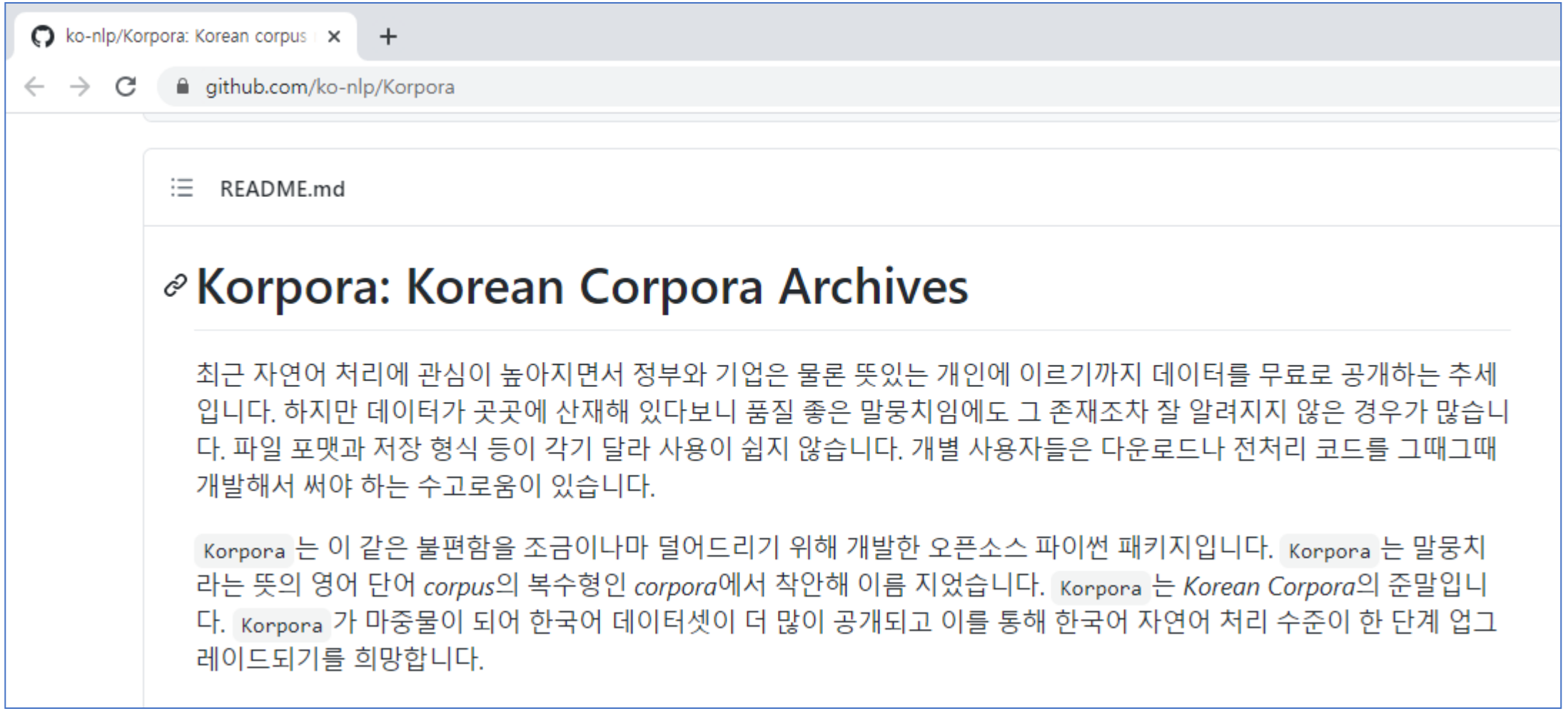
한국어 말뭉치(Corpus)

박경규

한국어 말뭉치

https://github.com/ko-nlp/Korpora



한국어 말뭉치

corpus_name	description	link
korean_chatbot_data	챗봇 트레이닝용 문답 페어	https://github.com/songys/Chatbot_data
kcbert	KcBERT 모델 학습용 댓글 데이터	https://github.com/Beomi/KcBERT
korean_hate_speech	한국어 혐오 데이터셋	https://github.com/kocohub/korean-hate-speech
korean_petitions	청와대 국민 청원	https://github.com/lovit/petitions_archive
kornli	Korean NLI	https://github.com/kakaobrain/KorNLUDatasets
korsts	Korean STS	https://github.com/kakaobrain/KorNLUDatasets
kowikitext	한국어 위키 텍스트	https://github.com/lovit/kowikitext/
namuwikitext	나무위키 텍스트	https://github.com/lovit/namuwikitext
naver_changwon_ner	네이버 x 창원대 개체명 인식 데이터셋	https://github.com/naver/nlp-challenge/tree/master/missions/ner
nsmc	NAVER Sentiment Movie Corpus	https://github.com/e9t/nsmc
question_pair	한국어 질문쌍 데이터셋	https://github.com/songys/Question_pair
modu_news	모두의 말뭉치: 신문	https://corpus.korean.go.kr
modu_messenger	모두의 말뭉치: 메신저	https://corpus.korean.go.kr
modu_mp	모두의 말뭉치: 형태 분석	https://corpus.korean.go.kr
modu_ne	모두의 말뭉치: 개체명 분석	https://corpus.korean.go.kr
modu_spoken	모두의 말뭉치: 구어	https://corpus.korean.go.kr
modu_web	모두의 말뭉치: 웹	https://corpus.korean.go.kr
modu_written	모두의 말뭉치: 문어	https://corpus.korean.go.kr
aihub_translation	한국어-영어 번역 말뭉치	https://aihub.or.kr/aidata/87
open_subtitles	영화 자막 한영 병렬 말뭉치	http://opus.nlpl.eu/OpenSubtitles-v2018.php
korean_parallel_koen_news	한국어-영어 병렬 말뭉치	https://github.com/jungyeul/korean-parallel-corpora

챗봇 문답 페어

챗봇 문답 페어

- text : 질문
- pair : 답변
- label : 일상다반사 0, 이별(부정) 1, 사랑(긍정) 2

```
[5]  from Korpora import Korpora

     Korpora.fetch("korean_chatbot_data", root_dir=DATA_PATH)
     corpus = Korpora.load("korean_chatbot_data")
```

```
[6]  corpus.train[0]
```

```
     LabeledSentencePair(text='12시 땡!', pair='하루가 또 가네요.', label=0)
```

```
[7]  corpus.train[0].text
```

```
     '12시 땡!'
```

```
[9]  corpus.get_all_texts()
```

```
     ['12시 땡!',
      '1지망 학교 떨어졌어',
      '3박4일 놀러가고 싶다',
      '3박4일 정도 놀러가고 싶다',
```

KcBERT 댓글 데이터

```
from Korpora import Korpora

Korpora.fetch("kcbert", root_dir=DATA_PATH)
corpus = Korpora.load("kcbert")
```

```
[kcbert] download kcbert-train.tar.gzaa: 100%|██████████| 2.10G/2.10G [01:17<00:00, 26.9MB/s]
[kcbert] download kcbert-train.tar.gzab: 100%|██████████| 2.10G/2.10G [00:22<00:00, 94.8MB/s]
[kcbert] download kcbert-train.tar.gzac: 671MB [00:07, 89.3MB/s]
Unzip tar. It needs a few minutes ... done

[kcbert] download kcbert-train.tar.gzaa: 100%|██████████| 2.10G/2.10G [00:22<00:00, 92.8MB/s]
[kcbert] download kcbert-train.tar.gzab: 100%|██████████| 2.10G/2.10G [00:27<00:00, 75.5MB/s]
[kcbert] download kcbert-train.tar.gzac: 671MB [00:16, 39.7MB/s]
Unzip tar. It needs a few minutes ...
```

```
corpus = Korpora.load("kcbert")
```

```
corpus.train[0]
```

- 2019년 1월 ~ 2020년 06월 사이에 작성된 댓글 많은 뉴스 기사들의 댓글과 대댓글을 모두 수집한 데이터, 15GB 이상
- Huggingface의 Transformers 라이브러리로 사용할 수 있으므로, 파일 다운로드는 필요하지 않습니다.

- Pretrain Dataset(정제 데이터) 다운로드
  https://www.kaggle.com/junbumlee/kcbert-pretraining-corpus-korean-news-comments

한국어 혐오 데이터셋

한국어 혐오 데이터셋

- text : 뉴스 댓글
- title/pair : 뉴스 제목
- gender_bias : 성적 차별 여부(True/False)
- bias : 차별 종류(종교 인종 나이 외모 등)
- hate : 특정 계층 혐오 여부(hate/none)

```
[4]  from Korpora import Korpora

     Korpora.fetch("korean_hate_speech", root_dir=DATA_PATH)
     corpus = Korpora.load("korean_hate_speech")
```

```
[5]  corpus.train[0]
```

```
     KoreanHateSpeechLabeledExample(text='(현재 호텔주인 심정) 아18 난 마른하늘에 날벼락맞고 호텔망하게생겼는데 누군 계속 추모받네....',
```

청와대 국민청원

청와대 국민청원

- text : 청원 내용
- category : 청원 범주
- num_agree : 청원 동의 수
- begin : 청원 시작일
- end : 청원 종료일
- title : 청원 제목

```
[9]  from Korpora import Korpora

     Korpora.fetch("korean_petitions", root_dir=DATA_PATH)
     corpus = Korpora.load("korean_petitions")
```

```
[10] corpus.train[0]
```

```
KoreanPetition(text="안녕하세요. 현재 사대, 교대 등 교원양성학교들의 예비교사들이 임용절벽에 매우 힘들어 하고 있는 줄로 압니다. 정부 부처에서는
```

KorNLI(Natural Language Inference)

KorNLI

- text : 문장
- pair : text와 쌍이 되는 문장
- label : text, pair 사이의 관계

```
[11]  from Korpora import Korpora

      Korpora.fetch("kornli", root_dir=DATA_PATH)
      corpus = Korpora.load("kornli")
```

```
[14]   corpus.multinli_train[0]
```

```
      LabeledSentencePair(text='개념적으로 크림 스키밍은 제품과 지리라는 두 가지 기본 차원을 가지고 있다.'
```

KorSTS(Semantic Textual Similarity)

KorSTS

- text : 문장
- pair : text와 쌍이 되는 문장
- label : text, pair 사이의 관계
- 기타 : 데이터 관련 추가 정보

```
[15] from Korpora import Korpora

     Korpora.fetch("korsts", root_dir=DATA_PATH)
     corpus = Korpora.load("korsts")
```

```
[16] corpus.train[0]
```

```
KorSTSExample(text='비행기가 이륙하고 있다.', pair='비행기가 이륙하고 있다.', label=5.0, genre='main-captions',
```

위키 텍스트

한국어 위키 텍스트

- text : 섹션 본문
- pair : 섹션 타이틀

```
[13]  from Korpora import Korpora

      Korpora.fetch("kowikitext", root_dir=DATA_PATH)
      corpus = Korpora.load("kowikitext")
```

```
[14]  corpus.train[0]
```

```
      SentencePair(text='외교부장₩n외교부장', pair=' = 분류:중화인민공화국의 외교부장 =')
```

나무 위키 텍스트

- text : 섹션 본문
- pair : 섹션 타이틀

```
[ ]  from Korpora import Korpora

     Korpora.fetch("namuwikitext", root_dir=DATA_PATH)
     corpus = Korpora.load("namuwikitext")
```

```
[ ]  corpus.train[0]
```

네이버 NER, NSMC 데이터

네이버 x 창원대 NER 데이터

```
[43]  from Korpora import Korpora

      Korpora.fetch("naver_changwon_ner", root_dir=DATA_PATH)
      corpus = Korpora.load("naver_changwon_ner")
```

```
[44]  corpus.train[0]
```

```
      WordTag(text='비토리오 양일 만에 영사관 감호 용퇴, 항룡 압력설 의심만 가율 ', words=['비토리오', '양일', '만에',
```

NAVER Sentiment Movie Corpus

```
[45]  from Korpora import Korpora

      Korpora.fetch("nsmc", root_dir=DATA_PATH)
      corpus = Korpora.load("nsmc")
```

```
[46]  corpus.train[0]
```

```
      LabeledSentence(text='아 더빙.. 진짜 짜증나네요 목소리', label=0)
```

한국어 질문쌍, 한영 병렬 말뭉치

한국어 질문쌍

```
[47]  from Korpora import Korpora

      Korpora.fetch("question_pair", root_dir=DATA_PATH)
      corpus = Korpora.load("question_pair")
```

```
[48]  corpus.train[0]
```

```
      LabeledSentencePair(text='1000일 만난 여자친구와 이별', pair='10년 연예의끝', label='1')
```

한영 병렬 말뭉치

```
[49]  from Korpora import Korpora

      Korpora.fetch("korean_parallel_koen_news", root_dir=DATA_PATH)
      corpus = Korpora.load("korean_parallel_koen_news")
```

```
[50]  corpus.train[0]
```

```
      SentencePair(text='개인용 컴퓨터 사용의 상당 부분은 "이것보다 뛰어날 수 있느냐?"', pair='Much of personal
```

모두의 말뭉치

https://corpus.korean.go.kr/

모두의 말뭉치

신문 말뭉치

종합지, 전문지, 인터넷 기반
신문 매체의 기사로 구성된
말뭉치입니다.

신청하기 ⊕


```
{
    "id": "NIRW1900000014",
    "metadata": {
        "year": "2019",
        "category": "신문 > 인터넷 기반 신문",
        ……
    },
    "document": [
        {
            "id": "NIRW1900000014.1",
            "metadata": {
                "date": "20120101",
                "topic": "사회",
                ……
            },
            "paragraph": [
                {
                    "id": "NIRW1900000014.1.1",
                    "form": "벤처정신 되살리자"
                },
                {
                    "id": "NIRW1900000014.1.5",
                    "form": "최근 취업 틈바구니에서 성공 벤처를 꿈꾸는 젊은이들은 물론이고 현장에서 물러나는 베이비붐 세대들의 창업 열기가 뜨겁다. 그 중에서도 혁신적인 아이디어와 기술력으로 승부하는 벤처 성공시대가 새롭게 열리고 있다. 말 그대로 벤처 수준을 넘어선 기업도 점차 늘어나고 있다."
                },
                ……
                …….
            ]
        }
    ]
}
```

KorQuAD 1.0

https://korquad.github.io/


```
{
   "version":"KorQuAD_v1.0_dev",
   "data":[
      {
         "paragraphs":[
            {
               "qas":[
                  {
                     "answers":[
                        {
                           "text":"1839",
                           "answer_start":0
                        }
                     ],
                     "id":"6548850-0-0",
                     "question":"바그너가 파우스트를 처음으로 읽은 년도는?"
                  },
                  ……..
               ],
               "context":"1839년 바그너는 괴테의 파우스트을 처음 읽고 그 내용에 마음이 끌려 이를 소재로 해서 하나의 교향곡을 쓰려는 뜻을 갖는다. …..
"
            },
         ]
      }
   ]
}
```

KorQuAD 2.0

https://korquad.github.io/


```
{
   "data": [
      {
         "context": "<!DOCTYPE html>\n<html>\n<head>\n<meta>\n<title>2016년 하계 올림픽 배구 남자 선수 명단 - 위키백과, 우리 모두의 백과사전
</title>
                     ......
                     <td>184cm</td>\n<td>78kg</td>\n<td>325cm</td>\n<td>310cm</td>
                     ......
                     \n<div></div>\n</div>\n\n\n\n</body>\n</html>\n",
         "qas": [
            {
               "question": "파스칼레 소틸레의 스파이크 높이는 몇 cm인가?",
               "id": "38824",
               "answer": {
                  "text": "<td>332cm</td>",
                  "html_answer_text": "<td>332cm</td>",
                  "answer_start": 10698,
                  "html_answer_start": 45900
               }
            }
            ……..
          ],
         "title": "2016년_하계_올림픽_배구_남자_선수_명단",
         "url": "https://ko.wikipedia.org/wiki/2016년_하계_올림픽_배구_남자_선수_명단",
         "raw_html": "<!DOCTYPE html>\n<html class=\"client-nojs\" dir=\"ltr\" lang=\"ko\">\n<head>........"
      }
   ]
}
```

말뭉치 사용 실습

korean_corpus.ipynb

Thank you

kgpark88@gmail.com