# (통합) 최종 결과물 제작

## 부스트캠프 2기 분석

| 트랙 | 번호 | 주제 | 항목 |
|---|---|---|---|
| CV | 18 | 상품번호 이미지가 들어오면 상품번호 인식(OCR) | 0. 수행 목표 1. 문제정의(시장 조사, 서비스 기획 주제, 서비스 기획 세분화) 2. 서비스 흐름도 - 동작 workflow(사용자가 입력부터 결과가 도출될 때까지 흐름도) 3. 서비스 요구사항 - 원하는 목표(상품번호 타이핑 시간보다 빠르게 OCR 인식) - 실제 목표로 하는 사항(5초이내 결과를 도출한다) 4. 서비스 시스템 구조 - 시스템 아키텍처 5. 데이터셋 설명 6. 모델 선정 - 성능 비교 테이블 7. 모델 평가 - Metric 8. 프로토타입 시현 9. Future Works - 발생한 문제점 - 더 해보고 싶은 것 10. 참여자 설명 |
| NLP | 13 | 라이브 커머스를 진행하는 호스트를 위한 실시간 채팅 매니지먼트 - 악성채팅 문구 처리 - 악성채팅 / 부정글 / FAQ 구분 - 채팅글을 바탕으로 한 실시간 시청자 반응 대시보드 구현 | 목차 0. 팀 소개 1. 프로젝트 소개 - 주제 선택 - 실제 문제 사례 - 프로젝트 목표: 어떤 기능을 만들 것인지 - 시스템 아키텍처 2. 모델링 - 활용 데이터셋 선택 및 데이터셋 여부 - 모델 백본 / 아키텍처 - 모델 실험 3. 서비스 구현 - 프론트엔드 (화면 구성과 기능) - 백엔드 4. 영상 |
| NLP | 12 | 악플 수집 서비스 | 1. 프로젝트 목적 & 개요 2. 팀 소개 3. 모델 선정 & 모델 평가(캐글 점수) 4. 플로우 차트 - 시스템 아키텍처 - 파이프라인 5. 데모 영상 6. 직접 사용하는 방법 7. 레퍼런스 |
| CV | 6 | 공항 물품 검색 자동화 | 1. 데이터셋 2. 모델 선정 3. 훈련 결과 |
| NLP | 5 | 어제 있었던 주요 뉴스의 요약된 내용을 오디오 형태로 듣는다 | 1. 팀 소개 2. 프로젝트 소개 - 주제 선정 이유 3. 모듈별 소개 - 크롤링: 뉴스 요약이기 때문에 뉴스를 어떻게 크롤링할 지에 관해 담아놓은 것 같다 - 클러스터링: 뉴스 묶기: 어떤 방식으로 했는지 → 모델 선정 - Summarization: 데이터셋, 모델 설명, 실제 A/B 테스트 진행 - Text to Speech: 실제 뉴스를 음성으로 바꾼다 - Serving: 서빙 4. 시연 영상 5. 향후 개선사항 - 각 항목별 아쉬웠던 점 6. QnA |
| CV | 4 | 올바른 손 씻기 교육을 위한 손씻기 단계 인식 모델 | 1. 팀 소개 2. 프로젝트 개요 3. 개발 목표: 현실적으로 원하는 서비스 형태 & 기능들 4. 데이터셋: 캐글 손씻기 데이터셋 5. 문제 정의 6. 모델 선정 배경 & 평가 7. 프로덕트 서빙 8. Future Works |

## 피드백 자료

⚠️ 현업자 피드백 특강 정리

- 시간이 더 있다면 어떤 기능(TO DO)을 추가해보고 싶다
  - 느낀점, 부족한 점, 보완할 점
- 성능에 대해 정량적 평가를 할 수 있는지 → 없다면 왜 할 수 없었는지
  - 추가로 성능이 왜 잘 나오지 않았는지에 대한 분석까지
- 플로우는 한 눈에 들어오기 쉽게하기
  - 복잡한 것은 적게
  - 강조될 내용만 부각해서 나타내기
  - 감출 것은 감추자
- 왜 그러한 데이터를 가져다 썼을까?
- 고민을 많이 한 느낌을 주면 좋을 것 같다
  - use case 가정
  - 기술적인 가정 → 애초에 안될 부분은 가정으로 걸러내자
- 끝나고 마무리하면서 아쉬운 점, 부족한 점, 더 해보고 싶은 점

## 부스트캠프에서 예시로 든 목차 구성

- Intro : 팀 소개 / 프로젝스 소개(문제 정의) / 개발 목표
- Model/Reasearch: 데이터셋 / 모델 / 연구 / 최종 적용 모델
- Product Serving: 전체 서비스 아키텍쳐/ 구현/ 데모
- Result/Conclusion: 시연 영상 / 후속 개발 및 연구 / 결과 및 고찰
- Appendix:도전적인 실험 / 레슨런 / 예상 Q&A / 팀원 개별 소개 등
- (필수) 별첨. 서비스 아키텍쳐 → 별첨은 뭐야?, 서비스 아키텍처는 꼭 들어갔으면 좋겠다는 뜻인 것 같다

제작 색상: 건웅, 경연, 동준, 민영

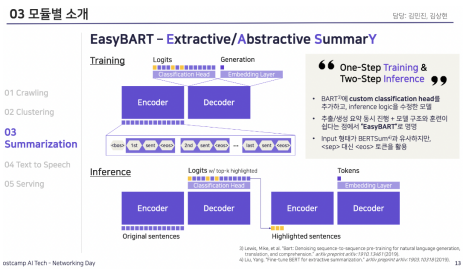
## 발표 ppt 내용 - Why? 적인 측면이 어느정도 잘 나올 수 있도록 …

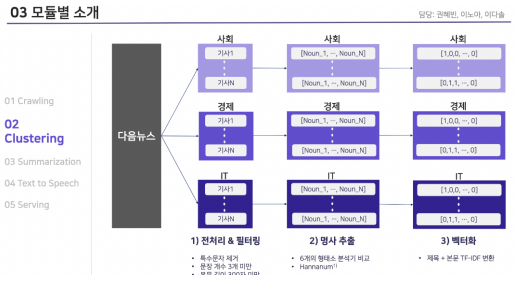
| 대분류 | 소분류 | 내용 | 비고 |
|---|---|---|---|
| 제목 | | 제목: 도서에 음악을 더하다(가제) #음악추천 #도서추천 #임베딩 #즐거움에즐거움을더하다 | ppt template 첫 화면 |
| 0. 팀 소개 | | 멤버 소개 - 이름, 캠퍼 ID, 이메일 - 각자 역할을 해쉬태그로 (#데이터 제작 #PM) | |
| 1. 프로젝트 소개 | 기획 배경 | 1. 웹 소설 시장 규모: 매년 웹소설 시장 규모가 커진다 → 유망한 시장 - 국내 웹소설 시장 규모 6000억 - 전업 주부부터 검사까지 그들은 왜 웹소설에 빠졌나? - 한국 콘텐츠 진흥원 웹소설 이용자 실태조사 - 2020년 기준 콘텐츠 산업조사 2. 음악 스트리밍 플랫폼 규모: 단독 음악 콘텐츠 추천으로는 시장 점유가 쉽지 않을 수도 있다는 점 - K팝 콘텐츠 뮤비 즐기며 유튜브로 음악을 '보기도' - 치열해지는 음악 세계 음악 스트리밍 시장 최강자는? 3. 부가적인 창출 이익 - 한국 콘텐츠 진흥원 웹소설 이용자 실태조사 유료 컨텐츠 결제 비율 1·2위 4. 현재 시중에 만들어진 다양한 음악 플레이리스트 - 유튜브에서 플레이리스트 검색시 다양한 동영상이 나온다 - 별도로 검색을 해야하는 불편함 존재 5. 읽는 콘텐츠 + 듣는 콘텐츠: 아직 존재하지 않음 / 새로운 형태의 플랫폼 강조 - 콘텐츠 원천 웹소설 플랫폼 특화 전쟁 뜨겁다 - 새로운 융합 컨텐츠 플랫폼으로 새로운 시장 개척 6. 카카오 페이지 + 전체적인 문화 컨텐츠 내용 7. 음원사이트 수익구조 | 한국 콘텐츠 진흥원 웹소설 이용자 실태조사 - p.29: 웹소설 시장 5년 만에 40배 이상 성장, 2018년 기준 규모 4000억 - p.141: 유료결제 콘텐츠 비율: 웹소설(46.8, 1위), 음악(35.8, 2위) - 어떤 형태의 기기에서 사용하는 지에 대한 정보도 같이 있음 2020년 기준 콘텐츠 산업조사 |
| | 문제 정의 | 1. 문제 정의: 책(웹소설)과 잘 어울릴만한 노래 플레이리스트를 만들어보자! | |
| | 개발 목표 | 1. 도서와 음악에 대한 임베딩 벡터 만들기 → 임베딩 벡터를 만든 이유 (1) 기존 머신러닝 추천 방법을 사용할 수 없는 이유, 딥러닝 방법을 적용할 수 없는 이유: 아마 거의 비슷한 이유로 추려질 것 같지만 하나씩 이유를 만들어보자 (2) 왜 임베딩을 만들려고 했는 지에 대한 내용이 담기면 좋을 것 같다 2. 자신이 보려는 콘텐츠를 검색하고 그에 대한 음악을 추천받는 형태의 서비스 (1) 검색 (2) 플레이리스트 기능 ※ 이에 대한 설명을 잘 꾸몄으면 좋겠다 | 1. 기존 추천 방식을 사용할 수 없는 이유 (1) Item-Based Recommendation (2) User-Based Recommendation (3) Matrix Factorization - interaction 정보의 부재, - MF도 결국 hidden representation을 만드는 것이기 때문이라고 생각한다 ※ 기타 다른 머신러닝 방식.. 2. 딥러닝 방법을 적용할 수 없는 이유 (1) AutoEncoder 기반 (2) Sequential 모델 기반 3. 하나의 도서와 음악 객체를 같은 공간에 표현 가능하다면 여러 유사도 지표를 통해 유사도를 통한 추천이 가능하지 않을까? |
| 2. Dataset | 도서 데이터셋 | 1. 교보문고를 크롤링한 이유 → 책 태그 및 소개 항목 많음 2. 데이터 칼럼: 간단하게 크롤링 방식 소개(선택) | 1. 감정 및 책이 갖고있는 분위기를 나타내는 태그 정보의 풍부 및 사용할 데이터가 다양했다. 2. yes24 → 전체 책 목록을 찾을 수 있어 먼저 책 목록 제작 → 교보 문고에서 해당 책 데이터 크롤링 |
| | 음악 데이터셋 | 1. 카카오 아레나 멜론 음악 데이터셋 (플레이리스트 단어는 가급적 제외) - 학술적 목적으로 사용한다dbf는 것을 명시! 2. 데이터 칼럼 | 카카오 아레나 멜론 음악 데이터셋 중 song_meta.json, genre_gn_all.json 데이터셋을 사용하였으며, 추후에 가사는 크롤링 |
| 3. Model | 임베딩 벡터 만들기 | 1. 무엇을 기반으로 임베딩 벡터를 만들 것인가? - 도서와 음악이 지닌 공통적인 정보: 글자 2. 어떻게 만들 것인가? → 이를 위해 도서와 음악에서 어떤 칼럼을 사용했는지도 같이 기록 (1) 키워드 방식 (2) 감정 정보 방식 (3) 내용 방식 ※ 왜 위와 같은 3가지 방식을 사용했는지 ※ 왜 합치는 방식을 이용했는지 | (1) 키워드 방식 : 책 태그, 멜론 노래 태그 정보 + 가사, 책 소개 글을 태그화 (3) 내용 방식 : 책 태그, 책 정보, 멜론 노래 태그, 장르, 가사 등을 kcelectra, ko-sbert 등을 이용해 각 칼럼의 768 차원 임베딩 벡터화 |
| | 내용 방식 | ※ 내용 방식으로 임베딩 벡터 만들기 ※ 예시 ※ (1) 모델 선정 이유: 방대한 데이터셋으로 사전학습된 모델/토크나이저, corpus - github에 한국어 모델 비교한 것 있으니 그 자료를 사용해도 좋을 것 같다 (2) 어떤 방식으로 임베딩 벡터를 만들었는지 - 어떤 column을 이용했는지() - 문장 임베딩 방식: 이런 용어가 정확히 존재하는지 잘모르겠다 - 나중에 풀링을 했다면 어떤 방식으로 풀링을 했는지 (3) 추천 방식: cosine similarity 등 | ※ 아래 3가지 내용은 필수라고 생각된다 (1) 모델 선정 이유 (2) 어떤 방식으로 임베딩 벡터를 만들었는지 (3) 추천 방식 다른 유사도 기준도 필요?? > 다른 유사도 기준은 왜 안써봤냐는 지적?? |
| | 키워드 방식 | ※ 키워드 방식으로 임베딩 벡터 만들기 | ※ 아래 3가지 내용은 필수라고 생각된다 (1) 모델 선정 이유 (2) 어떤 방식으로 임베딩 벡터를 만들었는지 (3) 추천 방식 |
| | 감정 정보 방식 | ※ 감정 정보 방식으로 임베딩 벡터 만들기 | ※ 아래 3가지 내용은 필수라고 생각된다 (1) 모델 선정 이유 (2) 어떤 방식으로 임베딩 벡터를 만들었는지 (3) 추천 방식 |
| | 최종 추천 방식 | 1. 합치는 방법에 대한 설명 - 겹치는 것을 최우선으로 - 최종 후보 음악들을 선별하는 과정: 몇 개의 후보군을 골랐는지 | |
| 4. Product Serving | 요구사항 정리 | 1. 프로덕트 요구사항 명세서 - 요구사항 간단하게 나열 - 즉, 1권의 책을 위한 음악을 추천해보자 / 사용자가 직접 검색을 통해 들어온다(예시) - 검색어를 입력하고 이와 연관된 도서들을 보여준다. - 하나의 도서를 클릭하였을 시 어울리는 음악 50 개를 플레이리스트로 제공한다. - 제공된 음악 유튜브 화면을 클릭 시 음악이 재생된다. - 구글 드라이브에서 정리한 요구사항 명세서와 목업 디자인을 캡처 화면으로 보여주었으면 좋겠다 (너무 간단하게 작성한 것처럼 티나지 않게 보여주기) 2. 실시간 방식 vs 배치 방식 - 실시간 방식이 아닌 이유 - 오른쪽 요구사항 형식을 참고해서 왜 배치 방식을 선택했는지 기록하는 것도 나쁘지 않을 것 같다 | 1. 서비스 요구사항 예시 (2기-CV 18조) - 정성적 요구사항: 상품번호 타이핑 시간보다 빠르게 OCR 인식 - 정량적 요구사항: 이미지를 잘라낸 시점부터 정보를 제공받기까지 5초 이내 |

| 대분류 | 소분류 | 내용 | 비고 |
|---|---|---|---|
| | 서비스 흐름도 | 서비스 흐름도 작성(선택) - 플로우 차트 | ※ 서비스 흐름도 예시 (2기-CV 18조) https://github.com/boostcampaitech2/final-project-level3-cv-18#2️⃣-서비스-흐름도 |
| | 서비스 아키텍처 | 서비스 아키텍처 작성 | ※ 서비스 아키텍처 예시1 (2기-CV 18조) https://github.com/boostcampaitech2/final-project-level3-cv-18#four-서비스-시스템-구조 ※ 예시2 발표 pdf: 22~28 page → 아래 보시면 2기 NLP 5조 발표자료를 참고하시면 됩니다! |
| | 화면 구성과 기능 | 화면 구성과 기능 작성 | ※ 예시 발표 pdf: 34~37 page |
| | 시연 영상 | 시연 영상 제작 (길이를 정해야함!) 영상 제작 전 시나리오 작성 필수! | |
| 5. Future Work | 아쉬운 점 (아쉬운 점이 아니라 다른 이름도 괜찮을 것 같다) | 미처 구현하지 못했거나 아쉬웠던 부분들 정리 ※ 예시 ※ 1. 데이터셋에서 아쉬웠던 부분 - 도서를 볼 때 어떤 음악을 들었는 지에 대한 부분 2. 모델 학습할 때 어려웠던 부분 - 키워드를 뽑아낼 때 - 감정을 선택할 때 - column에 대한 아쉬움 3. 도서를 읽을 때 어떤 음악을 들었는 지에 관한 정보가 존재하지 않다는 점 4. 프로덕트 서빙시 아쉬웠던 점 (예시) - Docker 이용 못함 - GCP 같은 클라우드 서비스를 사용하지 못함 5. 플레이리스트 제목과 썸네일 - 어떤 책에 대한 플레이리스트를 제공해줄 때, 책의 내용을 다시 연상할 수 있도록 하는 플레이 리스트 제목과 썸네일도 함께 제공해줬으면 어땠을까 하는 아쉬움 ex. 책에서 인상적인 문구 | 미처 구현하지 못했거나 아쉬웠던 부분들을 정리하지만 부족한 부분을 많이 드러낸다는 것이 아님 |
| | 도전하고 싶은 점 | 도전해보고 싶은 것 ※ 예시 ※ 1. 도서를 읽을 때 어떤 음악을 들었는지 에 관한 정보가 존재하지 않다는 점 - 실제 추천된 음악중 어떤 음악을 선택했는지에 관한 정보를 가져와서(실제 TEST) 추천 모델(MF, AutoEncoder, Sequential Model 등)을 사용해본다 - 평가 정보 사용 (Metric 등) 2. GCP 같은 클라우드 서비스를 사용 | 최대한 아쉬웠던 점들을 개선해나가는 방향으로 도전하고 싶은 부분들 작성 |
| Appendix | | Appendix에 포함시키고 싶은 것들 ※ 예시 ※ - 데이터셋 링크 - 모델 GitHub 주소 - reference | 2기 예시 - About Team - 모델 결과 예시 - 팀 내 역할 |

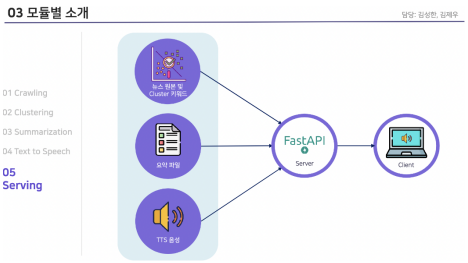
### 제작할 때

- 각 행은 1페이지를 가리키는 것이 아님 → 내용에 따라 여러페이지가 될 수 있음
- 이를 위해 표를 이용하여 각 페이지에 어떤 내용을 적을 것인지 기록할 필요가 있음
- 여러사람이 동시에 제작한다면 ppt template을 통일할 필요가 있음
- 최대한 깔끔하게: 색상 통일(ppt-template에 색상이 명시됨), 글자보다는 그림 위주로 ⇒ 예시




- 그림 적극활용




### 참고 발표 자료 링크

- 2기 CV 18조: https://github.com/boostcampaitech2/final-project-level3-cv-18#nine-future-works
- 2기 NLP 13조: https://docs.google.com/presentation/d/1pmBbn0NuesvNO-vzAuRdGg0Ph0q09odh/edit#slide=id.p1
- **2기 NLP 5조: https://github.com/boostcampaitech2/final-project-level3-nlp-05/blob/main/NLP05_귀가노니_ppt.pdf**
- 2기 CV 4조: https://github.com/boostcampaitech2/final-project-level3-cv-04/blob/main/docs/Networking Day 발표자료 최종본.pdf

## GitHub Repository 내용 → ppt 내용 작성 후 정리!

- 어떤 방식으로 만들지 정리하기 → 2기 CV 18조 예시
- README.md 내용
- 디렉토리 & 파일 구성 예시
  - 디렉토리 구성
    - 모델 불러오기 등을 각자 class로 작성해서 python 파일로 저장!
    - 모델 결과 만들 때 사용한 함수들 (utils) 정리!
    - front-end
    - back-end
  - notebook 파일: example 파일로 생각하면 될 것 같다
    - 모델을 불러오고 임베딩 벡터를 만드는 과정

## Reference

카카오 관련 bi

주식회사 카카오엔터테인먼트

카카오의 엔터테인먼트 비지니스 기업, Kakao Entertainment Corp

https://www.kakaoent.com/introduce/company

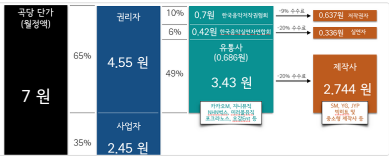
kakao ENTERTAINMENT

스트리밍 음원 수익 구조

스트리밍 음원 수익구조 2020

2018년 6월 20일 문화체육관광부가 '음원 전송사용료 징수규정 개정안'을 최종 승인했습니다. 여기서 말하는 '음원 전송사용료'는 '스트리밍 또는 다운로드 방식으로 음악을 재생할 때 작곡, 작사가, 실연자, 음반제작자 등 권리자가 받는 저작권료'로 규정되어 있습니다. 이후 스트리밍 음원에 대한 수익구조는 권리자의 수익 분배 비율이 60%에서 65% 인상되었죠.

https://www.pianocroquis.com/165




센텐스버트 문장 임베딩

점프 투 파이썬

점프 투 파이썬 오프라인 책(개정판) 출간 !! (2019.06) ** * [책 구입 안내](https://wikidocs.net/4321) 이 책은 파이썬 ...

https://wikidocs.net/156176


추천시스템 유사도 기법

추천시스템에서의 유사도 지표와 피드백 특징 연구.

CF알고리즘에서 메모리 기반의 kNN 알고리즘에서 주요 이슈로 기본적으로 cold-start가 있고, 그리고 또 동시 평가 아이템을 기준으로 유사도를 구하는 알고리즘의 특성상, 동시 평가 아이템이 적을 수 밖에 없는 평점 매트릭스에서 더 잘 유사도를 구하기 위한 연구들에 대해 알아본다. 주어진 평점 매트릭스는 늘 sparse하다. 그럴 수 밖에 없는 것이 아이템은 제한되어 있고, 사용자는 상대적

https://data-science-hi.tistory.com/150

$$\frac{\sum_{i \in I_{a,b}} (R_{a,i} - \overline{R}_a)(R}{\sqrt{\sum_{i \in I_{a,b}} (R_{a,i} - \overline{R}_a)^2} \sqrt{\sum_{i \in}}}$$

셀레니움 크롤링

https://velog.io/@mincho920/Python-X-Selenium-%EC%9C%A0%ED%8A%9C%EB%B8%8C-%ED%81%AC%EB%A1%A4%EB%A7%81

예스24 교보문고

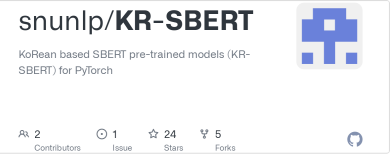
www.yes24.com http://www.kyobobook.co.kr/index.laf?OV_REFFER=https://www.google.com/

kr-sbert

GitHub - snunlp/KR-SBERT: KoRean based SBERT pre-trained models (KR-SBERT) for PyTorch

A pretrained Korean-specific Sentence-BERT model (Reimers and Gurevych 2019) developed by Computational Linguistics Lab at Seoul National University. We recommend Python 3.6 or higher and sentence-transformers v2.2.0 or higher. You can see the sentence '잠이 옵니다' is more similar to '졸음이 옵니다' (cosine similarity 0.65774536) than '기차

https://github.com/snunlp/KR-SBERT




sbert data augmentation

Augmented SBERT - Sentence-Transformers documentation

https://www.sbert.net/examples/training/data_augmentation/README.html