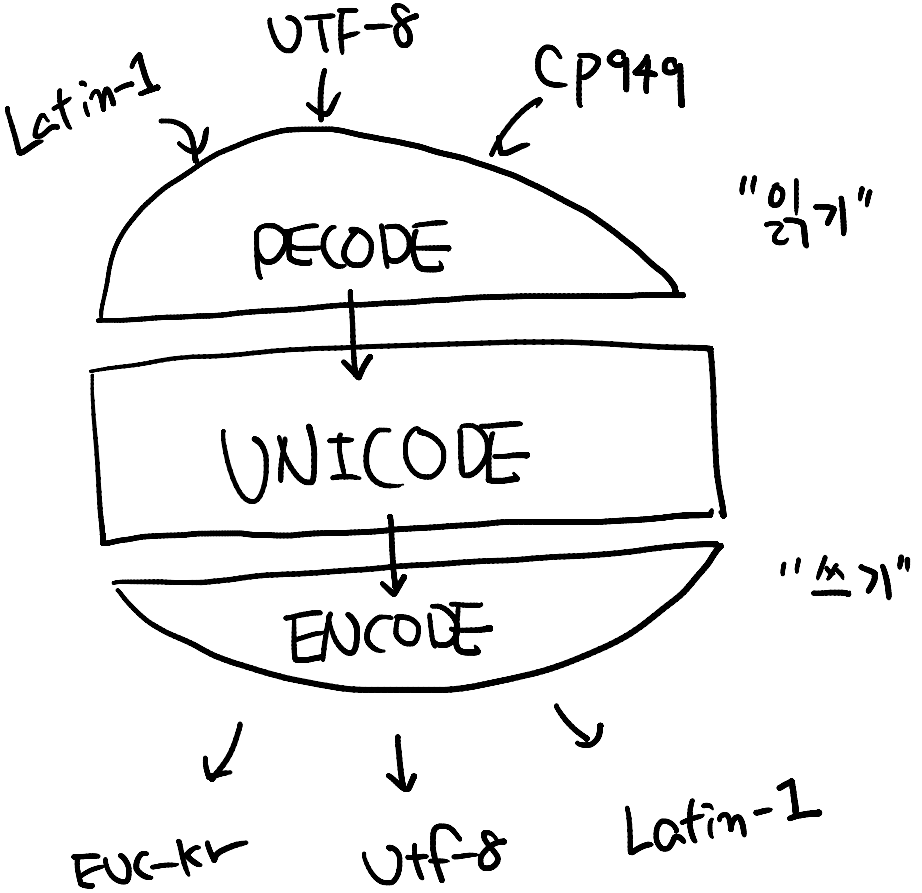
UnicodeEncodeError Traceback (most recent call last)

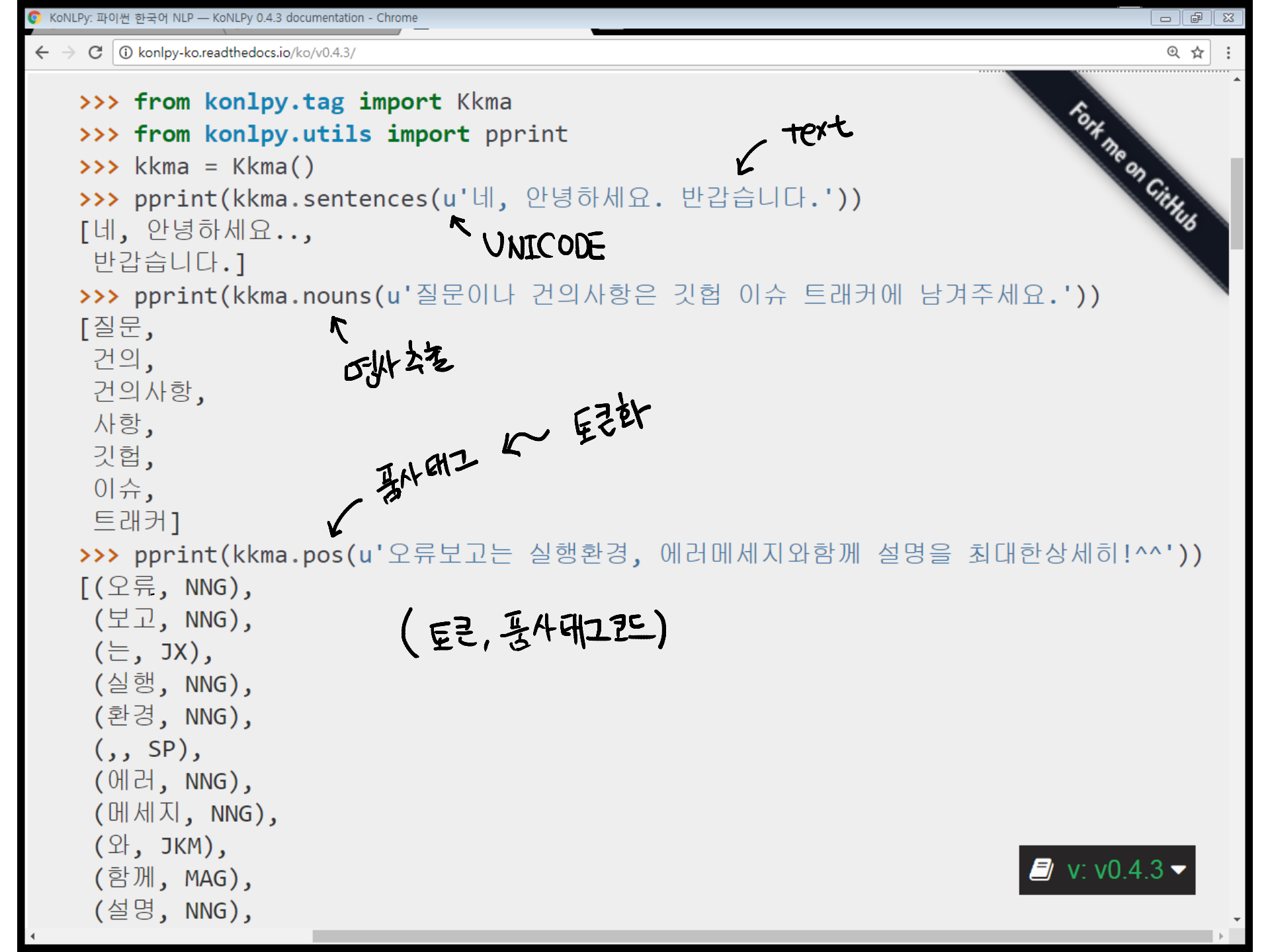
<ipython-input-75-7ebbb0b66347> in <module>()

----> 1 이름.encode('ascii')

UnicodeEncodeError: 'ascii' codec can't encode characters in position 0-2: ordinal not in range(128)

UNICODE SANDWICH






```
>>> 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 트래커에 남겨주세요.'))
[질문,
 건의,
 건의사항,
 사항,
 깃헙,
 이슈,
 트래커]
>>> pprint(kkma.pos(u'오류보고는 실행환경, 에러메세지와함께 설명을 최대한상세히!^^'))
[(오류, NNG),
 (보고, NNG),
 (는, JX),
 (실행, NNG),
 (환경, NNG),
 (,, SP),
 (에러, NNG),
 (메세지, NNG),
 (와, JKM),
 (함께, MAG),
 (설명, NNG),
```

데이터 — KoNLPy 0.4.3 documentation - Chrome

konlpy-ko.readthedocs.io/ko/v0.4.3/data/#corpora

말뭉치

nltk.corpus
gutenberg "CR"

다음의 말뭉치(corpus)를 사용할 수 있습니다:

1. `kolaw`: 한국 법률 말뭉치.
    - constitution.txt
2. `kobill`: 대한민국 국회 의안 말뭉치. 파일 ID는 의안 번호를 의미합니다.
    - 1809890.txt - 1809899.txt

KoNLPy에 포함된 말뭉치의 사용은 *corpus Package* 에서 더 자세하게 확인해볼 수 있습니다.

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

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

Related Pages

Home

- Previous page: 형태소 분석 및 품사 태깅
- Next page: 사용 예시

사전

사전은 대부분 *말뭉치* 를 이용해 구축되었으며 *형태소 분석 및 품사 태* 다.

Hannanum 시스템 사전

KAIST 말뭉치를 이용해 생성된 사전. (4.7MB)

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

```
...
나라경제        ncn
나라기획        nqq
나라기획회장    ncn
나라꽃  ncn
나라님  ncn
나라도둑        ncn
나라따르        pvg
나라링링프로덕션        ncn
나라말  ncn
```

위의 항목 중 하나라도 어긋나는 것이 있다면 제보 부탁드립니다.

라이센스

KoNLPy는 오픈소스 소프트웨어이며, 아래의 라이센스를 채택하고 있습니다:

- GPL v3 또는 그 이상 [2]

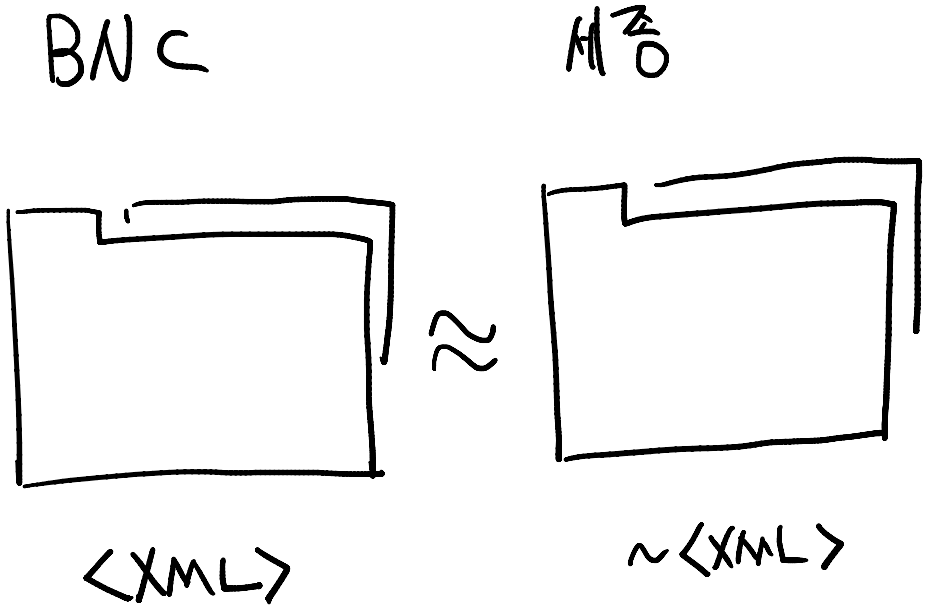
라이센스에 따라 자유롭게 코드를 이용하실 수 있으며, 연구에 KoNLPy를 사용하신 경우 아래 논문을 인용해주시기 바랍니다.

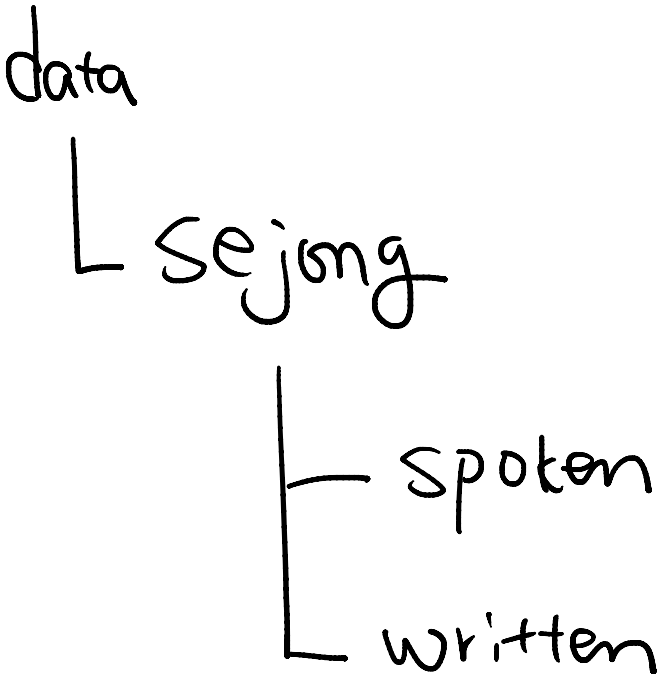
- 박은정, 조성준, "KoNLPy: 쉽고 간결한 한국어 정보처리 파이썬 패키지", 제 26회 한글 및 한국어 정보처리 학술대회 논문집, 2014.

BibTeX는 아래의 코드를 사용하시면 됩니다.:

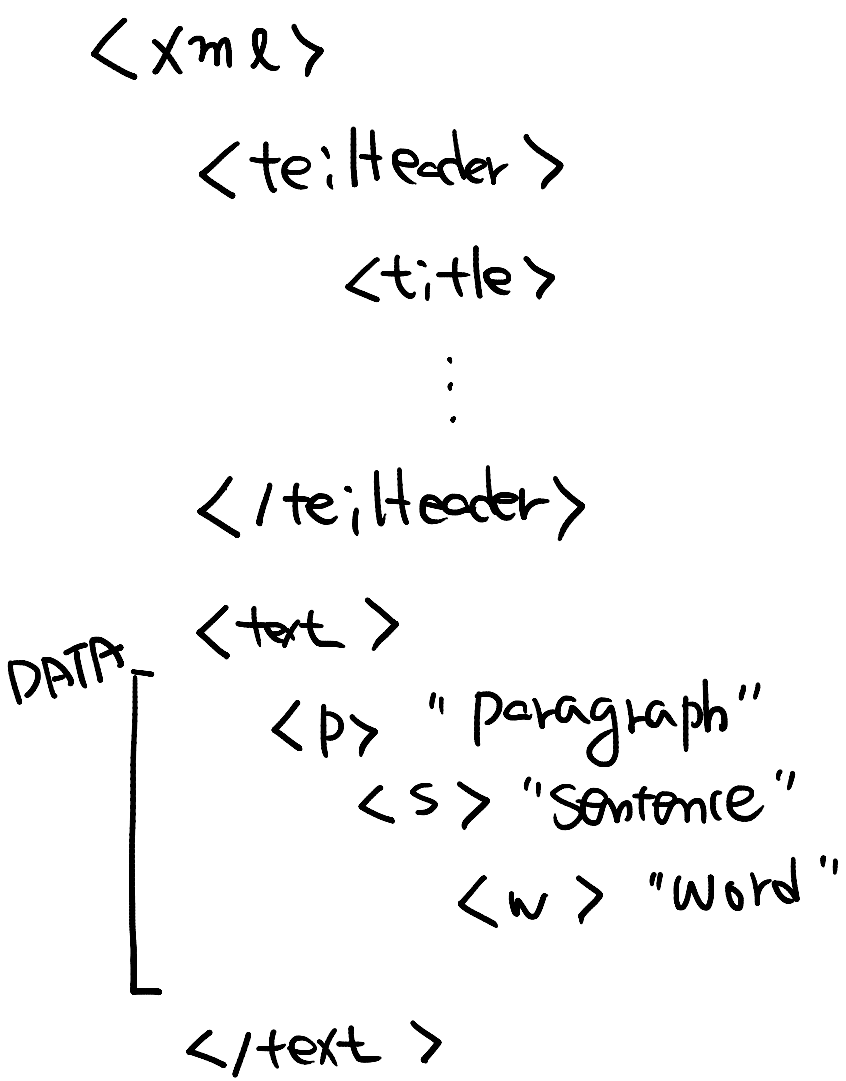
```
@inproceedings{park2014konlpy,
  title={KoNLPy: Korean natural language processing in Python},
  author={Park, Eunjeong L. and Cho, Sungzoon},
  booktitle={Proceedings of the 26th Annual Conference on Human & Cognitive Langua
  address={Chuncheon, Korea},
  month={October},
```

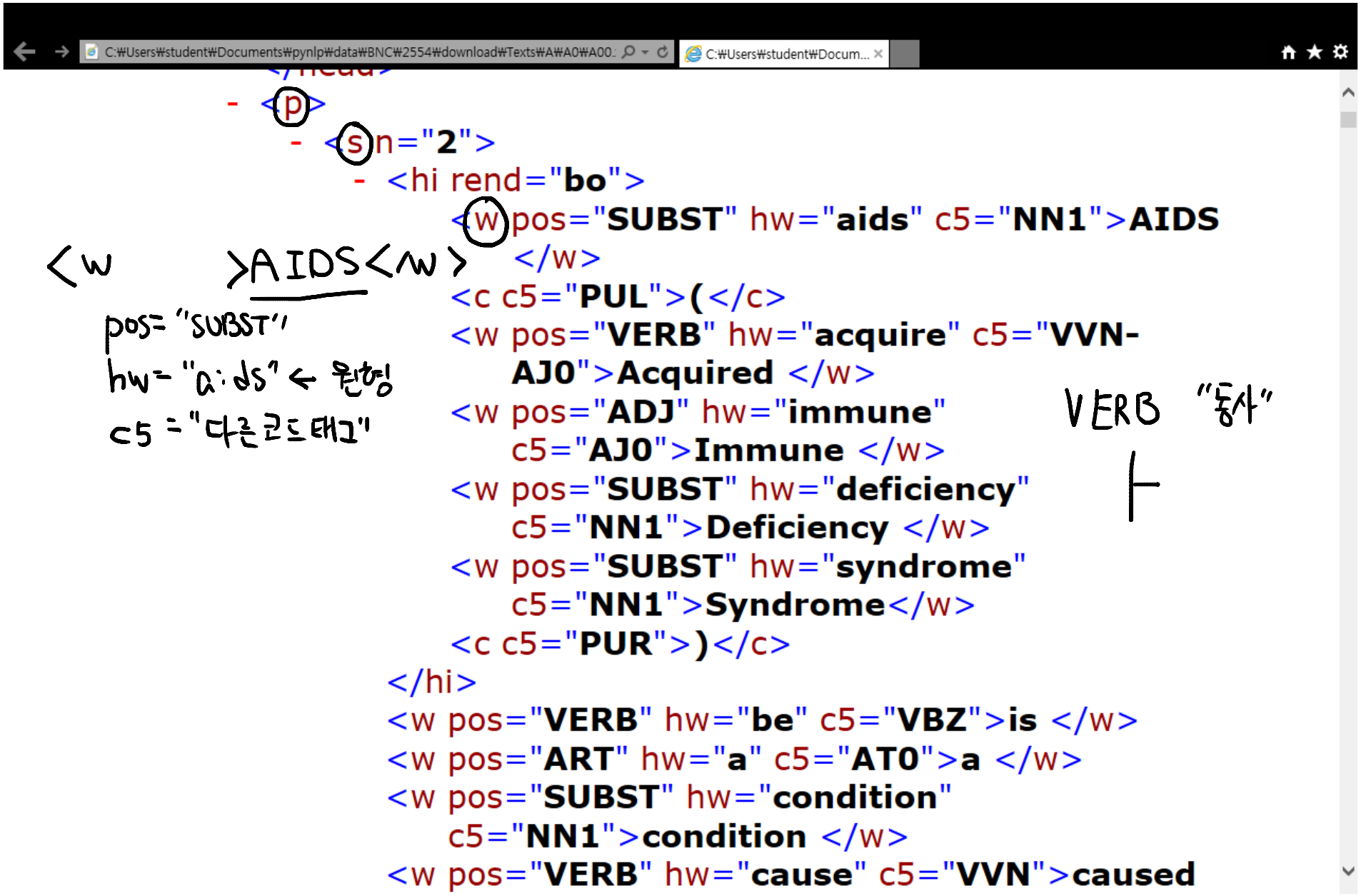
nltk.reader.BNCcorpusReader ≈ Sejong CorpusReader




data
sejong
spoken
written

BNC


<xml>
<te:Header>
<title>
</te:Header>
<text>
DATA
<p> "paragraph"
<s> "sentence"
<w> "word"
</text>


C:₩Users₩student₩Documents₩pynlp₩data₩BNC₩2554₩download₩Texts₩A₩A0₩A00..
C:₩Users₩student₩Docum...
- <p>
- <s n="2">
- <hi rend="bo">
<w pos="SUBST" hw="aids" c5="NN1">AIDS
</w>
<c c5="PUL">(</c>
<w pos="VERB" hw="acquire" c5="VVN-
AJ0">Acquired </w>
<w pos="ADJ" hw="immune"
c5="AJ0">Immune </w>
<w pos="SUBST" hw="deficiency"
c5="NN1">Deficiency </w>
<w pos="SUBST" hw="syndrome"
c5="NN1">Syndrome</w>
<c c5="PUR">)</c>
</hi>
<w pos="VERB" hw="be" c5="VBZ">is </w>
<w pos="ART" hw="a" c5="AT0">a </w>
<w pos="SUBST" hw="condition"
c5="NN1">condition </w>
<w pos="VERB" hw="cause" c5="VVN">caused
<w >AIDS</w>
pos="SUBST"
hw="aids" ← 원형
c5="다른코드태그"
VERB "동사"

Trusted | Python 3

File Edit View Insert Cell Kernel Widgets Help

In [92]:
```
from nltk.corpus.reader.bnc import BNCCorpusReader
```

In [93]:
```
bnc_reader = BNCCorpusReader(
    root='data/BNC/2554/download/Texts/',
    fileids=r'[A-Z]/₩w+/₩w+₩.xml'
)
```

패턴 → 정규식

In [95]:
```
bnc_reader.fileids()[:100]
```

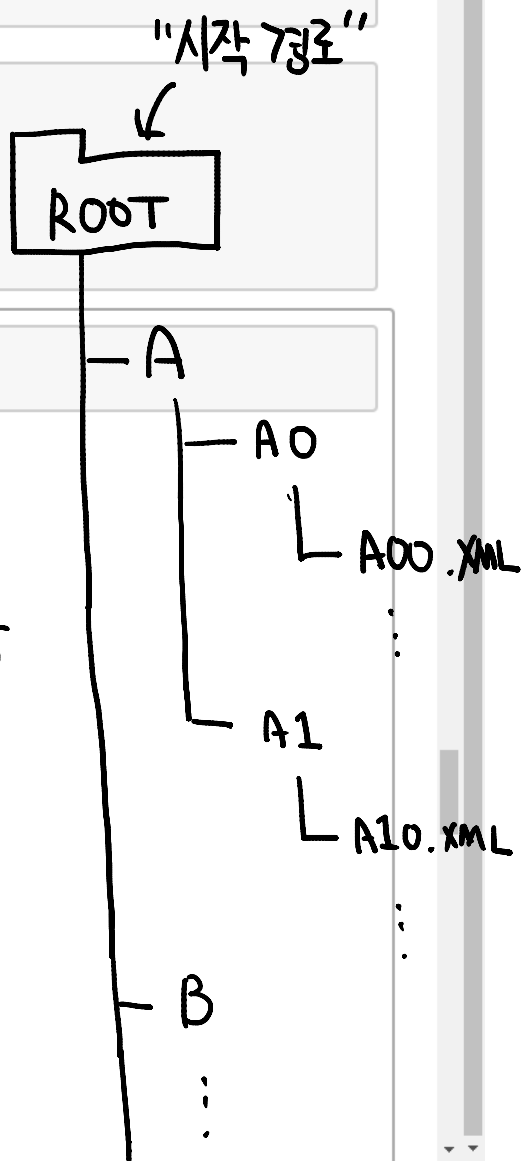
Out[95]:
```
['A/A0/A00.xml',
 'A/A0/A01.xml',
 'A/A0/A02.xml',
 'A/A0/A03.xml',
 'A/A0/A04.xml',
 'A/A0/A05.xml',
 'A/A0/A06.xml',
 'A/A0/A07.xml',
 'A/A0/A08.xml',
 'A/A0/A0A.xml',
 'A/A0/A0B.xml',
 'A/A0/A0C.xml',
 'A/A0/A0D.xml',
 'A/A0/A0E.xml',
 'A/A0/A0F.xml',
```

[A-Z]

\W+ : 아무글자 1자이상

\W+\.xml

e.g. A00.xml



Notebook saved | Trusted | Python 3

File Edit View Insert Cell Kernel Widgets Help

```
In [95]: bnc_reader.fileids()[:100]
```

... <w pos, hw, c5 > WHAT </w>

```
In [96]: bnc_reader.words('A/A0/A00.xml')
```

.words( )

```
Out[96]: ['FACTSHEET', 'WHAT', 'IS', 'AIDS', '?', 'AIDS', '(', ...]
```

```
In [99]: bnc_reader.words('A/A0/A00.xml')[5:]
```

```
Out[99]: ['AIDS', '(', 'Acquired', 'Immune', 'Deficiency', ...]
```

단어줄기 ≈ 원형 <w ...

```
In [98]: bnc_reader.words('A/A0/A00.xml', stem=True)[5:]
```

→ hw="Aquire" >

```
Out[98]: ['aids', '(', 'acquire', 'immune', 'deficiency', ...]
```

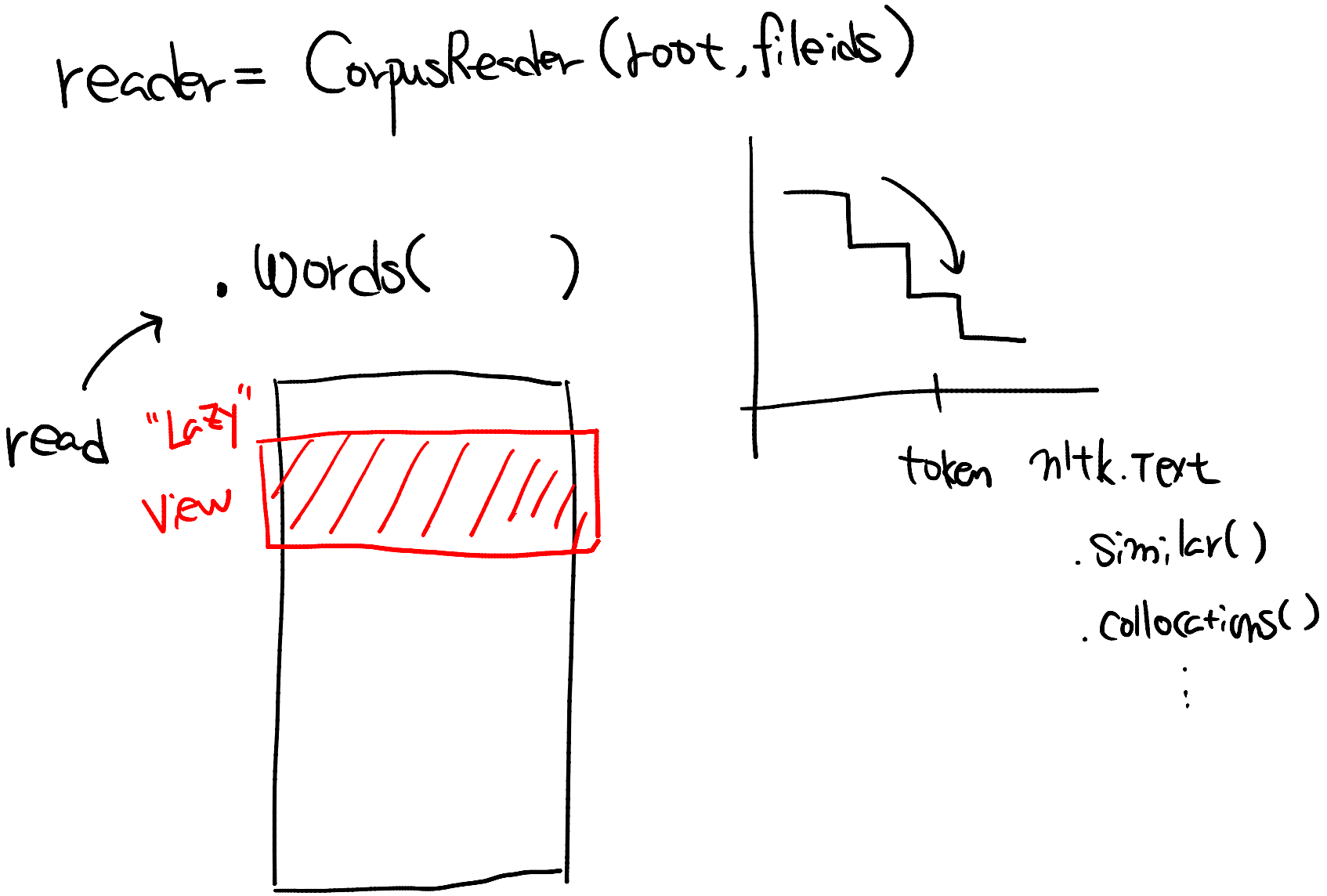
Aquired </w>

```
In [100]: bnc_reader.tagged_words('A/A0/A00.xml')
```

```
Out[100]: [('FACTSHEET', 'SUBST'), ('WHAT', 'PRON'), ...]
```

```
In [ ]:
```

<w pos="SUBST" ...> FACTSHEET </w>


reader = CorpusReader(root, fileids)
.words( )
read
"Lazy"
view
token
nltk.Text
.similar()
.collocations()

Trusted | Python 3

File Edit View Insert Cell Kernel Widgets Help

Code

In [102]:
```
bnc_text = nltk.Text(bnc_reader.words('A/A0/A00.xml'))
```

In [103]:
```
bnc_text.similar('AIDS')
```

```
hiv patients people global acet time unconditional only praise iron
ing
thursday
```

In [104]:
```
bnc_text.collocations()
```

```
Home Care; 840 7879; 081 840; South East; Health Authority; Maurice
Adams; Sue Lore; home care; Anthony Kasozi; equipment loans;
International Adviser; General Manager; Peter Johnson; HIV positiv
e;
Patrick Dixon; people ill; financial support; Jackie Sears; John
Creedy; Jonathan Mann
```

In [ ]:

4CM00003 - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

```
        </revisionDesc>
</teiHeader>
<text>
<u who="P1"><s n="00001">뭘 좀 올려야지.</s></u>
<u who="P2"><s n="00002">뭘 좀 올렸어.</s></u>
<u who="P1"><s n="00003">다시 돌려 앞으로.</s></u>
<u who="P2"><s n="00004">됐어.</s></u>
<timeLine><when id="T1"/><when id="T2"/></timeLine>
<u who="P1"><s n="00005">우리가족은 아빠 <anchor synch
<u who="P2"><s n="00006"><anchor synch="T1"/>목소리 외
<u who="P1"><s n="00007">이렇게 네 명인데,</s>
<s n="00008">원래 할아버지두 같이 살었는데,</s>
<s n="00009">할아버지는 내가 고삼 때 돌아가셨고,</s>
<s n="00010">음 아빠 엄마 나 오빤데,</s>
<s n="00011">소개할 거 가지두 없지만,</s>
<s n="00012">그냥 우리 가족은 전체적으로,</s>
<s n="00013">엄청 보수적이야.</s>
```

entence

raw speech → Text

USER "화자"

5CT_0013 - 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

aquired → aquire/동사 + ed/과거형

```
5CT_0013-0000100    </u>
5CT_0013-0000110    <u who="P2">
5CT_0013-0000120    <s n="00003">
5CT_0013-0000130    지하철.  지하철/NNG+./SF
5CT_0013-0000140    </s>
5CT_0013-0000150    </u>
5CT_0013-0000160    <u who="P1">
5CT_0013-0000170    <s n="00004">
5CT_0013-0000180    기차?    기차/NNG+?/SF
5CT_0013-0000190    </s>
5CT_0013-0000200    <s n="00005">
5CT_0013-0000210    아침에   아침/NNG+에/JKB
5CT_0013-0000220    몇       몇/MM
5CT_0013-0000230    시에     시/NNB+에/JKB
5CT_0013-0000240    타고     타/VV+고/EC
5CT_0013-0000250    가는데?  가/VV+는데/EF+?/SF
5CT_0013-0000260    </s>
```

토큰(token) 품사 + 토큰(token) 품사

원문(raw)

줄기 + 원형

타다

타고

탔어

⋮

형태소분석

CorpusReader

nltk/bnc.py at 2c1966bb469d8a279d82e257a10d6c9da7ebc8e0 · nltk/nltk · GitHub - Chrome

GitHub, Inc. [US] | https://github.com/nltk/nltk/blob/2c1966bb469d8a279d82e257a10d6c9da7ebc8e0/nltk/corpus/reader/bnc.py

nltk/corpus/reader/bnc.py

XMLCorpusReader

```
13
14  class BNCCorpusReader(XMLCorpusReader):
15      """Corpus reader for the XML version of the British N
16
17      For access to the complete XML data structure, use th
18      method.  For access to simple word lists and tagged w
19      ``words()``, ``sents()``, ``tagged_words()``, and ``t
20
21      You can obtain the full version of the BNC corpus at
22      http://www.ota.ox.ac.uk/desc/2554
23
24      If you extracted the archive to a directory called `B
```

nltk/bnc.py at 2c1966bb469d8a279d82e257a10d6c9da7ebc8e0 · nltk/nltk · GitHub - Chrome

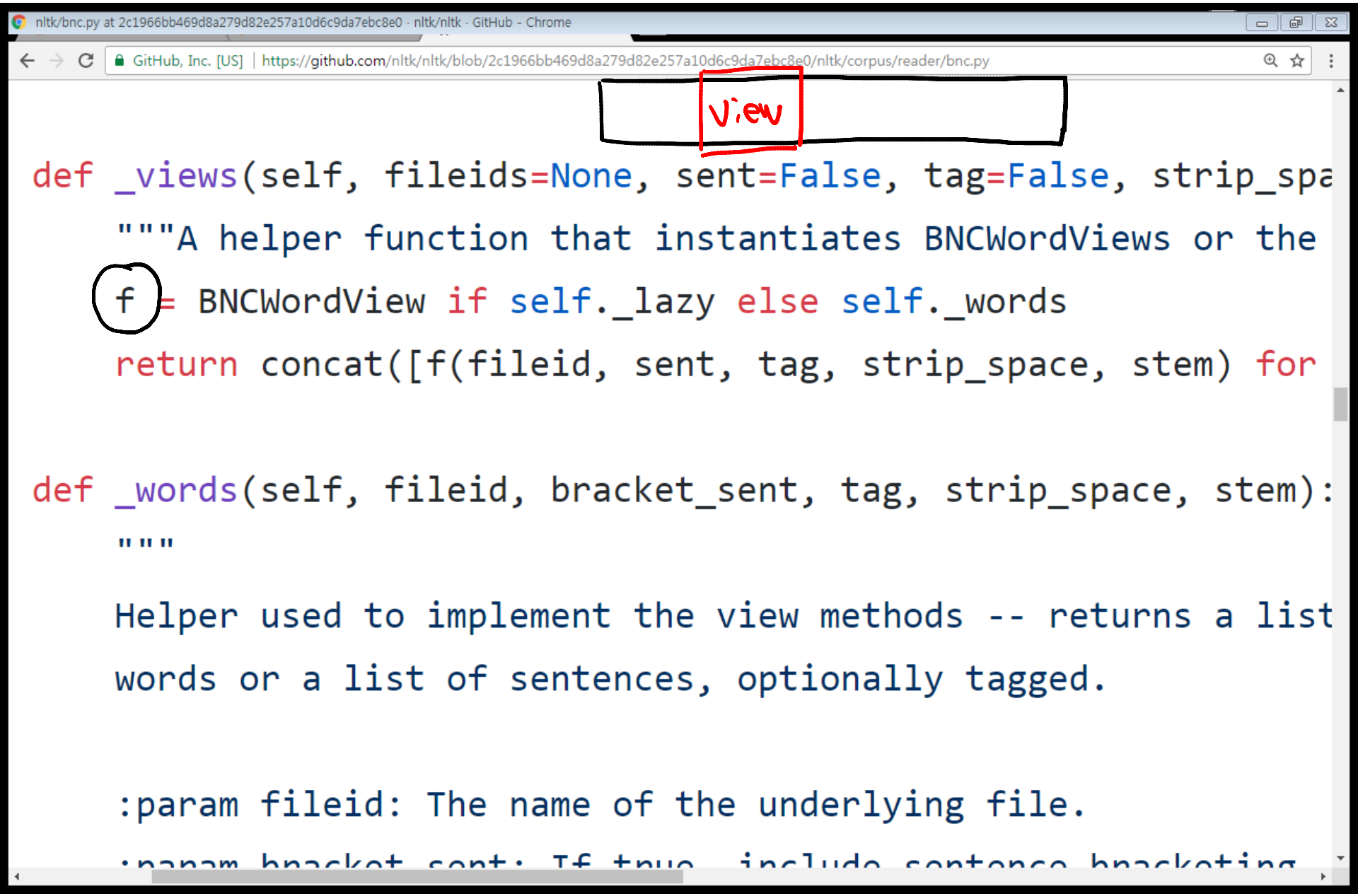
GitHub, Inc. [US] | https://github.com/nltk/nltk/blob/2c1966bb469d8a279d82e257a10d6c9da7ebc8e0/nltk/corpus/reader/bnc.py

```
def words(self, fileids=None, strip_space=True, stem=False):
    """
    :return: the given file(s) as a list of words
        and punctuation symbols.
    :rtype: list(str)

    :param strip_space: If true, then strip trailing spaces frc
        word tokens.  Otherwise, leave the spaces on the tokens
    :param stem: If true, then use word stems instead of word s
    """
    return self._views(fileids, False, None, strip_space, stem)
```

<w hw= > ... </w>

"private"


nltk/bnc.py at 2c1966bb469d8a279d82e257a10d6c9da7ebc8e0 · nltk/nltk · GitHub - Chrome
GitHub, Inc. [US] | https://github.com/nltk/nltk/blob/2c1966bb469d8a279d82e257a10d6c9da7ebc8e0/nltk/corpus/reader/bnc.py
View
def _views(self, fileids=None, sent=False, tag=False, strip_spa
"""A helper function that instantiates BNCWordViews or the
f = BNCWordView if self._lazy else self._words
return concat([f(fileid, sent, tag, strip_space, stem) for
def _words(self, fileid, bracket_sent, tag, strip_space, stem):
"""
Helper used to implement the view methods -- returns a list
words or a list of sentences, optionally tagged.
:param fileid: The name of the underlying file.

<xml>

```
        xmldoc = ElementTree.parse(fileid).getroot()
        for xmlsent in xmldoc.findall('.//s'):
            sent = []
            for xmlword in _all_xmlwords_in(xmlsent):
                word = xmlword.text
                if not word:
                    word = ""  # fixes issue 337?
                if strip_space or stem:
                    word = word.strip()
                if stem:
                    word = xmlword.get('hw', word)
                if tag == 'c5':
                    word = (word, xmlword.get('c5'))
                elif tag == 'pos':
                    word = (word, xmlword.get('pos', xmlword.g
                sent.append(word)
```

</xml>

[$t_1, t_2, \cdots$]

[$St_1, St_2, \cdots$]

[$(t_1, pos), \cdots$]

<s> … </s>

<w> … </w>

.text

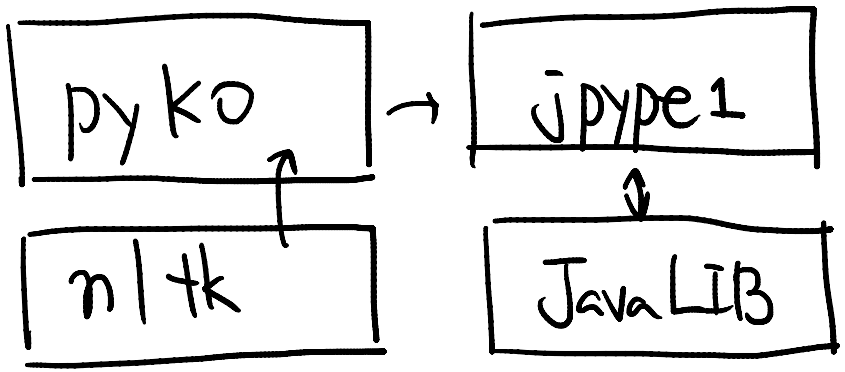
단어원형?

Install

1. jpype1

   a. JRE

tokenizer

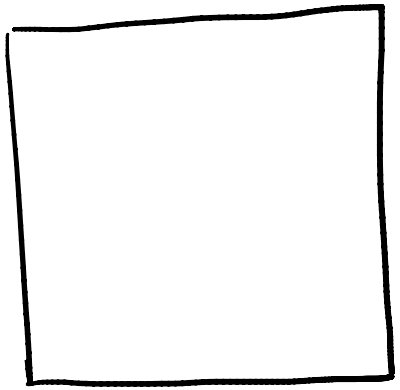


2. pip install pyko

Python

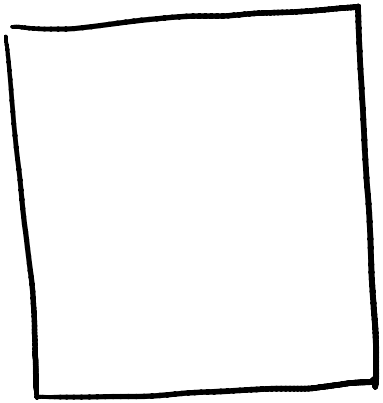
JVM ← install JRE8

PyPI



$ pip install

Conda



$ Conda install

Trusted | Python 3

File Edit View Insert Cell Kernel Widgets Help

Code

In [105]:
```
import pyko
```

nltk.reader.CorpusReader

In [106]:
```
sejong_reader = pyko.reader.SejongCorpusReader(
    root='data/sejong/',
    fileids='spoken/word_tag/.+W.txt', encoding='utf-16'
)
```

In [108]:
```
sejong_reader.fileids()[:10]
```

Out[108]:
```
['spoken/word_tag/5CT_0013.txt',
 'spoken/word_tag/5CT_0014.txt',
 'spoken/word_tag/5CT_0015.txt',
 'spoken/word_tag/5CT_0016.txt',
 'spoken/word_tag/5CT_0017.txt',
 'spoken/word_tag/5CT_0018.txt',
 'spoken/word_tag/5CT_0019.txt',
 'spoken/word_tag/5CT_0020.txt',
 'spoken/word_tag/5CT_0021.txt',
 'spoken/word_tag/5CT_0022.txt']
```

In [ ]:

Trusted | Python 3

File Edit View Insert Cell Kernel Widgets Help

Code

In [108]:
```
sejong_reader.fileids()[:10]
```

```
...
```

View

In [110]:
```
sejong_reader.words(tagged=False)
```

Out[110]:
```
['뭐', '타고', '가?', '지하철.', '기차?', '아침에', ...]
```

In [111]:
```
sejong_reader.words(tagged=True)
```

Out[111]:
```
[('뭐', (('뭐', 'NP'),)), ('타고', (('타', 'VV'), ('고', 'EC'))),
('가?', (('가', 'VV'), ('ㅏ', 'EF'), ('?', 'SF'))), ('지하철.',
(('지하철', 'NNG'), ('.', 'SF'))), ('기차?', (('기차', 'NNG'),
('?', 'SF'))), ('아침에', (('아침', 'NNG'), ('에', 'JKB'))), ...]
```

In [ ]:

타고 → 타/VV + 고/EC → (타고, ((타, VV), (고, EC))

Trusted | Python 3

File Edit View Insert Cell Kernel Widgets Help

Code

...

In [123]:
```
경로 = 'data/sejong/spoken/raw/4CM00003.txt'
```

In [124]:
```
from bs4 import BeautifulSoup
```

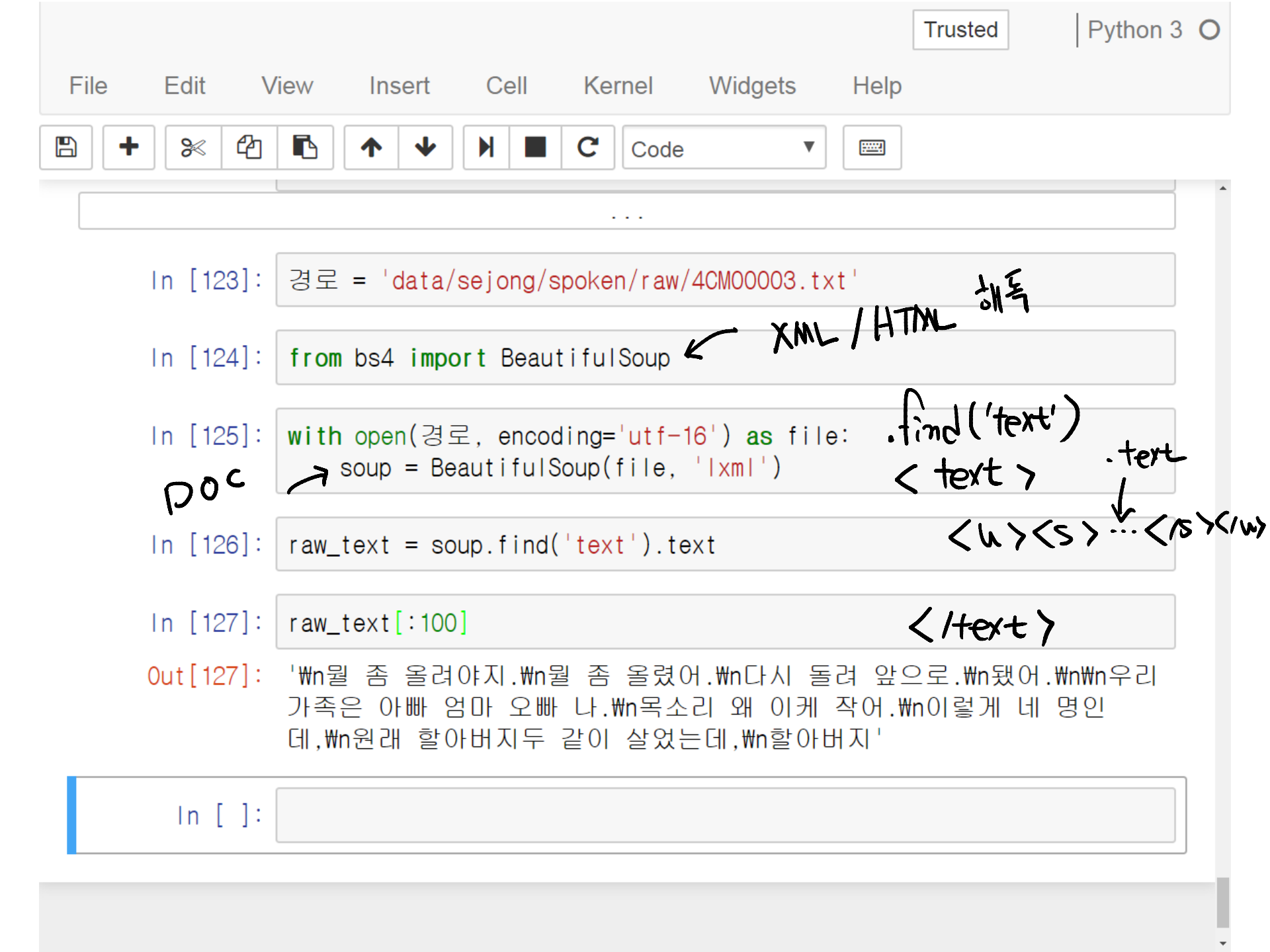
In [125]:
```
with open(경로, encoding='utf-16') as file:
    soup = BeautifulSoup(file, 'lxml')
```

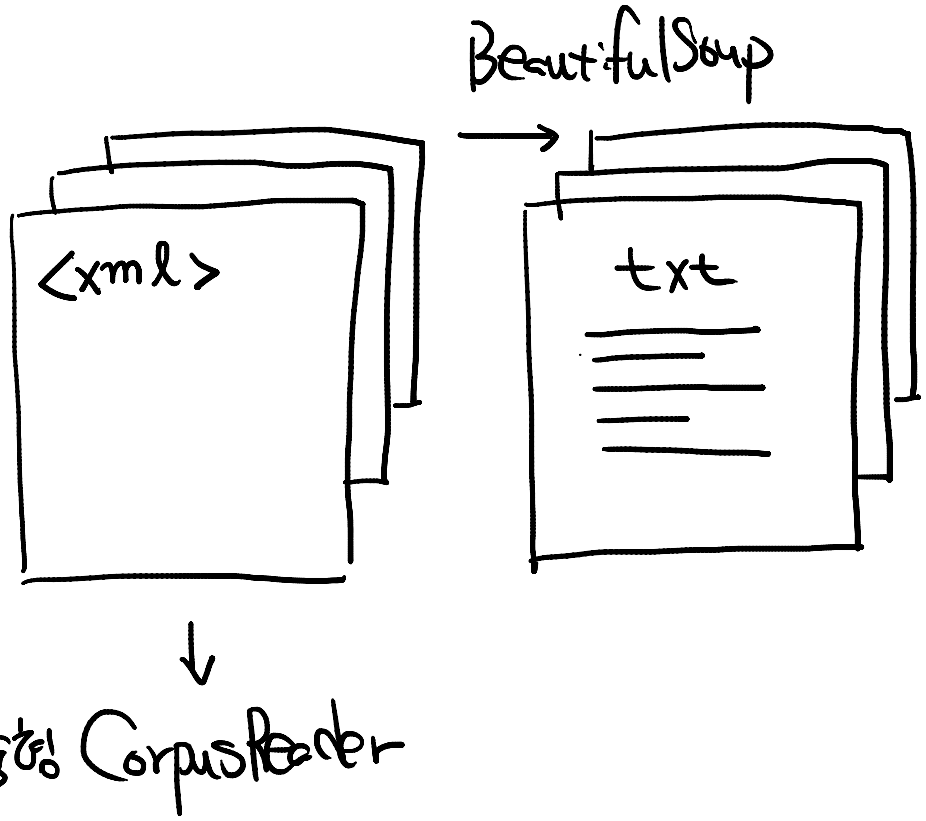
In [126]:
```
raw_text = soup.find('text').text
```

In [127]:
```
raw_text[:100]
```

Out[127]:
```
'₩n뭘 좀 올려야지.₩n뭘 좀 올렸어.₩n다시 돌려 앞으로.₩n됐어.₩n₩n우리 가족은 아빠 엄마 오빠 나.₩n목소리 왜 이게 작어.₩n이렇게 네 명인데,₩n원래 할아버지두 같이 살었는데,₩n할아버지'
```

In [ ]:





nltk.corpus

PlaintextCorpusReader

.words( )

word_tokenize

"언어 특수성"

Trusted | Python 3

File Edit View Insert Cell Kernel Widgets Help

Code

In [132]:
```
reader.fileids()
```

Out[132]:
```
['sejong_4CM00003.txt']
```

In [134]:
```
reader.words()[10:20]
```

Out[134]:
```
['앞으로', '.', '됐어', '.', '우리가족은', '아빠', '엄마', '오빠',
'나', '.']
```

Java LIB

In [135]:
```
처리기 = pyko.OpenKoreanTextProcessor()
```

언어적 특수성

In [136]:
```
처리기.tokenize('한국어를 처리합니다ㅋㅋ')
```

Out[136]:
```
['한국어', '를', '처리', '합니다', 'ㅋㅋ']
```

In [ ]:

Trusted | Python 3

File Edit View Insert Cell Kernel Widgets Help

Code

In [135]:
```
처리기 = pyko.OpenKoreanTextProcessor()
```

In [136]:
```
처리기.tokenize('한국어를 처리합니다ㅋㅋ')
```

Out[136]:
```
['한국어', '를', '처리', '합니다', 'ㅋㅋ']
```

In [137]:
```
reader = PlaintextCorpusReader(
    root='.', fileids='sejong_4CM00003.txt',
    word_tokenizer=처리기
)
```

↖ 넣기

processor.tokenize(Str)

X

In [139]:
```
reader.words()[20:30]
```

Out[139]:
```
['₩r₩n', '우리', '가족', '은', '아빠', '엄마', '오빠', '나', '.',
'₩r₩n']
```

In [ ]:

Notebook saved | Trusted | Python 3

File Edit View Insert Cell Kernel Widgets Help

Code

In [147]:
```
'seongjoo' in 영어사전.words()
```

Out[147]: False

In [150]:
```
단어들 = Series(t for t in tokens).str.lower()
```

In [153]:
```
고유단어들 = 단어들.drop_duplicates()
```

In [155]:
```
수록여부 = 고유단어들.isin(영어사전.words())
```

Out[155]:

0	False
1	True
2	True
3	True
4	False
5	False
6	False
7	True
8	True
9	True
12	True
13	False
14	True

Series.isin([$x_1, x_2, x_3, \cdots$])

$x_1 \rightarrow$ True

$x_{100} \rightarrow$ False

Trusted | Python 3

File Edit View Insert Cell Kernel Widgets Help

Code

```
In [150]: 단어들 = Series(t for t in tokens).str.lower()

In [153]: 고유단어들 = 단어들.drop_duplicates()

In [156]: 수록여부 = 고유단어들.isin(영어사전.words())

In [161]: 알파벳문자여부 = 고유단어들.str.isalpha()

In [162]: 고유단어들[~수록여부 & 알파벳문자여부]

Out[162]: 4          austen
          29         seemed
          36       blessings
          47          years
          63       youngest
          67      daughters
          102          died
          117      caresses
          124      supplied
          147        taylor
          150            mr
          166          fond
```

and or
True False False True

~ NOT

& AND

| or

^ xor

Trusted | Python 3

File Edit View Insert Cell Kernel Widgets Help

Code

In [166]:
```
en_stopwords = stopwords.words('english')
```

In [167]:
```
en_stopwords[:10]
```

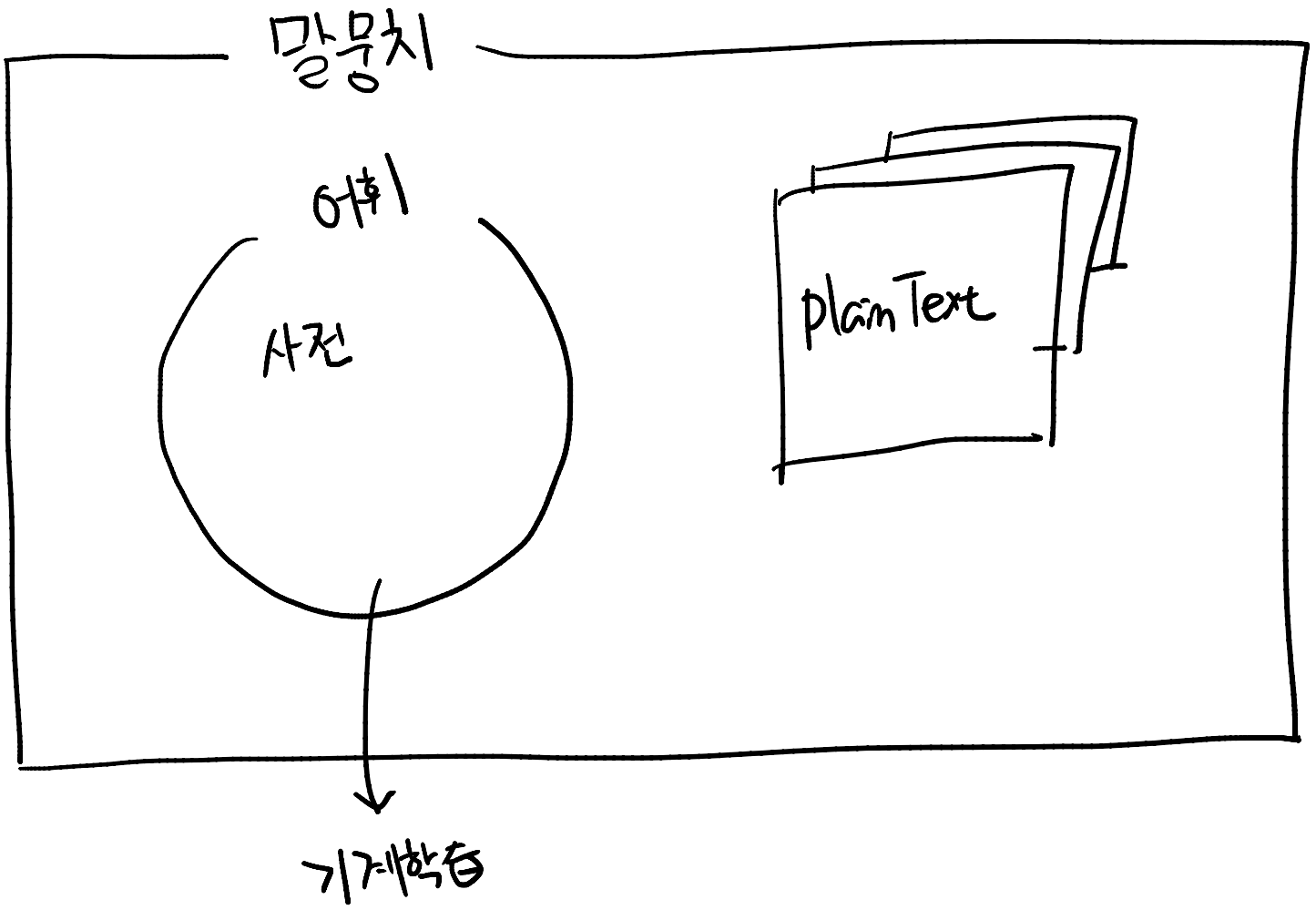
Out[167]:
```
['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'yo
u', "you're"]
```

영문자단어 & NOT

In [170]:
```
단어들[단어들.str.isalpha() & ~단어들.isin(en_stopwords)].value_coun
```

Out[170]:

mr	1153
emma	865
could	837
would	820
mrs	699
miss	599
must	567
harriet	506
much	486
said	484
one	452
weston	440
every	435


말뭉치
어휘
사전
plain Text
기계학습

Trusted | Python 3

File Edit View Insert Cell Kernel Widgets Help

Code

In [173]:
```
female_names = names.words('female.txt')
```

In [177]:
```
male_names = names.words('male.txt')
```

In [175]:
```
female_names = Series(t for t in female_names)
```

In [178]:
```
female_names[female_names.isin(male_names)]
```

Out[178]:

3	Abbey
5	Abbie
6	Abby
16	Addie
40	Adrian
45	Adrien
87	Ajay
111	Alex
119	Alexis
120	Alfie
122	Ali
138	Alix
146	Allie
154	Allyn
228	Andie