Project 소개:

경영대학 CDTB 데이터 활용 경진대회 with KU-BIG

Project Proposal(2020.11.05.)

Team T.O.P.-K

박시전 권형근 안수빈

차 례

고려대학교 경영대학
Center for Digital Transformation & Business

2020년도 데이터 활용 경진대회

1. 참가자격

1) 고려대학교 안암캠퍼스 소속 학부 및 대학원 재학생, 휴학생, 수료생
2) 개인 혹은 팀 단위(최대5명) 구성 (1인 이상의 경영대학(원) 재학생 필수)
3) 연구 지도에 참여하는 지도교수(서울캠퍼스 소속 전임교수에 한함) 1인이 있어 야함
4) 제한 사항
졸업예정자 (8월 졸업 예정 포함)
지도교수의 장기출장, 연구년 등 연구지도를 지속적으로 진행할 수 없는 경우 지원 불가

2. 데이터 분야(자유 주제)

ex. 교육/ 국토관리/ 공공행정/ 재정금융/ 산업고용/ 사회복지 등

3. 지원 서류 제출

가. 제출 기간 : 2020년 8월 18일~ 9월 4일 17시
나. 제출 방법 : KUBS_CDTB@korea.ac.kr, 이메일 제출 ※ 파일명 : 팀명_과제명
다. 제출 서류: 지원서, 활동계획서, 지도교수 확인서, 개인정보 동의서
- 포탈 공지사항 또는 Google 드라이브에서 양식 다운로드
* 포탈 공지사항에서 전체 경진대회 일정 참고 (메일 본문內 링크有)
* 신청양식은 메일본문에 있는 링크에서 다운받아서 메일로 제출

4. 팀당 지원 사항

가. 활동계획서 내용에 따라, 필요시 별도의 데이터 구입, 클라우드 서비스, 유료 app 및 SW지원 가능

나. 본선 대회 후 상장 및 시상금 수여
- 최우수 1팀 2,000,000원
- 우수 2팀 각 1,000,000원
- 장려 3팀 각 500,000원
- 참가 4팀 각 300,000원

다. 전문가 콜로퀴엄 제공

라. 모든 결과물 (지원서, 활동계획서, 보고서, 진행일지 등)은 대학혁신지원사업 및 경영대학(원)의 연구 · 평가보고용으로 활용될 예정임. (개인 연구 사용 불가)

CONTACT INFO 02-3290-5364 KUBS_CDTB@korea.ac.kr

1 대회 및 팀 소개

- 고려대학교 경영대학 주관 교내 경진대회: 1회
- 2020년 9월 ~ 2021년 1월(약 5개월)에 걸쳐 진행
- 자유주제; 현재 7팀이 본선에서 경쟁 중에 있음.
- 상금 규모:
  - 최우수 1팀: 200만원
  - 우수 2팀: 100만원
  - 장려 3팀: 50만원 ~~+ 참가 4팀: 각 30만원?~~
- 지원 규모:
  - 데이터 구입 비용 200만원 지원
  - 클라우드 서버비 100만원 지원
  - 유료 app 및 SW 비용 200만원 지원
  - 기타 활동 지원금 150만원 지원

1 대회 및 팀 소개

4주 간의 활동을 거쳐 개별 팀원들의 역할이 일정 수준 확정되었음.

+ 지도교수: 김상용*(경영대 마케팅 분과)*

박시전

팀장, 12기

- 팀 업무 총괄
- 각종 행정서류 담당
- 대외연락 담당 등

권형근

팀원, 11기

- 코드 개발 담당
- 리뷰 크롤링 전담
- 전산 관련 업무 지원

안수빈

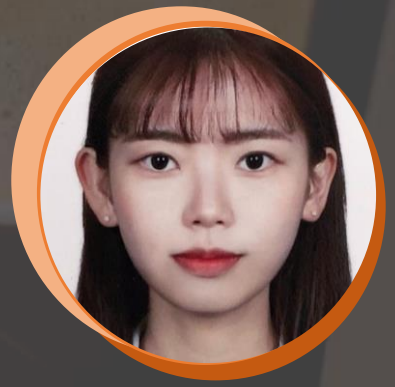
팀원, 11기

- 데이터 통계 처리 담당
- 상권 구획 업무 총괄
- 유동인구/점포위치

이은지

팀원, ~~누구세요?~~

- 선행연구 검토 , 데이터셋 수집 담당
- 점포별 매출 데이터 업무 지원



이가영

팀원, 12기

- 선행연구 검토 , 데이터셋 수집 담당
- 점포별 매출 데이터 업무 지원

1 대회 및 팀 소개

4주 간의 활동을 거쳐 개별 팀원들의 역할이 일정 수준 확정되었음.

+ 지도교수: 김상용*(경영대 마케팅 분과)*

박시전

팀장, 12기

- 팀 업무 총괄
- 각종 행정서류 담당
- 대외연락 담당 등

권형근

팀원, 11기

- 코드 개발 담당
- 리뷰 크롤링 전담
- 전산 관련 업무 지원

안수빈

팀원, 11기

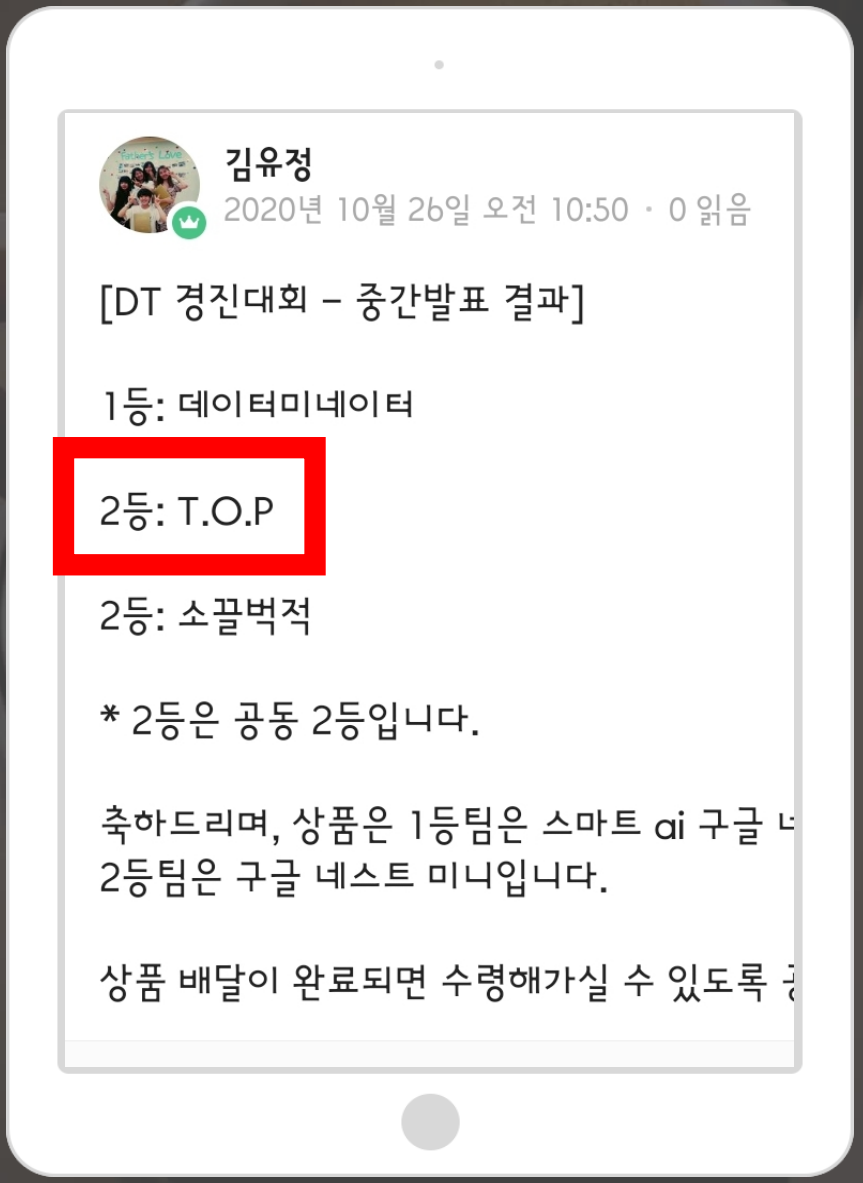
- 데이터 통계 처리 담당
- 상권 구획 업무 총괄
- 유동인구/점포위치

T.O.P.-K

Toward
Optimal
Provision
in **K**U-BIG

1 대회 및 팀 소개

1차 중간보고 결과: 2등 팀 T.O.P.



김유정
2020년 10월 26일 오전 10:50 · 0 읽음

[DT 경진대회 – 중간발표 결과]

1등: 데이터미네이터

2등: T.O.P

2등: 소끌벅적

* 2등은 공동 2등입니다.

축하드리며, 상품은 1등팀은 스마트 ai 구글 ㄴ
2등팀은 구글 네스트 미니입니다.

상품 배달이 완료되면 수령해가실 수 있도록 공

2 연구 주제 및 방향 소개

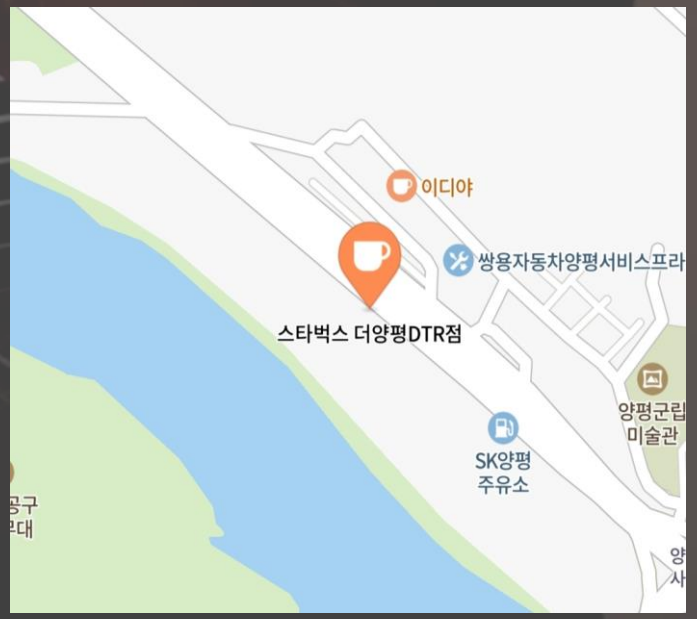
9월부터 진행해 온, T.O.P.팀의 연구 주제에 대한 간략한 소개

커피전문점을 중심으로

주제: 상권-브랜드 이미지 매칭을 통한 창업컨설팅 모형 개선

Motivated by: 스타벅스 더양평DTR점

BUT



2 연구 주제 및 방향 소개

9월부터 진행해 온, T.O.P.팀의 연구 주제에 대한 간략한 소개

커피전문점을 중심으로

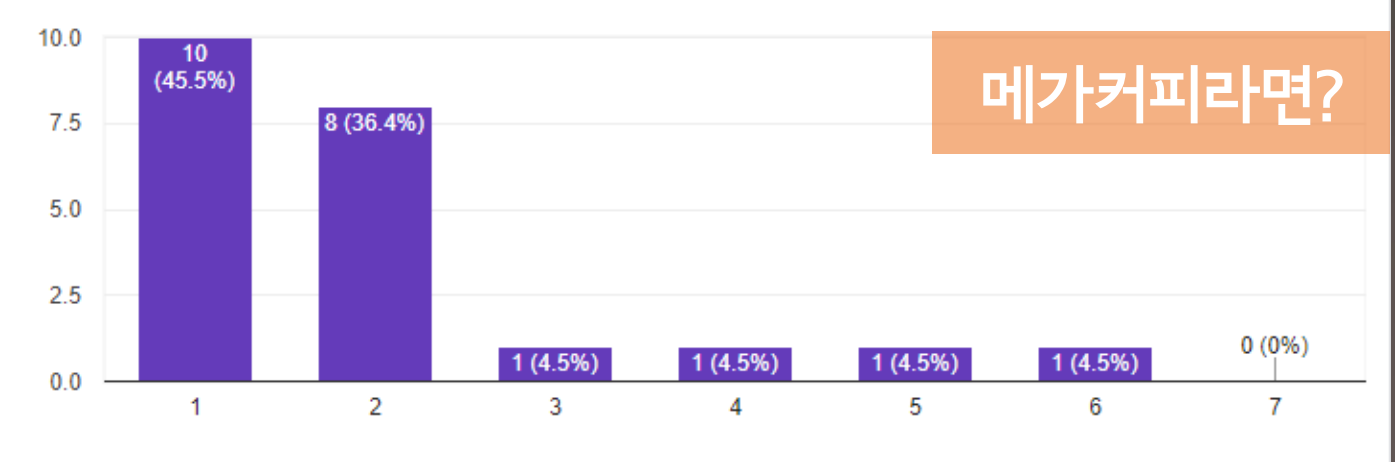
주제: 상권-브랜드 이미지 매칭을 통한 창업컨설팅 모형 개선

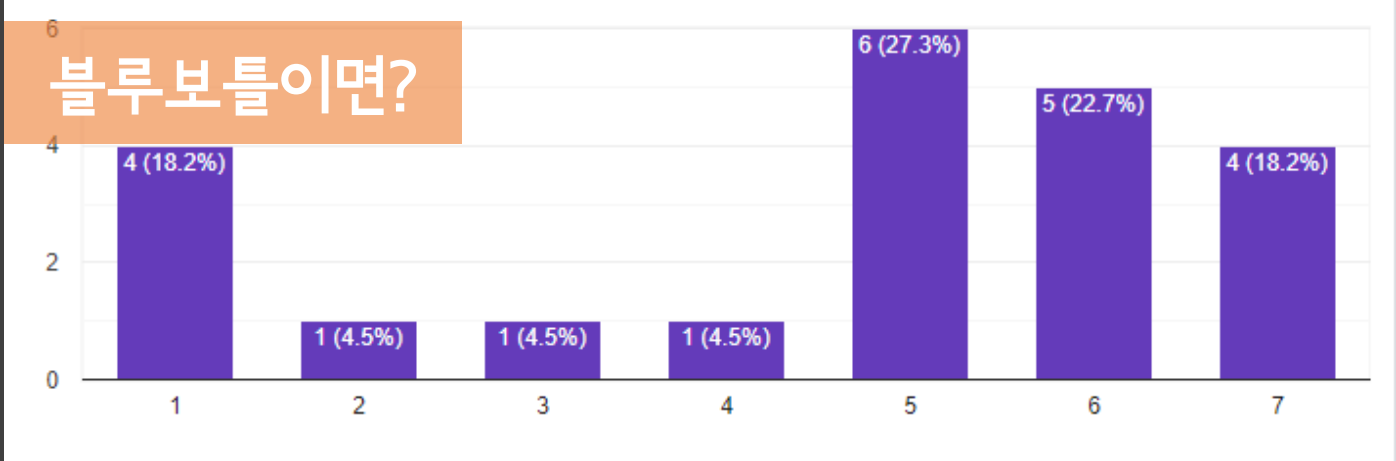
Motivated by: 스타벅스 더양평DTR점

여기가 스타벅스가 아니라 이디야였으면, 갔다? 안갔다?

2 연구 주제 및 방향 소개

9월부터 진행해 온, T.O.P.팀의 연구 주제에 대한 간략한 소개





브랜드에 따라 방문 의사가 전혀 다르다!

그럼, 상권의 이미지와 브랜드 이미지 간에 어떤 상관관계가 있지는 않을까?

2 연구 주제 및 방향 소개

지난 4주 간 다음과 같이 세 부분으로 나누어 연구 진행을 준비하였음.

커피전문점을 중심으로

주제: 상권-브랜드 이미지 매칭을 통한 창업컨설팅 모형 개선

상권 구획 정하기

- ✓ 상권의 정의: 상가업소 클러스터와 버스정류장 별 유동인구 클러스터의 교집합
- ✓ 버스정류장 별 승하차 인원 정보 및 위도 경도 정보, 지역별 상권 개수 등을 주요 정보로 활용

리뷰 크롤링하기

- ✓ 상권 구획이 완료되면, 해당 구역 내에 있는 모든 커피 전문점의 리뷰를 크롤링할 예정
- ✓ 주요 Reference: 네이버 지도 리뷰, 다음 지도 리뷰, 망고플레이트, 기타 가용한 대형 리뷰 밀집 사이트 전체

재무 데이터 찾기

- ✓ 카페 브랜드/점포별 재무제표 공유를 시도
- ✓ 상권 별로 어느 브랜드가 가장 많은 영업이익을 보고 있는지 분석하는 준거로 삼을 예정이었음.

3 업무 현황 및 계획 I. 리뷰 크롤링

(1) 연구 진척 상황

네이버, 다음 지도 각각에 대한 리뷰 크롤링 성공

NAVER MAP

좋아여
직원 불친절. 베이커리 맛없음. 사람 너무 많음. 역대 갔던 스벅 중 최악.
오늘은 날씨까지 참좋아 강건너 경치 구경하기 좋았어요—
한강 바로 앞에 위치해서 뷰가 끝내준다는 핫한 스타 벅스!! 밤이라 풍경은 못봤지만 통유리창으로 된 건 물이 인상적이었고 드라이브쓰루도 가능해서 ...
빵 맛있어요!
빵맛잇어요. 사람이 생각보다 아직도 많아요. 커피는 동일한 맛. 뷰 포인트는 진짜 일찍 안오면 없습니다
오픈시간 전에 가도 사람들이 ㅎㄷㄷ했지만 넓고 분위기 좋고 빵도 맛있었음
스타벅스는 어디든 맛있어요
2번째 방문인데, 역시나 아쉬운 부분이 많습니다. 매 장 건축이나 디자인은 새롭지만, 관리나 운영은 아쉽 습니다. 1) 매장 청결, 관리가 부족합니다. 사...
평일 낮인데도 사람이 많아요
뷰도 좋고~ 다 좋네요^^ 평일 오후 커피한잔의 여유
멋지네요
전망이좋네요
방문자가 많아서 정신 없어요 ㅠ 주차나 DT는 안내해주시는 분들이 많아 편합니다.
분위기 너무 좋더라고요.
풍경이 멋지고 빵이 맛있고 사람이 많아요
스벅은 진리. 사회적 거리두기 때문인가, 좌석배치도 드문드문. 한가로이 남한강뷰보며 힐

DAUM MAP

뷰 좋네요~평일 오후라 잘 있다 가요. 사람들 많지 않을 때로 골라서 와야하는 불편함이 있지만, 한적할 거 같은 시간에 또 오고 싶네요. 별적립 관련 문의했는데, 사소한 질문에도 여러 직원분들이 친절히 대해주시며 해결해 주었어요. 미소로 응대하는 모습들에 글을 남기게 되네요~^^
어느 스벅을 가도 똑같은 커피맛,매장 청결관리,친절함 스벅은 항상 평타이상은 쳐서 좋음.. 직원분이 추천 해주신 음료 처음 마셔 보는데 만족합니다. 거기서만 파는 '패션푸르트 칵테일' 음료 시원하고,달달하니 맛있네요. 가족들이랑 잘 마시고 갑니다~
스타벅스에서 볼 수 있는 모든것
조음 사람좀 빠지니까
평일아침에 방문. 일찍부터 사람들이 꽤 있으나, 주말처럼 미친듯이 많지는 않다. 통유리창의 뷰가 좋으며 꽤나 넓다.
좋아요
아주 좋았습니다아아아
사람없고 좋음
그냥 스벅임 아침에 오니까 사람없고 좋네, 빵도 맛남
난웃긴게 뭐만 했다하면 개떼들처럼 너도나도 로보트 처럼 하려는 미개한 국민성에 더 소름돋는다 지금 천지로 널린게 스타벅스 매장인데 굳이 한시간주차 한시간 웨이팅 기다려가면서 까지 커피를 마시려는 이유가 뭐임?거기 커피는 금가루라고 뿌려놨음?오히려 더 맛없다는데?용진씨도 이런 미개한 국민성을 마케팅으로 삼아 장사하고 있는거지

3 업무 현황 및 계획 I. 리뷰 크롤링

(2) 향후 연구 진행 방향(계획)

상권 내 점포 URL list-up

- ✓ Google, Naver, Daum 지도 API를 사용하여 자동으로 URL을 list-up 하는 코드를 작성
- ✓ 만약, 위 방법이 안될 경우 직접 점포별 URL을 검색하여 list-up
- ✓ Txt 파일 형식을 이용하여 list-up하고, 추후 Python으로 불러와 반복문을 수행

타 사이트에서 크롤링 시도

- ✓ 네이버, 다음 지도 리뷰 이외에도 유의미한 데이터 군집이 있는 사이트를 찾아 크롤링을 시도할 예정
- ✓ 페이지 구조가 다를 가능성이 높으므로, 코드를 변형해야할 가능성 높음.

자연어 처리 코드 짜기

- ✓ 수집한 데이터에 대한 분석 진행
- ✓ koNLPy 패키지의 twitter 모듈을 이용하여 형태소 분석
- ✓ 의미 있는 단어들을 도출하여 상권 이미지 분석 및 시각화
- ✓ 결과 저장 format은 이후에 나올 매출 데이터 format에 따라 결정

3 업무 현황 및 계획 II. 상권 구획

(1) 현재 상황은 어떠한가?

(1)유동인구 (2)상가 클러스터링 중 유동인구 클러스터링 성공

```
df = floating[['X', 'Y', 'f_sum']]
df.head()
```

버스정류장 좌표별(X, Y)
승하차 인원수 합계(f_sum)

	X	Y	f_sum
0	126.987750	37.569765	24294
1	126.996566	37.579183	100495
2	126.998340	37.582671	143669
3	126.987613	37.568579	34281
4	127.001744	37.586243	96051

```
#데이터 표준화_standardscaler
from sklearn.preprocessing import StandardScaler
standardScaler = StandardScaler()
print(standardScaler.fit(df))
df_scaled = standardScaler.transform(df)
```

```
StandardScaler(copy=True, with_mean=True, with_std=True)
```

```
df_scaled
```

```
array([[ 0.03186667,  0.3281433 ,  0.11380605],
       [ 0.13585299,  0.49769442,  2.48739752],
       [ 0.15677858,  0.5604919 ,  3.83222815],
       ...,
       [ 2.22633163,  0.09959713,  0.09686095],
       [ 2.1698711 ,  0.45054865, -0.57324933],
       [ 1.69259514,  0.03205169, -0.50774281]])
```

(1) 유동인구 데이터 전처리

(2) K-means clustering

```
[13] # 비계층적 군집 분석 k means model
     model = KMeans(n_clusters=10, random_state=0, algorithm='auto')
     # random_state=0 : seed 역할 (모델을 일정하게 생성 = 랜덤X)
     model.fit(df_scaled)
```

```
KMeans(algorithm='auto', copy_x=True, init='k-means++', max_iter=300,
       n_clusters=10, n_init=10, n_jobs=None, precompute_distances='auto',
       random_state=0, tol=0.0001, verbose=0)
```

```
[14] # 클러스터링(군집) 결과
     pred = model.predict(df_scaled)
     pred
```

```
array([3, 5, 5, ..., 0, 0, 0], dtype=int32)
```

3 업무 현황 및 계획 II. 상권 구획

(1) 현재 상황은 어떠한가?

(1)유동인구 (2)상가 클러스터링 중 유동인구 클러스터링 성공

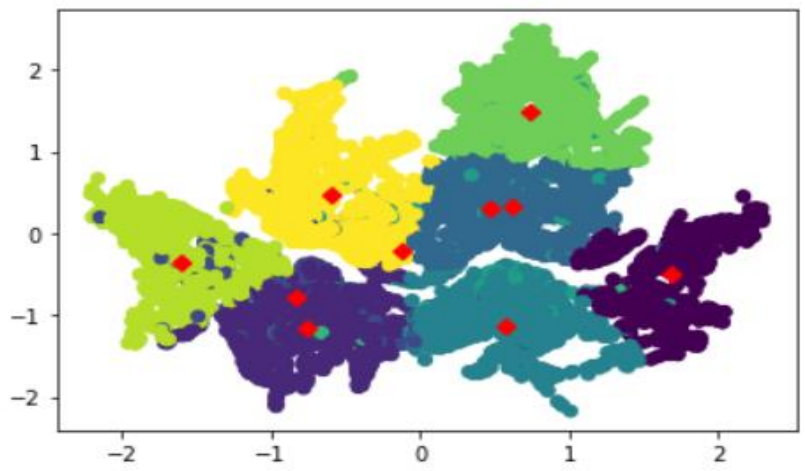
(3) 각 군집별 중앙값 계산

```
[15] # 군집별 중앙값
     centers = model.cluster_centers_
     centers
```

```
array([[ 1.68866526, -0.49599839, -0.15941029],
       [-0.75575234, -1.16069633, -0.2853896 ],
       [-0.83443187, -0.77439955,  1.7041186 ],
       [ 0.62580489,  0.32570036, -0.2236042 ],
       [ 0.57832853, -1.12766115, -0.1395578 ],
       [ 0.46398615,  0.30902592,  2.33713876],
       [-0.12387005, -0.20499211,  6.65993967],
       [ 0.74410048,  1.49243021, -0.27679294],
       [-1.59609965, -0.36106795, -0.303475  ],
       [-0.58897348,  0.48100911, -0.25741537]])
```

(4) 중앙값과 함께 시각화

```
[16] # 중앙값과 함께 시각화
     plt.scatter(x=df_scaled['X'], y=df_scaled['Y'], c=pred)
     plt.scatter(x=centers[:,0], y=centers[:,1], marker='D', c='r')
     plt.show()
```



3 업무 현황 및 계획 II. 상권 구획

(2) 향후 연구 진행 방향(계획)

유동인구 클러스터링 구체화

- ✓ 3차원 시각화 방법 고안
- ✓ Heuristic한 방법으로 적절한 군집의 수(k) 모색
- ✓ 클러스터링 결과 생성된 각 군집에 대한 위도&경도 범위 파악
- ✓ 최대&최소 위도와 최대&최소 경도로 클러스터 규정 및 csv 생성

=> 추후 상가업소 클러스터링과의 교집합을 도출하는 데에 이용

상가업소 클러스터링 진행

- ✓ 상가업소의 매출 데이터 혹은 대안 데이터 모색
- ✓ 상가업소 좌표(위도, 경도)와 매출 데이터(혹은 대안 데이터)로 클러스터링
- ✓ 생성된 각 군집의 위도&경도 범위 파악
- ✓ 3차원 시각화 및 적절한 군집 수 모색

=> 추후 유동인구 클러스터링과의 교집합을 도출하는 데에 이용

클러스터 교집합 도출

- ✓ 유동인구 클러스터링 결과와 상가업소 클러스터링 결과를 취합해, 중첩되는 지역 좌표의 최대/최소 값 계산
- ✓ 위 결과 도출된 좌표의 범위에 의해 생성된 지역을 “상권”으로 정의
- ✓ 정의된 모든 상권의 리스트를 csv형태로 저장
- ✓ 상권 리스트 csv를 대상으로 지도 위에 2차원(위도, 경도) 시각화

3 업무 현황 및 계획 III. 점포별 매출 데이터

(1) 현재 상황은 어떠한가?

인터넷에 검색해보거나,
안나오면 본사에 정보공유
요청 하면 되겠지 ?

뭐여 왜 공개
안해줌?

3 업무 현황 및 계획 III. 점포별 매출 데이터

(1) 현재 상황은 어떠한가?

주력 대안

서울시내 BC카드 사용 내역 정보 데이터(커피전문점 한정)

- ✓ 서울시 강남구, 서초구, 송파구, 마포구, 용산구의 카페업종 매출 집계 데이터
- ✓ 가맹점 지역(우편번호)에 따라 분류한 카드 사용 내역 포함 (but 개인정보보호법!)
- ✓ 데이터 집계 기간: 1개월 / 일 단위로 제공
- ✓ 주최 측에 데이터 구매를 신청한 상황

보조 데이터

권리금 또는 임대료

- ✓ 영업 중인 카페가 적어도 권리금 및 임대료를 감당할 수 있을 만큼은 매출을 내고 있다는 가정
- ✓ 온라인으로 조사해보았으나 부족한 점이 많아, 해당 구의 부동산에 방문하여 추가 자료 수집 시도함
- ✓ 점포별 매출을 지역적 정보와 연관 지어 볼 수 있는 보완적 데이터

3 업무 현황 및 계획 III. 점포별 매출 데이터

(1) 현재 상황은 어떠한가?

부동산 방문 면담 결과: 대안적 지표로서의 권리금 및 임대료의 문제점

권리금의 문제점

- ✓ 권리금은 매출보다는 상가에 대한 수요, 즉 지역 자체의 부동산 가격에 영향을 받음
- ✓ 카페 폐점 시 권리금을 받고 나가는 구조라서 권리금은 쉽게 공개되지 않는 데이터
- ✓ 점포가 권리금을 받지 않고 나가는 경우도 많음

  → 부동산은 추정치만 제시 가능

임대료의 문제점

- ✓ 동일 지역구 내에서도 상세 주소, 대로변과의 거리 등에 따라 임대료 차이가 큼 (ex. 강남구 대치동의 한티역 주변 / 대치역 주변)
- ✓ 동일 지역 내에서도 준공 시기, 건물 내 점포의 위치(층수, 가장자리에 있는지 등)에 따라서도 차이가 큼
- ✓ 지역별 임대료를 기준으로 지역의 가치를 추정하기 어려움

3 업무 현황 및 계획 III. 점포별 매출 데이터

(2) 향후 연구 진행 방향(계획)

현 시점, 가장 고민이 많은 지점

주제를 지속할 것인가?

- ✓ 우편번호 단위까지 세분화 되어있는 데이터를 본부 측에 구매요청한 상황
- ✓ 정상적으로 수령하게 되면, 동일한 우편번호 구역 내에서 개별 지점을 특정할 수 있는 방법에 대해 조금 더 고민해볼 예정(11월~)
  -> 권리금/임대료 이외에 또 다른 유효한 보조 데이터의 존재 여부 지속 탐색
  -> 김상용 교수님과의 면담 재요청 예정

주제를 변경할 것인가?

- ✓ 주제 변경의 가능성 또한 이제는 배제하기 어려움
- ✓ 기보유한 데이터셋을 최대로 활용할 수 있는 주제가 무엇이 있을지, 원점에서부터의 재검토 실시
  -> 필요 시 유시진 교수님과의 면담 재요청
- ✓ 최종 주제 변경 여부는 11월 12일까지 결정을 목표로 함

4 향후 일자별 진행목표

목표 달성을 위한 일자별 세부 계획

상가 목록 정리 (~11/26)

- ✓ 상가 목록과 상가별 리뷰 크롤링 사이트들의 list 작성을 완료하여, 12월 말부터 있을 본격적인 크롤링 작업의 기반 마련
- ✓ 이후 시험기간 관계로 활동 일시중지

상권 정의 (~11/19)

- ✓ 우선은 (버스 기준) 유동인구를 기준으로 한 지역 클러스터링과 상점 개수를 기준으로 한 지역 클러스터링의 교집합을 상권으로 정의
- ✓ 관련해 사회학과 신은경 교수님과 면담 예정

주제 확정 (~11/12)

- ✓ 11월 초 중 경영학과 김병조 교수님과 면담 예정
- ✓ 11/11 김상용 지도교수님과 면담 예정

=> 조언을 토대로, 데이터 부족의 해소 가능 여부 판단 후 주제 최종 확정

4 향후 일자별 진행목표

목표 달성을 위한 일자별 세부 계획

~12/23

2차 보고 자료 완성

보고서, 발표 영상 등
행정 업무 처리 완료

~ 12/30

Review 크롤링 완료

NLP 작업을 위한 Raw
Dataset 확보 완료

~ 1/6

NLP Preprocessing

모형 구축을 위한 리뷰
토큰화 완료

~ 1/13

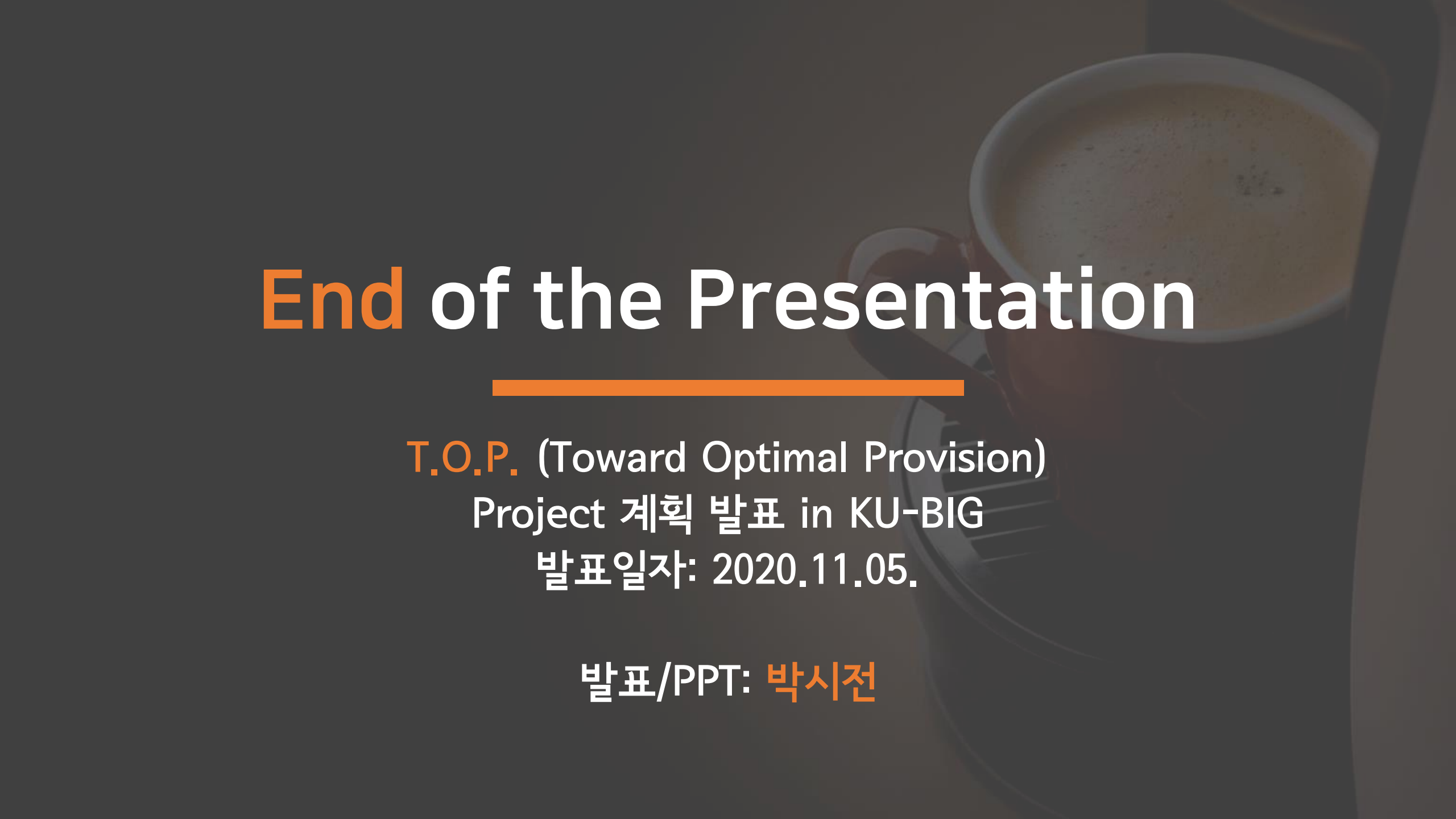
결론 도출 완료

토큰화된 데이터와 재무 데이터
등을 바탕으로 결론 도출 완료

~ 1/18

최종 발표

1/20일 있을 최종 발표를
위한 행정 업무 처리 완료



End of the Presentation

T.O.P. (Toward Optimal Provision)
Project 계획 발표 in KU-BIG
발표일자: 2020.11.05.

발표/PPT: 박시전