K디지털교육 실무과제 소개

2023.05.23

천재교육 AI 센터
데이터기획분석팀 전대일

Confidential Information

과제 1 – 유튜브 인강 자막 크롤링 및 텍스트 분석

인기있는 인터넷 강의
v.s.
인기가 없는 인터넷 강의

둘의 언어적 차이는 무엇일까?

기획
- 과목, 학년 선정
- 인기/비인기 강좌 선정
- 분석 범위(데이터 볼륨 등)
- DB 설계
- *유튜브로 한정*

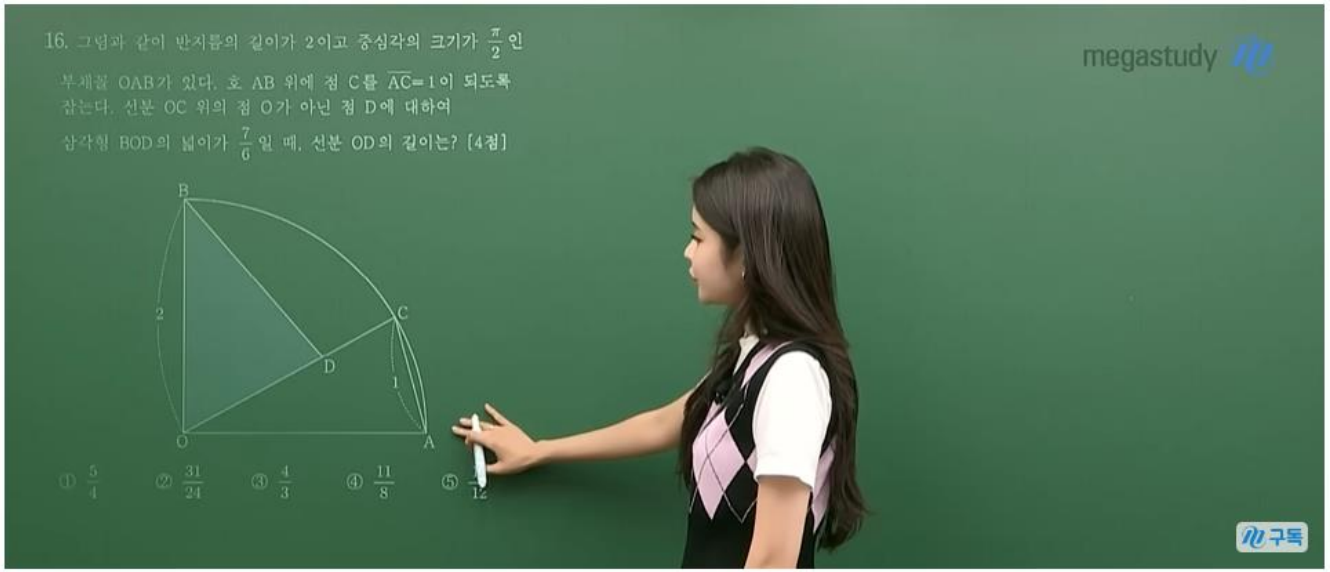
크롤링
- Selenium 활용
- Transcript 추출
- 기타 메타 데이터 추출 (조회수, 게시일 등)

정제 및 데이터화
- 맞춤법, 띄어쓰기 등
- DB 보관: 원본 데이터, 정제 데이터, 형태소 데이터
- 기초 EDA 데이터(길이, 개수 등)

분석/모델링
- 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)

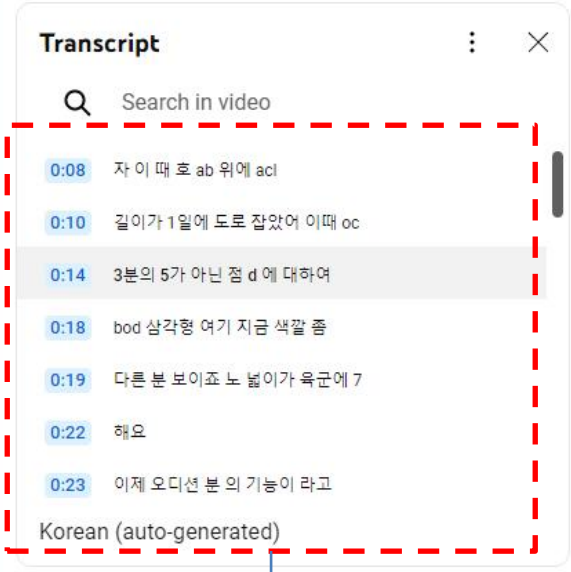
과제 1 – 유튜브 인강 자막 크롤링 및 텍스트 분석



[메가스터디] 수학 장미리쌤 - 2022년 고2 6월 학평 수학 해설 강의

theMEGASTUDY Subscribe

14K Share



All | From theMEGASTUDY | Lectures | Relat

기본 크롤링 대상
(이 외에도 필요한 메타 데이터 수집)

과제 2 – 재구매 상담 분류 모델 개발

Digital Transformation
All about Education, Chunjae Education.

< 제공될 상담 녹취록 예시 >

재구매 여부	전사 내용
O	네 안녕하세요 자 이제 들어왔어요 네 네 자 다혜씨 선생님이 오늘으
O	하세요 시영아 잘 지냈니 네 시영아 시간이 되게 빨리 지나가는 거
O	여보세요 여보세요 네 어머니 안녕하세요 아 네네네 혹시 잠깐 통화
	세요 준혁아 안녕하세요 네 네 밀크티 선생님이에요 네 아 우리 이번
	안녕하세요 어머니 선생님 네네 생각해 봤는데요 네네 네 어 그만하
O	하세요 어머니 네 안녕하세요 선생님 혹시 어머니 운전 중이세요 아
O	네 어머니 네네 안녕하세요 아 네네 우리 성원이가 지금 이제 십 이
O	우리 설이 안녕하세요 자 우리 설이 숙제 너무 너무 잘했네요 오늘은

1. 데이터 정제: 탐색 후 정제 이슈 파악 및 실행
2. 기본 분석: 재구매 성공 및 실패 녹취 corpus 에서 상대적으로 더 많이 등장하는 키워드 및 연결 어휘들(N-Gram)은?
3. 주요 관심 키워드별 빈도수에 따른 재구매 확률 예측 모델(Logistics Regression, 다수 키워드에 대한 모델 구현 자동화)
4. 아래 문서 분류 모델 중 최소 3종 성능 테스트(성능지표 선정) → 결과 비교 → 1개 최종 선택 후 성능 최적화 진행 → 보고서 작성

문서분류 모델
Naïve Bayesian Classification
Doc2Vec
Google BERT
Huggingface Transformer(GPT 요소기술)

- 모델 성능평가 지표를 무엇을 사용할지 고민하고, 왜 해당 지표를 선정했는지 이유를 결과보고서에 추가
- 세부 사항은 교육생들이 주도적으로 기획, 조사, 협의 및 의사결정을 진행함(멘토는 간접 지원)