Data Crawling

Ki Hyun Kim

nlp.with.deep.learning@gmail.com

Preprocessing Workflow

데이터(코퍼스) 수집
- 구입, 외주
- 크롤링을 통한 수집

정제
- Task에 따른 노이즈 제거
- 인코딩 변환

레이블링(Optional)
- 문장마다 또는 단어마다 labeling 수행

Tokenization
- 형태소 분석기를 활용하여 분절 수행

Subword Segmentation (Optional)
- 단어보다 더 작은 의미 단위 추가 분절 수행

Batchify
- 사전 생성 및 word2index 맵핑 수행
- 효율화를 위한 전/후처리

데이터 구입 및 외주의 한계

- 구입
  - 정제 및 레이블링이 완료된 양질의 데이터를 얻을 수 있음
  - 양이 매우 제한적
  - 구입처: 대학교, 한국전자통신연구원(ETRI), 플리토 등

- 외주
  - 수집, 정제 및 레이블링을 외주 줄 수 있음
  - 가장 높은 비용 → 양이 매우 제한적
  - 품질 관리를 위한 인력이 추가로 필요

최대 10만 단위

무료 공개 데이터

- 공개 사이트
  - **AI-HUB**
  - WMT competetion
  - Kaggle
  - OPUS (http://opus.nlpl.eu/)

최대 10만 단위

- 마찬가지로 양이 매우 제한적
- 한국어 코퍼스는 흔치 않음

Crawling

- 무한한 양의 코퍼스 수집 가능
  - 원하는 도메인 별로 수집 가능

- 하지만 품질이 천차만별이며, 정제 과정에 많은 노력 필요
  - e.g. 특수문자, 이모티콘, 노이즈, 띄어쓰기

Crawling

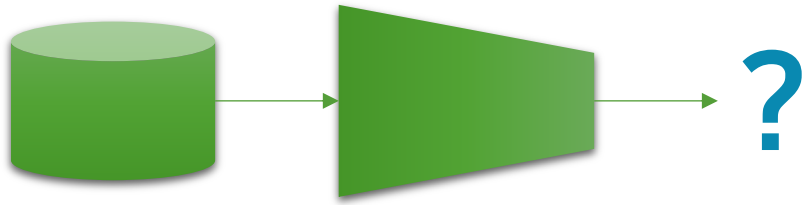
- 아직은 회색지대
  하지만 적법한 절차에 따른 크롤링이 필수

- robots.txt

```
$ wget https://www.ted.com/robots.txt
$ cat robots.txt
User-agent: *
Disallow: /latest
Disallow: /latest-talk
Disallow: /latest-playlist
Disallow: /people
Disallow: /profiles
Disallow: /conversations

User-agent: Baiduspider
Disallow: /search
Disallow: /latest
Disallow: /latest-talk
Disallow: /latest-playlist
Disallow: /people
Disallow: /profiles
```

http://www.robotstxt.org/



저작권이 존재하는 코퍼스로부터 학습한
모델과 그 생성물의 저작권은 누가 갖는가?

수집처

수집처	도메인	문체	수집 난이도	양방향	정제 난이도	비고
블로그	일반	대화체	낮음	X	최상	
N* 지식인	다양함	대화체	낮음	X	중간	
뉴스 기사	시사	문어체	낮음	O	낮음	문법 준수
Wikipedia	다양함	문어체	덤프 제공	O	낮음	
나무위키	다양함	문어체	낮음	X	낮음	
커뮤니티	다양함	대화체	중간	X	높음	클리앙 등
TED	다양함	대화체	낮음	O	낮음	
자막	일반	대화체	낮음	O	높음	