유튜브 인강 자막 크롤링 및 텍스트 분석

이병률, 이재영, 정우찬, 최난경

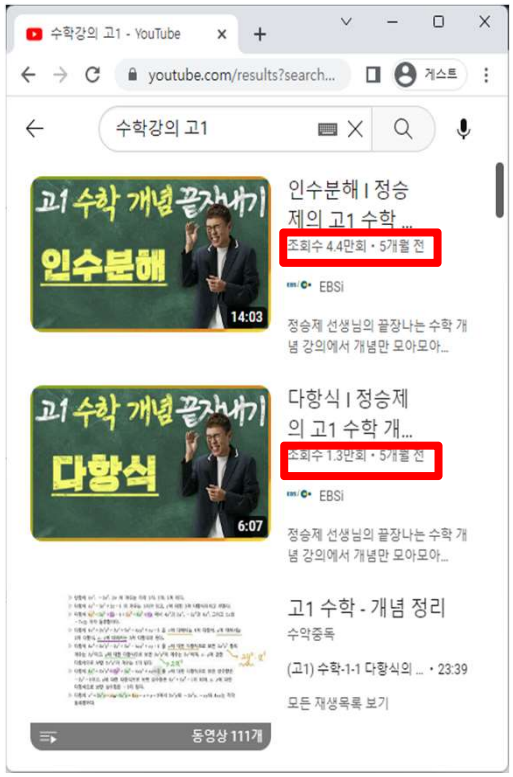
인기 콘텐츠 분석

- 인터넷 강의
- 인기 유무에 따른 언어적 차이가 무엇인가?

인기 콘텐츠 분석

- 인터넷 강의
- 인기 유무에 따른 언어적 차이가 무엇인가?

- 조회수, 게시일
  - 비율?
- 좋아요, 싫어요
  - 비율
- 댓글 감정 분석



인기 콘텐츠 분석

- 인터넷 강의
- 인기 유무에 따른 언어적 차이가 무엇인가?

- 형태소
  - Konlpy, soynlp
- 벡터화
  - Word2vec

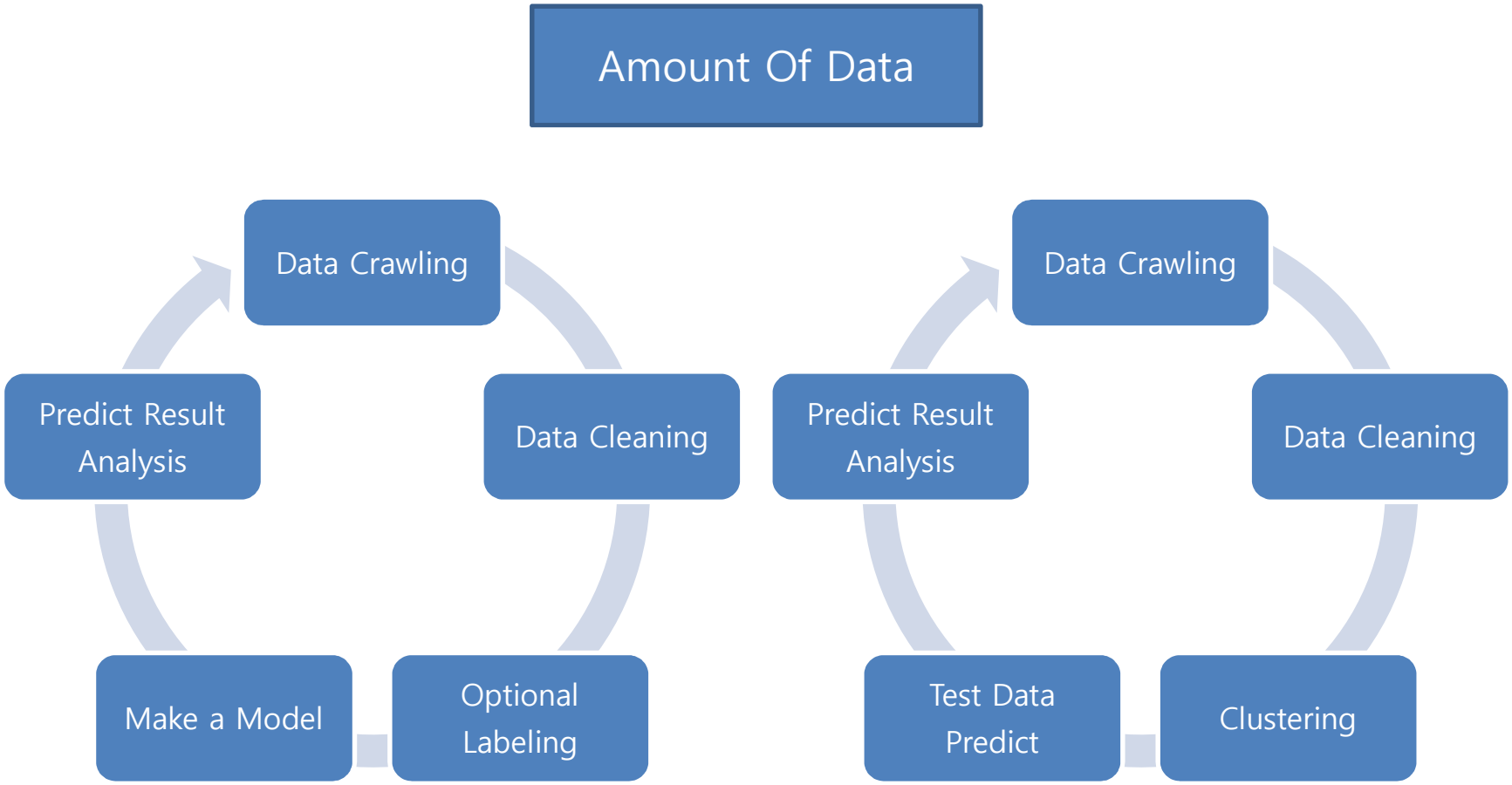
- 어순 중요?

빈출/분류모델/LDA/감정분석/ 등 모델 적합 및 분석

트랜스포머를 활용한 자연어 처리 (박해선 역, 한빛미디어)

Do it! BERT 와 GPT 로 배우는 자연어 처리 (이기창, 이지스퍼블리싱)

Process?


Amount Of Data
Data Crawling
Data Cleaning
Optional Labeling
Make a Model
Predict Result Analysis
Data Crawling
Data Cleaning
Clustering
Test Data Predict
Predict Result Analysis

Crawling

- DBMS?
  - MySQL, PostgreSQL, MariaDB
  - MongoDB
- 데이터 수집 방법
  - Crawling?
  - API?
  - 3rd party Libraries?
- 데이터 제공 방법
  - CSV -> DB -> API

대상

- 나이
  - 성인, 대학, 고등, 중등, 초등
- 과목
  - 국어, 영어, 수학, 과학, 사회, 예체능?

Ref

과제 1 – 유튜브 인강 자막 크롤링 및 텍스트 분석

Digital Transformation
All about Education, Chunjae Education.

인기있는 인터넷 강의
v.s.
인기가 없는 인터넷 강의

둘의 언어적 차이는 무엇일까?

기획	크롤링	정제 및 데이터화	분석/모델링
• 과목, 학년 선정 • 인기/비인기 강좌 선정 • 분석 범위(데이터 볼륨 등) • DB 설계 • *유튜브로 한정*	• Selenium 활용 • Transcript 추출 • 기타 메타 데이터 추출 (조회수, 게시일 등)	• 맞춤법, 띄어쓰기 등 • DB 보관: 원본 데이터, 정제 데이터, 형태소 데이터 • 기초 EDA 데이터(길이, 개수 등)	• 텍스트 분석 • 시각화 • Regression, Classification, Clustering or custom probabilistic model(1종 택일) • 보고서: 인기 강의 영상 제작을 위한 인사이트 제안

• 세부 사항은 교육생들이 주도적으로 기획, 조사, 협의 및 의사결정을 진행함(멘토는 간접 지원)

Confidential Information 1

과제 1 – 유튜브 인강 자막 크롤링 및 텍스트 분석

Digital Transformation
All about Education, Chunjae Education.

기본 크롤링 대상
(이 외에도 필요한 메타 데이터 수집)

Confidential Information 2

Ref

1. 프로젝트 명 : 인기 콘텐츠 분석
2. 프로젝트 과정
   ① 인기 콘텐츠 선정
   ② 데이터 수집 및 정제
   ③ 모델링 및 분석
3. 역량 및 지식
   - 파이썬
   - 자연어 처리에 대한 기본 지식
   - 기계학습 및 딥러닝 기초
4. 추천 교재
   - 딥 러닝을 이용한 자연어 처리 입문 (Team NLP, 위키독스)
   - 트랜스포머를 활용한 자연어 처리 (박해선 역, 한빛미디어)
   - Do it! BERT 와 GPT 로 배우는 자연어 처리 (이기창, 이지스퍼블리싱)
5. 프로젝트 적용 서비스 예시도

스크립트 스크래핑을 이용한 자료 수집

빈출/분류모델/LDA/감정분석/ 등 모델 적합 및 분석