Lecture 3: 코퍼스와 텍스트 데이터 수집

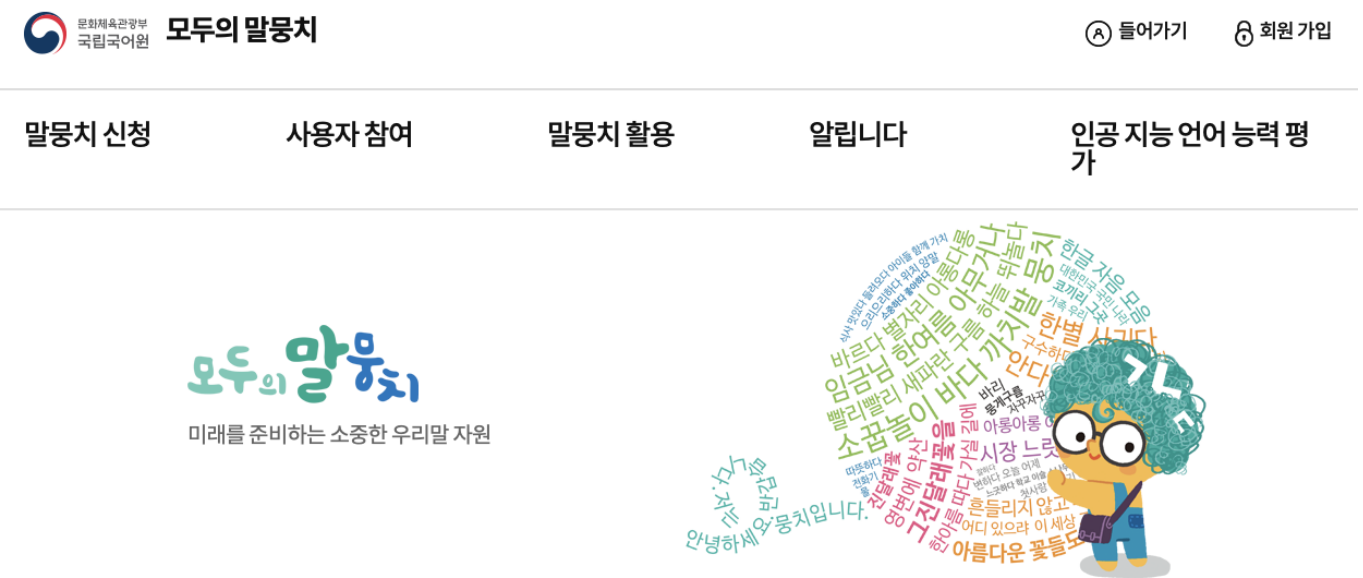
코퍼스 (말뭉치)

- 말뭉치 또는 코퍼스(영어: corpus, 복수형: corpora)는 자연언어 연구를 위해 특정한 목적을 가지고 언어의 표본을 추출한 집합이다.
- 컴퓨터의 발달로 말뭉치 분석이 용이해졌으며 분석의 정확성을 위해 해당 자연언어를 형태소 분석하는 경우가 많다.
- 확률/통계적 기법과 시계열적인 접근으로 전체를 파악한다.
- 언어의 빈도와 분포를 확인할 수 있는 자료이며, 현대 언어학 연구에 필수적인 자료이다.
- 인문학에 자연과학적 방법론이 가장 성공적으로 적용된 경우로 볼 수 있다.

> source: https://ko.m.wikipedia.org/wiki/말뭉치

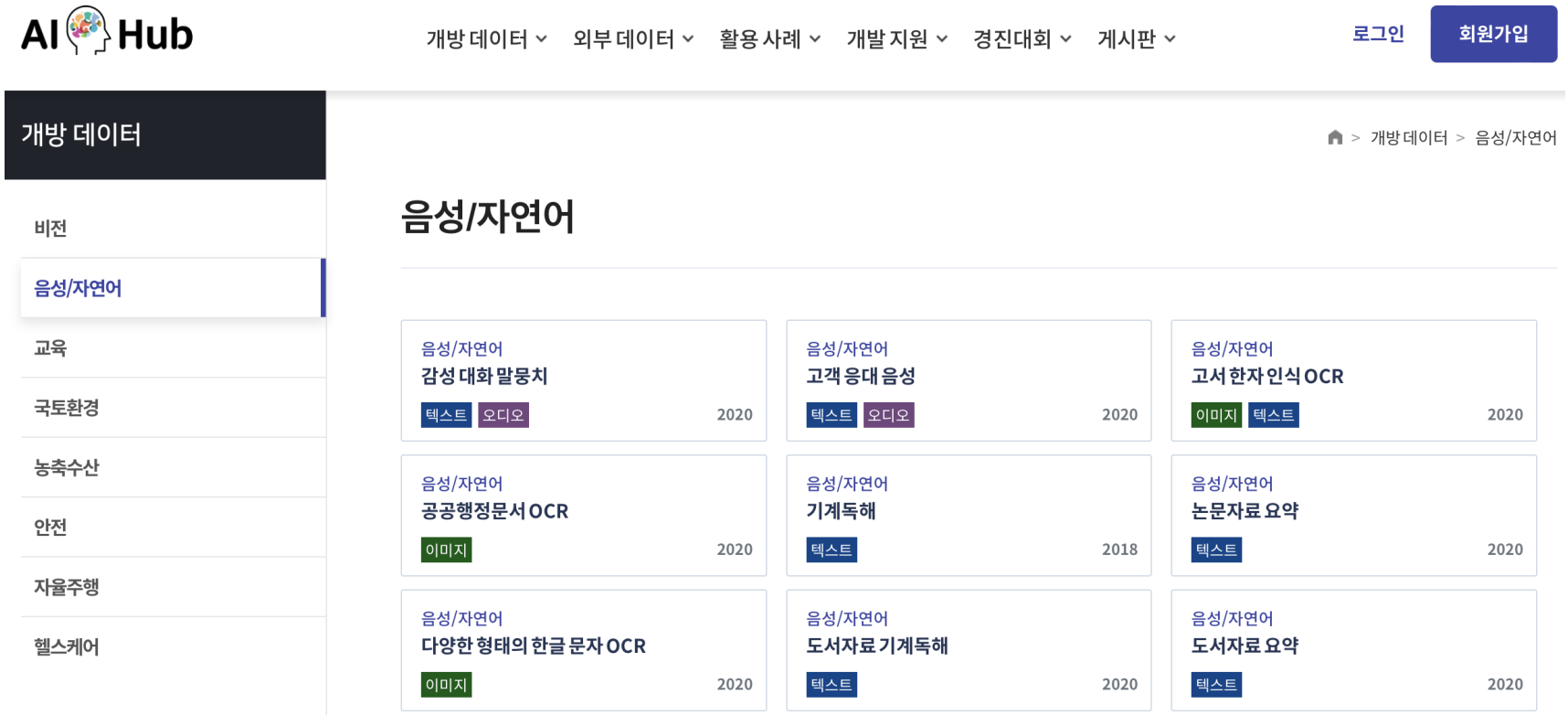
국립국어원 모두의 말뭉치

다양한 분석 말뭉치(형태소 분석과 구문 분석 말뭉치 등), 다양한 도메인의 말뭉치(문어, 신문, 구어, 웹), 자연어 추론을 위한 말뭉치(유사 문장) 등 다양한 데이터들이 체계적으로 구축되어 있다. 로그인, 메일 인증을 거쳐 데이터를 신청할 수 있고 다운로드 받기 위해서는 연구과제명과 수행기관, 약정 기간 등이 필수 입력 요소이다.

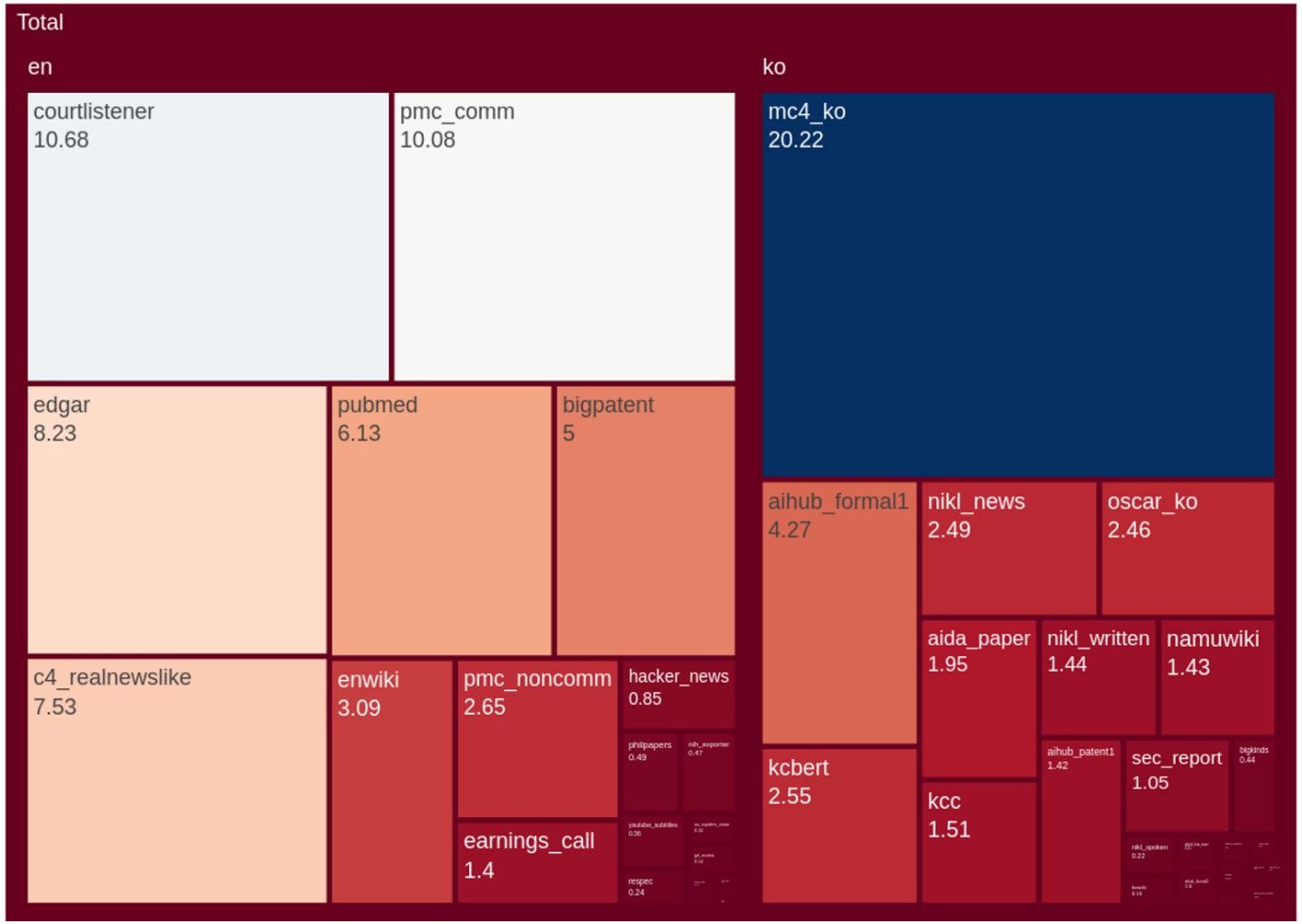


AIHub

텍스트와 음성 멀티모달까지 가장 광범위한 데이터, 로그인 및 사용 목적과 기간을 명시한 사용 신청서 작성 후 허가 메일이 오면 다운로드 가능



The eKorpkit Corpus



The eKorpkit Corpus is a large, diverse, multilingual (ko/en) language modelling dataset.

English: 258.83 GiB, Korean: 190.04 GiB, Total: 448.87 GiB

The eKorpkit Corpus

Name	Language	Size	Weight	# Docs	# Sents	# Words
mc4_ko	ko	90.76 GiB	20.22%	15,618,718	665,858,888	8,007,674,274
courtlistener	en	47.92 GiB	10.68%	3,489,298	335,079,871	8,324,277,457
pmc_comm	en	45.26 GiB	10.08%	51,276,102	297,884,818	7,365,607,900
edgar	en	36.94 GiB	8.23%	213,376	177,270,203	6,053,677,897
c4_realnewslike	en	33.79 GiB	7.53%	13,813,090	155,883,681	6,040,207,703
pubmed	en	27.51 GiB	6.13%	22,498,747	190,907,356	4,281,121,705
bigpatent	en	22.46 GiB	5.00%	1,244,053	2,488,106	4,613,882,925
aihub_formal1	ko	19.16 GiB	4.27%	1,073,944	93,148,022	1,993,574,713
enwiki	en	13.85 GiB	3.09%	6,200,658	129,066,417	2,400,717,561

Name	Language	Size	Weight	# Docs	# Sents	# Words
pmc_noncomm	en	11.88 GiB	2.65%	14,142,294	79,748,279	1,923,415,913
kcbert	ko	11.45 GiB	2.55%	82,990,213	82,990,213	1,088,177,367
nikl_news	ko	11.19 GiB	2.49%	4,104,534	42,527,395	1,138,897,337
oscar_ko	ko	11.05 GiB	2.46%	3,673,262	61,833,262	1,122,638,494
aida_paper	ko	8.77 GiB	1.95%	481,389	38,808,105	1,025,422,060
kcc	ko	6.80 GiB	1.51%	46,529,987	46,529,987	703,222,627
nikl_written	ko	6.45 GiB	1.44%	20,128	27,231,846	679,547,033
namuwiki	ko	6.43 GiB	1.43%	571,026	67,315,244	691,537,393
aihub_patent1	ko	6.40 GiB	1.42%	155,939	29,206,198	673,134,598
earnings_call	en	6.30 GiB	1.40%	159,380	32,391,491	1,160,525,933
sec_report	ko	4.70 GiB	1.05%	817,040	32,644,657	495,245,547

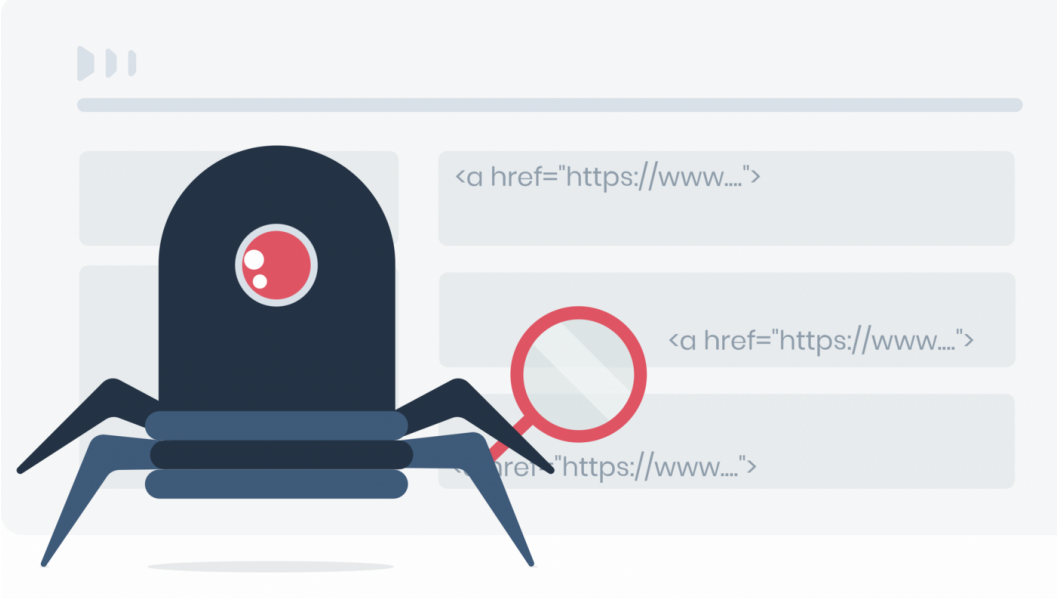
Name	Language	Size	Weight	# Docs	# Sents	# Words
hacker_news	en	3.80 GiB	0.85%	818,299	41,573,998	662,524,112
philpapers	en	2.19 GiB	0.49%	31,016	139,518	365,576,851
nih_exporter	en	2.10 GiB	0.47%	1,017,230	13,540,126	326,974,102
bigkinds	ko	1.99 GiB	0.44%	871,304	7,759,115	197,746,184
youtube_subtitles	en	1.61 GiB	0.36%	150,749	16,074,289	303,286,377
respec	en	1.08 GiB	0.24%	1,119,640	7,083,257	169,590,880
nikl_spoken	ko	1002.49 MiB	0.22%	25,614	19,042,013	116,067,432
kowiki	ko	715.39 MiB	0.16%	563,959	5,671,388	70,263,451
us_equities_news	en	714.16 MiB	0.16%	220,976	1,834,664	131,179,752
aihub_law_case	ko	689.96 MiB	0.15%	77,202	1,095,140	66,686,761
aihub_formal2	ko	650.03 MiB	0.14%	95,990	1,650,141	64,523,191

Name	Language	Size	Weight	# Docs	# Sents	# Words
gd_review	en	642.76 MiB	0.14%	1,929,910	6,733,680	112,977,678
aihub_patent2	ko	457.18 MiB	0.10%	147,674	1,879,909	46,045,036
enron_mail	en	428.36 MiB	0.09%	247,586	7,908,959	65,258,456
aihub_paper	ko	370.11 MiB	0.08%	98,344	1,802,883	35,556,261
kaist	ko	304.92 MiB	0.07%	11,157	1,926,901	30,929,508
reuters_financial	en	288.63 MiB	0.06%	101,055	1,983,069	49,495,061
aihub_book	ko	236.66 MiB	0.05%	180,001	1,201,956	23,052,720
aihub_koen_formal	ko	206.37 MiB	0.04%	1,350,000	1,350,000	20,659,619
aihub_koen_ssci	ko	186.49 MiB	0.04%	1,361,845	1,361,845	19,104,237
aihub_koen_sci	ko	164.42 MiB	0.04%	1,344,631	1,344,631	17,720,448
fomc	en	112.66 MiB	0.02%	2,822	950,620	18,640,148

Name	Language	Size	Weight	# Docs	# Sents	# Words
esg_report	ko	24.17 MiB	0.01%	15,561	119,031	2,488,545
aihub_law_kb	ko	9.99 MiB	0.00%	17,373	46,140	934,632
bok_minutes	ko	9.54 MiB	0.00%	163	33,027	918,203
pathobook	en	4.28 MiB	0.00%	28	33,603	648,221
English	en	258.83 GiB	57.66%			
Korean	ko	190.04 GiB	42.34%			
Total		448.87 GiB	100.00%			

텍스트 데이터 수집 개요

- 수집대상 : 웹페이지, SNS, 댓글, 음성, 비디오 등 텍스트 형태로 변환 가능한 모든 데이터
- 저장유형 : Plain Text, PDF, Table, XML, JSON
- 수집방법 : Web Crawling, API 호출, DB Query, Online Survey



텍스트 데이터 수집 유형 - 오프라인 데이터

- 수집방법
  - 온/오프라인 설문지
  - 음성 녹음, 비디오 촬영
- 장단점
  - 타게팅 대상에 대한 데이터 수집가능
  - 사람이 직접 수집해야함
  - 수집에 시간적, 공간적 제약이 큼

텍스트 데이터 수집 유형 - 자체 시스템

- 수집방법
  - 서비스 또는 사내 데이터베이스 활용
  - 사내 게시판, 유저 댓글, 업무 보고서, 내부문서
- 장단점
  - 기수집된 데이터를 빠르게 활용 가능
  - 소속기관/업체/서비스 내부 관련자가 아니면 접근이 어려움
  - 정보유출에 대한 위험이 큼

텍스트 데이터 수집 유형 - 웹크롤링

- 수집방법
  - 프로그래밍 언어를 활용해 웹페이지에 존재하는 대량의 정보를 반복 수집
- 장단점
  - 대량의 정보를 빠르게 수집할 수 있음
  - 데이터 수집과 함께 정규화된 데이터셋 구성 가능
  - 프로그래밍 언어 활용이 필요함
  - 개인정보 문제에 취약함

텍스트 데이터 수집 유형 - API 호출

- 수집방법
  - 프로그래밍 언어를 활용해 서비스에서 정식으로 제공하는 데이터를 수집
  - 네이버 API, 카카오 API, Reddit API, News API, SNS (Twitter, Facebook, Instagram)
- 장단점
  - 바로 활용할 수 있는 양질의 데이터를 얻을 수 있음
  - 수집할 수 있는 소스가 제한적임
  - 프로그래밍 언어 활용이 필요함

텍스트 데이터 수집 유형 - Dump

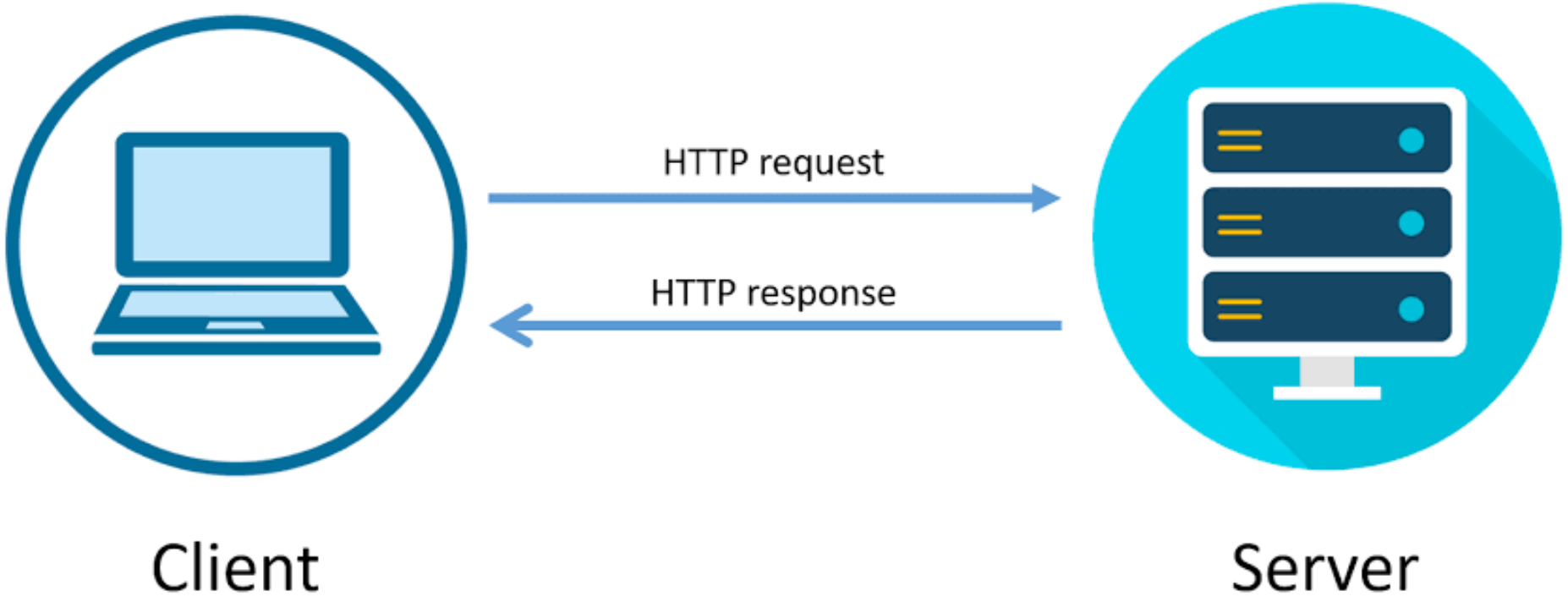
- 나무위키: 나무위키에서는 데이터베이스의 덤프 파일을 제공합니다. 나무위키를 크롤링하는 것은 서버에 많은 부담이 되기 때문에 되도록이면 DB 덤프를 받아주시기 바랍니다. 일반 사용자분들이라면 다운로드하실 필요가 없으며 이 파일은 빅 데이터 분석, 미러 제작, 소장 등의 목적으로 필요하신 분들만 받으시면 됩니다.
- 위키백과: 위키백과의 자료를 여러가지 용도로 이용하려는 사람들을 위해, 위키백과에서는 주기적으로 전체 문서를 묶어서 배포하고 있습니다.
- NCBI: The majority of NCBI data are available for downloading, either directly from the NCBI FTP site or by using software tools to download custom datasets.

텍스트 데이터 수집 유형 - Common Crawl

- Common Crawl 은 비영리 단체로 웹을 크롤링하고 아카이브와 데이터 세트를 대중에게로 자유롭게 제공하는 조직. Common Crawl의 웹 아카이브 는 2011 년 이후 수집 된 페타 바이트의 데이터로 구성. 일반적으로 매월 크롤링을 완료.
- Common Crawl은 Gil Elbaz 에 의해 설립. 조직의 크롤러는 nofollow 및 robots.txt 정책을 준수합니다. Common Crawl의 데이터 세트를 처리하기위한 오픈 소스 코드가 공개적으로 제공
- Amazon Web Services 2012 년에 공용 데이터 세트 프로그램을 통해 Common Crawl의 아카이브를 호스팅하기 시작했습니다.
- 이 조직은 .arc 파일과 함께 메타 데이터 파일과 크롤러의 텍스트 출력을 공개하기 시작.
- Common Crawl은 OpenAI의 GPT-3 언어모델을 훈련하는 데 사용되었습니다.

웹 서비스의 동작 원리

요청 (Request)과 응답 (Response)



HTTP Response Conent-Type

- HTML
- JSON/XML

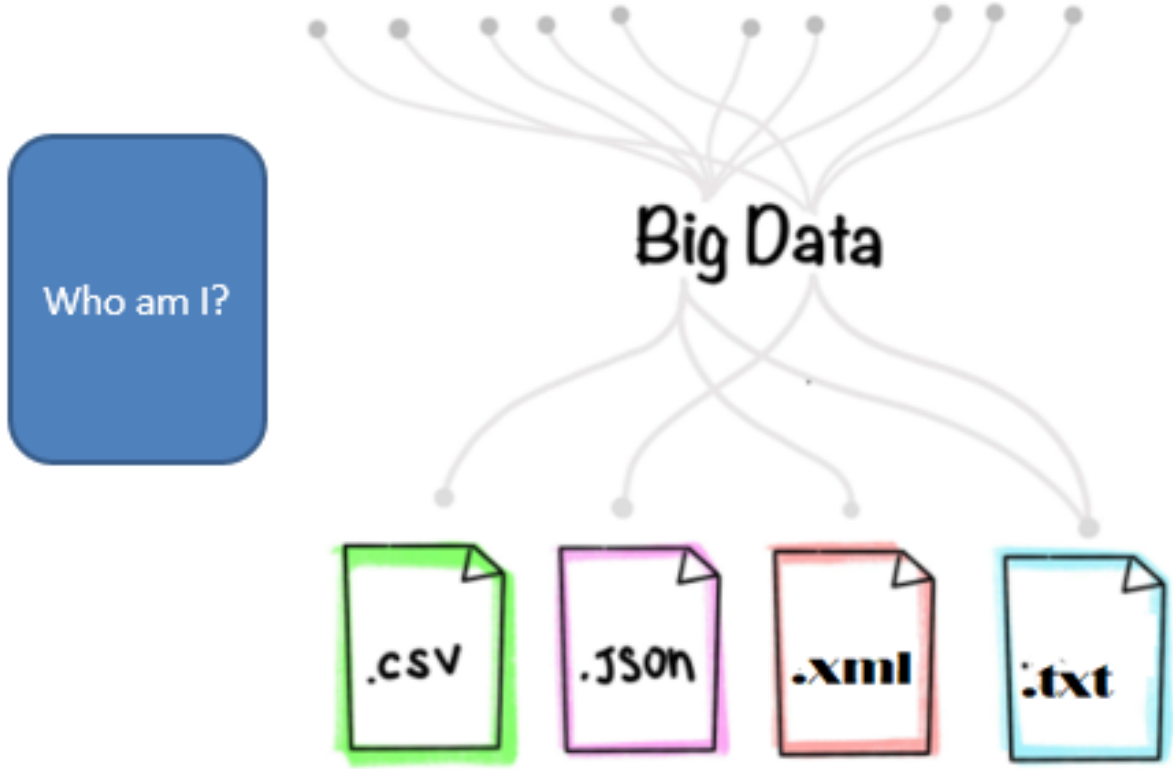
> Example

```
PUT / HTTP/1.1
Accept: application/json, */*;q=0.5
Accept-Encoding: gzip, deflate
Content-Type: application/json
Host: pie.dev

{
    "name": "John",
    "email": "john@example.org"
}
```

Parsing

- 파싱(parsing)은 문장이나 데이터문자열(html, json 등)에서 원하는 데이터를 분석하여 추출하는 기술
- 특정한 패턴과 규칙, 순서를 이용하여 자신이 필요로 하는 데이터를 추출해내는 작업



Open API vs. 웹 크롤링

Open API 활용

- 장점: 정제된 데이터(JSON, XML)을 받을 수 있음
- 단점: 제공하는 데이터만 받을 있음

웹 크롤링

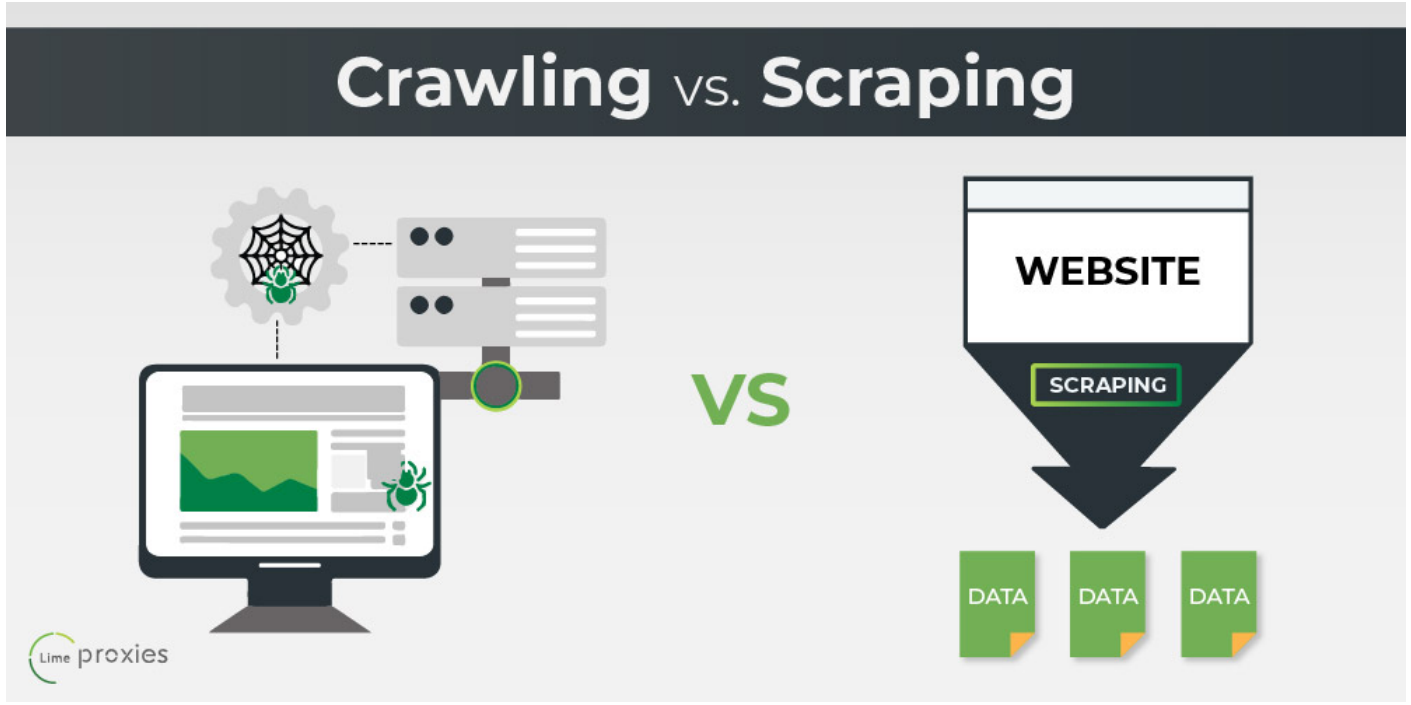
- 장점: API가 제공하지 않는 웹상에서 볼 수 있는 데이터를 모두 받을 수 있음
- 단점: 정제되지 않음(HTML), 상업적 이용에 제약이 있음

Crawling 이란?

- 정제되지 않은 웹페이지에서 필요한 데이터를 추출하는 행위
- 활용할 수 있는 데이터가 정리되어 올라가 있는 데이터(API, 파일형태)를 제외하고, 웹페이지에게 시된 자료를 가져오는 기술
- API를 통한 데이터 제공이 활성화 되고 있는 추세이나, 국내에서는 소극적인 API 제공으로 웹크롤 링을 통한 수집이 반드시 필요함

Web Crawling vs. Scrapping

- Crawling: 웹페이지의 정보를 추출하는 방법(=spider, bot), 단순히 하이퍼링크를 돌아다니며 웹페이지를 다운로드
- Scrapping: 웹페이지에서 필요한 정보만을 추출하는 방법, 다운로드한 웹페이지에서 필요한 부분만을 추출하고 저장



크롤링의 법적 문제

데이터 무단수집과 저작권 침해

- 웹크롤링은 원래 검색엔진 등의 인터넷 사이트에서 데이터를 최신 상태로 유지하기 위해 사용
- 웹크롤링을 활용하여 타사 컨텐츠를 무단 활용하는 것은 불법행위에 해당함
- 웹크롤링을 통한 과도한 요청은 대상 서비스 서버 운영과 서비스 관리에 안좋은 영향을 끼침

Robots.txt

- 웹사이트에 배치된 텍스트 파일로, 크롤링 접근권한에 대해 명시해 놓은 문서
- 웹크롤링은 Robots.txt 파일에서 허용하는 항목에 대해서만 수집 가능하며 그 외의 수집에 대한 책 임은 모두 본인에게 있음
- 수집이 허용되 있더라도 대상 서비스 운영에 피해를 주지 않는선 에서 필요한 만큼만 수집
- Robots.txt 파일이 없는 경우 서비스 관리자에 직접 허락을 구한 후 수집

정적/동적 웹 페이지

정적 웹페이지(static web page)

- 서버에 저장되어있는 HTML+CSS 파일 그대로 보여주는 방법
- 웹페이지에서 추가적인 통신 및 계산이 필요 없기 때문에 속도가 빠르고 서버에 부담이 적다는 장점이 있지만추가/수정/삭제 등 내용 변경이 필요할 때 HTML 자체를 수정해야 하기 때문에 번거롭다는 단점이 존재

동적 웹페이지(dynamic web page)

- 상황에 따라 서버에 저장되어있는 HTML에 데이터 추가/가공을 해서 보여주는 방법
- 한 페이지에서 상황/시간/사용자요청에 따라 다른 모습을 보여줄 수 있다는 장점이 있지만 상대적으로 보안에 취약하고 모습이 계속 변하기 때문에 검색 엔진 최적화(search engine optimazation, SEO)가 어렵다는 단점이 존재

요약

＼	정적 웹페이지	동적 웹페이지
특 징	서버에 저장되어 있는 그대로 html전송	요청 정보에 따라 html을 처리하여 전송
장 점	속도가 빠르다. 서버 부담이 적다.	상황에 맞게 변하는 모습 관리가 쉽다.
단 점	서비스가 한정적이다. 내용 변경이 어렵다.	보안에 취약하다 검색엔진최적화가 어렵다.
예 시	회사소개, 음식메뉴, 포트폴리오	블로그, 게시판, 날씨 정보

동적 웹페이지의 종류

Client Side Rendering (CSR)

> 자바스크립트에 데이터를 포함해서 보낸 후, 클라이언트에서 HTML을 완성하는 방법
> 브라우저는 웹서버로부터 다운받은 js파일을 실행
> 실행하는 동안 딜레이 발생하지만 이후에는 서버의 의존도가 낮아 빠른 화면이나 인터렉션 가능
> 단순히 뼈대만 있기 때문에 SEO에 취약

Server Side Rendering (SSR)

> 서버에서 HTML에 데이터를 끼워넣어 완성된 형태의 HTML을 보내주는 방법
> 브라우저에서 보는 파일을 만들어 내는 로직 파일을 서버에 올려서 실행
> 이미 DOM 구성이 다 된 파일을 브라우저가 받기 때문에 초기 구동 속도가 빠름
> 이미 내용이 다 차있기 때문에 검색 엔진들이 정보를 수집할때 정확한 정보를 가져갈 수 있어서 SEO에 좋음

동적 웹 페이지

case 1. 로그인을 해야만 접속 가능한 네이버 메일

case2. 보고 있는 위치에 출력 결과와 url이 계속 변하는 네이버 지도

case 3. 드래그를 아래로 내리면 계속 새로운 사진과 영상이 나타나는 인스타그램과 유튜브

정적/동적 수집

> 정적 페이지에서 정보를 수집 하느냐, 동적 페이지를 하느냐에 따라 다른 파이썬 패키지 사용

	정적 수집	**동적 수집**
사용 패키지	requests / urllib	selenium
수집 커버리지	정적 웹 페이지	정적/동적 웹 페이지
수집 속도	빠름 (별도 페이지 조작 필요 X)	상대적으로 느림
파싱 패키지	beautifulsoup	beautifulsoup / selenium

정적/동적 수집

정적 수집

> 정적 수집은 멈춰있는 페이지의 html을 requests 혹은 urllib 패키지를 이용해 가져와서 beautifulsoup 패키지로 파싱하여 원하는 정보를 수집

> 바로 해당 url의 html을 받아와서 수집하기 때문에 수집 속도가 빠르다는 특징이 있지만, 여기저기 모두 사용할 수 있는 범용성은 떨어짐

동적 수집

> 동적 수집은 계속 움직이는 페이지를 다루기 위해서 selenium 패키지로 chromdriver를 제어
> 특정 url로 접속해서 로그인을 하거나 버튼을 클릭하는 식으로 원하는 정보가 있는 페이지까지 도달
> 브라우저를 직접 조작하고 브라우저가 실행될때까지 기다려주기도 해야해서 그 속도가 느림