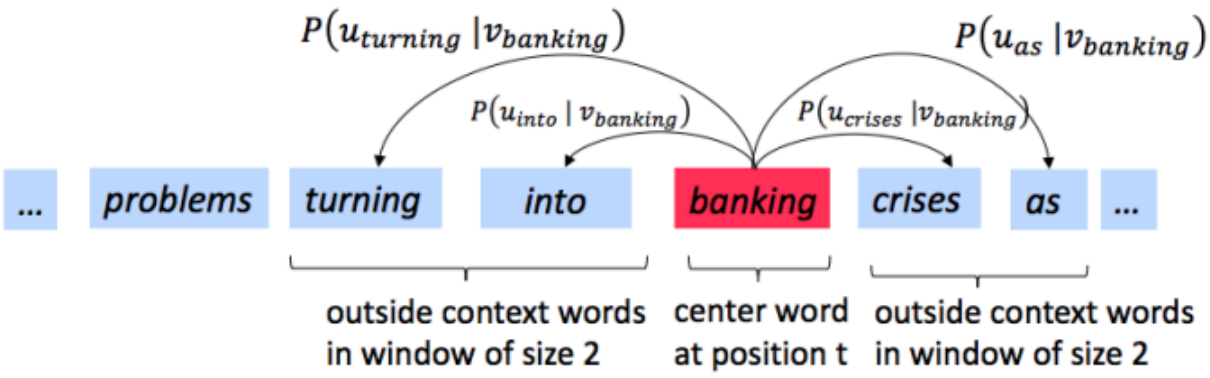
클린업 2주차 - Word2Vec for Korean



클린업 2주차에서는 Word2Vec 모델로 직접 한국어 임베딩을 생성해보겠습니다.

데이터는 위키피디아 한국어 데이터셋을 사용해볼게요!

이 데이터셋은 1.7GB 크기의 상당히 큰 데이터셋입니다.

위키피디아 데이터셋: https://ko-nlp.github.io/Korpora/en-docs/corpuslist/kowikitext.html

In [ ]:
```
import warnings
warnings.filterwarnings("ignore")
```

In [ ]:
```
import numpy as np
import os

def my_seed_everywhere(seed: int = 42):
    np.random.seed(seed) # numpy
    os.environ["PYTHONHASHSEED"] = str(seed) # os

my_seed = 42
my_seed_everywhere(my_seed)
```

데이터셋 다운로드

먼저 데이터 다운로드를 진행해줄게요. 데이터셋은 `Korpora`에서 다운로드 받을 수 있습니다

데이터 다운로드를 위해 중간에 빈칸이 뜨면 'yes'를 타이핑해주시면 됩니다

데이터 다운로드는 대략 1분 정도 걸립니다.

In [ ]:
```
from Korpora import Korpora
corpus = Korpora.load("kowikitext")
```

```
Korpora 는 다른 분들이 연구 목적으로 공유해주신 말뭉치들을
손쉽게 다운로드, 사용할 수 있는 기능만을 제공합니다.

말뭉치들을 공유해 주신 분들에게 감사드리며, 각 말뭉치 별 설명과 라이센스를 공유 드립니다.
해당 말뭉치에 대해 자세히 알고 싶으신 분은 아래의 description 을 참고,
해당 말뭉치를 연구/상용의 목적으로 이용하실 때에는 아래의 라이센스를 참고해 주시기 바랍니다.

# Description
Author : Hyunjoong Kim lovit@github
Repository : https://github.com/lovit/kowikitext
References :

한국어 위키피디아의 덤프 데이터를 바탕을 제작한 wikitext 형식의 텍스트 파일입니다.
학습 및 평가를 위하여 위키페이지 별로 train (99%), dev (0.5%), test (0.5%) 로 나뉘어져있습니다.

# License
CC-BY-SA 3.0 which kowiki dump dataset is licensed

[Korpora] Corpus `kowikitext` is already installed at C:\Users\Junho\Korpora\kowikitext\kowikitext_20200920.train.zip
[Korpora] Corpus `kowikitext` is already installed at C:\Users\Junho\Korpora\kowikitext\kowikitext_20200920.train
[Korpora] Corpus `kowikitext` is already installed at C:\Users\Junho\Korpora\kowikitext\kowikitext_20200920.test.zip
[Korpora] Corpus `kowikitext` is already installed at C:\Users\Junho\Korpora\kowikitext\kowikitext_20200920.test
[Korpora] Corpus `kowikitext` is already installed at C:\Users\Junho\Korpora\kowikitext\kowikitext_20200920.dev.zip
[Korpora] Corpus `kowikitext` is already installed at C:\Users\Junho\Korpora\kowikitext\kowikitext_20200920.dev
```

데이터 다운로드를 완료했습니다!

텍스트 전처리 (~15 mins)

이번에는 텍스트 전처리를 진행해볼게요~

`corpus`는 train, dev, test로 이루어져 있으며, 각각 `.`을 통해 접근할 수 있습니다.

In [ ]:
```
# 구조 확인하기
corpus.train[0]
```

Out[ ]:
```
SentencePair(text='외교부장\n외교부장', pair=' = 분류:중화인민공화국의 외교부장 =')
```

데이터는 튜플 `SentencePair`로 구성되어 있네요. 여기서 우리가 사용할 것은 train 데이터의 text이므로 분리해줍시다.

In [ ]:
```
train = []

for i in range(len(corpus.train)):
    if i % 100000 == 0:
```

```
    print(f"Split {i} 완료!")
  train.append(corpus.train[i].text)
```

```
Split 0 완료!
Split 100000 완료!
Split 200000 완료!
Split 300000 완료!
Split 400000 완료!
Split 500000 완료!
Split 600000 완료!
Split 700000 완료!
Split 800000 완료!
Split 900000 완료!
Split 1000000 완료!
Split 1100000 완료!
Split 1200000 완료!
Split 1300000 완료!
Split 1400000 완료!
Split 1500000 완료!
Split 1600000 완료!
Split 1700000 완료!
Split 1800000 완료!
Split 1900000 완료!
Split 2000000 완료!
Split 2100000 완료!
Split 2200000 완료!
Split 2300000 완료!
Split 2400000 완료!
Split 2500000 완료!
Split 2600000 완료!
Split 2700000 완료!
Split 2800000 완료!
Split 2900000 완료!
Split 3000000 완료!
Split 3100000 완료!
Split 3200000 완료!
Split 3300000 완료!
Split 3400000 완료!
Split 3500000 완료!
```

더 다루기 쉬운 리스트로 옮겨주었습니다.

랜덤하게 데이터 10개만 확인해볼게요

In [ ]:
```
import numpy as np
np.random.seed(42)

idx = np.random.randint(0, len(train), 10)
for i in idx:
    print(f"---------------------- {i}번째 데이터 ----------------------")
    print(train[i])
    print()
```

---------- 2219110번째 데이터 ----------
전장(mm) : 4,570(1991년~1995년)
전폭(mm) : 1,720(1991년~1995년)
전고(mm) : 1,405
축거(mm) : 2,520
윤거(전, mm) : 1,440
윤거(후, mm) : 1,430
승차정원 : 5명
연료탱크용량(l) : 60
변속기 : 수동 5단/자동 4단
서스펜션(전/후) : 맥퍼슨 스트럿/맥퍼슨 스트럿
제동장치(전/후) : 벤틸레이티드 디스크/드럼
구동형식 : 전륜구동
!구분
!2.0 DOHC
!2.0 SOHC
!1.8 SOHC
!1.8 LPG
!2.0 디젤
!연료
!배기량(cc)
!최고출력(ps/rpm)
!최대토크(kg*m/rpm)
!공차중량(kg)
!연비(km/l)

---------- 2768307번째 데이터 ----------
IAAF 이진일 프로필

---------- 2229084번째 데이터 ----------
정화(征和)는 전한(前漢) 무제(武帝)의 열번째 연호이다.
정화(政和)는 북송(北宋) 휘종(徽宗)의 네 번째 연호이다.
쇼와(正和)는 일본의 연호(元号) 중 하나이다.
조와(貞和)는 일본 남북조 시대의 연호(元号) 중 하나이다.

---------- 2356330번째 데이터 ----------
2010년 월드컵 선수 명단 표기를 보니 안드러가 아닌 안드레 였습니다. 문서 이동을 잘못했는데 복구가 안되는군요. --BSİ (Talk) 2013년 2월 23일 (토) 14:09 (KST)
: --Puzzlet Chung (토론) 2013년 2월 23일 (토) 15:07 (KST)

---------- 1692743번째 데이터 ----------
Tomoko Ninomiya's Web : 니노미야 도모코 공식 사이트
Kiss on Line : 고단샤 《키스》공식 사이트
노다메 왕국 : 고단샤 포털 사이트 〈MouRa(모우라)〉
애니메이션 "노다메 칸타빌레" 공식 사이트
노다메 칸타빌레 DS 공식 사이트

---------- 110268번째 데이터 ----------
2010년 MBC 《스타 오디션 위대한 탄생》 중국 예선편
2011년 12월 SBS 《가요대전》 (루한, 타오)

---------- 732180번째 데이터 ----------
비잔티움 황제 연대표
콤네노스 왕조

---------- 3200614번째 데이터 ----------
경희대는 5월 19일 인터넷을 뜨겁게 달궜던 이른바 '경희대 패륜녀'의 신원을 파악했다고 밝혔다. 학교 측은 이날 "화장실과 여학생 휴게실 복도의 CCTV 화면 등을 통해 신원을 확보했다"며 "해당 인물에게도 연락이 닿았으며, 정말 맞는지 본인을 상대로 최종적으로 사건을 확인할 계획"이라고 밝혔다. 이어 "오늘 중으로 학교 측의 공식 입장 발표가 있을 것이라는 일부 보도는 과장된 것"이라며 "현재 확실한 것은 신원을 파악했다는 정도"라고 재확인했다.
그러나 네티즌들의 비판 여론은 쉽게 가라앉지 않았다. 6월 4일에는 연세대학교에서 남학생이 미화원을 폭행한 사실까지 확인되면서 논란은 상당기간 진행되었다.이번엔 '연세대 패륜남'…미화원에 욕설·폭행까지 경향신문 2010년 06월 04일자
5월 20일 해당 여학생은 환경미화원을 찾아가 사과하였다.경희대 막말 여학생 '사과'…피해 환경미화원은 '용서' 경향신문 2010년 5월 20일 경희대에 의하면 해당 학생은 징계할 방침이지만 미화원 아주머니는 자식을 둔 어머니의 심정으로 징계를 원하지 않는 것으로 알려졌다. 사건은 유야무야 넘어갔다. 그러나 연세대학교 미화원 폭행 사건이 연이어 터지면서 쉽게 논란은 수그러들지 않았고 6월까지 인터넷 이슈거리가 되었다.

---------- 2234489번째 데이터 ----------
세어도

---------- 3097042번째 데이터 ----------
*

어떻게 전처리를 진행해줘야 할까요? 천천히 생각해보죠

1. 줄바꿈 제거

가장 먼저 줄바꿈을 없애주겠습니다. 이스케이프 문자 `\n`을 제거하면 됩니다.

In [ ]:
```
# 줄바꿈 전처리 예시
print("전처리 전: ", train[idx[0]])
print("전처리 후: ", train[idx[0]].replace("\n", ""))
```

```
전처리 전:  전장(mm) : 4,570(1991년~1995년)
전폭(mm) : 1,720(1991년~1995년)
전고(mm) : 1,405
축거(mm) : 2,520
윤거(전, mm) : 1,440
윤거(후, mm) : 1,430
승차정원 : 5명
연료탱크용량(l) : 60
변속기 : 수동 5단/자동 4단
서스펜션(전/후) : 맥퍼슨 스트럿/맥퍼슨 스트럿
제동장치(전/후) : 벤틸레이티드 디스크/드럼
구동형식 : 전륜구동
!구분
!2.0 DOHC
!2.0 SOHC
!1.8 SOHC
!1.8 LPG
!2.0 디젤
!연료
!배기량(cc)
!최고출력(ps/rpm)
!최대토크(kg*m/rpm)
!공차중량(kg)
!연비(km/l)
전처리 후:  전장(mm) : 4,570(1991년~1995년)전폭(mm) : 1,720(1991년~1995년)전고(mm) : 1,405축거(mm) : 2,520윤거(전, mm) : 1,440윤거(후, mm) : 1,430승차정원 : 5명연료탱크용량(l) : 60변속기 : 수동 5단/자동 4단서스펜션(전/후) : 맥퍼슨 스트럿/맥퍼슨 스트럿제동장치(전/후) : 벤틸레이티드 디스크/드럼구동형식 : 전륜구동!구분!2.0 DOHC!2.0 SOHC!1.8 SOHC!1.8 LPG!2.0 디젤!연료!배기량(cc)!최고출력(ps/rpm)!최대토크(kg*m/rpm)!공차중량(kg)!연비(km/l)
```

In [ ]:
```python
# 줄바꿈 제거
train = [i.replace("\n", "") for i in train]

np.random.seed(42)

idx = np.random.randint(0, len(train), 10)
for i in idx:
    print(f"---------------------- {i}번째 데이터 ----------------------")
    print(train[i])
    print()
```

```
---------------------- 2219110번째 데이터 ----------------------
전장(mm) : 4,570(1991년~1995년)전폭(mm) : 1,720(1991년~1995년)전고(mm) : 1,405축거(mm) : 2,520윤거(전, mm) : 1,440윤거(후, mm) : 1,430승차정원 : 5명연료탱크용량(l) : 60변속기 : 수동 5단/자동 4단서스펜션(전/후) : 맥퍼슨 스트럿/맥퍼슨 스트럿제동장치(전/후) : 벤틸레이티드 디스크/드럼구동형식 : 전륜구동!구분!2.0 DOHC!2.0 SOHC!1.8 SOHC!1.8 LPG!2.0 디젤!연료!배기량(cc)!최고출력(ps/rpm)!최대토크(kg*m/rpm)!공차중량(kg)!연비(km/l)

---------------------- 2768307번째 데이터 ----------------------
IAAF 이진일 프로필

---------------------- 2229084번째 데이터 ----------------------
정화(征和)는 전한(前漢) 무제(武帝)의 열번째 연호이다.정화(政和)는 북송(北宋) 휘종(徽宗)의 네 번째 연호이다.쇼와(正和)는 일본의 연호(元号) 중 하나이다.조와(貞和)는 일본 남북조 시대의 연호(元号) 중 하나이다.

---------------------- 2356330번째 데이터 ----------------------
2010년 월드컵 선수 명단 표기를 보니 안드러가 아닌 안드레 였습니다. 문서 이동을 잘못했는데 복구가 안되는군요. --<span style="color:#191970"><span style="font-size: 13pt; font-family: 'Lucida Handwriting', cursive;">BSİ</span></span> (<span style="color:#008B8B">Talk</span>) 2013년 2월 23일 (토) 14:09 (KST):  --Puzzlet Chung (토론) 2013년 2월 23일 (토) 15:07 (KST)

---------------------- 1692743번째 데이터 ----------------------
Tomoko Ninomiya's Web : 니노미야 도모코 공식 사이트Kiss on Line : 고단샤 《키스》공식 사이트노다메 왕국 : 고단샤 포털 사이트 〈MouRa(모우라)〉애니메이션 “노다메 칸타빌레” 공식 사이트노다메 칸타빌레 DS 공식 사이트

---------------------- 110268번째 데이터 ----------------------
2010년 MBC 《스타 오디션 위대한 탄생》 중국 예선편2011년 12월 SBS 《가요대전》 (루한, 타오)

---------------------- 732180번째 데이터 ----------------------
비잔티움 황제 연대표콤네노스 왕조

---------------------- 3200614번째 데이터 ----------------------
경희대는 5월 19일 인터넷을 뜨겁게 달궜던 이른바 '경희대 패륜녀'의 신원을 파악했다고 밝혔다. 학교 측은 이날 "화장실과 여학생 휴게실 복도의 CCTV 화면 등을 통해 신원을 확보했다"며 "해당 인물에게도 연락이 닿았으며, 정말 맞는지 본인을 상대로 최종적으로 사건을 확인할 계획"이라고 밝혔다. 이어 "오늘 중으로 학교 측의 공식 입장 발표가 있을 것이라는 일부 보도는 과장된 것"이라며 "현재 확실한 것은 신원을 파악했다는 정도"라고 재확인했다.그러나 네티즌들의 비판 여론은 쉽게 가라앉지 않았다. 6월 4일에는 연세대학교에서 남학생이 미화원을 폭행한 사실까지 확인되면서 논란은 상당기간 진행되었다.이번엔 '연세대 패륜남'…미화원에 욕설·폭행까지 경향신문 2010년 06월 04일자5월 20일 해당 여학생은 환경미화원을 찾아가 사과하였다.경희대 막말 여학생 '사과'…피해 환경미화원은 '용서' 경향신문 2010년 5월 20일 경희대에 의하면 해당 학생은 징계할 방침이지만 미화원 아주머니는 자식을 둔 어머니의 심정으로 징계를 원하지 않는 것으로 알려졌다. 사건은 유야무야 넘어갔다. 그러나 연세대학교 미화원 폭행 사건이 연이어 터지면서 쉽게 논란은 수그러들지 않았고 6월까지 인터넷 이슈거리가 되었다.

---------------------- 2234489번째 데이터 ----------------------
세어도

---------------------- 3097042번째 데이터 ----------------------
*
```

줄바꿈이 잘 제거된 것을 볼 수 있습니다.

2. html 및 css 태그 제거하기

두 번째로는 html 및 css 태그를 제거해줄게요.

이런 태그는 웹페이지를 꾸미기 위한 태그일 뿐 기사와 관련된 어떠한 의미도 가지고 있지 않습니다.

예를 들어 위의 예시에서 <₩span> 등의 태그입니다.

In [ ]:
```python
# html & css 태그 제거 전처리 예시
import re

print("전처리 전: ", train[idx[3]])
to_clean = re.compile("<.*?>")
print("전처리 후: ", re.sub(to_clean, "", train[idx[3]]))
```

```
전처리 전:  2010년 월드컵 선수 명단 표기를 보니 안드러가 아닌 안드레 였습니다. 문서 이동을 잘못했는데 복구가 안되는군요. --<span style="color:#191970"><span sty
le="font-size: 13pt; font-family: 'Lucida Handwriting', cursive;">BSİ</span></span> (<span style="color:#008B8B">Talk</span>) 2013년 2월 23일 (토) 14:09 (KS
T):  --Puzzlet Chung (토론) 2013년 2월 23일 (토) 15:07 (KST)
전처리 후:  2010년 월드컵 선수 명단 표기를 보니 안드러가 아닌 안드레 였습니다. 문서 이동을 잘못했는데 복구가 안되는군요. --BSİ (Talk) 2013년 2월 23일 (토) 14:09
(KST):  --Puzzlet Chung (토론) 2013년 2월 23일 (토) 15:07 (KST)
```

In [ ]:
```python
# html & css 태그 제거
to_clean = re.compile("<.*?>")
train = [re.sub(to_clean, "", i) for i in train]

np.random.seed(42)

idx = np.random.randint(0, len(train), 10)
for i in idx:
    print(f"---------------------- {i}번째 데이터 ----------------------")
    print(train[i])
    print()
```

```
---------------------- 2219110번째 데이터 ----------------------
전장(mm) : 4,570(1991년~1995년)전폭(mm) : 1,720(1991년~1995년)전고(mm) : 1,405축거(mm) : 2,520윤거(전, mm) : 1,440윤거(후, mm) : 1,430승차정원 : 5명연료탱크용
량(l) : 60변속기 : 수동 5단/자동 4단서스펜션(전/후) : 맥퍼슨 스트럿/맥퍼슨 스트럿제동장치(전/후) : 벤틸레이티드 디스크/드럼구동형식 : 전륜구동!구분!2.0 DOHC!2.0
SOHC!1.8 SOHC!1.8 LPG!2.0 디젤!연료!배기량(cc)!최고출력(ps/rpm)!최대토크(kg*m/rpm)!공차중량(kg)!연비(km/l)

---------------------- 2768307번째 데이터 ----------------------
IAAF 이진일 프로필

---------------------- 2229084번째 데이터 ----------------------
정화(征和)는 전한(前漢) 무제(武帝)의 열번째 연호이다.정화(政和)는 북송(北宋) 휘종(徽宗)의 네 번째 연호이다.쇼와(正和)는 일본의 연호(元号) 중 하나이다.조와(貞和)
는 일본 남북조 시대의 연호(元号) 중 하나이다.

---------------------- 2356330번째 데이터 ----------------------
2010년 월드컵 선수 명단 표기를 보니 안드러가 아닌 안드레 였습니다. 문서 이동을 잘못했는데 복구가 안되는군요. --BSİ (Talk) 2013년 2월 23일 (토) 14:09 (KST):  --P
uzzlet Chung (토론) 2013년 2월 23일 (토) 15:07 (KST)

---------------------- 1692743번째 데이터 ----------------------
Tomoko Ninomiya's Web : 니노미야 도모코 공식 사이트Kiss on Line : 고단샤 《키스》공식 사이트노다메 왕국 : 고단샤 포털 사이트 〈MouRa(모우라)〉애니메이션 “노다메
칸타빌레” 공식 사이트노다메 칸타빌레 DS 공식 사이트

---------------------- 110268번째 데이터 ----------------------
2010년 MBC 《스타 오디션 위대한 탄생》 중국 예선편2011년 12월 SBS 《가요대전》 (루한, 타오)

---------------------- 732180번째 데이터 ----------------------
비잔티움 황제 연대표콤네노스 왕조

---------------------- 3200614번째 데이터 ----------------------
경희대는 5월 19일 인터넷을 뜨겁게 달궜던 이른바 '경희대 패륜녀'의 신원을 파악했다고 밝혔다. 학교 측은 이날 "화장실과 여학생 휴게실 복도의 CCTV 화면 등을 통해 신
원을 확보했다"며 "해당 인물에게도 연락이 닿았으며, 정말 맞는지 본인을 상대로 최종적으로 사건을 확인할 계획"이라고 밝혔다. 이어 "오늘 중으로 학교 측의 공식 입장
발표가 있을 것이라는 일부 보도는 과장된 것"이라며 "현재 확실한 것은 신원을 파악했다는 정도"라고 재확인했다.그러나 네티즌들의 비판 여론은 쉽게 가라앉지 않았다. 6
월 4일에는 연세대학교에서 남학생이 미화원을 폭행한 사실까지 확인되면서 논란은 상당기간 진행되었다.이번엔 '연세대 패륜남'…미화원에 욕설·폭행까지 경향신문 2010년
06월 04일자5월 20일 해당 여학생은 환경미화원을 찾아가 사과하였다.경희대 막말 여학생 '사과'…피해 환경미화원은 '용서' 경향신문 2010년 5월 20일 경희대에 의하면 해
당 학생은 징계할 방침이지만 미화원 아주머니는 자식을 둔 어머니의 심정으로 징계를 원하지 않는 것으로 알려졌다. 사건은 유야무야 넘어갔다. 그러나 연세대학교 미화원
폭행 사건이 연이어 터지면서 쉽게 논란은 수그러들지 않았고 6월까지 인터넷 이슈거리가 되었다.

---------------------- 2234489번째 데이터 ----------------------
세어도

---------------------- 3097042번째 데이터 ----------------------
*
```

역시나 잘 제거되었네요

3. 특수 문자 및 영어 이외의 외국어 문자 제거

세 번째로는 특수 문자와 영어 이외의 외국어 문자들을 제거해줄게요. 이 문자들은 의미 파악에 필요가 없습니다.

아래의 정규식에서 의미하는 바는 다음과 같습니다.

- `\uAC00-\uD7A3` : 모든 한글 음절(가-힣)
- `0-9` : 모든 숫자
- `a-z` : 모든 알파벳 소문자
- `A-Z` : 모든 알파벳 대문자
- `\s` : 공백 이스케이프 문자

이 과정은 대략 40초 정도 걸립니다.

In [ ]:
```python
# html & css 태그 제거 전처리 예시
import re

print("전처리 전: ", train[idx[2]])
to_clean = re.compile("[^\uAC00-\uD7A30-9a-zA-Z\s]")
print("전처리 후: ", re.sub(to_clean, "", train[idx[2]]))
```

```
전처리 전:  정화(征和)는 전한(前漢) 무제(武帝)의 열번째 연호이다.정화(政和)는 북송(北宋) 휘종(徽宗)의 네 번째 연호이다.쇼와(正和)는 일본의 연호(元号) 중 하나이
다.조와(貞和)는 일본 남북조 시대의 연호(元号) 중 하나이다.
전처리 후:  정화는 전한 무제의 열번째 연호이다정화는 북송 휘종의 네 번째 연호이다쇼와는 일본의 연호 중 하나이다조와는 일본 남북조 시대의 연호 중 하나이다
```

In [ ]:
```python
# 특수 문자 및 영어 이외의 외국어 제외
import re

to_clean = re.compile("[^\uAC00-\uD7A30-9a-zA-Z\s]")
train = [re.sub(to_clean, "", i) for i in train]

np.random.seed(42)

idx = np.random.randint(0, len(train), 10)
```

```
for i in idx:
    print(f"---------------------- {i}번째 데이터 ----------------------")
    print(train[i])
    print
```

```
---------------------- 2219110번째 데이터 ----------------------
전장mm 45701991년1995년전폭mm 17201991년1995년전고mm 1405축거mm 2520윤거전 mm 1440윤거후 mm 1430승차정원 5명연료탱크용량l 60변속기 수동 5단자동 4단서스펜션전후 맥퍼슨 스트럿맥퍼슨 스트럿제동장치전후 벤틸레이티드 디스크드럼구동형식 전륜구동구분20 DOHC20 SOHC18 SOHC18 LPG20 디젤연료배기량cc최고출력psrpm최대토크kgmrpm공차중량kg연비kml
---------------------- 2768307번째 데이터 ----------------------
IAAF 이진일 프로필
---------------------- 2229084번째 데이터 ----------------------
정화는 전한 무제의 열번째 연호이다정화는 북송 휘종의 네 번째 연호이다쇼와는 일본의 연호 중 하나이다조와는 일본 남북조 시대의 연호 중 하나이다
---------------------- 2356330번째 데이터 ----------------------
2010년 월드컵 선수 명단 표기를 보니 안드러가 아닌 안드레 였습니다 문서 이동을 잘못했는데 복구가 안되는군요 BS Talk 2013년 2월 23일 토 1409 KST  Puzzlet Chung 토론 2013년 2월 23일 토 1507 KST
---------------------- 1692743번째 데이터 ----------------------
Tomoko Ninomiyas Web 니노미야 도모코 공식 사이트Kiss on Line 고단샤 키스공식 사이트노다메 왕국 고단샤 포털 사이트 MouRa모우라애니메이션 노다메 칸타빌레 공식 사이트노다메 칸타빌레 DS 공식 사이트
---------------------- 110268번째 데이터 ----------------------
2010년 MBC 스타 오디션 위대한 탄생 중국 예선편2011년 12월 SBS 가요대전 루한 타오
---------------------- 732180번째 데이터 ----------------------
비잔티움 황제 연대표콤네노스 왕조
---------------------- 3200614번째 데이터 ----------------------
경희대는 5월 19일 인터넷을 뜨겁게 달궜던 이른바 경희대 패륜녀의 신원을 파악했다고 밝혔다 학교 측은 이날 화장실과 여학생 휴게실 복도의 CCTV 화면 등을 통해 신원을 확보했다며 해당 인물에게도 연락이 닿았으며 정말 맞는지 본인을 상대로 최종적으로 사건을 확인할 계획이라고 밝혔다 이어 오늘 중으로 학교 측의 공식 입장 발표가 있을 것이라는 일부 보도는 과장된 것이라며 현재 확실한 것은 신원을 파악했다는 정도라고 재확인했다그러나 네티즌들의 비판 여론은 쉽게 가라앉지 않았다 6월 4일에는 연세대학교에서 남학생이 미화원을 폭행한 사실까지 확인되면서 논란은 상당기간 진행되었다이번엔 연세대 패륜남미화원에 욕설폭행까지 경향신문 2010년 06월 04일자5월 20일 해당 여학생은 환경미화원을 찾아가 사과하였다경희대 막말 여학생 사과피해 환경미화원은 용서 경향신문 2010년 5월 20일 경희대에 의하면 해당 학생은 징계할 방침이지만 미화원 아주머니는 자식을 둔 어머니의 심정으로 징계를 원하지 않는 것으로 알려졌다 사건은 유야무야 넘어갔다 그러나 연세대학교 미화원 폭행 사건이 연이어 터지면서 쉽게 논란은 수그러들지 않았고 6월까지 인터넷 이슈거리가 되었다
---------------------- 2234489번째 데이터 ----------------------
세어도
---------------------- 3097042번째 데이터 ----------------------
```

4. 짧은 텍스트 제거

어떤 텍스트들은 너무 짧기 때문에 필요가 없습니다. 어느 정도가 적당할 지, 위의 샘플 텍스트들을 보면서 판단해보겠습니다.

전처리가 꽤 진행되었으니 30개를 보도록 할게요.

In [ ]:
```
np.random.seed(42)

idx = np.random.randint(0, len(train), 30)
for i in idx:
    print(f"---------------------- {i}번째 데이터, 길이 = {len(train[i])} ----------------------")
    print(train[i])
    print()
```

---------- 2219110번째 데이터, 길이 = 241 ----------
전장mm 45701991년1995년전폭mm 17201991년1995년전고mm 1405축거mm 2520윤거전 mm 1440윤거후 mm 1430승차정원 5명연료탱크용량l 60변속기 수동 5단자동 4단서스펜션전후 맥퍼슨 스트럿맥퍼슨 스트럿제동장치전후 벤틸레이티드 디스크드럼구동형식 전륜구동구분20 DOHC20 SOHC18 SOHC18 LPG20 디젤연료배기량cc최고출력psrpm최대토크kgmrpm공차중량kg연비kml

---------- 2768307번째 데이터, 길이 = 12 ----------
IAAF 이진일 프로필

---------- 2229084번째 데이터, 길이 = 80 ----------
정화는 전한 무제의 열번째 연호이다정화는 북송 휘종의 네 번째 연호이다쇼와는 일본의 연호 중 하나이다조와는 일본 남북조 시대의 연호 중 하나이다

---------- 2356330번째 데이터, 길이 = 136 ----------
2010년 월드컵 선수 명단 표기를 보니 안드러가 아닌 안드레 였습니다 문서 이동을 잘못했는데 복구가 안되는군요 BS Talk 2013년 2월 23일 토 1409 KST Puzzlet Chung 토론 2013년 2월 23일 토 1507 KST

---------- 1692743번째 데이터, 길이 = 129 ----------
Tomoko Ninomiyas Web 니노미야 도모코 공식 사이트Kiss on Line 고단샤 키스공식 사이트노다메 왕국 고단샤 포털 사이트 MouRa모우라애니메이션 노다메 칸타빌레 공식 사이트노다메 칸타빌레 DS 공식 사이트

---------- 110268번째 데이터, 길이 = 54 ----------
2010년 MBC 스타 오디션 위대한 탄생 중국 예선편2011년 12월 SBS 가요대전 루한 타오

---------- 732180번째 데이터, 길이 = 18 ----------
비잔티움 황제 연대표콤네노스 왕조

---------- 3200614번째 데이터, 길이 = 591 ----------
경희대는 5월 19일 인터넷을 뜨겁게 달궜던 이른바 경희대 패륜녀의 신원을 파악했다고 밝혔다 학교 측은 이날 화장실과 여학생 휴게실 복도의 CCTV 화면 등을 통해 신원을 확보했다며 해당 인물에게도 연락이 닿았으며 정말 맞는지 본인을 상대로 최종적으로 사건을 확인할 계획이라고 밝혔다 이어 오늘 중으로 학교 측의 공식 입장 발표가 있을 것이라는 일부 보도는 과장된 것이라며 현재 확실한 것은 신원을 파악했다는 정도라고 재확인했다그러나 네티즌들의 비판 여론은 쉽게 가라앉지 않았다 6월 4일에는 연세대학교에서 남학생이 미화원을 폭행한 사실까지 확인되면서 논란은 상당기간 진행되었다이번엔 연세대 패륜남미화원에 욕설폭행까지 경향신문 2010년 06월 04일자5월 20일 해당 여학생은 환경미화원을 찾아가 사과하였다경희대 막말 여학생 사과피해 환경미화원은 용서 경향신문 2010년 5월 20일 경희대에 의하면 해당 학생은 징계할 방침이지만 미화원 아주머니는 자식을 둔 어머니의 심정으로 징계를 원하지 않는 것으로 알려졌다 사건은 유야무야 넘어갔다 그러나 연세대학교 미화원 폭행 사건이 연이어 터지면서 쉽게 논란은 수그러들지 않았고 6월까지 인터넷 이슈거리가 되었다

---------- 2234489번째 데이터, 길이 = 3 ----------
세어도

---------- 3097042번째 데이터, 길이 = 0 ----------

---------- 1570006번째 데이터, 길이 = 0 ----------

---------- 2003274번째 데이터, 길이 = 8 ----------
조문탁웅흔흔상니

---------- 1136074번째 데이터, 길이 = 56 ----------
더 가이즈 2016년 고윤발 역살랑살랑 기무라상 2015년 멀티 역뉴보잉보잉 2009년 순성 역

---------- 3009908번째 데이터, 길이 = 445 ----------
그가 영등포약사회 회장을 하고 있던 어느 날이었다 평소 알고 지내던 선배 한 명이 찾아와 구주제약이라는 회사를 약업계에 있는 몇 사람과 동업하고 있다며 약을 좀 팔아달라는 부탁을 했다 그는 차도 없이 혼자 영업을 하던 그 선배에게 차를 알선해주고 함께 약을 들고 영업을 돕기도 했다 그러다가 출자만 하는 형식으로 그는 구주제약의 동업에 참여하게 되었는데 1년 뒤 회사가 잘 안되어 설립 4년 만에 폐업 위기에 놓이게 되었다 1978년 그는 폐업 직전에 놓인 회사를 인수하여 제약회사 CEO로서 발돋움을 하게 된다 하지만 이전까지 주먹구구식으로 운영되던 부실한 회사를 다시 정상궤도에 올려놓기까지는 운영 중이던 3개의 약국을 하나씩 정리하며 메꿔야 하는 경제적인 어려움을 겪기도 했다 구주제약은 2013년 기준 직원 300명에 연 매출 600억 원이라는 성장을 기록하며 꾸준한 성장세를 보이고 있다

---------- 2272355번째 데이터, 길이 = 4 ----------
내용주

---------- 1239911번째 데이터, 길이 = 10 ----------
여자쿠바단거리달리기

---------- 278167번째 데이터, 길이 = 0 ----------

---------- 2138242번째 데이터, 길이 = 0 ----------

---------- 329365번째 데이터, 길이 = 61 ----------
이시가메 아키라 1985년 5월 20일 는 일본의 축구 선수이다 오미야 아르디자 더스파 구사쓰에서 활동하였다

---------- 3210548번째 데이터, 길이 = 918 ----------
로젠 로젠 메이든 시리즈를 만든 수수께끼의 천재 인형사 인형들에게는 아버님이라고 불리고 있다궁극의 소녀 앨리스를 만들기 위해 연금술로 로자 미스티카를 생성해 로젠 메이든을 만들었지만 아무도 앨리스가 되지 못한 것에 비관해 자취를 감춘다 인형사가 되기 전에는 철학자였으며 그 후 연금술로 불로불사를 얻어당대 시기에 그는 생제르맹 백작이라고 불리었으며 로젠 크로이츠와 무슨 관계가 있지 않았을까 의심이 가지만 진상은 모른다 이름이나 직함을 세월에 따라 수시로 바꾸어 몇십 년 몇백 년 동안 살아온 것으로 보인다 신쿠는 이름이나 모습은 아버님에게 중요하지 않다고 하였다지금도 n필드 어디엔가 있는 것 같지만 앨리스가 된 로젠 메이든 밖에 만날 생각밖에 없다고 한다 엔쥬 성우 오노 다이스케 최지훈애니메이션 오리지널 캐릭터로 2기에 등장한다 시라사키와 함께 인형 가게 영업을 한 인형사 청년 언제나 과묵하고 공방에서 로젠 메이든과 같이 태엽으로 움직이는 인형을 만들고 있다 앨리스 게임을 원하는 로젠의 의도를 이해할 수 없다고 한 준에게 인형사의 심정을 전했다그의 정체는 자칭 로젠의 제자로 스승을 뛰어넘는 인형을 만들기 위해 바라스이쇼를 만들어 신쿠 일행에게는 자신이 로젠이라고 속이고 앨리스 게임을 하도록 유도한 원흉 중 한 사람이다 인형사로서의 실력은 가히 로젠에 필적할 만한 정도로 그가 만든 바라스이쇼는 신쿠 일행과의 싸움에서 이겼지만 결국 바라스이쇼와 함께 빛에 휩싸여 소멸했다특별편에서 19세기 후반의 영국에서 시라사키와 함께 로젠 메이든을 찾으러 다녔던 장면이 나와 로젠과 같이 연금술이나 다른 무엇인가로 불로불사의 힘을 얻은 모양이다사실 엔쥬는 애니판 1기의 흥행으로 2기의 방영이 갑작스럽게 결정되자 애니판 제작자들이 진짜 로젠 메이든 7번 째 인형인 키리키쇼를 본따서 바라스이쇼를 만들어낸 인물로 급조한 캐릭터라는 평가가 많다

---------- 787201번째 데이터, 길이 = 0 ----------

---------- 1370455번째 데이터, 길이 = 0 ----------

---------- 1766891번째 데이터, 길이 = 0 ----------

---------- 327069번째 데이터, 길이 = 57 ----------
생일 1979년 2월 15일닉네임 prepix HAW춤 스타일 재즈와 힙합이 섞인 프리스타일단장 안무가

---------- 1825573번째 데이터, 길이 = 64 ----------
형 류승완 1973년 12월 15일 배우자 1990년 슬로바키아인 10살 연하딸 2020년 6월 20일

---------- 3344769번째 데이터, 길이 = 331 ----------
토론에 참여하는 사건 당사자 및 허가받은 제3자는 자신의 주장을 뒷받침하는 증거를 제시하는 것이 좋습니다 되도록이면 각 문서의 관련 역사 항목의 링크를 걸어주는 것이 중재위원의 직관적인 이해를 돕기 때문에 도움이 됩니다 빠른 이해를 위하여 주장은 간략하게 증거물도 축약하여 작성하는 것이 좋습니다 1000자의 비난보다는 5개의 잘 추려진 증거가 더욱 설득력 있을 수 있습니다당사자들은 요청 양식에 자신의 사용자명이 적힌 정해진 문단에만 주장과 증거를 작성할 수 있습니다 상대방의 증거나 주장을 훼손해서는 안됩니다 상대방의 주장과 증거에 대한 반박 또한 자신에게 정해진 문단에만 제시해야 합니다

```
---------------------- 791743번째 데이터, 길이 = 491 ----------------------
불교 일반에서는 축생뿐 아니라 모든 유정을 출생 형태에 따라 난생태생습생화생의 4생으로 분류하기도 한다난생은 알껍질로부터 생겨나는 유정류를 말하는데 축생의 경우 거위공작앵무새기러기 등과 같은 조류가 난생에 해당한다태생은 탯집으로부터 생겨나는 유정류를 말하는데 축생의 경우 코끼리말소돼지양나귀 등과 같은 포유류가 태생에 해당한다습생은 습기로부터 생겨나는 유정류를 말하는데 축생의 경우 벌레누에나방모기노래기지네 등과 같은 벌레와 곤충류가  습생에 해당한다화생은 알껍질탯집습기에 의지하지 않고 생겨나는 유정류를 말하는데 감관을 모두 갖추어 수족이나 마디마디의 결함 없이 신체가 단박에 생겨나는 것을 말한다 없다가 홀연히 있기 때문에 화생이라 한다 구사론 제8권에 따르면 축생의 경우 용이나 게로다 가루다 즉 금시조 등이 화생에 해당한다 이에 비해 장아함경 제19권 30 세기경 5 용조품에 따르면 용과 게로다금시조는 화생만 있는 것이 아니며 난생태생습생화생의 네 출생 유형 모두가 있다

---------------------- 103355번째 데이터, 길이 = 407 ----------------------
아느 시잉성우  히사카와 아야일  정혜원한라그의 어머니 누군가에 의해 아카츠키로 끌려갔다 소식 및 생사는 확인 불명 라그에 의하면 여제와 얼굴이 비슷한것 같다사브리나 메리성우  타니 이쿠코일  성선녀한현재는 캠벨 리트스에서 살고 있다 예전에 라그의 집의 부근에서 살고 있던 여성 고슈의 편지로써 배달된 라그를 키웠다 라그에게는 또 한사람의 어머니로 존경받고 있다네리네로 남매성우  나마타메 히토미네리 마츠오카 유키일네로  정혜옥한네로지기 페퍼와는 의형제 관계 누나 네리는 키리에의 마을에 살고 있다 아픈 네로는 지기에게 편지를 남겨 사망했고 네리는 지기가 네로를 버렸다고 생각하고 레터비가된 지기를 원망하고 있었지만 라그의 심탄에 의해서 네로와 지기의 진심을 알아 오해가 풀린다 그 다음부터 지기와 네리의 교류가 부활한 모양이다

---------------------- 3381524번째 데이터, 길이 = 429 ----------------------
상관 분석  Correlation analysis 또는 상관관계 또는 상관은 확률론과 통계학에서 두 변수간에 어떤 선형적 또는 비선형적 관계를 갖고 있는지를 분석하는 방법이다우리말샘 상관 상관관계 상관분석 등 두 변수는 서로 독립적인 관계이거나 상관된 관계일 수 있으며 이때 두 변수간의 관계의 강도를 상관관계Correlation Correlation coefficient라 한다 상관분석에서는 상관관계의 정도를 나타내는 단위로 모상관계수로 를 사용하며 표본 상관 계수로 r 을 사용한다상관관계의 정도를 파악하는 상관 계수 Correlation coefficient는 두 변수간의 연관된 정도를 나타낼 뿐 인과관계를 설명하는 것은 아니다 두 변수간에 원인과 결과의 인과관계가 있는지에 대한 것은 회귀분석을 통해 인과관계의 방향 정도와 수학적 모델을 확인해 볼 수 있다

---------------------- 1262752번째 데이터, 길이 = 565 ----------------------
CD  생명은 아름다워  태양 노크  지금 말하고 싶은 누군가가 있어  봄망초 필 무렵  계기  태양에게 설득당해  욕망의 리인카네이션  슬픔 잊는 법 Threefold choice 작사 아키모토 야스시  작곡편곡 후루카와 타카히로  저체온의 키스 작사 아키모토 야스시  작곡 나카타니 아츠코  편곡 타노우에 요이치  머나먼 부탄 작사 아키모토 야스시  작곡 츠키다 타다시  편곡 haj  포피팟파파 작사 아키모토 야스시  작곡편곡 Akira Sunset haj  제복을 벗고 작별을 작사 아키모토 야스시  작곡편곡 후루카와 타카히로  우울과 풍선껌 작사 아키모토 야스시  작곡 HIROTOMO DrLilcom  편곡 APAZZI  재기 중 작사 아키모토 야스시  작곡 후쿠다 타카시  편곡 TATOO  노기자카의 시DVD  사람은 왜 달리는가  성급한 달팽이  왼쪽 가슴의 용기  몇 번째 푸른 하늘인가  당케 쇤  그런 바보같은  마음의 약 13  13일의 금요일  선풍기  그 앞의 출구  말 없는 라이온  내가 안 가면 누가 가  데코핑  한숨의 메서드
```

음... 적어도 길이가 50은 되어야 하나의 문장이라고 여길 수 있겠군요.

길이가 50 미만인 텍스트들은 모두 제거해주겠습니다.

In [ ]:
```python
# 길이가 50 미만인 텍스트 모두 제거
print("전처리 전: 길이 =", len(train))

for i in range(len(train)):
    if len(train[i]) < 50:
        train[i] = ""

train = [i for i in train if i != ""]

print("전처리 후: 길이 =", len(train))
```

```
전처리 전: 길이 = 3510734
전처리 후: 길이 = 1790990
```

데이터가 대략 절반으로 줄었네요. 랜덤하게 30개의 데이터를 뽑아서 확인해보겠습니다.

In [ ]:
```python
np.random.seed(42)

idx = np.random.randint(0, len(train), 30)
for i in idx:
    print(f"---------------------- {i}번째 데이터, 길이 = {len(train[i])} ----------------------")
    print(train[i])
    print()
```

---------- 121958번째 데이터, 길이 = 102 ----------
 width7 scopecol 연도 width34 scopecol 작품 width34 scopecol 부문 width9 scopecol 결과 width9 scopecol 출처

---------- 671155번째 데이터, 길이 = 110 ----------
사우스조지아 사우스샌드위치 제도의 문장은 1985년 영토가 창설되면서 부여되었다 문장은 사우스조지아섬에있는 순록 두 무리의 순록이다 표어는 Leo Terram Propriam Protegat이다국장

---------- 131932번째 데이터, 길이 = 55 ----------
본점 충청남도 천안시 서북구 직산읍 거리막길 33서울지점 서울특별시 중구 정동길 14 정동 오송빌딩

---------- 1414414번째 데이터, 길이 = 432 ----------
 앤티가 바부다세인트존스 VC 버드 국제공항 아루바오라녜스타트 퀸 비어트릭스 국제공항 바하마프리포트 그랜드 바하마 국제공항나소 린든 핀들링 국제공항 바베이도스 그레이트 레이 아담스 국제공항 케이맨 제도그랜드 케이먼 오웬 로버츠 국제공항 도미니카 공화국푼타카나 푼타카나 국제공항산티아고데로스카바예로스 시바오 국제공항 산토도밍고 라스 아메리카 국제공항 자메이카몬티 상스터 국제공항 푸에르토리코산후안 루이스 무노즈 마린 국제공항 세인트키츠네비스세인트 키츠 로버트 L 브래드쇼 국제공항 세인트루시아뷰포트 하와노리아 국제공항 신트마르턴신트마르턴 프린세스 줄리아나 국제공항 터크스 케이커스 제도프로비던셜스 프로비던셜스 국제공항 터크스 케이커스 제도프로비던셜스 프로비던셜스 국제공항 버진 아일랜드세인트 토마스 시릴 E 킹 공항세인트 크로 헨리 E 로이슨 공항

---------- 259178번째 데이터, 길이 = 277 ----------
2017년 7월 21일 홍일표 의원의 아들 서울동부지법 홍성균 판사가 지하철에서 몰카를 찍다가 시민의 신고를 받고 출동한 경찰에 의해 현장에서 검거되었다 압수한 휴대폰에서 여성의 신체 사진이 발견되었으나 본인은 동영상 앱이 저절로 작동하여 촬영한것 같다며 혐의를 부인했다 성폭력범죄 처벌 등에 관한 특례법 위반 혐의로 벌금 300만원에 기소되어 법원에서 확정되었다 초범이고 피해자와 합의한 점 등을 고려했다고 검찰은 양형이유를 설명하였다 이후 홍성균 판사는 판사를 그만두고 변호사 개업을 하였다

---------- 1692743번째 데이터, 길이 = 401 ----------
사맛있는 만두님이 백학모와 사유니차니님의 위키피디아 스토리까지 모두 표절하시고 사용자맛있는 만두재미로 떠나는 위키 테스트나도 몰랐던 나의 대한 진실위키공부용백과공주미션1사토맛있는 만두에 내용이 있습니다 이건 차단신청이 되는건가요 아님 너무 섣불리 판단한건가요강주 2020년 4월 17일 금 1900 KST 글쎄요 백차단을 읽어보니 차단 사유에는으로 뜨는데 맛있는 만두님 같은 경우에는 백저작권 위반으로 책과 지침을 어긴 것 같네요 정확하지 않을 수 있지만 차단 가능 할 듯 합니다 2020년 4월 17일 금 1904 KST 차단은 최수의 수단이 되어야 합니다 그냥 저작자를 유니차니 님이라고 표시해 주시고 당사자가 거부하면 다시 경고를 주세요 Motoko C K 토론 2020년 4월 17일 금 1911 KST

---------- 110268번째 데이터, 길이 = 660 ----------
1938년에 현 마쓰다의 옛 이름을 딴 도요 공업 축구단을 창단 J리그 전신인 JSL에서 뛰기 시작했다1971년에 마쓰다 SC 도요로 팀명을 변경하고 JSL이 J리그로 바뀌던 1992년에 마쓰다를 비롯한 히로시마 지역 47개 기업이 출자한 산프레체 히로시마로 팀명을 변경한다J리그 참가 이후 1994년 J리그 챔피언결정전 준우승 1995 1996 1999년에는 천황배 준우승을 기록하는 등 초창기만 해도 좋은 성적을 기록하며 J1에 잔류했으나 2002시즌 15위를 기록하며 J2리그로 추락한다그러나 2003년 J2리그 2위를 기록하며 1년만에 J리그에 복귀했으나 2007년 J리그 교체전에서 교토 상가 FC에게 패하여 다시 J2리그로 강등되었다 2008년 가시마 앤틀러스를 꺾고 슈퍼컵 정상에 올랐고 J2리그 1위를 기록하여 다시 J1으로 승격했다 재승격한 첫 해 4위를 기록하였으나 3위팀 감바 오사카의 천황배 우승으로 AFC 챔피언스리그 2010 출전권을 승계하게 되었다2010년 AFC 챔피언스리그에 처음 출전하였으나 같은 H조에 속한 포항 스틸러스와 애들레이드 유나이티드에 밀려 32강 조별 예선에서 탈락하였다 리그에서는 작년보다 못한 7위에 올랐고 이듬해에도 7위를 기록하였다2012 시즌 창단 후 처음으로 J리그 정상에 오른데 이어 2013 시즌에도 우승을 차지하였다

---------- 732180번째 데이터, 길이 = 95 ----------
엑스선 결정학은 단백질의 구조를 결정하는 데 사용되는 가장 일반적인 방법이다 단백질의 구조를 보는 데 높은 해상도를 제공하지만 단백질의 형태에 대한 정보는 제공하지 않는다

---------- 1103462번째 데이터, 길이 = 151 ----------
 colspan12핑둥 선 colspan1 aligncenter 번호 colspan1 aligncenter 이름 colspan1 aligncenter 중국어 이름 colspan1 aligncenter Station Name colspan1 aligncenter 다른 노선

---------- 137337번째 데이터, 길이 = 713 ----------
판교 분기점 등대한민국 고속도로 나들목과 관련지어 삭제 토론 필요성 제기 Jonsoh 토론 2008년 8월 9일 토 0343 KST분기점과 관련된 문서는 대표 문서로 분류되어 있지만 예를 들어 신갈 분기점이나 호법 분기점 노오지 분기점 등과 같은 교차로 성격이 짙은 문서의 경우도 저명성에 문제나 영향을 받을 의무가 없습니다 지식지기 토론 2008년 8월 9일 토 0446 KST및 I110 IRTC1015 2008년 8월 9일 토 1028 KST 2008년 8월 9일 토 1031 KSTFx120님 말씀에 동의합니다 RedmosQ 토론 2008년 8월 10일 일 0125 KST 도로 문서의 구조화에 도움을 줍니다 StarLight 토론 2008년 8월 11일 월 1126 KST또는 단일 문서로 존재하기에 문서 내용이 토막글 수준에서 발전될 가능성이 없어 보입니다 나들목 문서와 같이 하나로 묶어서 다루던지 아예 고속도로 문서에 소주제로 병합하는 것이 좋다고 생각합니다 현재의 분류에는 역시 강력히 반대 CLAWtheUltimate talks 2008년 8월 24일 일 0508 KST교차로의 의미가 크므로 없앨필요가 거의 없다고 봅니다 이 의견을 작성한 사용자는 황도준 토론이나 서명을 남기지 않아 다른 사용자가 추가하였습니다삭제는 해야 하지만 널리 알려진 분기점을 제외하곤 지울 필요가 없습니다 윤성현 토론 2008년 8월 27일 수 0728 KST

---------- 999890번째 데이터, 길이 = 321 ----------
체 K 위의 ntimes n 정사각 행렬 MinoperatornameMatnK에 대하여 다음 조건들이 서로 동치이다M은 대각화 가능 행렬이다M의 고유 공간들의 차원들의 합이 n이다pM 0이 되는 최소차 일계수 다항식 p in Kx의 차수가 k라고 할 때 p는 k개의 서로 다른 중복되지 않는 근들을 갖는다체 K 위의 ntimes n 정사각 행렬 MinoperatornameMatnK에 대하여 다음 조건을 만족시키는 행렬은 대각화 가능 행렬이다고유 다항식 chiMx detxM in Kx은 n개의 서로 다른 중복되지 않는 근을 갖는다이는 충분 조건이지만 필요 조건이 아니다

---------- 1570006번째 데이터, 길이 = 58 ----------
성공회 계열 성공회의 교황주의 en카리스마 운동 계열 토론토 축복 운동 en루터교 계열 하우게 운동 en

---------- 1136074번째 데이터, 길이 = 159 ----------
크리박급 프리깃 1척 운용상황은 불명북한 남포항에 3000t급 러시아 호위함 조선일보 20071107 0055구글어스 좌표 38 43 07N 125 23 44E나진급 호위함 2척서호급 호위함 1척 쌍동선 선체의 독자적 설계로 밀 MI14PL헬기 탑재가 가능하지만 거의 운용되지 않는다

---------- 912756번째 데이터, 길이 = 64 ----------
꿈의 섬 은 중국의 걸 그룹 SNH48의 열두 번째 미니 음반이다 타이틀 곡의 뮤직 비디오는 모리셔스에서 촬영됐다

---------- 175203번째 데이터, 길이 = 209 ----------
1998년 11월 17일 아르보렐리우스는 교황 요한 바오로 2세에 의해 스톡홀름의 주교로 임명되었으며 같은 해 12월 29일 스톡홀름 교구장 후베르투스 브란덴부르크 주교에 의해 주교 서품을 받았다 그는 브란덴부르크의 뒤를 이어 스웨덴의 유일한 가톨릭 교구인 스톡홀름 교구장이 되면서 종교개혁 이후 첫 번째 스웨덴인 가톨릭 주교이자 스칸디나비아의 두 번째 가톨릭 주교가 되었다

---------- 1239911번째 데이터, 길이 = 218 ----------
재판관 이공현직업수행의 자유 및 평등권 침해에 관하여는 다수의견과 뜻을 같이 했으나 결정주문의 형식에 관하여 이 사건 심판대상 규정들에 대해 단순위헌 결정을 한다고 하더라도 민영 미디어렙의 수의 증가나 그들 상호간의 경쟁 등이 방송의 공공성 공익성 등을 결정적으로 훼손시키고 법치국가적 법적 안정성을 심각하게 저해하는 결과를 가져올 것으로 예상되지 않는다고 하여 단순위헌을 선고를 주장하였다

---------- 278167번째 데이터, 길이 = 158 ----------
return name Mindanaotop 1073bottom 456left 11816right 12664image Philippines location map Mindanaosvgimage1 Philippines relief location map Mindanaosvg

---------- 41090번째 데이터, 길이 = 1679 ----------
원작에서는 코우사카 코우스케를 서술자로 하는 1인칭 시점으로 내용이 전개되며 서술자는 독자의 존재를 인식하고 있다소설 내 여동생이 이렇게 귀여울 리가 없어 제4권 12쪽 14쪽에서 서술자 쿄우스케 자신은 독자에게 내용을 서술하면서 종종 여동생과의 불화 관계를 강조하고 공감을 부탁하며 자신은 여동생 키리노를 싫어하고 있고 그처럼 키리노도 자신을 싫어할 것이라고 설명한다소설 내 여동생이 이렇게 귀여울 리가 없어 제6권 13쪽에서 쿄우스케가 자신이 싫어하고 있을 터인 키리노를 위하여 분투하는 이유는 복잡한데 쿄우스케 자신도 잘 설명하기 어려운 감정소설 내 여동생이 이렇게 귀여울 리가 없어 1권 243쪽에서 오빠으로서의 의무감 한번 한 약속은 지켜야 된다는 생각소설 내 여동생이 이렇게 귀여울 리가 없어 제2권 303쪽 314쪽에서 질투의 반대소설 내 여동생이 이렇게 귀여울 리가 없어 제3권 296쪽 297쪽에서 자기 만족을 위한 절개소설 내 여동생이 이렇게 귀여울 리가 없어 제5권 287쪽에서 등이 그 이유들로 뽑히고 있다쿄우스케는 비록 여동생을 매우 싫어하지만 여동생을 소중한 가족으로 생각하며 오빠으로서 지켜줘야 할 대상으로써 감싸고 있다고 밝힌다 저자 후시미는 쿄우스케의 이러한 행동에 대한 동기에 대하여 키리노는 싫어하지만 여동생은 좋아한다라는 식으로 설명했다한편으로 서술자 쿄우스케는 여동생의 본심을 붙잡아두고 있는데 쿄우스케의 입장에서는 오빠를 싫어하는 것처럼 보이는 키리노가 실은 오빠에게 어떤 감정을 품고 있지는 않은가에 대해서 명확히 밝혀지지 않은 채 이야기가 진행된다 이것의 진상에 관해서는 집필 개시 시점부터 후시미와 편집자 미키 사이에서 견해가 갈리고 있었는데 미키가 키리노는 일명 츤데레라기보다는 애당초 오빠를 진심으로 싫어하고 있다는 해석에 바탕을 두고 의견을 냈지만 후시미는 키리노가 내심으로는 처음부터 오빠에 대해 호감을 가지고 있었다는 식으로 해석할 수 있다는 묘사를 의식하여 제1권을 써내려갔다 그 후의 전개는 당초의 구상과는 상관이 없는 빠태로의 관계성의 변화도 이루어져 실제로 키리노가 갖고 있는 감정이 어떤 것인지는 다른 등장 인물과의 인간 관계와도 엮여있는 등소설 내 여동생이 이렇게 귀여울 리가 없어 제5권 217쪽에서소설 내 여동생이 이렇게 귀여울 리가 없어 제7권에서 작품 속의 수수께끼 중 하나로 남아 있다저자 후시미는 이 작품에서 묘사하고 있는 내용은 어디까지나 믿을 수 없는 화자 로 설정된 쿄우스케의 주관에 따른 것이며 진실과는 크게 동떨어져있음을 설명했는데 독자가 작품에서 화자의 오해가 들어가 있는 내용을 따라 읽는 것으로 재미를 느낄 수 있는 부분으로서 넣었다고 이러한 내용 전개의 취지를 설명했다Gs Festival Vol19 60쪽에서 편집자 미키는 미움이라는 감정은 좋아한다는 것의 반대가 아닌 상대에 대한 관심의 한 갈래로 다양한 감정이 들어간 개념이라고 설명하여 작품 속에서 묘사되는 것은 보통은 서로를 싫어하지만 가끔은 서로를 걱정해주는 남매의 유대라고 밝혔다만화판과 애니메이션판에서는 원작에서 자세히 묘사하던 서술자의 심리 묘사를 배제하는 한편 원작 속에서는 밝혀지지 않았던 키리노의 표정의 변화와 키리노의 시점 등도 묘사하고 있다 애니메이션판의 각본을 맡은 구라타 히데유키는 키리노는 겉으로는 말하지 않지만 내심으로는 오빠에게 홀딱 반한 상태라고 해석하여 원작의 독자와 애니메이션의 시청자도 그

렇게 해석하는 경향이 있다

---------- 329365번째 데이터, 길이 = 265 ----------
무대의상 stage costume은 무대 위에서 공연자가 착용하는 옷을 의미한다 무대분장stage performance 가운데서도 중요한 부분을 차지한다 의상계획은 작품에 따라서는 출연하는 배우와 연출가가 상담하여 결정하는 경우도 있다 그러나 등장인물이 많은 작품 혹은 고전극과 같은 특수한 스타일의 작품인 경우에는 의상 디자이너에 의하여 통일적인 계획을 수립하는 일이 필요하다무대의상은 역사적인 모습을 묘사할 수 있고 등장인물의 특정한 면을 과장하기 위해 사용할 수도 있다

---------- 1113396번째 데이터, 길이 = 2970 ----------
2011년 프로 야구 드래프트 회의에서는 노무라 유스케메이지 대학 후지오카 다카히로도요 대학와 함께 대학 빅3로 불리며 주목을 받았다 산케이 스포츠 2012년 11월 22일 숙부 하라 다쓰노리가 감독을 맡고 있는 요미우리 자이언츠가 단독 지명한다는 소문이 파다했으나2011 드래프트 총결산 슈칸 베이스볼2011년 11월 14일자 베이스볼 매거진사 p67 잡지 204421114 홋카이도 닛폰햄 파이터스도 1순위로 지명해 추첨 결과 닛폰햄이 교섭권을 획득했다 드래프트 후에는 조부인 하라 미쓰구가 닛폰햄의 사전 인사없이 강행한 지명을 인권 유린이라고 비난하는 발언을 했다 Sponichi Annex스포츠 닛폰 2011년 11월 10일11월 7일에는 닛폰햄의 지명 인사에 동석하여 진로에 대해서는 눈앞의 일 뿐만 아니라 그 이후의 야구 인생을 생각하여 판단하고 싶다고 말했으나 Sponichi Annex스포츠 닛폰 2011년 11월 8일 21일에는 닛폰햄의 입단을 거부하겠다는 의사를 밝혔다 요미우리 신문 2011년 11월 21일 사회인 야구와 일본 국내 독립 리그 해외의 프로 리그 등에 진출할 경우 2년간 지명받는 것이 불가능하기 때문에 프로아마 그 어떤 구단에도 소속하지 않은 상태에서 대학 재수를 하기로 결심했다 닛칸 스포츠 2011년 11월 4일 도카이 대학의 졸업 연기 제도를 이용해서 그대로 대학에 남고 Number Web분게이슌주 2012년 10월 22일 도카이 대학의 연습 시설을 사용하여 자유 연습을 하는 등 다음 해 이후의 드래프트 지명을 기다리게 됐다 입단 거부의 이유로는 닛폰햄으로부터 여러 이야기를 듣고 영광으로 생각했다 그러나 그 이상으로 나 자신이 어렸을 때부터 가지고 있던 꿈이 컸었다라고 말했고 닛칸 스포츠 2011년 11월 21일 결정적인 이유였던 어렸을 적부터의 꿈에 대해서는 말로하는 것은어렵지만 어떨까요 이해해주세요라고 말했다 산케이 신문 2011년 11월 21일이에 대해 닛폰햄은 이 정도로 포기할 거면 애초부터 지명하지도 않았다 교섭 기간이 거의 끝날 때까지 계속 설득할 생각이다라고 했지만주니치 신문 2011년 11월 22일자 교섭 기간 마지막 날인 2012년 3월 31일에 스가노의 영입을 포기하겠다는 입장을 발표했다 니혼케이자이 신문 2012년 3월 31일 그 후 한때는 이듬해 드래프트에서 재지명할 가능성도 있다고 공언했지만일본의 야구 협약에는 한번 입단 거부를 한 선수는 재지명하려면 본인의 동의가 필요라고 명시돼있지만 진학과 그 외의 사유로 인해 그 선수가 다시 취학했을 경우에는 그에 해당하지 않는다라고도 적혀 있어 재수한 걸로 해서 도카이 대학에 계속 학생으로 있는 스가노는 재취학에 해당하기 때문에 본인의 동의가 없어도 재지명이 가능하다고 닛폰햄측은 주장하고 있었다주니치 신문 2012년 4월 1일자 나중에는 그 방침을 바꿔서 닛폰햄의 야마다 마사오 단장이 우리로서는1위는 그 해에 가장 위력적인 선수이다 1년간의 공백이 있었음을 감안할 때 스가노가제일 위력적인 투수는 아니라고 본다라는 이유를 들어 지명을 안하겠다는 입장을 밝혔고 실제로 오타니 쇼헤이를 1순위로 지명했다닛폰햄은 2012년 드래프트에서 메이저 리그를 지망한다고 공언했던 오타니 쇼헤이를 1순위로 강행 지명해 오타니를 설득하는 데 성공하면서 입단이 성사됐다 닛칸 스포츠 2012년 10월 16일 동시에 그 해 드래프트에서 의중이 있던 구단 이외의 구단이 교섭권을 획득한 경우에 대해 요코이 히토키 도카이 대학 감독이 미국에 갈 거라고 본다 야구 유학도 포함해서 본인도 여러 모로 흥미가 있는 것 같다라고 말하는가 하면 닛칸 스포츠 2012년 10월 16일 스가노 본인도 작년하고 또다시 같은 일이 벌어지기라도 한다면 뭐라고 해야 할까 일본에서 야구를 하기 싫어질지도 모르겠다라고 발언한 것과 요코하마 DeNA 베이스타스로부터 발송된 조사서의 수취 거부한 적도 있는 등와시다 야스시 Sports Graphic Number 제814호 분게이슌주 p82 85 잡지 268541025 데일리 스포츠 2012년 10월 20일 이러한 스가노의 행동에 대해 요코하마 DeNA 베이스타스의 다카다 시게루 단장과 나카하타 기요시 감독 도호쿠 라쿠텐 골든이글스의 호시노 센이치 감독 등으로부터 비판을 받았다 Sponichi Annex스포츠 닛폰 2012년 10월 17일 데일리 스포츠 2012년 10월 18일도카이 대학의 졸업연기제도를 이용해서 계속 다니는 재수생이었지만 재수 생활하는 동안에는 대학 야구 규정에 따라 대외 경기 출전이 불가능하기 때문에 도카이 대학 야구부를 연습 거점으로 삼아 공던지기나 팀 동료를 상대로 한 홍백전 또는 시트 타격에 등판하는 등 실전 감각으로부터 멀어지지 않도록 애썼다 다만 홍백전이나 시트 타격에서의 등판은 긴장감이나 중압감이 있는 점에서 실전과는 다르기 때문에 공백에 대한 우려가 있었다 Sponichi Annex스포츠 닛폰 2012년 10월 25일 스가노 자신은 대학 야구뿐만 아니라 고교 야구나 일본 프로 야구 미국 메이저 리그 등을 적극적으로 관전한 것이 야구 선수로서 많은 도움이 됐다고 한다 2012년 1월 말부터 한 달 반 동안 미국 애리조나주에서 체류하여 메이저 리그 선수들을 비롯한 여러 경기의 운동 선수들이 모이는 시설에서 천웨이인에게서 왕젠민을 소개받아 조언을 받기도 했다슈칸 베이스볼 2018년 10월 1일자 p112113그러다가 10월 25일에 열린 드래프트 회의에서 당초에는 DeNA나 라쿠텐도 지명 가능성을 공언했으나 결국은 요미우리의 단독 1순위 지명을 받아 예전부터 열망하던 요미우리와의 교섭권을 획득했다 드래프트 회의 후에는 숙부이자 요미우리 감독이기도 한 하라 다쓰노리가 도카이 대학을 방문해서 등번호 19번과 이름이 들어간 요미우리 유니폼을 건네주었다 SANSPOCOM 2012년 10월 25일 같은 해 11월 21일에 입단 계약을 맺었고 마이니치 신문 2012년 11월 21일 11월 23일에 개최된 팬 감사 데이에서 입단을 공식 발표했다 산케이 스포츠 2012년 11월 23일

---------- 787201번째 데이터, 길이 = 610 ----------
즉위 전에 경왕에 봉해졌다1594년 막낀꿍이 후 레 왕조의 찐뚱에게 패하여 사명부로 달아났고 명나라에 칭신했다 막낀코안은 부를 거느리고 타이응우옌을 근거지로 삼아 지켰으며 백부 막낀꿍과 회합했다1621년 막낀꿍이 경왕 막낀코안에게 양위했고 이후 대자현에서 칭제하여 연호를 롱타이로 개원했고 막낀꿍을 태상황으로 높였다1623년 찐뚱이 죽고 찐짱과 찐쑤언이 쟁립하자 막낀코안이 기회를 틈타 군사를 거느리고 남쪽을 쳤으며 동시에 각 로의 세력을 선동하여 병사를 일으켜 반란하도록 했다 병사가 가림현에 이르렀으나 8월에 찐짱의 공격을 받아 패했고 전군이 몰살당했다 막낀코안은 겨우 몸만 건져 산림 가운데로 도망가 까오방으로 돌아갔다1625년 찐짱이 까오방을 함락해 막낀꿍과 황태자 등을 잡아 죽였으며 막낀코안과 그의 차자 막낀부는 명나라로 도망했다 같은 해 사람을 후 레 왕조로 보내 항복을 표하였고 막 왕조의 연호를 제거하고 후 레 왕조의 정삭을 받들었다 명나라의 압력 아래 후 레 왕조는 막낀코안을 태위 통국공에 봉하였으며 까오방으로 돌아오는 것을 허락했다 이후 막 왕조의 군사 실력은 더욱 쇠약해졌으며 아울러 후 레 왕조에게 조공해야 했다1638년 막낀코안이 병으로 죽자 막낀부가 뒤를 이었다

---------- 1370455번째 데이터, 길이 = 75 ----------
크리스찬 베일윌렘 대포자레드 레토조쉬 루카스사만다 마티스맷 로스빌 세이지클로에 세비니카라 세이무어저스틴 서룩스기네비어 터너리즈 위더스푼

---------- 1766891번째 데이터, 길이 = 716 ----------
1979년 한의사 지산 임달규19311988그의 생애는 대전 중구문화원 역사인물 소개를 참조 도서관 명칭은 지산도서관 에 의해 혜화학원 설립1980년 문교부로부터 혜화학원 정관 승인 대학 제 104211063호1981년 개교 제1회 입학식 1호관현 창학관 준공1985년 제1회 학사학위 수여식 거행1989년 종합대학교로 승격 대학원 과정 증설1990년 교표 변경 상징물 제정1991년 천안한방병원 개원1992년 청주한방병원 개원1993년 기숙사 완공1999년 혜화 신문화 선포식2004년 둔산캠퍼스 개설 및 둔산한방병원 개원2010년 설립자 지산 임달규 선생 국민훈장국민교육발전 유공추서2010년 개교 30주년 30주년기념관 준공2011년 교육과학기술부 선정 정부재정지원제한대학정부 43개 사립대학 재정지원 제한 연합뉴스 2011년 9월 5일 작성2014년 융합과학관산학협력관 준공2014년 군사산업정보대학원 폐지2014년 상담대학원 신설 보건스포츠대학원 명칭변경 보건의료대학원2015년 교육부 대학구조개혁평가 결과 D등급대학구조개혁평가 결과교육부 재정 지원 제한 장학금까지 영향 중앙일보 2015년 9월 1일 작성2016년 교육부 대학구조개혁평가 2차년도 이행점검 결과 재정지원제한 완전 해제이홍하 설립 4개 모두 최하위재정지원 전면 제한 한국대학신문 2017년 9월 4일 작성2018년 대전한방병원을 둔산한방병원으로 통합2019년 서울한방병원 개원서울특별시 송파구 문정동

---------- 327069번째 데이터, 길이 = 115 ----------
단위 격자란 각각의 자신을 평행이동시킨 것에서 결정을 표현하는 것이 가능한 최소단위이다 단위격자 중에 격자점이 정점을 단순 단위 격자라고 하고 반대로 쓸데 없는 간격이 가장 적은 것을 최대 충전 구조라 한다

---------- 1247617번째 데이터, 길이 = 112 ----------
Face A Fais que ton rve soit plus long que la nuit 1532 Face B Fais que ton rve soit plus long que la nuit 1525

---------- 791743번째 데이터, 길이 = 98 ----------
비제이 노박케빈 코리건J B 스무브아시프 맨드비아제이 나이두프레드 멜라메드엘리야스 커레시제프 그로스만안소니 맹가노메간 폭스존 C 라일리제이슨 맨트조카스세이드 바드레야리즈원 맨지

---------- 103355번째 데이터, 길이 = 135 ----------
MLS컵 1997는 메이저 리그 사커 1997의 최종 우승 팀을 가리는 축구 경기로 1997년 10월 26일에 워싱턴 DC의 로버트 F 케네디 메모리얼 스타디움에서 열렸다 DC 유나이티드가 콜로라도 래피즈에 21 승리를 거두며 우승을 차지했다

---------- 1284372번째 데이터, 길이 = 305 ----------
수소화악티늄은 섭씨 300도에서 삼염화악티늄과 칼륨에 전자가 첨가되었을 때 생성되고 그것의 구조는 대응하는 란타넘의 수소화물인 LaH2의 구조로 추론되었다 이 반응에서 수소는 어떻게 생성되었는지는 밝혀지지 않았다악티늄과 염산을 섞은 용액과 인산수소화나트륨NaH2PO4을 섞으면 하얀색인 인산악티늄의 반수화물AcPO405H2O이 생성되고 섭씨 1400도에서 몇 분동안 황화수소 기체로 수산염악티늄을 가열하면 검정색 황화악티늄Ac2S3이 생성된다 이것은 섭씨 1000도에서 황화수소와 이황화탄소의 혼합물과 산화악티늄을 반응시켰을 때 생성될 수도 있다

---------- 1262752번째 데이터, 길이 = 145 ----------
연극 명성황후 1995년 2000년까지 150회 출연 코러스에서부터 독일공사 4인의외상 등 여러 역할을 맡음015B 뮤직비디오 21C모노리스 1997년브라운 아이즈 뮤직비디오 With coffee 2001년다이나믹듀오 뮤직비디오 고백go back 2005년

---------- 184779번째 데이터, 길이 = 67 ----------
2006년 오버 더 레인보우 OST Start2012년 OK PUNK Ugly2012년 OK PUNK OK PUNK

5. 한국어 데이터 포함 비율

아직 전처리가 조금 부족한 것 같습니다. 한국어 데이터셋을 학습시켜야 하지만 온전한 영어 문장도 있기 때문이죠.

따라서 아래의 언어 탐지 함수가 한국어라고 판단한 데이터들만 남겨주겠습니다.

In [ ]:
```
# 영어 알파벳 모으기
import string

en = []
```

```
lower = string.ascii_lowercase
lower = [lower[i:i+1] for i in range(len(lower))]

upper = string.ascii_uppercase
upper = [upper[i:i+1] for i in range(len(upper))]

en = lower + upper
print(en)
```

```
['a', 'b', 'c', 'd', 'e', 'f', 'g', 'h', 'i', 'j', 'k', 'l', 'm', 'n', 'o', 'p', 'q', 'r', 's', 't', 'u', 'v', 'w', 'x', 'y', 'z', 'A', 'B', 'C', 'D', 'E',
'F', 'G', 'H', 'I', 'J', 'K', 'L', 'M', 'N', 'O', 'P', 'Q', 'R', 'S', 'T', 'U', 'V', 'W', 'X', 'Y', 'Z']
```

In [ ]:
```
# 한국어 포함 비율 판단 함수
def ko_prob(text):
    en_count = sum(1 for char in text if char in en)
    return 1 - (en_count / len(text))
```

In [ ]:
```
# 한국어 포함 비율 판단 예시
np.random.seed(42)

idx = np.random.randint(0, len(train), 30)
for i in idx:
    print(f"---------------------- {i}번째 데이터 ----------------------")
    print(train[i])
    print(f"한국어 포함 비율: {ko_prob(train[i]):.2f}")
    print()
```

---------- 121958번째 데이터 ----------
width7 scopecol 연도 width34 scopecol 작품 width34 scopecol 부문 width9 scopecol 결과 width9 scopecol 출처
한국어 포함 비율: 0.36

---------- 671155번째 데이터 ----------
사우스조지아 사우스샌드위치 제도의 문장은 1985년 영토가 창설되면서 부여되었다 문장은 사우스조지아섬에있는 순록 두 무리의 순록이다 표어는 Leo Terram Propriam Protegat이다국장
한국어 포함 비율: 0.77

---------- 131932번째 데이터 ----------
본점 충청남도 천안시 서북구 직산읍 거리막길 33서울지점 서울특별시 중구 정동길 14 정동 오송빌딩
한국어 포함 비율: 1.00

---------- 1414414번째 데이터 ----------
앤티가 바부다세인트존스 VC 버드 국제공항 아루바오라녜스타트 퀸 비어트릭스 국제공항 바하마프리포트 그랜드 바하마 국제공항나소 린든 핀들링 국제공항 바베이도스 그레이트 레이 아담스 국제공항 케이맨 제도그랜드 케이먼 오웬 로버츠 국제공항 도미니카 공화국푼타카나 푼타카나 국제공항산티아고데로스카바예로스 시바오 국제공항 산토도밍고 라스 아메리카 국제공항 자메이카몬티 상스터 국제공항 푸에르토리코산후안 루이스 무노즈 마린 국제공항 세인트키츠네비스세인트 키츠 로버트 L 브래드쇼 국제공항 세인트루시아뷰포트 하와노리아 국제공항 신트마르턴신트마르턴 프린세스 줄리아나 국제공항 터크스 케이커스 제도프로비던셜스 프로비던셜스 국제공항 터크스 케이커스 제도프로비던셜스 프로비던셜스 국제공항 버진 아일랜드세인트 토마스 시릴 E 킹 공항세인트 크로 헨리 E 로이슨 공항
한국어 포함 비율: 0.99

---------- 259178번째 데이터 ----------
2017년 7월 21일 홍일표 의원의 아들 서울동부지법 홍성균 판사가 지하철에서 몰카를 찍다가 시민의 신고를 받고 출동한 경찰에 의해 현장에서 검거되었다 압수한 휴대폰에서 여성의 신체 사진이 발견되었으나 본인은 동영상 앱이 저절로 작동하여 촬영한것 같다며 혐의를 부인했다 성폭력범죄 처벌 등에 관한 특례법 위반 혐의로 벌금 300만원에 기소되어 법원에서 확정되었다 초범이고 피해자와 합의한 점 등을 고려했다고 검찰은 양형이유를 설명하였다 이후 홍성균 판사는 판사를 그만두고 변호사 개업을 하였다
한국어 포함 비율: 1.00

---------- 1692743번째 데이터 ----------
사맛있는 만두님이 백학모와 사유니차니님의 위키피디아 스토리까지 모두 표절하시고 사용자맛있는 만두재미로 떠나는 위키 테스트나도 몰랐던 나의 대한 진실위키공부용백과공주미션1사토맛있는 만두에 내용이 있습니다 이건 차단신청이 되는건가요 아님 너무 섣불리 판단한건가요강주 2020년 4월 17일 금 1900 KST 글쎄요 백차단을 읽어보니 차단 사유에는으로 뜨는데 맛있는 만두님 같은 경우에는 백저작권 위반으로 책과 지침을 어긴 것 같네요 정확하지 않을 수 있지만 차단 가능 할 듯 합니다 2020년 4월 17일 금 1904 KST 차단은 최수의 수단이 되어야 합니다 그냥 저작자를 유니차니 님이라고 표시해 주시고 당사자가 거부하면 다시 경고를 주세요 Motoko C K 토론 2020년 4월 17일 금 1911 KST
한국어 포함 비율: 0.96

---------- 110268번째 데이터 ----------
1938년에 현 마쓰다의 옛 이름을 딴 도요 공업 축구단을 창단 J리그 전신인 JSL에서 뛰기 시작했다1971년에 마쓰다 SC 도요로 팀명을 변경하고 JSL이 J리그로 바뀌던 1992년에 마쓰다를 비롯한 히로시마 지역 47개 기업이 출자한 산프레체 히로시마로 팀명을 변경한다J리그 참가 이후 1994년 J리그 챔피언결정전 준우승 1995 1996 1999년에는 천황배 준우승을 기록하는 등 초창기만 해도 좋은 성적을 기록하며 J1에 잔류했으나 2002시즌 15위를 기록하며 J2리그로 추락한다그러나 2003년 J2리그 2위를 기록하며 1년만에 J리그에 복귀했으나 2007년 J리그 교체전에서 교토 상가 FC에게 패하여 다시 J2리그로 강등되었다 2008년 가시마 앤틀러스를 꺾고 슈퍼컵 정상에 올랐고 J2리그 1위를 기록하여 다시 J1으로 승격했다 재승격한 첫 해 4위를 기록하였으나 3위팀 감바 오사카의 천황배 우승으로 AFC 챔피언스리그 2010 출전권을 승계하게 되었다2010년 AFC 챔피언스리그에 처음 출전하였으나 같은 H조에 속한 포항 스틸러스와 애들레이드 유나이티드에 밀려 32강 조별 예선에서 탈락하였다 리그에서는 작년보다 못한 7위에 올랐고 이듬해에도 7위를 기록하였다2012 시즌 창단 후 처음으로 J리그 정상에 오른데 이어 2013 시즌에도 우승을 차지하였다
한국어 포함 비율: 0.95

---------- 732180번째 데이터 ----------
엑스선 결정학은 단백질의 구조를 결정하는 데 사용되는 가장 일반적인 방법이다 단백질의 구조를 보는 데 높은 해상도를 제공하지만 단백질의 형태에 대한 정보는 제공하지 않는다
한국어 포함 비율: 1.00

---------- 1103462번째 데이터 ----------
colspan12핑둥 선 colspan1 aligncenter 번호 colspan1 aligncenter 이름 colspan1 aligncenter 중국어 이름 colspan1 aligncenter Station Name colspan1 aligncenter 다른 노선
한국어 포함 비율: 0.28

---------- 137337번째 데이터 ----------
판교 분기점 등대한민국 고속도로 나들목과 관련지어 삭제 토론 필요성 제기 Jonsoh 토론 2008년 8월 9일 토 0343 KST분기점과 관련된 문서는 대표 문서로 분류되어 있지만 예를 들어 신갈 분기점이나 호법 분기점 노오지 분기점 등과 같은 교차로 성격이 짙은 문서의 경우도 저명성에 문제나 영향을 받을 의무가 없습니다 지식지기 토론 2008년 8월 9일 토 0446 KST및 I110 IRTC1015 2008년 8월 9일 토 1028 KST 2008년 8월 9일 토 1031 KSTFx120님 말씀에 동의합니다 RedmosQ 토론 2008년 8월 10일 일 0125 KST 도로 문서의 구조화에 도움을 줍니다 StarLight 토론 2008년 8월 11일 월 1126 KST또는 단일 문서로 존재하기에 문서 내용이 토막글 수준에서 발전될 가능성이 없어 보입니다 나들목 문서와 같이 하나로 묶어서 다루던지 아예 고속도로 문서에 소주제로 병합하는 것이 좋다고 생각합니다 현재의 분류에는 역시 강력히 반대 CLAWtheUltimate talks 2008년 8월 24일 일 0508 KST교차로의 의미가 크므로 없앨필요가 거의 없다고 봅니다 이 의견을 작성한 사용자는 황도준 토론이나 서명을 남기지 않아 다른 사용자가 추가하였습니다삭제는 해야 하지만 널리 알려진 분기점을 제외하곤 지울 필요가 없습니다 윤성현 토론 2008년 8월 27일 수 0728 KST
한국어 포함 비율: 0.90

---------- 999890번째 데이터 ----------
체 K 위의 ntimes n 정사각 행렬 MinoperatornameMatnK에 대하여 다음 조건들이 서로 동치이다M은 대각화 가능 행렬이다M의 고유 공간들의 차원들의 합이 n이다pM 0이 되는 최소차 일계수 다항식 p in Kx의 차수가 k라고 할 때 p는 k개의 서로 다른 중복되지 않는 근들을 갖는다체 K 위의 ntimes n 정사각 행렬 MinoperatornameMatnK에 대하여 다음 조건을 만족시키는 행렬은 대각화 가능 행렬이다고유 다항식 chiMx detxM in Kx은 n개의 서로 다른 중복되지 않는 근을 갖는다이는 충분 조건이지만 필요 조건이 아니다
한국어 포함 비율: 0.74

---------- 1570006번째 데이터 ----------
성공회 계열 성공회의 교황주의 en카리스마 운동 계열 토론토 축복 운동 en루터교 계열 하우게 운동 en
한국어 포함 비율: 0.90

---------- 1136074번째 데이터 ----------
크리박급 프리깃 1척 운용상황은 불명북한 남포항에 3000t급 러시아 호위함 조선일보 20071107 0055구글어스 좌표 38 43 07N 125 23 44E나진급 호위함 2척서호급 호위함 1척 쌍동선 선체의 독자적 설계로 밀 MI14PL헬기 탑재가 가능하지만 거의 운용되지 않는다
한국어 포함 비율: 0.96

---------- 912756번째 데이터 ----------
꿈의 섬 은 중국의 걸 그룹 SNH48의 열두 번째 미니 음반이다 타이틀 곡의 뮤직 비디오는 모리셔스에서 촬영됐다
한국어 포함 비율: 0.95

---------- 175203번째 데이터 ----------
1998년 11월 17일 아르보렐리우스는 교황 요한 바오로 2세에 의해 스톡홀름의 주교로 임명되었으며 같은 해 12월 29일 스톡홀름 교구장 후베르투스 브란덴부르크 주교에 의해 주교 서품을 받았다 그는 브란덴부르크의 뒤를 이어 스웨덴의 유일한 가톨릭 교구인 스톡홀름 교구장이 되면서 종교개혁 이후 첫 번째 스웨덴인 가톨릭 주교이자 스칸디나비아의 두 번째 가톨릭 주교가 되었다
한국어 포함 비율: 1.00

---------- 1239911번째 데이터 ----------
재판관 이공현직업수행의 자유 및 평등권 침해에 관하여는 다수의견과 뜻을 같이 했으나 결정주문의 형식에 관하여 이 사건 심판대상 규정들에 대해 단순위헌 결정을 한다고 하더라도 민영 미디어렙의 수의 증가나 그들 상호간의 경쟁 등이 방송의 공공성 공익성 등을 결정적으로 훼손시키고 법치국가적 법적 안정성을 심각하게 저해하는 결과를 가져올 것으로 예상되지 않는다고 하여 단순위헌을 선고를 주장하였다
한국어 포함 비율: 1.00

---------- 278167번째 데이터 ----------
return name Mindanaotop 1073bottom 456left 11816right 12664image Philippines location map Mindanaosvgimage1 Philippines relief location map Mindanaosvg
한국어 포함 비율: 0.25

---------- 41090번째 데이터 ----------
원작에서는 코우사카 코우스케를 서술자로 하는 1인칭 시점으로 내용이 전개되며 서술자는 독자의 존재를 인식하고 있다소설 내 여동생이 이렇게 귀여울 리가 없어 제4권 12

쪽 14쪽에서 서술자 쿄우스케 자신은 독자에게 내용을 서술하면서 종종 여동생과의 불화 관계를 강조하고 공감을 부탁하며 자신은 여동생 키리노를 싫어하고 있고 그처럼 키리노도 자신을 싫어할 것이라고 설명한다소설 내 여동생이 이렇게 귀여울 리가 없어 제6권 13쪽에서 쿄우스케가 자신이 싫어하고 있을 터인 키리노를 위하여 분투하는 이유는 복잡한데 쿄우스케 자신도 잘 설명하기 어려운 감정소설 내 여동생이 이렇게 귀여울 리가 없어 1권 243쪽에서 오빠으로서의 의무감 한번 한 약속은 지켜야 된다는 생각소설 내 여동생이 이렇게 귀여울 리가 없어 제2권 303쪽 314쪽에서 질투의 반대소설 내 여동생이 이렇게 귀여울 리가 없어 제3권 296쪽 297쪽에서 자기 만족을 위한 절개소설 내 여동생이 이렇게 귀여울 리가 없어 제5권 287쪽에서 등이 그 이유들로 뽑히고 있다쿄우스케는 비록 여동생을 매우 싫어하지만 여동생을 소중한 가족으로 생각하며 오빠으로서 지켜줘야 할 대상으로써 감싸고 있다고 밝힌다 저자 후시미는 쿄우스케의 이러한 행동에 대한 동기에 대하여 키리노는 싫어하지만 여동생은 좋아한다라는 식으로 설명했다한편으로 서술자 쿄우스케는 여동생의 본심을 붙잡아두고 있는데 쿄우스케의 입장에서는 오빠를 싫어하는 것처럼 보이는 키리노가 실은 오빠에게 어떤 감정을 품고 있지는 않은가에 대해서 명확히 밝혀지지 않은 채 이야기가 진행된다 이것의 진상에 관해서는 집필 개시 시점부터 후시미와 편집자 미키 사이에서 견해가 갈리고 있었는데 미키가 키리노는 일명 츤데레라기보다는 애당초 오빠를 진심으로 싫어하고 있다는 해석에 바탕을 두고 의견을 냈지만 후시미는 키리노가 내심으로는 처음부터 오빠에 대해 호감을 가지고 있었다는 식으로 해석할 수 있다는 묘사를 의식하여 제1권을 써내려갔다 그 후의 전개는 당초의 구상과는 상관이 없는 빠태로의 관계성의 변화도 이루어져 실제로 키리노가 갖고 있는 감정이 어떤 것인지는 다른 등장 인물과의 인간 관계와도 엮여있는 등소설 내 여동생이 이렇게 귀여울 리가 없어 제5권 217쪽에서소설 내 여동생이 이렇게 귀여울 리가 없어 제7권에서 작품 속의 수수께끼 중 하나로 남아 있다저자 후시미는 이 작품에서 묘사하고 있는 내용은 어디까지나 믿을 수 없는 화자 로 설정된 쿄우스케의 주관에 따른 것이며 진실과는 크게 동떨어져있음을 설명했는데 독자가 작품에서 화자의 오해가 들어가 있는 내용을 따라 읽는 것으로 재미를 느낄 수 있는 부분으로서 넣었다고 이러한 내용 전개의 취지를 설명했다Gs Festival Vol19 60쪽에서 편집자 미키는 미움이라는 감정은 좋아한다는 것의 반대가 아닌 상대에 대한 관심의 한 갈래로 다양한 감정이 들어간 개념이라고 설명하여 작품 속에서 묘사되는 것은 보통은 서로를 싫어하지만 가끔은 서로를 걱정해주는 남매의 유대라고 밝혔다만화판과 애니메이션판에서는 원작에서 자세히 묘사하던 서술자의 심리 묘사를 배제하는 한편 원작 속에서는 밝혀지지 않았던 키리노의 표정의 변화와 키리노의 시점 등도 묘사하고 있다 애니메이션판의 각본을 맡은 구라타 히데유키는 키리노는 겉으로는 말하지 않지만 내심으로는 오빠에게 홀딱 반한 상태라고 해석하여 원작의 독자와 애니메이션의 시청자도 그렇게 해석하는 경향이 있다
한국어 포함 비율: 0.99

---------- 329365번째 데이터 ----------
무대의상 stage costume은 무대 위에서 공연자가 착용하는 옷을 의미한다 무대분장stage performance 가운데서도 중요한 부분을 차지한다 의상계획은 작품에 따라서는 출연하는 배우와 연출가가 상담하여 결정하는 경우도 있다 그러나 등장인물이 많은 작품 혹은 고전극과 같은 특수한 스타일의 작품인 경우에는 의상 디자이너에 의하여 통일적인 계획을 수립하는 일이 필요하다무대의상은 역사적인 모습을 묘사할 수 있고 등장인물의 특정한 면을 과장하기 위해 사용할 수도 있다
한국어 포함 비율: 0.89

---------- 1113396번째 데이터 ----------
2011년 프로 야구 드래프트 회의에서는 노무라 유스케메이지 대학 후지오카 다카히로도요 대학와 함께 대학 빅3로 불리며 주목을 받았다 산케이 스포츠 2012년 11월 22일 숙부 하라 다쓰노리가 감독을 맡고 있는 요미우리 자이언츠가 단독 지명한다는 소문이 파다했으나2011 드래프트 총결산 슈칸 베이스볼2011년 11월 14일자 베이스볼 매거진사 p67 잡지 204421114 홋카이도 닛폰햄 파이터스도 1순위로 지명해 추첨 결과 닛폰햄이 교섭권을 획득했다 드래프트 후에는 조부인 하라 미쓰구가 닛폰햄의 사전 인사없이 강행한 지명을 인권 유린이라고 비난하는 발언을 했다 Sponichi Annex스포츠 닛폰 2011년 11월 10일11월 7일에는 닛폰햄의 지명 인사에 동석하여 진로에 대해서는 눈앞의 일 뿐만 아니라 그 이후의 야구 인생을 생각하여 판단하고 싶다고 말했으나 Sponichi Annex스포츠 닛폰 2011년 11월 8일 21일에는 닛폰햄의 입단을 거부하겠다는 의사를 밝혔다 요미우리 신문 2011년 11월 21일 사회인 야구와 일본 국내 독립 리그 해외의 프로 리그 등에 진출할 경우 2년간 지명받는 것이 불가능하기 때문에 프로아마 그 어떤 구단에도 소속하지 않은 상태에서 대학 재수를 하기로 결심했다 닛칸 스포츠 2011년 11월 4일 도카이 대학의 졸업 연기 제도를 이용해서 그대로 대학에 남고 Number Web분게이슌주 2012년 10월 22일 도카이 대학의 연습 시설을 사용하여 자유 연습을 하는 등 다음 해 이후의 드래프트 지명을 기다리게 됐다 입단 거부의 이유로는 닛폰햄으로부터 여러 이야기를 듣고 영광으로 생각했다 그러나 그 이상으로 나 자신이 어렸을 때부터 가지고 있던 꿈이 컸었다라고 말했고 닛칸 스포츠 2011년 11월 21일 결정적인 이유였던 어렸을 적부터의 꿈에 대해서는 말로하는 것은어렵지만 어떨까요 이해해주세요라고 말했다 산케이 신문 2011년 11월 21일이에 대해 닛폰햄은 이 정도로 포기할 거면 애초부터 지명하지도 않았다 교섭 기간이 거의 끝날 때까지 계속 설득할 생각이다라고 했지만주니치 신문 2011년 11월 22일자 교섭 기간 마지막 날인 2012년 3월 31일에 스가노의 영입을 포기하겠다는 입장을 발표했다 니혼케이자이 신문 2012년 3월 31일 그 후 한때는 이듬해 드래프트에서 재지명할 가능성도 있다고 공언했지만일본의 야구 협약에는 한번 입단 거부를 한 선수는 재지명하려면 본인의 동의가 필요라고 명시돼있지만 진학과 그 외의 사유로 인해 그 선수가 다시 취학했을 경우에는 그에 해당하지 않는다라고도 적혀 있어 재수한 걸로 해서 도카이 대학에 계속 학생으로 있는 스가노는 재취학에 해당하기 때문에 본인의 동의가 없어도 재지명이 가능하다고 닛폰햄측은 주장하고 있었다주니치 신문 2012년 4월 1일자 나중에는 그 방침을 바꿔서 닛폰햄의 야마다 마사오 단장이 우리로서는1위는 그 해에 가장 위력적인 선수이다 1년간의 공백이 있었음을 감안할 때 스가노가제일 위력적인 투수는 아니라고 본다라는 이유를 들어 지명을 안하겠다는 입장을 밝혔고 실제로 오타니 쇼헤이를 1순위로 지명했다닛폰햄은 2012년 드래프트에서 메이저 리그를 지망한다고 공언했던 오타니 쇼헤이를 1순위로 강행 지명해 오타니를 설득하는 데 성공하면서 입단이 성사됐다 닛칸 스포츠 2012년 10월 16일 동시에 그 해 드래프트에서 의중이 있던 구단 이외의 구단이 교섭권을 획득한 경우에 대해 요코이 히토키 도카이 대학 감독이 미국에 갈 거라고 본다 야구 유학도 포함해서 본인도 여러 모로 흥미가 있는 것 같다라고 말하는가 하면 닛칸 스포츠 2012년 10월 16일 스가노 본인도 작년하고 또다시 같은 일이 벌어지기라도 한다면 뭐라고 해야 할까 일본에서 야구를 하기 싫어질지도 모르겠다라고 발언한 것과 요코하마 DeNA 베이스타스로부터 발송된 조사서의 수취 거부한 적도 있는 등와시다 야스시 Sports Graphic Number 제814호 분게이슌주 p82 85 잡지 268541025 데일리 스포츠 2012년 10월 20일 이러한 스가노의 행동에 대해 요코하마 DeNA 베이스타스의 다카다 시게루 단장과 나카하타 기요시 감독 도호쿠 라쿠텐 골든이글스의 호시노 센이치 감독 등으로부터 비판을 받았다 Sponichi Annex스포츠 닛폰 2012년 10월 17일 데일리 스포츠 2012년 10월 18일도카이 대학의 졸업연기제도를 이용해서 계속 다니는 재수생이었지만 재수 생활하는 동안에는 대학 야구 규정에 따라 대외 경기 출전이 불가능하기 때문에 도카이 대학 야구부를 연습 거점으로 삼아 공던지기나 팀 동료를 상대로 한 홍백전 또는 시트 타격에 등판하는 등 실전 감각으로부터 멀어지지 않도록 애썼다 다만 홍백전이나 시트 타격에서의 등판은 긴장감이나 중압감이 있는 점에서 실전과는 다르기 때문에 공백에 대한 우려가 있었다 Sponichi Annex스포츠 닛폰 2012년 10월 25일 스가노 자신은 대학 야구뿐만 아니라 고교 야구나 일본 프로 야구 미국 메이저 리그 등을 적극적으로 관전한 것이 야구 선수로서 많은 도움이 됐다고 한다 2012년 1월 말부터 한 달 반 동안 미국 애리조나주에서 체류하여 메이저 리그 선수들을 비롯한 여러 경기의 운동 선수들이 모이는 시설에서 천웨이인에게서 왕젠민을 소개받아 조언을 받기도 했다슈칸 베이스볼 2018년 10월 1일자 p112113그러다가 10월 25일에 열린 드래프트 회의에서 당초에는 DeNA나 라쿠텐도 지명 가능성을 공언했으나 결국은 요미우리의 단독 1순위 지명을 받아 예전부터 열망하던 요미우리와의 교섭권을 획득했다 드래프트 회의 후에는 숙부이자 요미우리 감독이기도 한 하라 다쓰노리가 도카이 대학을 방문해서 등번호 19번과 이름이 들어간 요미우리 유니폼을 건네주었다 SANSPOCOM 2012년 10월 25일 같은 해 11월 21일에 입단 계약을 맺었고 마이니치 신문 2012년 11월 21일 11월 23일에 개최된 팬 감사 데이에서 입단을 공식 발표했다 산케이 스포츠 2012년 11월 23일
한국어 포함 비율: 0.96

---------- 787201번째 데이터 ----------
즉위 전에 경왕에 봉해졌다1594년 막낀꿍이 후 레 왕조의 찐뚱에게 패하여 사명부로 달아났고 명나라에 칭신했다 막낀코안은 부를 거느리고 타이응우옌을 근거지로 삼아 지켰으며 백부 막낀꿍과 회합했다1621년 막낀꿍이 경왕 막낀코안에게 양위했고 이후 대자현에서 칭제하여 연호를 롱타이로 개원했고 막낀꿍을 태상황으로 높였다1623년 찐뚱이 죽고 찐짱과 찐쑤언이 쟁립하자 막낀코안이 기회를 틈타 군사를 거느리고 남쪽을 쳤으며 동시에 각 로의 세력을 선동하여 병사를 일으켜 반란하도록 했다 병사가 가림현에 이르렀으나 8월에 찐짱의 공격을 받아 패했고 전군이 몰살당했다 막낀코안은 겨우 몸만 건져 산림 가운데로 도망가 까오방으로 돌아갔다1625년 찐짱이 까오방을 함락해 막낀꿍과 황태자 등을 잡아 죽였으며 막낀코안과 그의 차자 막낀부는 명나라로 도망했다 같은 해 사람을 후 레 왕조로 보내 항복을 표하였고 막 왕조의 연호를 제거하고 후 레 왕조의 정삭을 받들었다 명나라의 압력 아래 후 레 왕조는 막낀코안을 태위 통국공에 봉하였으며 까오방으로 돌아오는 것을 허락했다 이후 막 왕조의 군사 실력은 더욱 쇠약해졌으며 아울러 후 레 왕조에게 조공해야 했다1638년 막낀코안이 병으로 죽자 막낀부가 뒤를 이었다
한국어 포함 비율: 1.00

---------- 1370455번째 데이터 ----------
크리스찬 베일윌렘 대포자레드 레토조쉬 루카스사만다 마티스맷 로스빌 세이지클로에 세비니카라 세이무어저스틴 서룩스기네비어 터너리즈 위더스푼
한국어 포함 비율: 1.00

---------- 1766891번째 데이터 ----------
1979년 한의사 지산 임달규19311988그의 생애는 대전 중구문화원 역사인물 소개를 참조 도서관 명칭은 지산도서관 에 의해 혜화학원 설립1980년 문교부로부터 혜화학원 정관 승인 대학 제 104211063호1981년 개교 제1회 입학식 1호관현 창학관 준공1985년 제1회 학사학위 수여식 거행1989년 종합대학교로 승격 대학원 과정 증설1990년 교표 변경 상징물 제정1991년 천안한방병원 개원1992년 청주한방병원 개원1993년 기숙사 완공1999년 혜화 신문화 선포식2004년 둔산캠퍼스 개설 및 둔산한방병원 개원2010년 설립자 지산 임달규 선생 국민훈장국민교육발전 유공추서2010년 개교 30주년 30주년기념관 준공2011년 교육과학기술부 선정 정부재정지원제한대학정부 43개 사립대학 재정지원 제한 연합뉴스 2011년 9월 5일 작성2014년 융합과학관산학협력관 준공2014년 군사산업정보대학원 폐지2014년 상담대학원 신설 보건스포츠대학원 명칭변경 보건의료대학원2015년 교육부 대학구조개혁평가 결과 D등급대학구조개혁평가 결과교육부 재정 지원 제한 장학금까지 영향 중앙일보 2015년 9월 1일 작성2016년 교육부 대학구조개혁평가 2차년도 이행점검 결과 재정지원제한 완전 해제이홍하 설립 4개 모두 최하위재정지원 전면 제한 한국대학신문 2017년 9월 4일 작성2018년 대전한방병원을 둔산한방병원으로 통합2019년 서울한방병원 개원서울특별시 송파구 문정동
한국어 포함 비율: 1.00

---------- 327069번째 데이터 ----------
단위 격자란 각각의 자신을 평행이동시킨 것에서 결정을 표현하는 것이 가능한 최소단위이다 단위격자 중에 격자점이 정점을 단순 단위 격자라고 하고 반대로 쓸데 없는 간격이 가장 적은 것을 최대 충전 구조라 한다
한국어 포함 비율: 1.00

---------- 1247617번째 데이터 ----------
Face A Fais que ton rve soit plus long que la nuit 1532 Face B Fais que ton rve soit plus long que la nuit 1525
한국어 포함 비율: 0.30

---------- 791743번째 데이터 ----------
비제이 노박케빈 코리건J B 스무브아시프 맨드비아제이 나이두프레드 멜라메드엘리야스 퀴레시제프 그로스만안소니 맹가노메간 폭스존 C 라일리제이슨 맨트조카스세이드 바드레야리즈원 맨지
한국어 포함 비율: 0.97

---------- 103355번째 데이터 ----------
MLS컵 1997는 메이저 리그 사커 1997의 최종 우승 팀을 가리는 축구 경기로 1997년 10월 26일에 워싱턴 DC의 로버트 F 케네디 메모리얼 스타디움에서 열렸다 DC 유나이티드가 콜로라도 래피즈에 21 승리를 거두며 우승을 차지했다
한국어 포함 비율: 0.94

```
---------------------- 1284372번째 데이터 ----------------------
수소화악티늄은 섭씨 300도에서 삼염화악티늄과 칼륨에 전자가 첨가되었을 때 생성되고 그것의 구조는 대응하는 란타넘의 수소화물인 LaH2의 구조로 추론되었다 이 반응에서 수소는 어떻게 생성되었는지는 밝혀지지 않았다악티늄과 염산을 섞은 용액과 인산수소화나트륨NaH2PO4을 섞으면 하얀색인 인산악티늄의 반수화물AcPO405H2O이 생성되고 섭씨 1400도에서 몇 분동안 황화수소 기체로 수산염악티늄을 가열하면 검정색 황화악티늄Ac2S3이 생성된다 이것은 섭씨 1000도에서 황화수소와 이황화탄소의 혼합물과 산화악티늄을 반응시켰을 때 생성될 수도 있다
한국어 포함 비율: 0.94

---------------------- 1262752번째 데이터 ----------------------
연극 명성황후 1995년  2000년까지 150회 출연  코러스에서부터 독일공사 4인의외상 등 여러 역할을 맡음015B 뮤직비디오 21C모노리스 1997년브라운 아이즈 뮤직비디오 With coffee 2001년다이나믹듀오 뮤직비디오 고백go back 2005년
한국어 포함 비율: 0.88

---------------------- 184779번째 데이터 ----------------------
2006년 오버 더 레인보우 OST  Start2012년 OK PUNK  Ugly2012년 OK PUNK  OK PUNK
한국어 포함 비율: 0.55
```

In [ ]:
```python
# 일부 데이터 전처리(5만 개)에 대한 시간 측정
import time

np.random.seed(42)

start = time.time()
idx = np.random.randint(0, len(train), 50000)
for i in idx:
  ko_prob(train[i])

end = time.time()
print(f"걸린 시간: {end - start:.4f}초")
```

```
걸린 시간: 7.7792초
```

이 전처리 과정은 **5~10분** 정도 소요됩니다.

In [ ]:
```python
# 한국어 포함 비율이 낮은 데이터 제거
print("전처리 전: 길이 =", len(train))

start = time.time()
for i in range(len(train)):
    prob = ko_prob(train[i])
    if prob < 0.7:
        train[i] = ""

train = [i for i in train if i != ""]

end = time.time()
print(f"걸린 시간: {end - start:.4f}초")
print("전처리 후: 길이 =", len(train))
```

```
전처리 전: 길이 = 1790990
걸린 시간: 314.5338초
전처리 후: 길이 = 1563464
```

최종 데이터를 보도록 할까요? 마찬가지로 30개의 데이터를 샘플링하겠습니다.

In [ ]:
```python
# 최종 전처리 데이터 샘플링
np.random.seed(42)

idx = np.random.randint(0, len(train), 30)
for i in idx:
    print(f"---------------------- {i}번째 데이터 ----------------------")
    print(train[i])
    print()
```

---------- 121958번째 데이터 ----------
2000년 11월 22일경 김옥경은 조부이자 애국지사인 김용원의 유적비 건립 현황을 확인하기 위해서 은평공원대전광역시 서구 월평동을 찾았다 그러나 비문 정면에 새겨진 생애비높이 14m 폭 18m와 휘호비높이 43m 폭 14m에는 김씨의 조부가 아닌 이돈직이라는 알 수 없는 사람의 휘호와 공적이 새겨져 있었다 현장에는 김용원의 흉상이 없었고 정작 김씨 조부의 생애와 휘호는 각각 뒷면에 새겨져 있었음이 드러났다 조사를 해보니 은평공원 휘호비생애비와 원정동 김용원 선생 공적비 효평동 이돈직 공적비 대덕구 비래동산무궁화동산 비문에도 이인구의 조부 이돈직의 부풀려진 공적이 적혀있는 것으로 확인됐다는 2003년 12월 독립투사의 공적비가 변조된 사연을 보도했다 보도에서 대전지역 대표적 독립운동가 중 한 명으로 꼽히는 김용원 선생의 독립운동 행적에 무명의 이돈직 끼워넣기로 독립운동가를 만들려는 첫 시도라고 고발했다대전애국지사숭모회 등이 대전지역 곳곳에 세운 이인구 계룡건설 명예회장의 조부인 고 이돈직 비문에 새겨진 항일운동 행적과 애국지사 김용원 선생과 함께 독립운동을 했다는 비문 등은 확인되지 않았고 무리하게 끼워넣어진 것이라고 내용이다 특히 대전시로부터 지원을 받아 은평공원월평공원에 세워진 이돈직 생애비와 휘호비는 당초 사업 목적에도 맞지 않는데다 휘호비의 경우 불법 조형물로 확인됐다고 지적했다이에 대해 이 명예회장 등은 지난 2004년 4월 와 MBC를 상대로 모두 16억원 6억원 2심에서는 12억원 3억원의 손해배상을 청구하는 소송을 제기했다 그러나 법원은 이인구 계룡건설산업 명예회장 조부의 항일운동 행적은 확인되지 않은 것이라는 보도에 대해 허위 내용으로 볼 수 없다고 판단했다2006년 10월 대전고등법원 제 2민사부는 계룡건설 이인구 명예회장 등이 조부의 반일 항일투쟁 경력을 조작한 것처럼 허위사실을 보도해 명예를 훼손하고 계룡건설과 의 사회적 가치를 저하시키고 조부의 명예를 훼손시켰다며 등을 상대로 제기한 손해배상청구 소송에 대해 원심 판결 그대로 이유 없다며 기각했다 MBC에 대해서도 3000만원을 지급하고 정정보도 하라는 원심을 깨고 계룡건설 측의 청구를 기각했다 원고 측은 고등법원 판결 후 상고를 제기하지 않아 소송이 종결됐다재판부는 판결문에서 이인구 명예회장 조부가 반일 항일투쟁을 하고 애국지사 김용원과 독립운동을 하였다는 원고 측의 주장을 입증할 만한 실증적인 자료를 찾아보기 어려운 점 등에 비춰볼 때 기사 내용은 진실에 부합하거나 진실로 믿을 만한 상당한 이유가 있다고 판시했다재판부는 또 해당 단체가 국고를 지원받아 엉뚱한 사람의 공적비를 세웠다는 보도에 대해서도 대전애국지사숭모회가 대전시로부터 보조금을 신청하면서 계룡건설 이 명예회장 조부에 대해서는 사업 내용에 포함시키지 않았다며 진실에 부합된다고 덧붙였다이돈직을 기리는 기념비는 2003년 당시에만 무궁화동산 은평공원 효평동 이렇게 3곳이 있었다 2019년 현재 월평공원 휘호비와 생애비 무궁화동산 공적비와 휘호비는 철거된 상태다 힘있는 후손에 의해 조상의 공적이 부풀려지는 대표적인 사례다

---------- 671155번째 데이터 ----------
내 혼은 꽃비 되어 2013육우당 일기 미간행 육우당 일기의 일부는 추모집 내 혼은 꽃비 되어에 일부 수록 소개되었다

---------- 131932번째 데이터 ----------
하이디 대한민국의 가수하이디 대한민국의 걸그룹하이레이디 대한민국의 걸그룹멤버 엣지리더 보컬 바니랩 보컬 하양서브 보컬 세연메인 보컬하이레이디 대한민국의 걸그룹멤버 씨에라 서아란 이사랑 하연

---------- 1414414번째 데이터 ----------
영화 문서를 주로 편집하다보면 줄거리 단락에 스포일러 경고 틀을 넣어주시는 분을 종종 보게 됩니다 하지만 저는 스포일러 경고 틀의 실효성을 전혀 알지 못하겠군요 기준도 애매한 스포일러를 왜 경고해야 하는지 말이에요 영화에 대해서 알지 못한 채로 영화를 접하는 사람들에게는 영화에 대한 정보 그 자체가 스포일러가 될 수도 있지 않습니까 줄거리를 다루는 매체 문서에 스포일러가 있다는 것이 전재되는 것은 당연한거죠 최소한 줄거리 단락에서는 넣지 않아주셨으면 좋겠습니다 지침에 나와있기도 하구요 위키백과스포 애니메이션이나 소설 문서에서도 마찬가지 닭살튀김 토론 2009년 2월 25일 수 1814 KST동의합니다 저는 스포일러를 머리에 달아주는 것만으로 충분하다고 생각합니다 관련 틀을 전부 없애는 것도 생각해볼 수 있겠군요 정한솔 2009년 2월 26일 목 0021 KST

---------- 259178번째 데이터 ----------
아크에인절스Archangels품계는 팔품이지만 실제로는 천사들 중 가장 높고 위대한 천사들로 집단이나 민족 국가를 이끄는 사명을 지니고 있다

---------- 110268번째 데이터 ----------
문경지치로 일컬어지는 문제경제 시기의 전한은 내실을 다지는 데에 주력하였고 이를 바탕으로 무제 시기에는 활발한 정복 활동을 펼쳤다 무제는 장건위청곽거병을 등용하여 변방의 흉노를 물리치고 동월과 남월 또 조선을 정복하여 전성기를 이루었다

---------- 732180번째 데이터 ----------
전기회로에서 실제 전자의 흐름은 음극에서 양극으로 진행한다 그러나 최초 정의한 전류의 흐름은 실제 전자의 운동과 다르게 양극에서 음극인 양전하의 흐름으로 알려졌다 이처럼 실제 전류가 흐르는 방향이 정반대로 정의한 까닭은 전류의 흐름을 발견할 당시 과학자들이 전자의 존재를 몰랐기 때문이다장요한 외 공저 기초회로이론 학문사 6163쪽 정공 양이온처럼 양전하의 이동으로 말미암아 발생한 전류의 방향은 양전자의 이동 방향과 같은데 양전하가 이동할 때나 음전하가 이동할 때 만들어진 전류에 현상적인 차이는 없으므로 옛부터 전류의 방향을 양전하의 흐름으로 통일하였다장준성 외 공저 고등학교 물리 I 지학사 2002년 교육과학기술부 검정 p981830년대 마이클 패러데이는 아래 그림처럼 전해전도 실험을 하였다패러데이는 이 실험을 통해 전해질의 전도를 통해 축적한 은의 양을 측정하여 전류의 이동을 입증하였으며 현대 SI 단위를 정의하기 전까지 전류의 단위 1 암페어는 1초 동안 0001118 그램의 은을 축적한 전류의 세기로 정의했었다 또한 패러데이는 계속하여 새로운 은 원자를 제공하는 은막대를 양극anode은 원자가 축적되는 강철 쪽을 음극Cathode로 정의하고 전류가 양극에서 음극으로 흐른다고 보았다 이때문에 전류가 실제로는 전자의 흐름이라는 게 밝혀진 오늘날에도 전류의 방향은 실제 전자의 운동과는 반대로 여전히 양극에서 음극으로 흐른다고 정의한다

---------- 1103462번째 데이터 ----------
해당 직업에 지원하려면 자신이 어떤 프렌차이즈 사업 영역에 관심이 있어야 한다 호텔경영 외식산업과 같은 과를 졸업했을 경우 해당 직업을 지원하는데 유리한 조건을 갖추게 된다 또한 세법 가맹사업법 상법 부동산거래 실무와 같은 지식 역시 필요하다

---------- 137337번째 데이터 ----------
흔히 자동 판매기에서 판매하는 상품 또는 서비스는 다음과 같다음료 물 탄산음료 이온 음료 커피 우유담배라면스낵아이스크림샐러드과일피자맥주어묵성 관련 제품콘돔 생리대 속옷 여성용 팬티 등휴대전화 충전대중 교통의 승차권 또는 교통카드시설의 입장권도서요리할 수 있는 즉석 식품이처럼 자동 판매기에서 판매하는 상품들은 보존이 쉬운 것들이 대부분이다 그러니까 무인 편의점 역할을 하는 것을 기능하게 되는 것이 자동 판매기로 분류되는 것으로 규정된다 병원에 구비된 자판기도 미니 편의점 역할을 구실하는 사례가 있다

---------- 999890번째 데이터 ----------
2015년 FIVB 배구 월드그랑프리는 2015년 6월 26일부터 8월 2일까지 미국 오마하에서 열린 23번째 국제 여자 대회이다

---------- 1136074번째 데이터 ----------
CDDVD판 CD판 이벤트 회장 한정판의 CD 2종의 총 4개의 혀아로 발매타이틀곡의 의상은 후지타 니콜이 프로듀스했다미소의 프롤로그는 Best Friend 이래로 모든 멤버가 부르게 되는 노래이다 절대 히로인은 2016월 4일에 가입한 타카무라 유카 카도가키 히카루 카와구치 유리나 이지마 미카 이누즈카 시오리 나카자토 메구무의 6명에 의한 노래이다

---------- 912756번째 데이터 ----------
호메로스의 오디세이Homers Odyssey 는 서구세계의 가장 오래된 이야기의 하나이며 스릴러의 초기 원형으로 간주된다 영웅 오디세우스는 트로이 전쟁 이후 귀향길에 오르지만 아내 페넬로페와 만나기 위해 험한 모험을 겪어야 했다 몽테크리스토 백작은 복수에 관한 스릴러이다 드라큘라는 고딕 초자연적 스릴러이다 존 버컨의 1915년작 39계단은 스파이 스릴러 장르를 확립한 작품으로 여겨진다 추운 나라에서 온 스파이The Spy Who Came in from the Cold 존 러카레이는 냉전시대 첩보원들의 세계를 배경으로 한 작품이다 본 아이덴티티는 현대적인 스타일로 쓰여진 최초의 스릴러중 하나이다 람보는 현대 액션 소설의 아버지로 간주되고 있다로버트 러들럼 에릭 앰블러 데이비드 모렐 David Morrell 프레더릭 포사이스 댄 브라운 톰 클랜시 마이클 크라이튼 더글러스 프레스턴 Douglas Preston 링컨 차일드 Lincoln Child 이언 플레밍 존 그리셤 앨리스터 매클레인 Alistair MacLean 등이 유명하다

---------- 175203번째 데이터 ----------
초 위왕   기원전 329년은 중국 초나라의 제36대 군주재위 기원전 339년 기원전 329년이다 이름은 상이다 초 선왕의 아들이다

---------- 1239911번째 데이터 ----------
운비닐늄 또는 에카라듐은 미발견된 120번 원소의 이름이다 원자 번호는 120 원소 기호는 Ubn이다알칼리 토금속이다

---------- 278167번째 데이터 ----------
김종현 1995년 6월 8일 은 예명 JR으로 잘 알려진 대한민국의 가수이다 플레디스 엔터테인먼트 소속이며 뉴이스트의 리더 메인댄서 메인래퍼이다 2017년 PRODUCE 101 시즌2에 참가하였으며 최종 순위 14위를 차지하였다 이후 2018년까지 유닛인 뉴이스트 W로 활동하였다 20192020년 현재 뉴이스트로 활동 중이다

---------- 41090번째 데이터 ----------
감입곡류는 저작권 침해로 삭제 감입곡류하천은 이동 요청합니다 Hwimale 토론 2018년 7월 10일 화 1139 KST특수차이21862897 문서 이동을 확인 완료 표시합니다 메이 토론 2018년 8월 6일 월 2150 KST

---------- 329365번째 데이터 ----------
지방도 제721호선은 전라북도 남원시 도통동 갈치삼거리에서 완주군 상관면 마치리를 잇는 전라북도의 지방도이다

---------- 1113396번째 데이터 ----------
딤미들은 종교 세력이 개입된 것이 아니라면 독자적인 법적 체계를 갖춘 법정을 운용할 수 있었고 대중의 질서에 치명적인 해악을 끼치는 행위에 대해 처벌할 수 있었다 그러나 1819세기의 오스만 제국에서는 딤미도 대개 무슬림 법정에 참가했다 출석이 강제적인 경우 뿐 아니라 딤미 공동체 내의 사업상 거래와 재산을 기록하기 위한 목적도 있었다 무슬림을 대상으로 한 범죄나 딤미 혹은 딤미의 가족에게 저지른 행위에 대한 재판도 열었다 대개는 결혼과 이혼 상속 등의 문제가 샤리아에 따라 판결됐다 재판 전 맹세 서약을 하기도 했는데 무슬림의 맹세와도 동급으로 간주됐으며 사전에 딤미의 종교적 상황에 맞춰 맹세문을 다듬는 것을 허용했다alQattan 1999

---------- 787201번째 데이터 ----------
카를로스 모야 욤파르트 1976년 8월 27일은 전 세계 랭킹 1위였던 스페인의 은퇴한 테니스 선수이다 그는 1998년 프랑스 오픈에서 우승했으며 1997년 호주 오픈에서 준우승했다 2004년에는 데이비스 컵 스페인 국가대표로 출전하여 스페인의 우승에 기여하기도 했다그는 ATP 레벨의 대회에서 500 경기 이상 승리를 거둔 4명의 현역 선수 중 한 명이다 나머지 선수는 레이튼 휴이트와 앤디 로딕으로 이 선수들은 2009년 US 오픈 직전까지 각각 514승과 505승을 거두었다 마지막 한 명은 로저 페더러로 664승 을 거두

었다 모야는 총 573승 으로 4명의 선수 중 2번째로 많은 기록을 갖고 있다

---------- 1370455번째 데이터 ----------
con vow 각각 자음consonants 또는 모음vowel 표의 드러내기 상태를 지정합니다 기본 상태는 보여주지 않으며collapsed 아무 값이나 입력하면 펼쳐있는 상태plain로 고정됩니다constate vowstate 각각 자음 표와 모음 표의 드러내기 상태를 지정하는데 펼쳐져 있지만 닫을 수 있는expanded 등의 상태를 직접 지정할 수 있습니다

---------- 327069번째 데이터 ----------
주연 코바야시 카오루는 1993년 언덕 위의 해바라기 이후 16년 만에 민영 방송 연속극 주연을 맡았다2010년 4월 23일에는 모든 이야기가 담긴 DVD가 발매되었으며 MBS에서는 2010년과 2011년에 재방송을 실시하였다캐치프레이즈는 허기도 마음도 채워드립니다이다9화까지 부록 형식으로 매 회 그 회의 게스트 캐릭터가 다시 등장해 그 회의 부제인 요리를 해설하는 코너가 있다심야식당 시즌2는 MBS에서는 2011년 10월 13일부터 목요 심야 드라마 시간대2455 2525에 TBS에서는 2011년 10월 18일부터 화요 심야시간대2455 2525에 방송된다 화수는 시즌1에 이어 11화부터 시작 첫회는 제66회 문화청 예술 참가 작품으로써 방송되었다

---------- 1247617번째 데이터 ----------
산울림 13집은 그룹 산울림의 열세 번째 정규 음반이자 마지막 정규 음반이다 1984 이후 사업 등으로 산울림 활동을 멈춘 두 동생 멤버 김창훈과 김창익이 13년 만에 다시 재결성하여 만든 음반이다 이후 산울림은 2007년경 14집 준비를 위해 재결성을 선언했으나 드러머이자 막내인 김창익이 2008년 캐나다에서 사망하면서 산울림은 해체를 선언 이 음반은 마지막 정규 음반이 되었다

---------- 791743번째 데이터 ----------
파리정치대학Institut dtudes Politiques de Paris은 프랑스의 사회과학 중심의 최상위 명문 그랑제콜이다 줄여서 시앙스포Sciences Po라고 불리기도 한다 입학률이 813로 입학이 굉장히 까다로운 학교로 알려져 있으며 소수 정예의 우수한 학생들만 선발한다 프랑스 전현직 대통령 국회의원 외교관 등 정계 주요인사를 가장 많이 배출한 프랑스 고위 엘리트 양성 대학으로국제관계 및 정치분야에서 프랑스 최고 동문 네트워크를 자랑한다영 Just Landed 정치권에서는 에마뉘엘 마크롱 대통령 프랑수아 올랑드 대통령 자크 시라크 대통령과 프랑수아 미테랑 대통령 등 프랑스 국내외 30명의 대통령 31명의 국무총리 21명의 외무부장관 및 유엔 사무총장을 배출했으며 프랑스 국회의원과 외교관의 거의 대부분은 이 학교 동문 출신들이다 그랑제콜 특성상 소수정예의 작은 학교 규모에도 불구하고 2020년 QS 세계 대학 순위에서 정치 및 국제관계 분야 세계 2위로 평가되며 해당 분야에 있어 세계 최고 수준으로 평가받고 있다

---------- 103355번째 데이터 ----------
흐로닝언에 나타난 최초의 인류 흔적은 기원전 3000년까지 거슬러 올라간다 고대 로마 제국 갈리아 벨기카 주둔군의 숙영지가 이 지역에 설치되었던 것이 역사에 처음 기록된 부분이다 기원 후 1000년 무렵 흐로닝언은 wooden church가 있는 곳으로 알려졌다 1024년 위트레흐트 주교령이 되었고 흐로닝은 Villa Cruoninga 혹은 Green River라고 불렸다 이 시기 도시는 급속히 발전하였고 사람들은 당시 전 유럽에서 그랬듯 주교에게서 완전한 도시의 독립을 얻길 바랐다 1284년 흐로닝언은 한자 동맹에 가입하였다 그리고 1405년 흐로닝은 마침내 자치 정부를 꾸리게 됐다 같은 시기 둘로 지은 집들이 지어지고 도시는 Hunze강을 통한 해상 무역 덕에 더 빨리 성장해갔다 국제 무역이 더 긴급해지면서 도시는 국제적인 무역에 영향을 미치던 Hanze그룹에 가입하게 됐다

---------- 1284372번째 데이터 ----------
에미시 또는 에비스 또는 에조는 일본 혼슈의 간토 지방 도호쿠 지방과 홋카이도 지역에 살면서 일본인야마토 민족에 의해 이민족시 되었던 민족집단을 일컫는 말이다 시대에 따라 그 지칭범위가 다른데 일반적으로 근세의 에조는 특히 아이누민족을 일컫는다 에조는 일본 동부 북부 지역 뿐만 아니라 쿠릴 열도 사할린 심지어는 캄차카 지방까지 정착해 살았다

---------- 1262752번째 데이터 ----------
ANAPG153 레이다 F5E 타이거1ANAPG157 레이다 F5F 타이거1ANAPG159 레이다 KF5EF 타이거2 타이거1에 비해 탐색거리 2배 증가ANAPG63 레이시온 공대공레이다 F15CDANAPG63v1v3 공대공공대지 F15Kv1기계식 펄스도플러방식 최대 161km F15 F15SG v3 AESA방식 1500개소자 v2는 알라스카배치 F15기체에만 시험적으로 소량 장착 너무 비쌈 소자 1500개 사각형모양 v3는 F15CD미군기체에만 현재 장착중이다 소자수는 1500개 다각형모양 v4는 v3이후 나온것으로 82v1일것으로 추정된다ANAPG65 레이시온 FA18CD AV8B F4F ICE 다중모드 화력통제 레이다ANAPG66 노스럽 그러먼 F16AB 95km F4EJ 호크200단좌모델 소형 화력통제 레이다 AIM120 공대공미사일에도 직경 15cm 공간에 휴즈사에 의해 축약해 넣어졌다ANAPG67 록히드마틴LM F20 FCK1GD53105km TA50v4시제기 80km 소형 화력통제 레이다 정확하게는 ANAPG67은 LMMSS제작으로 LM과 동일회사는 아니라고함관계회사인듯v4는 옵션으로 SAR 기능을 추가할 수 있다ANAPG68 노스럽 그러먼 F16CD ANAPG66 개량 F16K블록32초기형 115km KF16v5v7 129km 가장최신형으론 SAR기능이 추가된 AGP68v9가 있으며 이스라엘F16에 미국의 압력으로 ELM2032대신 68v9가 장착되었다ANAPG70ANAPG63SAR 레이시온 F15CDE ANAPG63을 개량해 NCTR표적식별기능을 추가함 ANAPG63 v1v2 보다는 하위의 레이다이다 지형추적기능의 공대지모드 자동포착수색범위 185km26m급 목표식별 지상매핑 수색범위 92km 공대공 185km 해상수색범위 최대 37km10m급 목표식별ANAPG71 레이시온 F14D ANAWG9 개량ANAPG73 레이시온 FA18CDEF ANAPG65 개량ANAPG76 노스럽 그러먼 F4 팬텀 2000 이스라엘 F4 업그레이드용ANAPG77 노스럽 그러먼 FA22 AESA 방식 모든 방향에 대해 탐색할 수 있도록 여러개로 되어 기수뿐 아니라 동체 다른부분에도 장착되며 초기 2001년경 12세대 소자의 경우 1500개 였으나 이후 2000개를 거쳐 2200개 까지 증가하였다 원형모양ANAPG79 레이시온 FA18EF수퍼호넷 1100개 모듈 AESA 방식 실전배치 AESA중에서 최초로 공대지 공대공 동시추적이 가능해진 레이다 ANAPG73 RUG 3 의 개량형 ANAPG77의 기술이 적용되었음 2005년에 4세대소자로 테스트RACRRaytheon Advanced Combat Radar AESA 레이시온APG79를 바탕으로 개발 F16 및 FA18ACD 를 위해 개발중에 있으며 경쟁사의 SABR보다 빠른 개발진척도를 보이고 있다 KF16에 이미 제안되었으며 한국공군의 장교와 실무자들에게 기밀브리핑까지 마쳤다ANAPG80 노스럽 그러먼 F16EF 1000개 모듈 AESA 방식 ANAPG68의 주요부분을 재설계원형모양 ANAPG68v7 의 거의 2배의 공대공탐지거리를 제공한다ANAPG81 노스럽 그러먼 F35ABC ANAPG77에서 소자만 1200개로 줄인 축소형레이다 3세대소자 원형모양 아직 개발중인 F35와 달리 81레이다는 이미 실전배치 가능한 수준으로 개발되어있으나 82레이다보다 숫자가 하나 적어 마케팅상의 문제를 토로중임SABR확장식 고속빔레이다 노스럽 그러먼 ANAPG81을 변형한 형태이며 81의 다운그레이드형으로 여겨지기도 하며 ANAPG80에 기반하고있다 F16의 레이다개량시장에 대한 수출형모델로 아직 개발중에 있다 14kg 의 중량 기존 F16의 구조 전력 냉각장치 개조없이 탑재가능 가격도 기존 AESA보다 낮고 기존기계식레이다와 동일가격판매를 선언했다 KFX나 F50 에도 장착가능하다고 2009 서울에어쇼ADEX에서 언급하였으나 한국에 대한 기술이전은 매우 제한될것으로 복수의 미국현지 소식통에 의해 확인되고 있다 TA50 FA50 에의 장착도 긍정적으로 검토되고 있다 SABR레이다의안테나와 백엔드 부분은 F16보다 크거나 작은 항공기를 위해 크기조정이 가능하게 설계되어있다 미국내 F16시장도 바라보고있다 ANAPG80보다 적은 공간 전력냉각장치로 설치될 수 있다고 홍보되고 있다 노스럽 그러먼이 미공군훈련기도입사업인 TX사업용으로 M346에 장착을 위해 이탈리아와 접촉중이다 오래도록 미국정부의 수출허가가 떨어지지 않다가 2010년 2월2일 수출상담용 기술정보 제공허가가 떨어졌다ANAPG82 레이시온 F15SE에 탑재가 예상되는 AESA레이다 APG63v3의 안테나와 APG79의 신형 backend processor 를 통합했다고 함 2009년 9월 F15E에도 장착할 레이다로 ANAPG82v1이 선정되었으며 2014년 초기운용능역 확보예정

---------- 184779번째 데이터 ----------
35px M38 간선도로 프나티슈스코예 조지프카 679 km35px A350 간선도로 조지프카 알마티 933 km35px M36 간선도로 316 km해당 도로는 경유하는 도시가 미상이다

---------- 1396025번째 데이터 ----------
제2차 후크고지 전투는 한국 전쟁 기간 중인 1952년 11월 18일부터 19일 사이에 영연방 제1사단과 중공군이 후크고지라는 지역에서 맞붙은 전투이다 후크고지는 이전 달에도 중공군과 격전을 치른 지역이었고 이는 후크고지가 전략적 요충지임을 보여준다 중공군은 이러한 전략적 요충지를 점령하기 위해 2번째로 후크고지를 공격했지만 중화기 사격과 효과적인 반격으로 중공군의 공격은 실패로 끝났다

---------- 1470485번째 데이터 ----------
부에노스아이레스 대사관브라질리아 대사관 리우데자네이루 총영사관 쿠리치바 영사관오타와 대사관 토론토 총영사관아바나 대사관멕시코시티 대사관리마 대사관워싱턴 DC 대사관 시카고 총영사관 뉴욕 총영사관 샌프란시스코 총영사관

---------- 989436번째 데이터 ----------
첫 출마 당시부터 아카시 해협 대교의 건설을 제창한 하라 겐자부로의 호방한 캐릭터에 의한 연설은 꿈같은 이야기로 붙잡혀 하라켄 별명이 많은 야유를 받기도 했지만 1986년에 드디어 착공을 실현했다 착공 후에도 연설에서 하라켄을 떨어뜨리면 다리도 떨어진다라는 문구를 이용하는 등 자신의 실적을 어필했다노동대신 재임 중 당시 이례적으로 관공서의 주 5일 근무제를 처음으로 제창했다

이 정도면 전처리가 잘 진행된 것 같습니다!

토큰화 (~50 mins)

다음 단계는 토큰화입니다. 토큰화를 위한 다양한 패키지들이 있습니다.

여기서는 `konlpy`, `nltk`, `spaCy` 중 고민해보겠습니다.

세 패키지 모두 자연어처리를 위한 패키지입니다.

`바른`으로 토큰화하는 경우, 가장 좋은 성능을 보이지만 설정 과정이 복잡할 수 있으므로 생략하겠습니다.

In [ ]:
```
# spacy를 위한 모듈 불러오기
!python -m spacy download ko_core_news_sm
```

```
Collecting ko-core-news-sm==3.7.0
  Downloading https://github.com/explosion/spacy-models/releases/download/ko_core_news_sm-3.7.0/ko_core_news_sm-3.7.0-py3-none-any.whl (14.7 MB)
     ---------------------------------------- 0.0/14.7 MB ? eta -:--:--
     ---------------------------------------- 0.0/14.7 MB 330.3 kB/s eta 0:00:45
     ---------------------------------------- 0.2/14.7 MB 1.8 MB/s eta 0:00:08
     - -------------------------------------- 0.7/14.7 MB 5.2 MB/s eta 0:00:03
     ---- ----------------------------------- 1.5/14.7 MB 8.6 MB/s eta 0:00:02
     ------ --------------------------------- 2.6/14.7 MB 11.6 MB/s eta 0:00:02
     ---------- ----------------------------- 4.2/14.7 MB 15.8 MB/s eta 0:00:01
     --------------- ------------------------ 6.1/14.7 MB 19.5 MB/s eta 0:00:01
     --------------------- ------------------ 8.2/14.7 MB 22.9 MB/s eta 0:00:01
     ---------------------------- ---------- 10.6/14.7 MB 32.8 MB/s eta 0:00:01
     ---------------------------------- ----- 13.0/14.7 MB 43.7 MB/s eta 0:00:01
     ---------------------------------------  14.7/14.7 MB 50.1 MB/s eta 0:00:01
     ---------------------------------------- 14.7/14.7 MB 40.9 MB/s eta 0:00:00
Requirement already satisfied: spacy<3.8.0,>=3.7.0 in c:\users\junho\anaconda3\envs\clean-up-week2\lib\site-packages (from ko-core-news-sm==3.7.0) (3.7.2)
Requirement already satisfied: spacy-legacy<3.1.0,>=3.0.11 in c:\users\junho\anaconda3\envs\clean-up-week2\lib\site-packages (from spacy<3.8.0,>=3.7.0->ko-cor
e-news-sm==3.7.0) (3.0.12)
Requirement already satisfied: spacy-loggers<2.0.0,>=1.0.0 in c:\users\junho\anaconda3\envs\clean-up-week2\lib\site-packages (from spacy<3.8.0,>=3.7.0->ko-cor
e-news-sm==3.7.0) (1.0.4)
Requirement already satisfied: murmurhash<1.1.0,>=0.28.0 in c:\users\junho\anaconda3\envs\clean-up-week2\lib\site-packages (from spacy<3.8.0,>=3.7.0->ko-core-
news-sm==3.7.0) (1.0.7)
Requirement already satisfied: cymem<2.1.0,>=2.0.2 in c:\users\junho\anaconda3\envs\clean-up-week2\lib\site-packages (from spacy<3.8.0,>=3.7.0->ko-core-news-s
m==3.7.0) (2.0.6)
Requirement already satisfied: preshed<3.1.0,>=3.0.2 in c:\users\junho\anaconda3\envs\clean-up-week2\lib\site-packages (from spacy<3.8.0,>=3.7.0->ko-core-news
-sm==3.7.0) (3.0.6)
Requirement already satisfied: thinc<8.3.0,>=8.1.8 in c:\users\junho\anaconda3\envs\clean-up-week2\lib\site-packages (from spacy<3.8.0,>=3.7.0->ko-core-news-s
m==3.7.0) (8.2.2)
Requirement already satisfied: wasabi<1.2.0,>=0.9.1 in c:\users\junho\anaconda3\envs\clean-up-week2\lib\site-packages (from spacy<3.8.0,>=3.7.0->ko-core-news-
sm==3.7.0) (0.9.1)
Requirement already satisfied: srsly<3.0.0,>=2.4.3 in c:\users\junho\anaconda3\envs\clean-up-week2\lib\site-packages (from spacy<3.8.0,>=3.7.0->ko-core-news-s
m==3.7.0) (2.4.8)
Requirement already satisfied: catalogue<2.1.0,>=2.0.6 in c:\users\junho\anaconda3\envs\clean-up-week2\lib\site-packages (from spacy<3.8.0,>=3.7.0->ko-core-ne
ws-sm==3.7.0) (2.0.10)
Requirement already satisfied: weasel<0.4.0,>=0.1.0 in c:\users\junho\anaconda3\envs\clean-up-week2\lib\site-packages (from spacy<3.8.0,>=3.7.0->ko-core-news-
sm==3.7.0) (0.3.4)
Requirement already satisfied: typer<0.10.0,>=0.3.0 in c:\users\junho\anaconda3\envs\clean-up-week2\lib\site-packages (from spacy<3.8.0,>=3.7.0->ko-core-news-
sm==3.7.0) (0.9.0)
Requirement already satisfied: smart-open<7.0.0,>=5.2.1 in c:\users\junho\anaconda3\envs\clean-up-week2\lib\site-packages (from spacy<3.8.0,>=3.7.0->ko-core-n
ews-sm==3.7.0) (5.2.1)
Requirement already satisfied: tqdm<5.0.0,>=4.38.0 in c:\users\junho\anaconda3\envs\clean-up-week2\lib\site-packages (from spacy<3.8.0,>=3.7.0->ko-core-news-s
m==3.7.0) (4.65.0)
Requirement already satisfied: requests<3.0.0,>=2.13.0 in c:\users\junho\anaconda3\envs\clean-up-week2\lib\site-packages (from spacy<3.8.0,>=3.7.0->ko-core-ne
ws-sm==3.7.0) (2.31.0)
Requirement already satisfied: pydantic!=1.8,!=1.8.1,<3.0.0,>=1.7.4 in c:\users\junho\anaconda3\envs\clean-up-week2\lib\site-packages (from spacy<3.8.0,>=3.7.
0->ko-core-news-sm==3.7.0) (2.5.3)
Requirement already satisfied: jinja2 in c:\users\junho\anaconda3\envs\clean-up-week2\lib\site-packages (from spacy<3.8.0,>=3.7.0->ko-core-news-sm==3.7.0) (3.
1.3)
Requirement already satisfied: setuptools in c:\users\junho\anaconda3\envs\clean-up-week2\lib\site-packages (from spacy<3.8.0,>=3.7.0->ko-core-news-sm==3.7.0)
(68.2.2)
Requirement already satisfied: packaging>=20.0 in c:\users\junho\anaconda3\envs\clean-up-week2\lib\site-packages (from spacy<3.8.0,>=3.7.0->ko-core-news-sm==
3.7.0) (23.2)
Requirement already satisfied: langcodes<4.0.0,>=3.2.0 in c:\users\junho\anaconda3\envs\clean-up-week2\lib\site-packages (from spacy<3.8.0,>=3.7.0->ko-core-ne
ws-sm==3.7.0) (3.3.0)
Requirement already satisfied: numpy>=1.19.0 in c:\users\junho\anaconda3\envs\clean-up-week2\lib\site-packages (from spacy<3.8.0,>=3.7.0->ko-core-news-sm==3.
7.0) (1.26.4)
Requirement already satisfied: annotated-types>=0.4.0 in c:\users\junho\anaconda3\envs\clean-up-week2\lib\site-packages (from pydantic!=1.8,!=1.8.1,<3.0.0,>=
1.7.4->spacy<3.8.0,>=3.7.0->ko-core-news-sm==3.7.0) (0.6.0)
Requirement already satisfied: pydantic-core==2.14.6 in c:\users\junho\anaconda3\envs\clean-up-week2\lib\site-packages (from pydantic!=1.8,!=1.8.1,<3.0.0,>=1.
7.4->spacy<3.8.0,>=3.7.0->ko-core-news-sm==3.7.0) (2.14.6)
Requirement already satisfied: typing-extensions>=4.6.1 in c:\users\junho\anaconda3\envs\clean-up-week2\lib\site-packages (from pydantic!=1.8,!=1.8.1,<3.0.0,>
=1.7.4->spacy<3.8.0,>=3.7.0->ko-core-news-sm==3.7.0) (4.9.0)
Requirement already satisfied: charset-normalizer<4,>=2 in c:\users\junho\anaconda3\envs\clean-up-week2\lib\site-packages (from requests<3.0.0,>=2.13.0->spacy
<3.8.0,>=3.7.0->ko-core-news-sm==3.7.0) (2.0.4)
Requirement already satisfied: idna<4,>=2.5 in c:\users\junho\anaconda3\envs\clean-up-week2\lib\site-packages (from requests<3.0.0,>=2.13.0->spacy<3.8.0,>=3.
7.0->ko-core-news-sm==3.7.0) (3.4)
Requirement already satisfied: urllib3<3,>=1.21.1 in c:\users\junho\anaconda3\envs\clean-up-week2\lib\site-packages (from requests<3.0.0,>=2.13.0->spacy<3.8.
0,>=3.7.0->ko-core-news-sm==3.7.0) (2.1.0)
Requirement already satisfied: certifi>=2017.4.17 in c:\users\junho\anaconda3\envs\clean-up-week2\lib\site-packages (from requests<3.0.0,>=2.13.0->spacy<3.8.
0,>=3.7.0->ko-core-news-sm==3.7.0) (2024.2.2)
Requirement already satisfied: blis<0.8.0,>=0.7.8 in c:\users\junho\anaconda3\envs\clean-up-week2\lib\site-packages (from thinc<8.3.0,>=8.1.8->spacy<3.8.0,>=
3.7.0->ko-core-news-sm==3.7.0) (0.7.9)
Requirement already satisfied: confection<1.0.0,>=0.0.1 in c:\users\junho\anaconda3\envs\clean-up-week2\lib\site-packages (from thinc<8.3.0,>=8.1.8->spacy<3.
8.0,>=3.7.0->ko-core-news-sm==3.7.0) (0.1.4)
Requirement already satisfied: colorama in c:\users\junho\anaconda3\envs\clean-up-week2\lib\site-packages (from tqdm<5.0.0,>=4.38.0->spacy<3.8.0,>=3.7.0->ko-c
ore-news-sm==3.7.0) (0.4.6)
Requirement already satisfied: click<9.0.0,>=7.1.1 in c:\users\junho\anaconda3\envs\clean-up-week2\lib\site-packages (from typer<0.10.0,>=0.3.0->spacy<3.8.0,>
=3.7.0->ko-core-news-sm==3.7.0) (8.1.7)
Requirement already satisfied: cloudpathlib<0.17.0,>=0.7.0 in c:\users\junho\anaconda3\envs\clean-up-week2\lib\site-packages (from weasel<0.4.0,>=0.1.0->spacy
<3.8.0,>=3.7.0->ko-core-news-sm==3.7.0) (0.16.0)
Requirement already satisfied: MarkupSafe>=2.0 in c:\users\junho\anaconda3\envs\clean-up-week2\lib\site-packages (from jinja2->spacy<3.8.0,>=3.7.0->ko-core-ne
ws-sm==3.7.0) (2.1.3)
✓ Download and installation successful
You can now load the package via spacy.load('ko_core_news_sm')
```

In [ ]:
```
import nltk
nltk.download('punkt')
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     C:\Users\Junho\AppData\Roaming\nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

Out[ ]: True

In [ ]:
```
# 토큰화 예시 비교
from konlpy.tag import Okt
from konlpy.tag import Kkma
from nltk import word_tokenize
import spacy
import time

okt = Okt()
```

```
kkma = Kkma()
nlp = spacy.load("ko_core_news_sm")
```

In [ ]:
```
s1 = time.time()
print(okt.morphs(train[0]))
e1 = time.time()
print(f"okt: {e1 - s1:.4f}초")

s1 = time.time()
print(kkma.morphs(train[0]))
e1 = time.time()
print(f"kkma: {e1 - s1:.4f}초")

s1 = time.time()
print(word_tokenize(train[0]))
e1 = time.time()
print(f"word_tokenize: {e1 - s1:.4f}초")

s1 = time.time()
doc = nlp(train[0])
print([token.text for token in doc])
e1 = time.time()
print(f"spacy: {e1 - s1:.4f}초")
```

```
['1931년', '서울시', '에서', '태어나', '경기', '중학교', '1982년', '4월', '12일', '자', '매일경제', '서울', '고등학교', '1981년', '4월', '25일', '자', '동아일보', '1956년', '서울대학교', '법학', '과를', '나온', '후', '1956년', '제', '8회', '고등고시', '사법', '과', '에서', '합격', '하였다', '1958년', '서울', '지방검찰청', '검사', '에', '임용', '되었다', '김세', '권', '은', '두산', '그룹', '창업', '주인', '박두병', '딸', '인', '박', '용언', '과', '결혼', '했다', '김세', '권', '과', '박', '용언', '은', '아들', '은', '1970년', '대', '봉제', '업', '으로', '성장한', '태흥', '의', '창업', '주', '권태', '흥', '의', '딸', '권', '혜경', '과', '결혼', '한', '김형일', '일', '경', '산업', '개발', '부회장', '으로', '1990년', '대', '초반', '대한민국', '에', '게스', '폴로', '등', '을', '수입', '해', '유명', '세', '를', '탔던', '기업가', '다', '딸', '김희정', '의', '남편', '은', '최원', '현', '케이', '씨엘', '대표', '변호사', '다', '박승직', '박두병', '박용곤', '박정원', '재벌', '가', '4', '대', '33', '명', '결혼', '스토리', '대검찰청', '차장', '으로', '재직', '하던', '1986년', '2월', '전국', '검사', '장', '회의', '를', '주재', '하면서', '국법', '질서', '확립', '을', '위', '한', '검찰', '의', '과제', '라는', '주제', '로', '토의', '를', '하여', '국법', '질서', '와', '사회', '기강', '확립', '경', '제', '도약', '의', '전기', '가', '될', '수', '있는', '현재', '의', '국내외', '경', '제', '여건', '을', '최대한', '유지', '활용', '할', '수', '있도록', '경제', '질서', '교란', '사범', '엄단', '민생', '을', '불안하게', '하는', '반', '윤리', '적', '인', '강력', '사범', '에', '대한', '단호', '한', '응징', '사회', '적', '신분', '과', '지위', '의', '고하', '를', '가리지', '않', '는', '엄정', '공평한', '법', '집행', '을', '1986년', '도', '검찰', '권', '행사', '의', '방향', '으로', '정', '했다', '1986년', '2월', '25일', '자', '동아일보']
okt: 2.7675초
['1931', '년', '서울시', '에서', '태어나', '경기중학교', '1982', '년', '4', '월', '12', '일자', '매일', '경제', '서울', '고등학교', '1981', '년', '4', '월', '25', '일자', '동아', '일보', '1956', '년', '서울대', '학교', '법', '학과', '를', '나오', 'ㄴ', '후', '1956', '년', '저', '의', '8', '회', '고등', '고시', '사', '법과', '에서', '합격', '하', '였', '다', '1958', '년', '서울', '지방', '검찰청', '검사', '에', '임용', '되', '었', '다', '김', '세권', '은', '두', '산', '그룹', '창업주', '이', 'ㄴ', '박두병', '딸', '이', 'ㄴ', '박용', '언', '과', '결혼', '하', '었', '다', '김', '세권', '과', '박용', '언', '은', '아들', '은', '1970', '년대', '봉제업', '으로', '성장', '하', 'ㄴ', '태흥', '의', '창업주', '권태', '흥', '의', '딸', '권혜경', '과', '결혼', '하', 'ㄴ', '김', '형', '일', '일경', '산업', '개발', '부회장', '으로', '1990', '년대', '초반', '대한민국', '에', '게스폴', '로', '등', '을', '수입', '하', '어', '유명세', '를', '타', '었', '더', 'ㄴ', '기업가', '다', '따', 'ㄹ', '김희정', '의', '남편', '은', '최', '원', '현', '케이', '씨', '엘', '대표', '변호사', '다박', '승직', '박두', '병', '박용', '골', 'ㄴ', '박', '정원', '재벌', '가', '아', '4', '대', '33', '명', '결혼', '스토리', '대', '검찰청', '차장', '으로', '재직', '하', '더', 'ㄴ', '1986', '년', '2', '월', '전국', '검사장', '회의', '를', '주재', '하', '면서', '국법', '질서', '확립', '을', '위하', 'ㄴ', '검찰', '의', '과제', '이', '라는', '주제', '로', '토의', '를', '하여', '국법', '질서', '와', '사회', '기강', '확립', '경제', '도', '약', '의', '전기', '가', '되', 'ㄹ', '수', '있', '는', '현재', '의', '국내외', '경제', '여건', '을', '최대한', '유지', '활용', '하', 'ㄹ', '수', '있', '도록', '경제', '질서', '교란', '사범', '엄단', '민생', '을', '불안', '하', '게', '하', '는', '반', '윤리적', '이', 'ㄴ', '강력', '사범', '에', '대하', 'ㄴ', '단호', '하', 'ㄴ', '응징', '사회적', '신분', '과', '지위', '의', '고하', '를', '가리', '지', '않', '는', '엄정', '공평', '하', 'ㄴ', '법', '집행', '을', '1986', '년도', '검찰권', '행사', '의', '방향', '으로', '정하', '었', '다', '1986', '년', '2', '월', '25', '일자', '동아', '일보']
kkma: 3.8420초
['1931년', '서울시에서', '태어나', '경기중학교1982년', '4월', '12일자', '매일경제', '서울고등학교1981년', '4월', '25일자', '동아일보', '1956년', '서울대학교', '법학과를', '나온', '후', '1956년', '제8회', '고등고시', '사법과에서', '합격하였다', '1958년', '서울지방검찰청', '검사에', '임용되었다김세권은', '두산그룹', '창업주인', '박두병', '딸인', '박용언과', '결혼했다', '김세권과', '박용언은', '아들은', '1970년대', '봉제업으로', '성장한', '태흥의', '창업주', '권태흥의', '딸', '권혜경과', '결혼한', '김형일', '일경산업개발', '부회장으로', '1990년대', '초반', '대한민국에', '게스폴로', '등을', '수입해', '유명세를', '탔던', '기업가다', '딸', '김희정의', '남편은', '최원현', '케이씨엘', '대표변호사다박승직', '박두병', '박용곤', '박정원', '재벌가', '4대', '33명', '결혼', '스토리대검찰청', '차장으로', '재직하던', '1986년', '2월', '전국', '검사장', '회의를', '주재하면서', '국법질서', '확립을', '위한', '검찰의', '과제라는', '주제로', '토의를', '하여', '국법질서와', '사회기강', '확립', '경제도약의', '전기가', '될', '수', '있는', '현재의', '국내외', '경제여건을', '최대한', '유지', '활용할', '수', '있도록', '경제질서교란사범', '엄단', '민생을', '불안하게', '하는', '반윤리적인', '강력사범에', '대한', '단호한', '응징', '사회적', '신분과', '지위의', '고하를', '가리지', '않는', '엄정', '공평한', '법집행을', '1986년도', '검찰권', '행사의', '방향으로', '정했다1986년', '2월', '25일자', '동아일보']
word_tokenize: 0.0339초
['1931년', '서울시에서', '태어나', '경기중학교1982년', '4월', '12일자', '매일경제', '서울고등학교1981년', '4월', '25일자', '동아일보', '1956년', '서울대학교', '법학과를', '나온', '후', '1956년', '제8회', '고등고시', '사법과에서', '합격하였다', '1958년', '서울지방검찰청', '검사에', '임용되었다김세권은', '두산그룹', '창업주인', '박두병', '딸인', '박용언과', '결혼했다', ' ', '김세권과', '박용언은', '아들은', '1970년대', '봉제업으로', '성장한', '태흥의', '창업주', '권태흥의', '딸', '권혜경과', '결혼한', '김형일', '일경산업개발', '부회장으로', '1990년대', '초반', '대한민국에', '게스폴로', '등을', '수입해', '유명세를', '탔던', '기업가다', '딸', '김희정의', '남편은', '최원현', '케이씨엘', '대표변호사다박승직', '박두병', '박용곤', '박정원', '재벌가', '4대', '33명', '결혼', '스토리대검찰청', '차장으로', '재직하던', '1986년', '2월', '전국', '검사장', '회의를', '주재하면서', '국법질서', '확립을', '위한', '검찰의', '과제라는', '주제로', '토의를', '하여', '국법질서와', '사회기강', '확립', '경제도약의', '전기가', '될', '수', '있는', '현재의', '국내외', '경제여건을', '최대한', '유지', '활용할', '수', '있도록', '경제질서교란사범', '엄단', '민생을', '불안하게', '하는', '반윤리적인', '강력사범에', '대한', '단호한', '응징', '사회적', '신분과', '지위의', '고하를', '가리지', '않는', '엄정', '공평한', '법집행을', '1986년도', '검찰권', '행사의', '방향으로', '정했다1986년', '2월', '25일자', '동아일보']
spacy: 0.3171초
```

토큰화 예시에서 OKT가 시간 대비 제일 괜찮은 결과를 주는 것 같습니다.

따라서 OKT로 토큰화 해줄게요

이 작업은 상당히 오래 걸리는 작업입니다 (20만개 기준 약 40분 소요)

rand_idx 부분의 코드에서 원하는 데이터만큼 변경해주면 됩니다.

In [ ]:
```
import numpy as np

tokens = []

np.random.seed(42)

rand_idx = np.random.randint(0, len(train), 200000)

start = time.time()

for i, idx in enumerate(rand_idx):
    if i % 1000 == 0:
        print(f"{i}번째 데이터 처리 중...")
        t1 = time.time()
        print(f"걸린 시간: {t1 - start:.2f}초")
    tokens.append(okt.morphs(train[idx]))

end = time.time()
print(f"걸린 시간: {end - start:.2f}초")
```

0번째 데이터 처리 중...
걸린 시간: 0.00초
1000번째 데이터 처리 중...
걸린 시간: 8.23초
2000번째 데이터 처리 중...
걸린 시간: 16.75초
3000번째 데이터 처리 중...
걸린 시간: 25.47초
4000번째 데이터 처리 중...
걸린 시간: 33.78초
5000번째 데이터 처리 중...
걸린 시간: 42.13초
6000번째 데이터 처리 중...
걸린 시간: 49.74초
7000번째 데이터 처리 중...
걸린 시간: 57.60초
8000번째 데이터 처리 중...
걸린 시간: 67.25초
9000번째 데이터 처리 중...
걸린 시간: 75.11초
10000번째 데이터 처리 중...
걸린 시간: 84.26초
11000번째 데이터 처리 중...
걸린 시간: 92.45초
12000번째 데이터 처리 중...
걸린 시간: 100.51초
13000번째 데이터 처리 중...
걸린 시간: 107.80초
14000번째 데이터 처리 중...
걸린 시간: 115.54초
15000번째 데이터 처리 중...
걸린 시간: 122.74초
16000번째 데이터 처리 중...
걸린 시간: 129.97초
17000번째 데이터 처리 중...
걸린 시간: 137.40초
18000번째 데이터 처리 중...
걸린 시간: 145.27초
19000번째 데이터 처리 중...
걸린 시간: 154.98초
20000번째 데이터 처리 중...
걸린 시간: 163.73초
21000번째 데이터 처리 중...
걸린 시간: 171.31초
22000번째 데이터 처리 중...
걸린 시간: 179.98초
23000번째 데이터 처리 중...
걸린 시간: 188.61초
24000번째 데이터 처리 중...
걸린 시간: 197.66초
25000번째 데이터 처리 중...
걸린 시간: 214.03초
26000번째 데이터 처리 중...
걸린 시간: 222.74초
27000번째 데이터 처리 중...
걸린 시간: 230.69초
28000번째 데이터 처리 중...
걸린 시간: 240.01초
29000번째 데이터 처리 중...
걸린 시간: 247.39초
30000번째 데이터 처리 중...
걸린 시간: 255.50초
31000번째 데이터 처리 중...
걸린 시간: 263.79초
32000번째 데이터 처리 중...
걸린 시간: 271.47초
33000번째 데이터 처리 중...
걸린 시간: 279.57초
34000번째 데이터 처리 중...
걸린 시간: 287.53초
35000번째 데이터 처리 중...
걸린 시간: 295.77초
36000번째 데이터 처리 중...
걸린 시간: 304.19초
37000번째 데이터 처리 중...
걸린 시간: 313.11초
38000번째 데이터 처리 중...
걸린 시간: 321.74초
39000번째 데이터 처리 중...
걸린 시간: 330.76초
40000번째 데이터 처리 중...
걸린 시간: 339.20초
41000번째 데이터 처리 중...
걸린 시간: 347.47초
42000번째 데이터 처리 중...
걸린 시간: 355.31초
43000번째 데이터 처리 중...
걸린 시간: 364.12초
44000번째 데이터 처리 중...
걸린 시간: 371.66초
45000번째 데이터 처리 중...
걸린 시간: 382.07초
46000번째 데이터 처리 중...
걸린 시간: 391.00초
47000번째 데이터 처리 중...
걸린 시간: 400.42초
48000번째 데이터 처리 중...
걸린 시간: 409.96초
49000번째 데이터 처리 중...
걸린 시간: 419.89초
50000번째 데이터 처리 중...
걸린 시간: 428.95초
51000번째 데이터 처리 중...

걸린 시간: 437.96초
52000번째 데이터 처리 중...
걸린 시간: 448.38초
53000번째 데이터 처리 중...
걸린 시간: 459.80초
54000번째 데이터 처리 중...
걸린 시간: 469.43초
55000번째 데이터 처리 중...
걸린 시간: 479.05초
56000번째 데이터 처리 중...
걸린 시간: 487.73초
57000번째 데이터 처리 중...
걸린 시간: 498.34초
58000번째 데이터 처리 중...
걸린 시간: 507.16초
59000번째 데이터 처리 중...
걸린 시간: 518.16초
60000번째 데이터 처리 중...
걸린 시간: 534.29초
61000번째 데이터 처리 중...
걸린 시간: 556.27초
62000번째 데이터 처리 중...
걸린 시간: 582.10초
63000번째 데이터 처리 중...
걸린 시간: 607.19초
64000번째 데이터 처리 중...
걸린 시간: 630.42초
65000번째 데이터 처리 중...
걸린 시간: 652.37초
66000번째 데이터 처리 중...
걸린 시간: 675.02초
67000번째 데이터 처리 중...
걸린 시간: 698.18초
68000번째 데이터 처리 중...
걸린 시간: 706.69초
69000번째 데이터 처리 중...
걸린 시간: 716.04초
70000번째 데이터 처리 중...
걸린 시간: 726.90초
71000번째 데이터 처리 중...
걸린 시간: 736.07초
72000번째 데이터 처리 중...
걸린 시간: 745.99초
73000번째 데이터 처리 중...
걸린 시간: 755.56초
74000번째 데이터 처리 중...
걸린 시간: 764.90초
75000번째 데이터 처리 중...
걸린 시간: 773.58초
76000번째 데이터 처리 중...
걸린 시간: 782.28초
77000번째 데이터 처리 중...
걸린 시간: 791.01초
78000번째 데이터 처리 중...
걸린 시간: 800.91초
79000번째 데이터 처리 중...
걸린 시간: 809.96초
80000번째 데이터 처리 중...
걸린 시간: 822.04초
81000번째 데이터 처리 중...
걸린 시간: 831.41초
82000번째 데이터 처리 중...
걸린 시간: 841.09초
83000번째 데이터 처리 중...
걸린 시간: 851.15초
84000번째 데이터 처리 중...
걸린 시간: 861.61초
85000번째 데이터 처리 중...
걸린 시간: 870.58초
86000번째 데이터 처리 중...
걸린 시간: 880.20초
87000번째 데이터 처리 중...
걸린 시간: 890.13초
88000번째 데이터 처리 중...
걸린 시간: 899.31초
89000번째 데이터 처리 중...
걸린 시간: 909.80초
90000번째 데이터 처리 중...
걸린 시간: 919.35초
91000번째 데이터 처리 중...
걸린 시간: 929.31초
92000번째 데이터 처리 중...
걸린 시간: 938.96초
93000번째 데이터 처리 중...
걸린 시간: 951.33초
94000번째 데이터 처리 중...
걸린 시간: 961.73초
95000번째 데이터 처리 중...
걸린 시간: 971.95초
96000번째 데이터 처리 중...
걸린 시간: 982.58초
97000번째 데이터 처리 중...
걸린 시간: 993.01초
98000번째 데이터 처리 중...
걸린 시간: 1002.98초
99000번째 데이터 처리 중...
걸린 시간: 1012.93초
100000번째 데이터 처리 중...
걸린 시간: 1022.47초
101000번째 데이터 처리 중...
걸린 시간: 1032.99초
102000번째 데이터 처리 중...
걸린 시간: 1044.03초

103000번째 데이터 처리 중...
걸린 시간: 1055.17초
104000번째 데이터 처리 중...
걸린 시간: 1065.32초
105000번째 데이터 처리 중...
걸린 시간: 1075.07초
106000번째 데이터 처리 중...
걸린 시간: 1086.79초
107000번째 데이터 처리 중...
걸린 시간: 1098.26초
108000번째 데이터 처리 중...
걸린 시간: 1109.50초
109000번째 데이터 처리 중...
걸린 시간: 1119.78초
110000번째 데이터 처리 중...
걸린 시간: 1129.74초
111000번째 데이터 처리 중...
걸린 시간: 1140.80초
112000번째 데이터 처리 중...
걸린 시간: 1150.85초
113000번째 데이터 처리 중...
걸린 시간: 1161.21초
114000번째 데이터 처리 중...
걸린 시간: 1172.41초
115000번째 데이터 처리 중...
걸린 시간: 1183.56초
116000번째 데이터 처리 중...
걸린 시간: 1194.75초
117000번째 데이터 처리 중...
걸린 시간: 1204.74초
118000번째 데이터 처리 중...
걸린 시간: 1215.66초
119000번째 데이터 처리 중...
걸린 시간: 1226.09초
120000번째 데이터 처리 중...
걸린 시간: 1238.77초
121000번째 데이터 처리 중...
걸린 시간: 1248.56초
122000번째 데이터 처리 중...
걸린 시간: 1260.84초
123000번째 데이터 처리 중...
걸린 시간: 1272.78초
124000번째 데이터 처리 중...
걸린 시간: 1285.50초
125000번째 데이터 처리 중...
걸린 시간: 1297.48초
126000번째 데이터 처리 중...
걸린 시간: 1308.37초
127000번째 데이터 처리 중...
걸린 시간: 1320.35초
128000번째 데이터 처리 중...
걸린 시간: 1333.47초
129000번째 데이터 처리 중...
걸린 시간: 1344.53초
130000번째 데이터 처리 중...
걸린 시간: 1356.14초
131000번째 데이터 처리 중...
걸린 시간: 1366.52초
132000번째 데이터 처리 중...
걸린 시간: 1377.17초
133000번째 데이터 처리 중...
걸린 시간: 1388.41초
134000번째 데이터 처리 중...
걸린 시간: 1399.00초
135000번째 데이터 처리 중...
걸린 시간: 1409.83초
136000번째 데이터 처리 중...
걸린 시간: 1421.70초
137000번째 데이터 처리 중...
걸린 시간: 1432.20초
138000번째 데이터 처리 중...
걸린 시간: 1442.74초
139000번째 데이터 처리 중...
걸린 시간: 1455.00초
140000번째 데이터 처리 중...
걸린 시간: 1466.37초
141000번째 데이터 처리 중...
걸린 시간: 1478.19초
142000번째 데이터 처리 중...
걸린 시간: 1489.86초
143000번째 데이터 처리 중...
걸린 시간: 1500.94초
144000번째 데이터 처리 중...
걸린 시간: 1511.51초
145000번째 데이터 처리 중...
걸린 시간: 1523.92초
146000번째 데이터 처리 중...
걸린 시간: 1535.82초
147000번째 데이터 처리 중...
걸린 시간: 1546.60초
148000번째 데이터 처리 중...
걸린 시간: 1558.55초
149000번째 데이터 처리 중...
걸린 시간: 1571.65초
150000번째 데이터 처리 중...
걸린 시간: 1583.84초
151000번째 데이터 처리 중...
걸린 시간: 1598.24초
152000번째 데이터 처리 중...
걸린 시간: 1610.43초
153000번째 데이터 처리 중...
걸린 시간: 1622.07초
154000번째 데이터 처리 중...

```
걸린 시간: 1635.14초
155000번째 데이터 처리 중...
걸린 시간: 1648.20초
156000번째 데이터 처리 중...
걸린 시간: 1661.08초
157000번째 데이터 처리 중...
걸린 시간: 1674.73초
158000번째 데이터 처리 중...
걸린 시간: 1687.17초
159000번째 데이터 처리 중...
걸린 시간: 1699.55초
160000번째 데이터 처리 중...
걸린 시간: 1711.50초
161000번째 데이터 처리 중...
걸린 시간: 1723.04초
162000번째 데이터 처리 중...
걸린 시간: 1735.05초
163000번째 데이터 처리 중...
걸린 시간: 1748.48초
164000번째 데이터 처리 중...
걸린 시간: 1761.10초
165000번째 데이터 처리 중...
걸린 시간: 1772.81초
166000번째 데이터 처리 중...
걸린 시간: 1785.60초
167000번째 데이터 처리 중...
걸린 시간: 1796.67초
168000번째 데이터 처리 중...
걸린 시간: 1809.27초
169000번째 데이터 처리 중...
걸린 시간: 1821.30초
170000번째 데이터 처리 중...
걸린 시간: 1833.51초
171000번째 데이터 처리 중...
걸린 시간: 1847.85초
172000번째 데이터 처리 중...
걸린 시간: 1859.22초
173000번째 데이터 처리 중...
걸린 시간: 1873.28초
174000번째 데이터 처리 중...
걸린 시간: 1886.13초
175000번째 데이터 처리 중...
걸린 시간: 1898.80초
176000번째 데이터 처리 중...
걸린 시간: 1912.08초
177000번째 데이터 처리 중...
걸린 시간: 1925.28초
178000번째 데이터 처리 중...
걸린 시간: 1937.56초
179000번째 데이터 처리 중...
걸린 시간: 1951.18초
180000번째 데이터 처리 중...
걸린 시간: 1963.69초
181000번째 데이터 처리 중...
걸린 시간: 1975.36초
182000번째 데이터 처리 중...
걸린 시간: 1988.04초
183000번째 데이터 처리 중...
걸린 시간: 2000.37초
184000번째 데이터 처리 중...
걸린 시간: 2012.07초
185000번째 데이터 처리 중...
걸린 시간: 2025.89초
186000번째 데이터 처리 중...
걸린 시간: 2038.22초
187000번째 데이터 처리 중...
걸린 시간: 2050.85초
188000번째 데이터 처리 중...
걸린 시간: 2064.26초
189000번째 데이터 처리 중...
걸린 시간: 2076.81초
190000번째 데이터 처리 중...
걸린 시간: 2089.27초
191000번째 데이터 처리 중...
걸린 시간: 2102.17초
192000번째 데이터 처리 중...
걸린 시간: 2114.65초
193000번째 데이터 처리 중...
걸린 시간: 2128.30초
194000번째 데이터 처리 중...
걸린 시간: 2141.57초
195000번째 데이터 처리 중...
걸린 시간: 2154.35초
196000번째 데이터 처리 중...
걸린 시간: 2166.51초
197000번째 데이터 처리 중...
걸린 시간: 2182.37초
198000번째 데이터 처리 중...
걸린 시간: 2195.98초
199000번째 데이터 처리 중...
걸린 시간: 2209.55초
걸린 시간: 2221.68초
```

In [ ]:

```
# import numpy as np

# tokens = []

# np.random.seed(42)

# start = time.time()

# for i in range(500000):
#     if i % 10000 == 0:
```

```
#         print(f"{i}번째 데이터 처리 중...")
#         t1 = time.time()
#         print(f"걸린 시간: {t1 - start:.2f}초")
#     tokens.append(okt.morphs(train[i]))

# end = time.time()
# print(f"걸린 시간: {end - start:.2f}초")
```

In [ ]:
```
total_tokens = sum(len(token) for token in tokens)
print(f"토큰의 총 개수: {total_tokens}")
```

```
토큰의 총 개수: 27041920
```

저희가 학습할 총 토큰 수는 약 2700만 개이네요!

Word2Vec 학습하기 (~20 mins)

토큰화까지 완료되었기 때문에 Word2Vec 모델을 학습시켜볼게요!

`gensim` 패키지에서 불러올 수 있습니다.

모델 파라미터는 아래와 같이 설정해주었습니다.

이 외에도 다양한 파라미터가 존재합니다. 이는 아래의 사이트를 참고해주세요

Word2Vec: https://radimrehurek.com/gensim/models/word2vec.html

- `sentences` = `tokens` - 학습할 데이터
- `vector_size` = `128` - 임베딩 벡터의 차원
- `window` = `5` - 좌우 context 사이즈
- `min_count` = `3` - Embedding에 추가할 때 필요한 단어 최소 등장 빈도
- `sg` = `1` - {0, 1}의 값. 0이면 CBOW로 학습
- `negative` = `5` - Negative Sampling할 노이즈 데이터 수

모델 구성하기

In [ ]:
```
from gensim.models import Word2Vec

# 모델 구성하기
model = Word2Vec(vector_size=128, window=5, min_count=3,
                 sg=1, negative=5)
```

먼저 첫 번째로 Word2Vec 임베딩을 만들기 위해서는 `build_vocab` 메서드로

단어 집합을 만들어줘야합니다. 아래와 같이 만들어줄 수 있습니다.

단어 집합 만들기

In [ ]:
```
# 단어 집합 만들기
t = time.time()

model.build_vocab(tokens, progress_per=10000)

print('Time to build vocab: {} mins'.format(round((time.time() - t) / 60, 2)))
```

```
Time to build vocab: 0.12 mins
```

단어 집합의 크기를 확인해볼게요!

In [ ]:
```
print(f"단어 집합 크기 확인: {model.wv.vectors.shape}")
```

```
단어 집합 크기 확인: (200523, 128)
```

`200523 X 128`의 임베딩 행렬이 만들어졌군요!

하지만 이 임베딩 행렬은 단지 초기값일 뿐입니다. 즉, 학습이 필요하다는 말이죠.

Embedding 만들기

이제 임베딩을 학습해줄게요. 이 작업은 약 **10~15분** 정도 걸릴 수 있습니다.

In [ ]:
```
t = time.time()

model.train(tokens, total_examples=model.corpus_count, epochs=10)

print('Time to train the model: {} mins'.format(round((time.time() - t) / 60, 2)))
```

```
Time to train the model: 11.48 mins
```

Embedding 확인하기

이제 만든 임베딩을 확인해볼까요?

학습된 임베딩은 `model.wv`에 저장되어 있습니다.

이 인스턴스는 KeyedVectors로, 접근할 수 있습니다.

유사도 분석

In [ ]:
```
# KeyedVectors 할당하기
word_vectors = model.wv
```

먼저 양의 관계에 있는 단어를 알아보겠습니다.

In [ ]:
```
# 긍정적인 단어 및 유사도 출력 (1)
word_vectors.most_similar(positive=["컴퓨터"])
```

Out[ ]:
```
[('하드웨어', 0.7588070631027222),
 ('소프트웨어', 0.7550681233406067),
 ('VAX', 0.7519944906234741),
 ('마이크로컴퓨터', 0.7400373816490173),
 ('CAD', 0.7320318222045898),
 ('IIGS', 0.7269949316978455),
 ('시분할', 0.7160534858703613),
 ('랩톱', 0.7152823805809021),
 ('Sprite', 0.7152731418609619),
 ('PVM', 0.7141070365905762)]
```

In [ ]:
```
# 긍정적인 단어 및 유사도 출력 (2)
word_vectors.most_similar(positive=["파이썬"])
```

Out[ ]:
```
[('자이썬', 0.8166624307632446),
 ('스몰토크', 0.8069655299186707),
 ('스크립팅', 0.7988269925117493),
 ('포트란', 0.7984868884086609),
 ('인터프리터', 0.797805905342102),
 ('런타임', 0.7949533462524414),
 ('Qt', 0.7938246130943298),
 ('wxWidgets', 0.7935881018638611),
 ('어셈블리어', 0.7932769060134888),
 ('액션스크립트', 0.7899165153503418)]
```

In [ ]:
```
# 긍정적인 단어 및 유사도 출력 (3)
word_vectors.most_similar(positive=["남자"])
```

Out[ ]:
```
[('여자', 0.820977509021759),
 ('마루운동', 0.6516976952552795),
 ('트램폴린', 0.6494662761688232),
 ('슈퍼헤비급', 0.633794903755188),
 ('슬라럼', 0.6256519556045532),
 ('EABA', 0.6239640116691589),
 ('장애물경주', 0.6229270100593567),
 ('로맨틱코미디', 0.6201909780502319),
 ('차서울', 0.6168308258056641),
 ('miyandau', 0.6140758991241455)]
```

단어 사이의 관계 또한 모델이 파악할 수 있을까요?

In [ ]:
```
# 단어 사이의 관계 (1)
word_vectors.most_similar(positive=["이탈리아", "요리"], topn=5)
```

Out[ ]:
```
[('아오스타', 0.711064338684082),
 ('폴렌타', 0.7022109627723694),
 ('코테키노', 0.6940657496452332),
 ('마르살라', 0.6791427731513977),
 ('피에몬테', 0.6693055629730225)]
```

In [ ]:
```
# 단어 사이의 관계 (1)
word_vectors.most_similar(positive=["한국", "요리"], topn=5)
```

Out[ ]:
```
[('후리카케', 0.6655272841453552),
 ('20100623', 0.6623979806900024),
 ('자폰', 0.6530624628067017),
 ('화과자', 0.65095055103302),
 ('음식', 0.6428438425064087)]
```

음.. 확실히 학습이 덜 되었기 때문에 결과가 잘 안 나오기도 하는군요

이 문제는 부족한 데이터를 더 학습하면 완화될 것입니다.

그 다음으로 단어 간의 의미를 분석하는 일을 해보겠습니다.

예를 들어, `사람`과 `나이`에 대한 임베딩이 주어지고, 여기서 `노인`의 임베딩을 빼면 어떤 결과가 나올까요?

과연 모델은 이들 사이의 관계를 파악할 수 있을까요?

In [ ]:
```
# 단어 간 의미 분석 (1)
word_vectors.most_similar(positive=["사람", "나이"], negative=["노인"], topn=5)
```

Out[ ]:
```
[('살의', 0.5787782073020935),
 ('이루어질수', 0.551463782787323),
 ('높임말', 0.5246351361274719),
 ('부르겠다고', 0.5226143598556519),
 ('햇수', 0.5183002948760986)]
```

In [ ]:
```
# 단어 간 의미 분석 (2)
word_vectors.most_similar(positive=["달", "목성"], negative=["지구"], topn=5)
```

Out[ ]:
```
[('질량', 0.5193728804588318),
 ('104일', 0.48612725734710693),
 ('시직경', 0.481829971075058),
 ('273일', 0.48083242774009705),
 ('천왕성', 0.48079437017440796)]
```

In [ ]:
```
# 단어 간 의미 분석 (3)
word_vectors.most_similar(positive=["노동", "구조"], negative=["사장"], topn=5)
```

Out[ ]:
```
[('불안정하고', 0.5937744379043579),
 ('mita', 0.5789293646812439),
 ('하부구조', 0.5718237161636353),
 ('Structural', 0.5711135268211365),
 ('Federalism', 0.5646846890449524)]
```

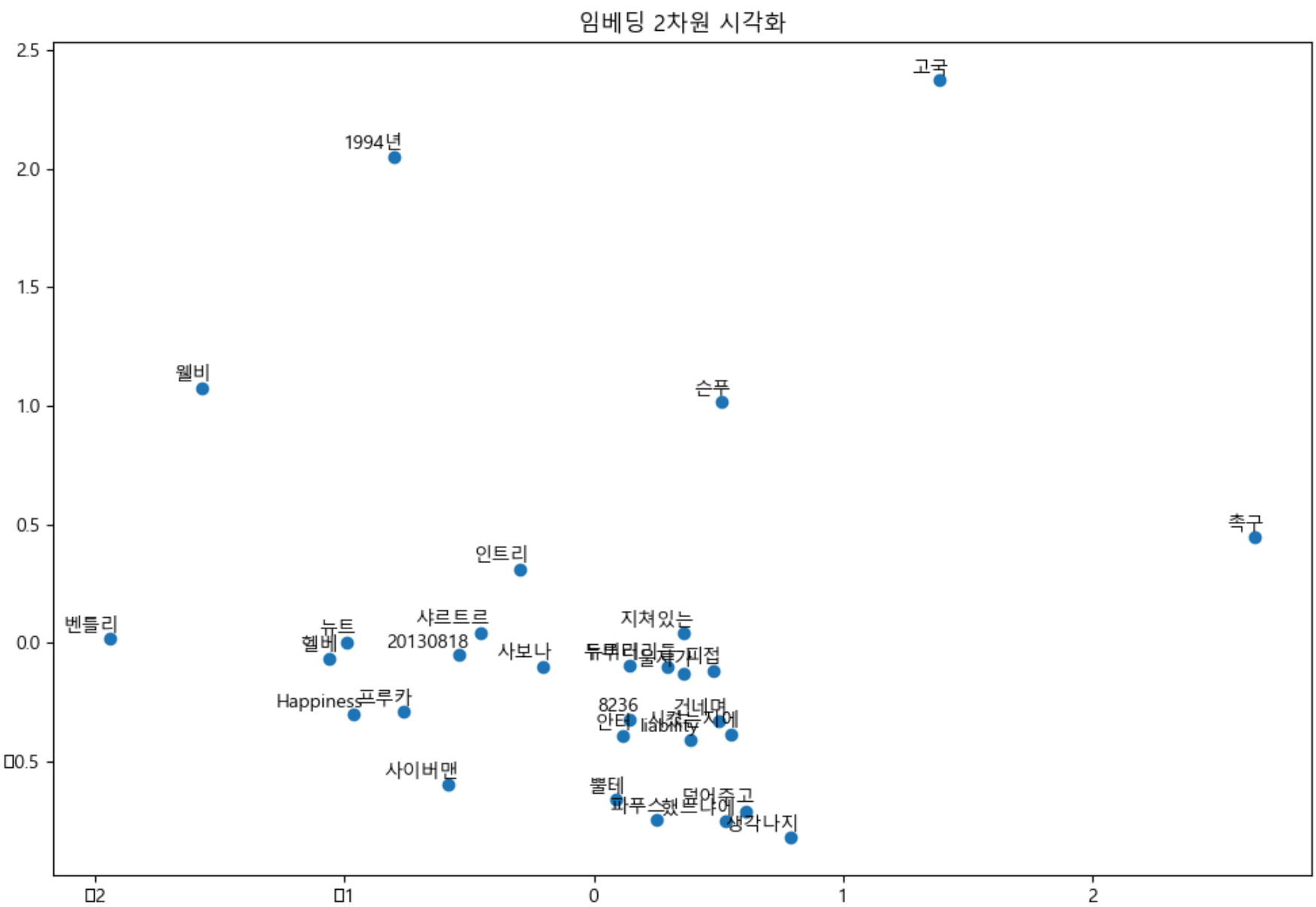
임베딩 시각화

마지막으로 임베딩을 시각화해보도록 하겠습니다. 2차원 시각화를 위해서 PCA로 차원 축소해줄게요!

그리고 너무 많은 단어가 있기 때문에, 30개만 샘플링할게요.

In [ ]:
```python
from sklearn.decomposition import PCA

np.random.seed(42)

rand_idx = np.random.randint(0, word_vectors.vectors.shape[0], 30)

vocabs = []
for idx in rand_idx:
    vocabs.append(word_vectors.index_to_key[idx])
word_vectors_list = [word_vectors[v] for v in vocabs]

pca = PCA(n_components=2)
xys = pca.fit_transform(word_vectors_list)
```

In [ ]:
```python
import matplotlib.pyplot as plt

xs = xys[:, 0]
ys = xys[:, 1]

# 한글 폰트를 위함 (만약 맑은고딕이 없으면 다운로드 후 이용할 수 있습니다)
plt.rc('font', family='Malgun Gothic')

plt.figure(figsize=(12, 8))
plt.scatter(xs, ys)
for x, y, vocab in zip(xs, ys, vocabs):
    plt.annotate(vocab, xy=(x, y), xytext=(5, 2), textcoords='offset points', ha='right', va='bottom')
plt.title('임베딩 2차원 시각화')
plt.show()
```



이렇게 실습을 마치도록 하겠습니다!