데이터

고려대학교 KOREA UNIVERSITY 정보대학 컴퓨터학과

공공데이터 활용

DATA 공공데이터포털 .GO.KR

https://www.data.go.kr/guide/guide/guide.do

https://www.data.go.kr/guide/guide/introduce.do

HTTP

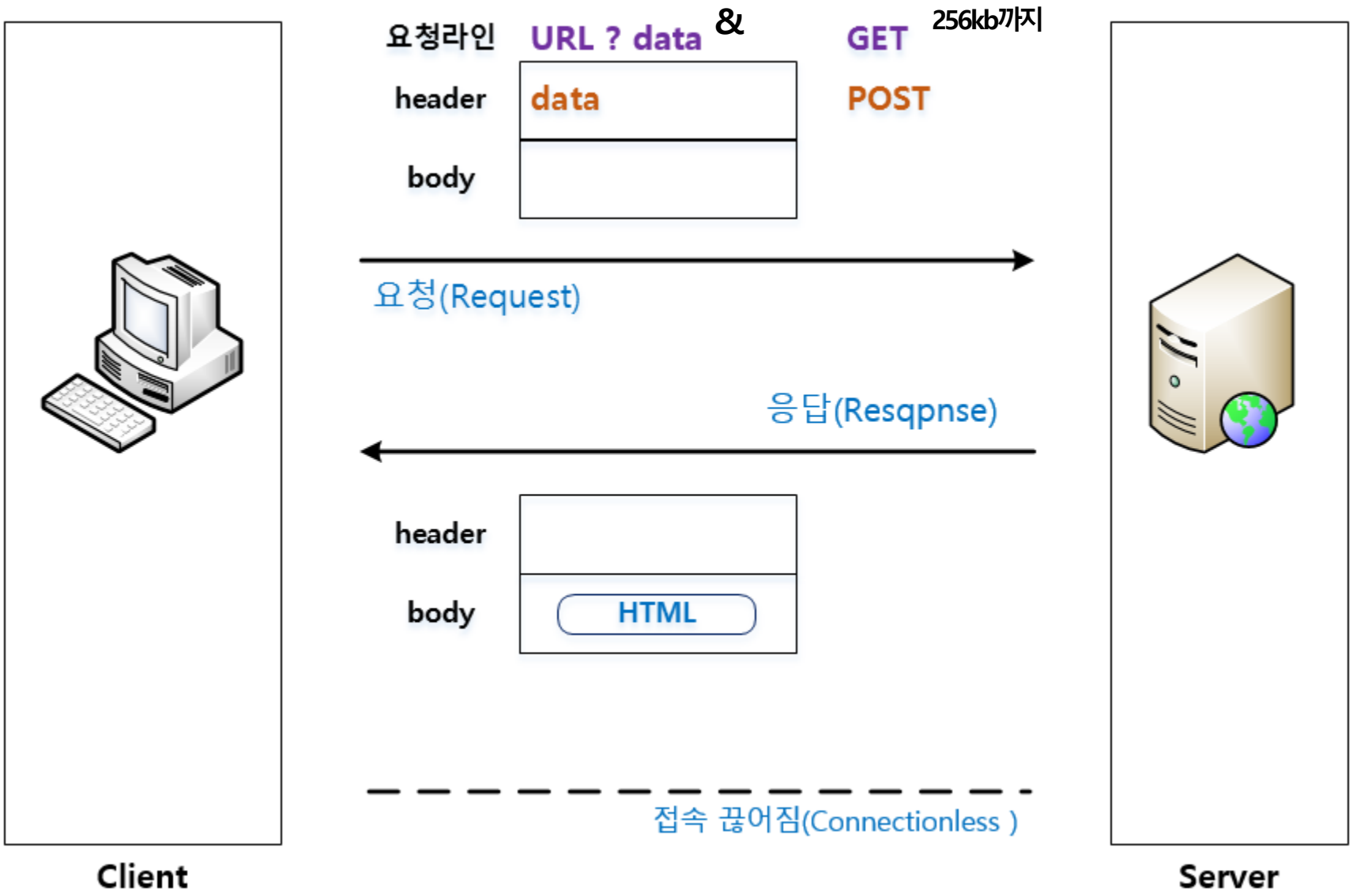
HyperText Transfer Protocol

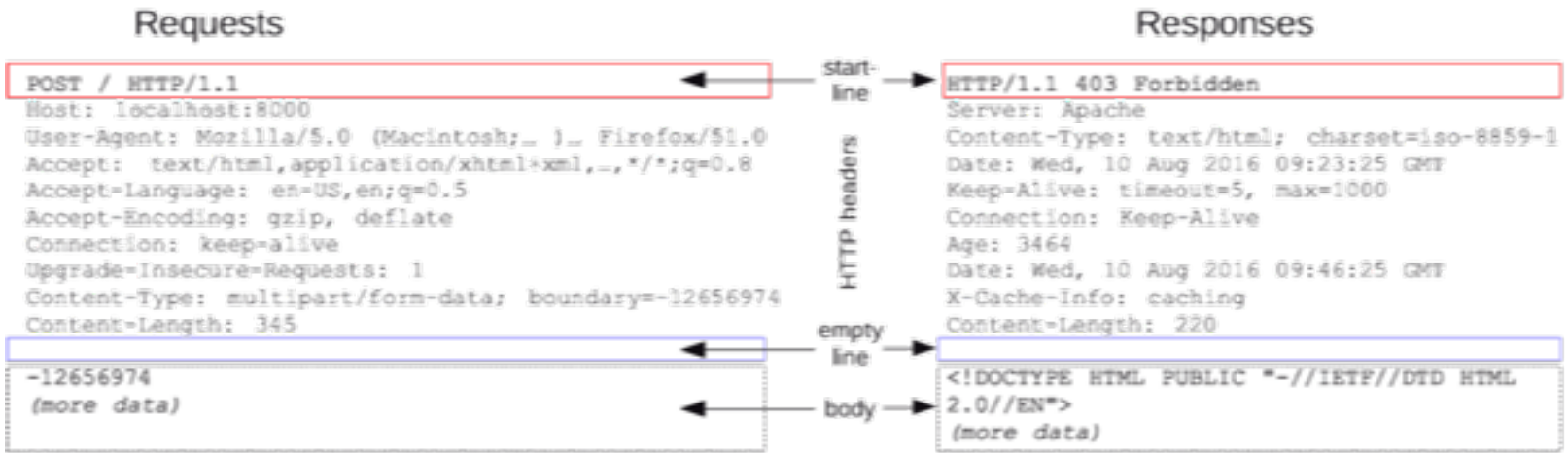
- Plain Language (Text)
- Human Readable

- Stateless
- Connectionless

> Cookie
> Session


HTTP Request
World Wide Web
or Internet
HTTP Response


요청라인
URL ? data
&
GET
256kb까지
header
data
POST
body
요청(Request)
응답(Resqpnse)
header
body
HTML
접속 끊어짐(Connectionless)
Client
Server



https://developer.mozilla.org/ko/docs/Web/HTTP/Messages

https://www.w3.org/Protocols/rfc2616/rfc2616-sec14.html

http://www.iana.org/assignments/http-status-codes/http-status-codes.xhtml

Web Browser에서 확인

Headers, ….

http://d2.naver.com/helloworld/59361

Python HTTP Library

Boilerplate Code
(상용구 코드)

requests

http 관련 library

http://docs.python-requests.org/en/master/user/quickstart/

Data 분류

http://www.dbguide.net/db.db?cmd=view&boardUid=186812&boardConfigUid=9&categoryUid=216&boardIdx=152&boardStep=1

정형 vs 반정형 vs 비정형

- RDBMS의 테이블들
- 스프레드시트 등

- 이진 파일 형태: 동영상, 이미지
- 스크립트 파일 형태: 소셜 데이터의 텍스트

Cf) NoSQL

정형 vs 반정형 vs 비정형

- URL 형태로 존재 – HTML
- 오픈 API 형태로 제공 - XML, JSON
- 로그형태 - 웹로그, IOT에서 제공하는 센서 데이터

Python Data library

scraping
parsing

Crawling vs Scraping vs Pashing

정형

- RDBMS의 테이블들
- CSV, 스프레드시트(XLS, XLSX)

DATA 공공데이터포털 .GO.KR

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$

https://pandas.pydata.org/pandas-docs/stable/generated/pandas.read_csv.html

반정형

DATA 공공데이터포털 .GO.KR

- URL 형태로 존재 – HTML
- 오픈 API 형태로 제공 - XML, JSON
- ~~로그형태 - 웹로그, IOT에서 제공하는 센서 데이터~~
- 그 외

XML

lxml

vs Beautifulsoup (with lxml)

http://lxml.de/

http://lxml.de/elementsoup.html

JSON

https://docs.python.org/3/library/json.html

HTML
: 일부 Data만 Parsing

Regular Expression vs Beautifulsoup

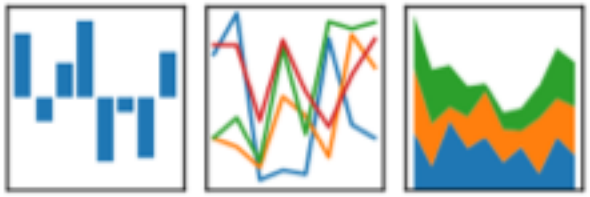
https://docs.python.org/3/howto/regex.html

https://www.crummy.com/software/BeautifulSoup/bs4/doc/

그 외

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



https://pandas.pydata.org/pandas-docs/stable/api.html#input-output