Preprocessing Pipeline

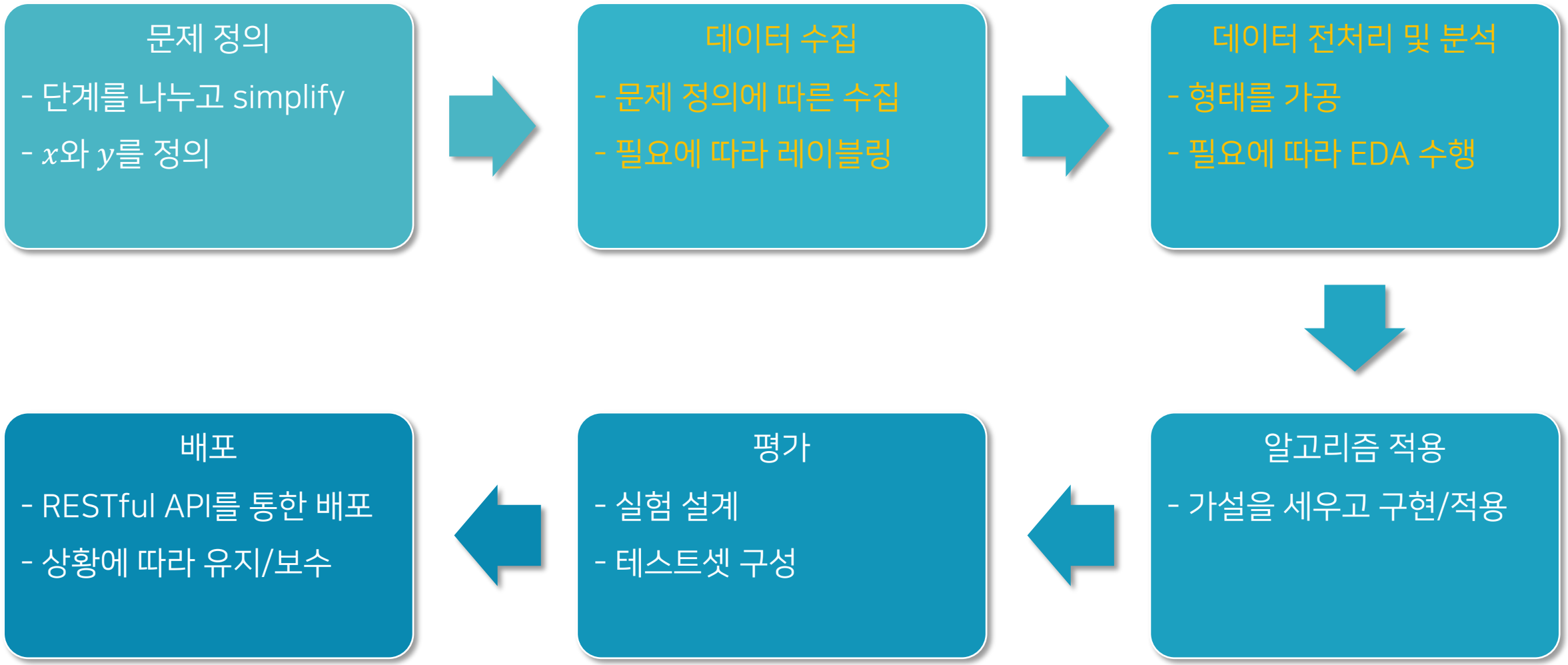
Ki Hyun Kim

nlp.with.deep.learning@gmail.com

전처리의 늪에 오신 것을 환영합니다.

- 가장 재미없고(?) 반복적인 끝이 없는 작업
- 하지만 가장 중요 – 어쩌면 모델링만큼

NLP Project Workflow



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

말뭉치(Corpus)란?

- 자연어처리를 위한 문장들로 구성된 데이터셋

- 복수표현: Corpora

- 포함된 언어 숫자에 따라,
  - Monolingual Corpus
  - Bi-lingual Corpus
  - Multilingual Corpus

- Parallel Corpus: 대응되는 문장 쌍이 labeling 되어 있는 형태

English	Korean
I love to go to school.	나는 학교에 가는 것을 좋아한다.
I am a doctor.	나는 의사 입니다.

Service Pipeline

정제

- 학습 데이터와 같은 방식의 정제 수행

Tokenization

- 학습 데이터와 같은 방식의 분절 수행

Subword Segmentation

- 학습 데이터로부터 얻은 '모델'을 활용하여 똑같은 분절 수행

Batchify

- 학습 데이터로부터 얻은 사전에 따른 word2index 맵핑 수행

Prediction

- 모델에 넣고 추론 수행
- 필요에 따라 search 수행 (자연어 생성)

Detokenization (Optional)

- 사람이 읽을 수 있는 형태로 변환(index2word)
- 분절 복원