

Portfolio

1. 앱 인기도 요인 분석
2. 관광객 유형별 맞춤 추천 시스템
3. FIFA 선수 추천 시스템
4. Autoencoder(논문 분석)
5. Stereotype Bias(논문 분석)
6. 제품 불량률 예측 모형 개발
7. CTR Prediction 모형 개발
8. EC21R&C internship
9. ECG 심전도 데이터를 활용한 LHV 예측 모델
10. 데이터 기반 코로나19 예측 공모전(우수상)

01. 앱 인기도 요인 분석

성균관대학교 통계 학회 P-SAT 프로젝트
2018.04 ~ 2018.06

Why

어플리케이션 시장의 꾸준한 성장
해당 시장에서의 수익성에 대한 기대가 상승
그러나 대부분의 수익이 상위권에 해당하는 소수에게 한정

Objective

상위 어플의 성공 요인분석
- 어플리케이션 시장 진출을 위한 인사이트 제공

Data

미국 애플 앱스토어 상위 250개 어플 데이터(Kaggle)



김정민 양지연 이재일 임소영 한대룡



Algorithm & Method

Linear regression, Multinomial Regression을 이용해 상위 어플의 성공 요인 해석
Text Mining을 이용해 앱의 Description 변수 정제
계층적 군집 분석을 이용해 키워드들로 군집을 나눈 후 각 군집의 의미 해석
Linear regression 모델에 군집 변수를 추가하여 최종적으로 성공 요인 해석

주제 선정 배경 데이터 전처리 다중선행회귀 로지스틱회귀 다항로지

결과해석

Book보다 Game일 때
→ 50위 안에 들 오즈 0.015배
→ Book 보다 Game일 때 50위 안에 들 확률 ↓

Seller가 0보다 1일 때
→ 50위 안에 들 오즈 1.668배
→ Seller 0보다 1일 때 50위 안에 들 확률 ↑

RatingsAllVersions이 한단위 증가할 때
→ 50위 안에 들 오즈 2.351배
→ Rating 수 증가할 때 50위 안에 들 확률 ↑

주제 선정 배경 데이터 전처리 다중선행회귀 로지스틱회귀 다항로지

해석하기 : 그룹1 대비 그룹4에 속할 추정오즈 (가격)

	Price	Price
grp2	TRUE	0.9665711
grp3	TRUE	0.9465647
grp4	TRUE	0.9125053

* 다른 변수가 고정되어 있을 때,
가격이 한 단위 증가할 때
그룹1이 아니라 그룹4에 속할 오즈가 0.91배 높다.
→ 가격이 올라갈수록 그룹1에 속할 확률이 높다

복습 & 피드백 텍스트마이닝 데이터 분석 결과 해석 결론

텍스트마이닝 5. 시각화: Wordcloud / 연관어 탐색

```

> # Find Associated Terms ##
> findAssoc(dtm_s_app, ctrl, "million", .25)
$term
download facebook popular
0.31 0.25 0.20

> #findAssoc(dtm_s_app, ctrl, "battle", .25)
$term
card monster power down team
quest fight fantasy
0.36 0.30 0.29 0.26 0.25 0.24
0.24 0.23 0.22
  
```

모든 Description에 대해서
단어들 간의 상관관계 파악

복습 & 피드백 텍스트마이닝 데이터 분석 결과 해석 결론

클러스터링 (계층적 군집분석)

```

## Hierarchical Clustering ##
dtm_s_app_cluster <- as.matrix(dtm_s_app_sparse)
# calculate the distance between terms
distance <- dist(dtm_s_app_cluster, method="euclidean")
# distance
App_JC <- hclust(distance, method = "ward.D")
plot(App_JC)

# Draw dendrogram with red borders around the 3 clusters
rect.hclust(App_JC, k=3, border="red")
App_JC_cut <- cutree(App_JC, k=3) # cut tree into 3 clusters
App_JC_cut
  
```

군집1 : battle, addict, creat
군집2 : adventure, join, feature, experi
군집3 : download, million

복습 & 피드백 텍스트마이닝 데이터 분석 결과 해석 결론

클러스터링 의미 파악

	Cluster1	Cluster2	Cluster3
N	3	4	2
Terms	battle addict create	adventure join feature experience	download million
Example	"Battle legendary monsters and lead your kingdom to victory! Join your friends to create an army and defeat epic bosses."	"Join millions of players worldwide in this exciting city-building adventure game."	"#1 FREE GAME in countries on the AppStore with over 3 Million Downloads!"
Label	Battle & create	adventure	worldwide

복습 & 피드백 텍스트마이닝 데이터 분석 결과 해석 결론

클러스터링 기존 데이터에 결합
: Cluster Score Matrix를 app_game 데이터와 결합

```

## Add a Score matrix to the original data ##
app_game <- (bind(app_game, score))
str(app_game)

> str(app_game)
'data.frame': 299 obs. of 16 variables:
 $ Rank      : int 1 2 3 4 5 6 7 8 9 10 ...
 $ Price     : num 0 0 0 0 0 0 0 0 0 0 ...
 $ Seller    : factor w/ 2 levels "0","1": 2 2 2 1 2 2 2 2 ...
 $ Size      : num 3.79 3.57 1.86 3.6 3.49 ...
 $ StarRating : num 4.5 4.4 4.5 4.4 4.5 4.4 4.5 4 ...
 $ RatingAllVersions : num 11.98 11.48 8.39 11.49 9.37 ...
 $ StarCurrentVersion : num 4.5 4.4 4.5 4.5 4.5 4.5 4.4 4.5 ...
 $ RatingCurrentVersion : num 9.08 10.08 7.77 8.81 8.4 ...
 $ Free      : factor w/ 2 levels "0","1": 2 2 2 2 2 1 2 2 2 ...
 $ releaseQuarter : int 7 4 5 9 12 10 12 19 7 10 ...
 $ monthDiff : int 0 0 2 0 1 0 0 1 1 ...
 $ Cluster1  : num 1 1 3 0 1 0 0 0 2 ...
 $ Cluster2  : num 0 3 2 1 1 0 0 0 2 ...
 $ Cluster3  : num 0 0 0 0 0 0 0 0 0 ...
  
```

Result

기존에 데이터에서 주어진 변수들을 다양하게 전처리 하여 새로운 변수 창출
Description 변수를 이용해 어플의 기능과 특징을 고려하는 설명변수 추가
→ 앱의 성공 요인에 영향을 미칠 수 있는 다양한 원인을 고려하여 모델 구상 후 해석

Conclusion

전망 있는 어플리케이션 시장 성공 요인에 대한 인사이트 도출
다양한 전처리를 통해 정보 손실을 최소화할 수 있도록 노력
비정형 텍스트 데이터를 적절히 처리해 모델링 변수로 추가

02. 관광객 유형별 맞춤 추천 시스템

성균관대학교 통계 학회 P-SAT 프로젝트 / Co-Deep Learning
2018.10 ~ 2018.12

Why

한국을 방문하는 외래 관광객 증가
현재 존재하는 여행 추천 시스템들의 구조적인 문제

Objective

여행객의 최고 만족도를 위한 추천 시스템 개발

- 유형별 만족도 요인 분석
- 여행지 추천
- 예산 예측

Data

2017년 외래 관광객 실태조사 데이터 (문화 체육 관광부 / 한국 관광 공사)



Github URL:

<https://github.com/JiYeon9610/-Tourist-recommendation-system>

Algorithm & Method

PAM Clustering을 통한 관광객 유형별 군집화

Multinomial logistic regression을 이용한 군집 별 만족도 요인 분석

네트워크 분석을 이용한 군집 별 여행지 추천

Random Forest, XG boost를 이용한 여행 예산 예측

Anova test, K - Fold cross validation, Grid search를 통한 파라미터 튜닝

Result

관광객 유형을 군집화 / 군집 별 만족도 요인 분석을 통한 제언 /

군집 별 적정 여행지 추천

여행 예산 최종 Accuracy : 53.45%

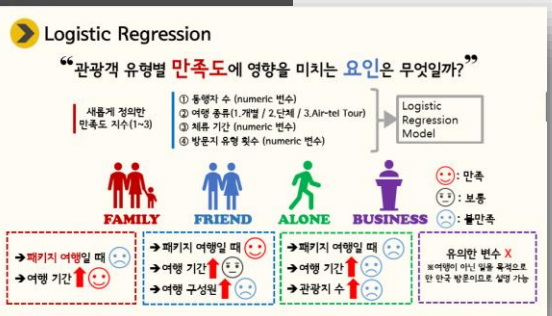
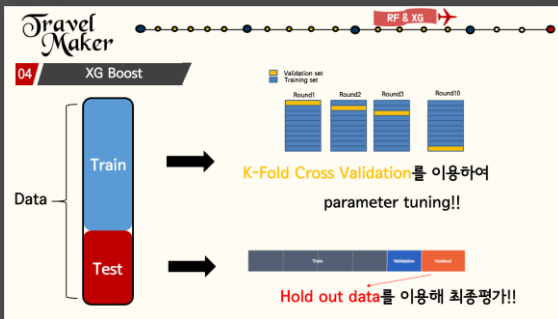
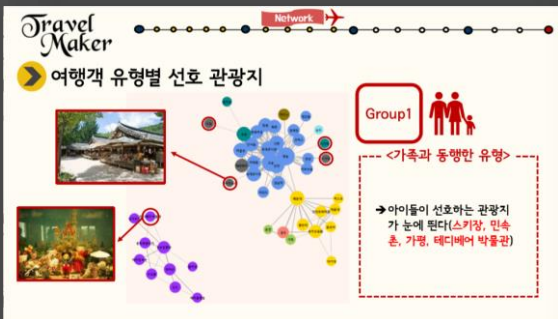
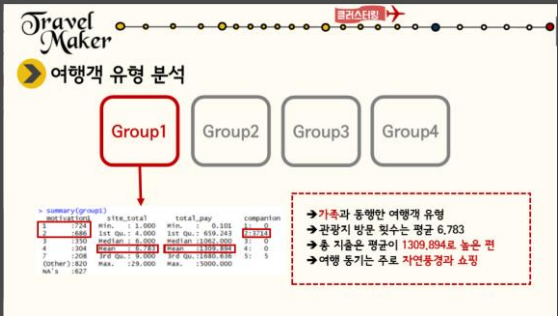
(4개의 카테고리 분류임을 고려하였을 때 나쁘지 않은 결과)

Conclusion

여러 모델을 다양하게 사용한 분석

만족도를 최대화 시키는 방향으로 여행의 전반적인 부분을 제시해주는 알고리즘 개발

어플리케이션으로 구현 가능성



03.FIFA 선수 추천 시스템

성균관대학교 '데이터사이언스와 Python' 기말 팀 프로젝트
2019.05 ~ 2019.06

Why

게임 시장의 지속적인 규모 확장

지속적인 수익을 내기 위해서 새로운 유저들의 유입이 필수적

하지만 진입장벽으로 유입 어려움이 발생

Objective

새로운 유저들을 위해 포지션별 선수들의 대략적인 능력치 인사이트 제공
개인별로 선호하는 능력치를 고려해 적절한 선수를 추천해주는 알고리즘

Data

FIFA19 게임 선수 데이터(Kaggle에서 수집)



Algorithm & Method

데이터 전처리를 통해 세부적으로 나뉘어져 있는 변수들의 범주를 알기 쉽게 압축하여 분류

- (포지션 변수(범주 27개) → 4가지 분류, 능력치 변수(범주 30개) → 6가지 분류)

다양한 시각화를 통해 능력치별, 포지션별로 발견할 수 있는 인사이트를 핵심적으로 요약
개인의 선수 능력치 선호도를 반영하여 선수를 추천해주는 프로그램 알고리즘 작성

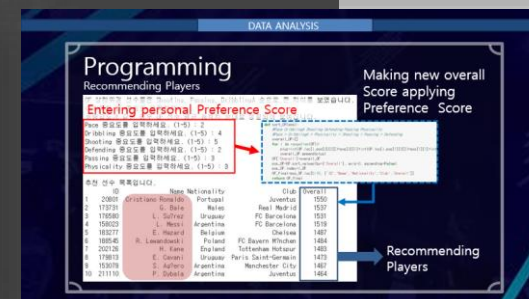
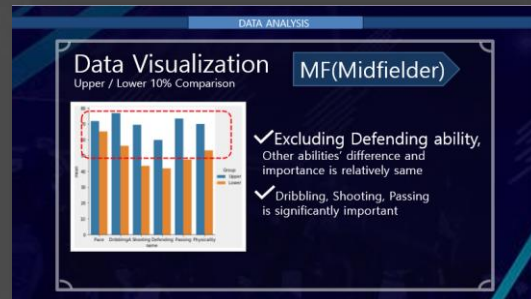
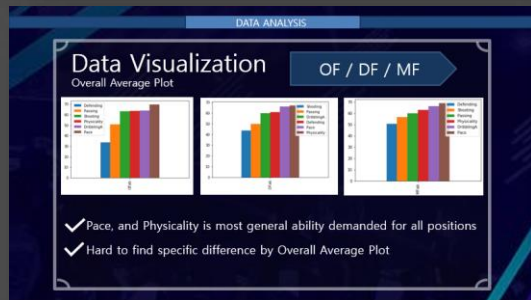
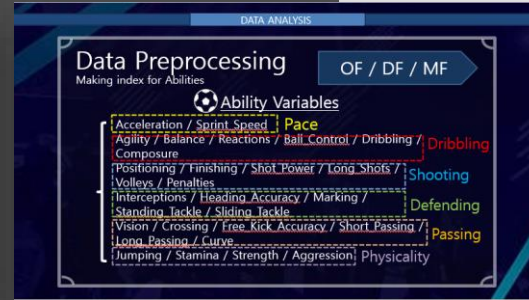
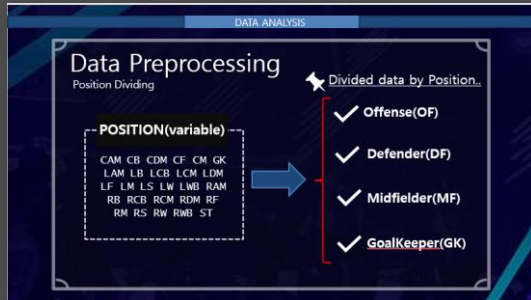
Result

개인별 선수 추천 프로그램 작성

- 포지션 선택지를 입력하면 포지션별 일반적인 능력치 가이드라인 제시
- 사용자의 선호 능력치를 고려해 가중치 점수를 매겨 적절한 선수 추천

Conclusion

FIFA를 플레이해보지 않은 유저들도 쉽게 접근할 수 있도록 포괄적인 인사이트 제공
개인의 개별적인 취향을 반영한 선수 추천 가능



04.AutoEncoder 논문 분석

성균관대학교 '응용머신러닝' 중간 개인 프로젝트
2019.06 ~ 2019.07

Why

비선형적 데이터의 차원 축소에 대해 고민
딥러닝 AI를 적용한 차원 축소 모델 탐구

What I learned

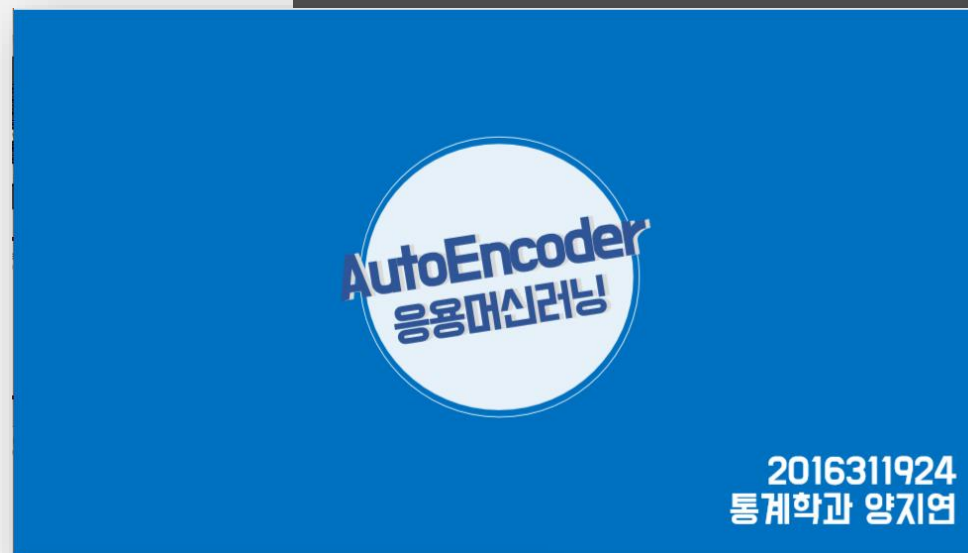
차원 축소의 목적과 필요성에 대해 공부
AutoEncoder 모델의 실제 application 사례 (음성 데이터)
딥러닝 모델의 근본적인 한계점 함께 탐구

Conclusion

비선형 데이터에 적용 가능한 차원 축소 모델
다양하게 파생되어 발전한 AutoEncoder 모델에 대한 근본적인 이해

Github URL:

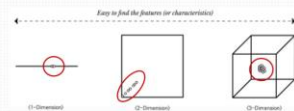
https://github.com/JiYeon9610/Autoencoder_study



01 AutoEncoder 기본 구성

차원 축소를 하는 이유: 차원의 저주 (Curse of Dimensionality)

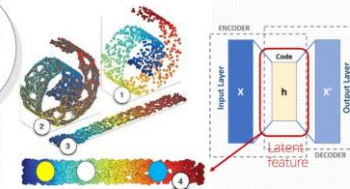
- 고차원 공간의 데이터를 분석하거나 측정할 때 저차원 공간에서는 나타나지 않았던 여러 문제들이 발생하는 것
- 머신 러닝에서는 일반적으로 차원이 증가할 경우 기하급수적인 데이터가 요구되는 현상



01 AutoEncoder 기본 구성

Keyword

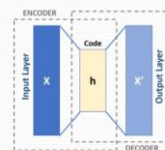
- Unsupervised Learning (비지도 학습)
- Manifold Learning (다원 축소)
- Generative model (생성 모델)



02 구체적인 모델 및 활용 방식

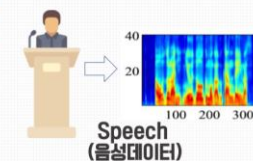
AutoEncoder 모델 종류

- Sparse AutoEncoder
- Denoising AutoEncoder
- Contractive AutoEncoder
- Variational AutoEncoder



03 Application

Speech Enhancement Based on Deep Denoising Autoencoder (DDAE)



Speech Enhancement 3가지 목표

- Noise Reduction (소음 축소)
- Speech Distortion (연설 왜곡)
- Perceptual Evaluation of Speech Quality (음성 품질 평가)

05. Stereotype Bias 논문 분석

성균관대학교 '딥러닝 입문' 기말 팀 프로젝트
2019.11 ~ 2019.12

Summarize

텍스트 데이터에 내제되어 있을 수 있는 Stereotype Bias에 대한 논문
사람이 작성한 데이터이기 때문에 Stereotype이 문서에 반영
주관이 개입되어 있는 데이터를 사용할 시 분석의 편향성 우려
Bias를 배제할 수 있는 다양한 해결책 필요

Discussion

데이터 안에 내제된 Stereotype 자체에 분석 의미를 부여
시대별 데이터를 이용하여 Stereotype Bias가
시간 흐름에 따라 어떻게 변화하는지를 파악하는 분석 방향 제안

arXiv:1605.06083v1 [cs.CL] 19 May 2016

Stereotyping and Bias in the Flickr30K Dataset

Emiel van Miltenburg
Vrije Universiteit Amsterdam
emiel.van.miltenburg@vu.nl

Abstract

An unstated assumption behind the crowdsourced descriptions of the images in the Flickr30K dataset (Young et al., 2014) is that they “focus only on the information that can be obtained from the image alone” (Hodosh et al., 2013, p. 859). This paper presents some evidence against this assumption, and provides a list of biases and unwarranted inferences that can be found in the Flickr30K dataset. Finally, it considers methods to find examples of these, and discusses how we should deal with stereotype-driven descriptions in future applications.

Keywords: image annotation, stereotypes, bias, Flickr30K

1. Introduction

The Flickr30K dataset (Young et al., 2014) is a collection of over 30,000 images with 5 crowdsourced descriptions each. It is commonly used to train and evaluate neural network models that generate image descriptions (e.g. (Vinyals et al., 2015)). An unstated assumption behind the dataset is that the descriptions are based on the images, and nothing else. Here are the authors (about the Flickr30K dataset, a subset of Flickr):

“By asking people to describe the people, objects, scenes and activities that are shown in a picture without giving them any further information about the context in which the picture was taken, we were able to obtain conceptual descriptions that focus only on the information that can be obtained from the image alone.” (Hodosh et al., 2013, p. 859)

What this assumption overlooks is the amount of *interpretation* or *recontextualization* carried out by the annotators. Let us take a concrete example. Figure 1 shows an image from the Flickr30K dataset.




Figure 1: Image 8061007 from the Flickr30K dataset.

This image comes with the five descriptions below. All but the first one contain information that cannot come from the image alone. Relevant parts are highlighted in bold:

1. A blond girl and a blond man with his arms crossed are standing inside looking at each other.
2. A worker is being scolded by her boss in a stern lecture.
3. A manager talks to an employee about job performance.
4. A hot, blond girl getting criticized by her boss.
5. Sonic employees talking about work.

2. Stereotype-driven descriptions

Stereotypes are ideas about how other (groups of) people commonly behave and what they are likely to do. These ideas guide the way we talk about the world. I distinguish two kinds of verbal behavior that result from stereotypes: (i) linguistic bias, and (ii) unwarranted inferences. The former is discussed in more detail by Beukeboom (2014), who defines linguistic bias as “a systematic asymmetry in word choice as a function of the social category to which the target belongs.” So this bias becomes visible through the *distribution* of terms used to describe entities in a particular

“The Flickr30K data also contains examples where annotators judge the subjects of the images on their looks. E.g. description #4 above calling the girl in the image hot. Analyzing this judgemental language goes beyond the scope of this paper.”

Bias Paper Introduction

➤ Stereotype-driven descriptions in future application

Collection of over 30,000 images with 5 crowdsourced descriptions each


Descriptions are also subjective, interpreted and recontextualized

Salient parts to the average annotator

Subjectivity result stereotypical descriptions.

➔ Language models trained on this data may propagate harmful stereotypes.

Flickr 30K dataset



1. A blond girl and a blond man with his arms crossed are standing inside looking at each other.
2. A worker is being scolded by her boss in a stern lecture.
3. A manager talks to an employee about job performance.
4. A hot, blond girl getting criticized by her boss.
5. Sonic employees talking about work.

Bias Paper Method

➤ Verbal behavior that result from stereotypes



Additional assumptions about the world

“Linguistic bias” “Unwarranted inferences”

1. Activity → ‘manager is talking about job performance’
2. Ethnicity → ‘African-American’
3. Event → making assumptions on the event
4. Goal → explaining the why of the situation
5. Relation → ‘Parents/older people with children’
6. Status/occupation
→ relatively general (college student → student)
→ very specific (worker → graphics designer)

Bias Paper Method

➤ How to detect stereotype-driven descriptions?



Scolding and laughing: a toddler with fire engine airplane mobile and other toys on the floor

A smiling child sitting in a white baby bouncer surrounded by toys.

in an bouncy seat with toys surrounding him.

in an activity chair in a child's playroom.

sitting in a toy

A young woman in a blue and white checked shirt with a absent.

A young child eating a snack while wearing a checked shirt.

A small boy is concentrating on food in a dish.

A little boy is painting a plate.

A young boy eating food.

Bias Paper Discussion

‘How about doing research intended for deriving Stereotype?’

Stereotype → beneficial to interpret human description

Research on the purpose of stereotype itself

research having enough time term → indicator of change of perception

06. 제품 불량률 예측 모형 개발

2018 삼성 SDS 공모전
2018.07~ 2018.08

Why

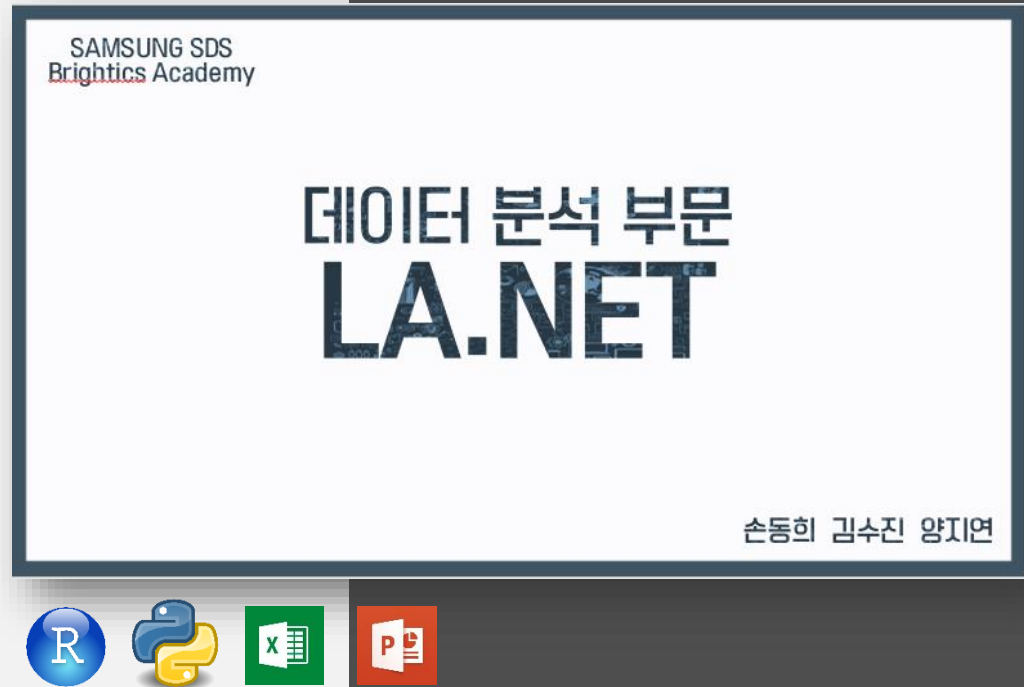
제품 공정에서 발생한 데이터를 이용하여 빅데이터 모형 개발
구축된 모델로 의사 결정을 하는 것이 시간 효율적
높은 정확도를 가진 모형 개발의 필요성

Objective

부가적인 설명이 주어지지 않은 86개의 연속형 설명 변수들
연속형 변수인 불량률 Y 를 정확하게 예측해내는 모형을 개발

Data

삼성 SDS 공모전 제공



Algorithm & Method

다양한 시각화를 통해 정보를 알 수 없는 변수들에 대한 인사이트 확보

SMOTE Oversampling을 통해 범주형으로 치환한 Y변수의 imbalance 해결

XGBoost Feature Selection을 이용해 중요한 X변수들 추출

Kernel PCA를 이용해 중요도가 비교적 낮은 변수들을 차원 축소해 데이터에 추가

Bayesian Optimization 이용해 XGBoost 최종 예측 모델 완성

Result

적절한 차원 축소로 86개 설명 변수 전처리

적절한 파라미터 조정을 통한 최적의 XGBoost 예측 모델 완성

Conclusion

다양한 시각화를 통해 변수 특성 파악

정보 손실을 최소화하여 압축하는 다양한 차원 축소 방법 사용

효율적인 파라미터 조정을 통해 모델 학습 시간 축소

목차

1. 제품 불량률 예측

1. 데이터 탐색
2. Oversampling
3. Data Partitioning
4. Feature Selection
5. Feature Extraction
6. Prediction
7. 한계 및 향후 개선 방향



1. 제품 불량률 예측

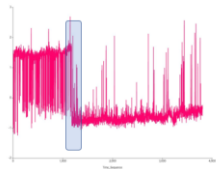
1-2. Oversampling

SMOTE를 이용한 oversampling

- ▶ 연속된 Y의 값을 50이상의 것들을 1로 변환하여 class를 나타내는 변수를 생성
- ▶ 적은 class(50이상)의 observation 값들의 k-nearest-neighbors를 추출
- ▶ 이를 통해 무작위로 유사한 값들을 뽑아내어 변수를 oversampling

1. 제품 불량률 예측

1-3. Data Partitioning

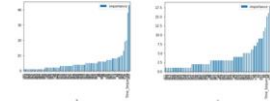


데이터 탐색 파트에서 설명했듯
1100을 기준으로 전 후 데이터에서
x 변수의 추세가 다른 경향을 보임

✓ 1100에서 y와 correlation이 있던 x 변수가
1100후의 y 값과는 correlation이 적어졌다 등등

1. 제품 불량률 예측

1-4. Feature Selection



1100을 기준으로 나눈 첫번째 train 데이터,
두번째 train 데이터에 대해 각각 최적의 xgboost 모델 선정
최적의 모델 선정을 위해 Bayesian Optimization을 사용
최적의 모델의 importance plot를 통해 변수 선별

1. 제품 불량률 예측

1-5. Feature Extraction

Importance plot으로 어느정도 변수를 걸렀지만
50개가 넘는 변수들의 처리가 필요

→ 삭제하기에는 너무 많은 정보를 버리게 되는 상황

▶ Kernel PCA를 사용해
나머지 변수들을 뽑아냄

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \xrightarrow{\text{feature extraction}} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix} = f \left(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \right)$$

1. 제품 불량률 예측

1-6. Prediction

XGBoost

- ✓ Regularization
- ✓ 빠른 속도
- ✓ 유연성
- ✓ NA 자동처리
- ✓ Post Pruning

Model의 residual에 새로운 트리를 fit
이 과정에서 Gradient descent를 이용해
loss function (RMSE, classification
error 등)을 최소화

07. CTR Prediction 모형 개발

2019 IGAWorks BIG DATA Competition
2019.12 ~ 2020.02

Why

오디언스 별 개인화 된 행동을 관찰하여 광고 클릭 여부를 예측
예측 결과를 반영하여 RTB 경매 입찰가 결정
빠르고 정확한 CTR Prediction 모형의 필요성

Objective

정확한 클릭 확률을 추정하는 모형을 개발
학습과 평가 과정에서 시간을 최소화

Data

IGAWorks 제공 데이터

Train.csv(학습기간 노출 로그 데이터)

Audience_profile.csv(오디언스 관련 정보 모음)



Github URL:

https://github.com/JiYeon9610/IGAWorks_competition

Algorithm & Method

Information gain을 고려하여 설명 변수의 범주를 줄이는 feature selection 진행
AutoEncoder 모델을 이용해 오디언스의 변수를 5개 컬럼으로 차원 축소
정제한 오디언스 데이터와 train 데이터를 합쳐 분석 데이터 완성
XDeepFM 모델의 파라미터 조정을 통해 최종 모형 결정

Result

최종 데이터를 이용해 광고를 클릭할 확률을 예측
최종 logloss: 0.24834

Conclusion

주어진 데이터의 특징에 최적화된 모델 사용하여 예측 시간 단축
고차원 범주형 데이터의 정보 손실을 최소화하며 차원 축소
다양한 파라미터 조정을 통해 보다 정교한 모델링 진행

01 분석 방향 및 시각화


데이터 특성

Train 데이터
각 bidding 별 정보.
예) 광고 타입, 광고주 아이디, 미디어 아이디, 기기 통신사, 광고 아이디, 광고사 아이디, 기기 연결망식 등

설명 변수 개수
총 22개

- media_name에는 총 6810개의 범주가 있다.
- 소매자 정보 유무에 따른 클러스터 차이가 크지 않다.

범주형 변수를 one-hot-encoding했을 때 데이터의 대다수가 0인 **Sparse 데이터**이다.



01 분석 방향 및 시각화

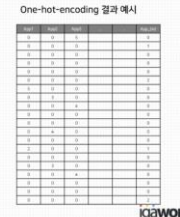
데이터 특성

Audience 데이터
각 소비자 별 정보.
예) 성별, 연령, 결혼 여부, 사용자가 앱에 부여한 범용

변수 개수
총 8개

cate_code 변수
사용한 앱과 범용에 따른 변수이다.
A 10이상의 사용에 대하여 1점을 부여한 것으로 점수와 범용을 분리한다.
이를 one-hot-encoding한 결과는 우측과 같다.

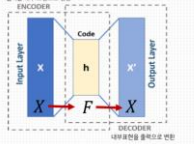
cate_code 변수를 one-hot-encoding했을 때 데이터의 대다수가 0인 **Sparse 데이터**이다.



01 분석 방향 및 시각화

모델링 - AutoEncoder

입력을 낮은 차원으로 변환
한정된 정보로 재구성



AutoEncoder란?
입력값과 출력값을 같게 만든다

장점

- 1) 비지도학습 + 지도학습
y를 설정하지 않고도 x를 예측할 수 있다 → **비지도학습**
예측한 x값을 실제 x값과 비교할 수 있다 → **지도학습**
- 2) data-specific이다
- 3) 입력 데이터에 대하여 자동적으로 학습

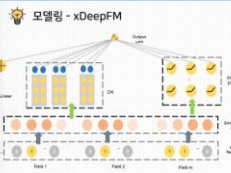
Why AutoEncoder?
속도를 줄이기 위해 데이터 차원을 줄이면서 동시에 정보 손실을 최소화 할 수 있는 방법이기 때문에 우리 데이터에 적합함

Algorithm

1. 입력값 x로부터 중요한 정보를 압축해 높은 차원 값
2. 차원을 이용해서 또 다시 x값을 예측
3. 원본 입력값 x와 비교

01 분석 방향 및 시각화

모델링 - xDeepFM



xDeepFM이란?
extreme Deep Factorization Machine
deepFM에서 개선된 모델로 명시적인 방식과 액티브한 feature interaction을 계산한다

장점

- 1) 범주형과 연속형을 다룰 수 있다
- 2) 다차원의 sparse data에 대해서도 feature interaction 계산이 가능하다

Why xDeepFM?
우리 데이터는 범주형과 연속형을 모두 가진 다차원의 sparse data이므로 기존의 FM, deepFM에 대해서도 성능이 높고 성능이 좋은 xdeepFM을 사용

02 전처리

Train/Test data - Feature Selection

변수	범주 개수(Train)	범주 개수(Test)
ad_id	17%	15%
campaign_id	166%	96%
advertiser_id	877%	231%
placement_type	4%	4%
media_id	5815%	5815%
media_domain	162%	7%
publisher_id	3939%	3939%
publisher_name	1222%	1222%
advertisement_id	30%	28%

아래의 변수를 제외하고 나머지는 feature selection을 통하여 범주의 개수를 줄였다

- media_id, publisher_id, publisher_name: 범주 개수가 너무 많으므로 불가능, 원시 데이터 사용
- placement type, device_cat: 이미 범주 개수 적으므로 할 필요 없음, 원시 데이터 사용

02 전처리

최종데이터

Audience data	Train/Test data
age gender marry cate_answer1-5 cate_like1-5	column별 feature selection hour(A1Z) day(요일)

NA 처리

- Audience data와 Train/Test data에 'device_fa' 변수를 사용하여 각각 merge함
- Audience data에는 있지만 Train/Test data에는 없는 cate_answer와 cate_like1-5의 NA는 원시 데이터에서 category에 대하여 모두 0을 찍은 device_fa와 오묘한 코드 결과 값으로 채워넣었다
- gender, marry, age의 경우 각각 'F', 'C', 'old'로 새로운 범주를 만들어 NA값을 채워주었다

03 모델링

고차원의 sparse 데이터에 강한 딥러닝 기반의 xDeepFM을 최종 모델로 선택


xDeepFM

- extreme DeepFM (DeepFM에서 기원함)
- 명시적인 방법, 벡터 단위로 피쳐 상호작용 산출
- 고차원 수준의 sparse feature를 학습 시킬 수 있다
- 범주형, 연속형 자료를 모두 다룰 수 있다

DeepFM

- Linear part, FM layer, Hidden layer(DNN)으로 구성
- 명시적인 방법, 벡터 단위로 피쳐 상호작용 산출

batch size



최종 모델

- xDeepFM은 DNN 기반의 모델
- batch size와 epoch를 튜닝하면서 최적의 파라미터 조합 찾음

03 모델링

batch_size	epoch	train_log_loss	validation_log_loss	test_log_loss
128	4	0.2445	0.2452	0.237
128	10	0.2424	0.2435	0.2361
128	13	0.2451	0.2494	0.2364
128	15	0.2421	0.2446	0.2367
256	10	0.2430	0.2499	0.2354
256	15	0.2380	0.2408	0.2347

Grid Search를 통하여 최적의 파라미터를 찾음
test_log_loss란 비교하면 batch size 256, epoch 10이 가장 좋으나, 오버피팅 되는 경향이 있어, 최종 모델은 batch size 128 epoch 10으로 설정

최종 모델

batch_size	epoch	리더보드
128	10	0.24834

08.EC21R&C internship

2020.01 ~ 2020.06

Projects



농맞춤_aT해외시장맞춤조사 프로젝트(2020.01~2020.04)

2020 미래예측 화장품 트렌드 분석(2020.04~2020.06)

CSF 중국전문가포럼 프로젝트_대외경제정책연구원(2020.01~2020.02) -보조

이머릭스 프로젝트_대외경제정책연구원(2020.01~2020.02) -보조

Business Role

R, Python 모델링 코드 작성 및 효율화를 위한 기존 코드 개선

Text Mining 관련 새로운 인사이트를 위한 논문 스터디

프로젝트 과정 기록 및 보고를 위한 매뉴얼 작성

Modeling

- TF/DF 빈도 분석, N-gram 빈도 분석
- LDA, CTM, DTM 토픽 모델링 / 토픽 트래킹 모델
- 키워드 상관 분석 / 연관분석 / 네트워크 분석
- LSTM, word2vec, 문장유사도 분석, 미래신호탐지기법

빅데이터 솔루션 토픽랭크(Topic Rank) 알고리즘

빅데이터 소개

뉴스빅데이터 분석 서비스, 빅데이터(BigKind): 빅데이터(BigKind)는 통합된 검색, 집계, 지역별 검색, 방송사 등을 포함한 최대 규모의 기사DB에 빅데이터 분석 기술을 접목한 전문 서비스로 뉴스 분석 서비스 누구나 무료로 이용할 수 있는 서비스이며, 초원가인 사 분석데이터 다운로드, 개인화 서비스 등 더 다양한 서비스 이용 가능

서비스 개념도

뉴스수집시스템, 분석시스템, 저장시스템 등으로 구성돼 있으며, 저장된 뉴스 분석 결과를 국민, 언론사, 학계, 스타트업 등이 활용할 수 있는 뉴스빅데이터 분석서비스 '빅데이터(BigKind)'로 제공.



지원내용

뉴스 속 인물 장소기반 개체명 간 관계도, 주요 인물의 발언 내용, 키워드 트렌드, 연관어, 공공데이터와 뉴스 통합 등 분석 서비스

제공되는 정보 분류

언론사 분류: 서울, 경기 등 7개 지역별 54개 언론사의 뉴스를 집계
유형분류: 총 93개 상세 분류 정보를 3단계로 제공
사건(사고) 분류: 총 58개 분류 정보를 3단계로 제공

TextRank를 이용한 문서요약

이러한 PageRank 알고리즘을 응용한 것이 바로 TextRank이다. TextRank는 PageRank의 용도가 넓은 웹 사이트는 다른 웹 사이트로 부터 링크를 받는다는 점에 착안하여 문서 내의 문장(or 단어)을 이용하여 문장의 Ranking을 계산하는 알고리즘이다.



TextRank 식은 아래와 같다.

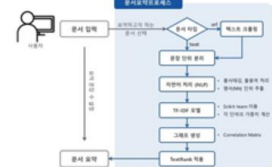
$$TR(V_i) = (1 - d) + d \sum_{j \in V} \frac{1}{|V_j|} \frac{1}{|V_i|} \frac{1}{|V_j|} \frac{1}{|V_i|} \frac{1}{|V_j|} \frac{1}{|V_i|}$$

TR(V_i) 문장 또는 단어(V_i)에 대한 TextRank 값

web: 웹상 단어(V_i) / |V_i| 단어 개수

d: damping factor, PageRank에서 웹 서핑을 하는 사람이 한달 페이지를 방문하지 못하고 리플라이자로 이동하는 확률이며, TextRank에서도 그 값을 그대로 사용(0.85로 설정)

TextRank TR(V_i)를 계산 한 뒤 높은 값을 선택함



20.04.10 미래예측 코드 작성 현황

별라스 작성

기존 미래예측 분석 과정

-Input: 분석하고자 하는 DB와 / MainWordList 데이터 / 미래예측 대상의 키워드 데이터
-Output: 키워드 키워드, TFIDF값, 총등(문장) / DF분석 키워드, 총등(문장)

-추가적인 Excel 데이터의 직접 과정
DOE/DOV 데이터 및 분석 결과에 대해
통계분석도 / 통계를 분석하여 분석 결과에 대해
미래예측 결과표를 위한 구조 및 분석 결과에 대해

-과정에 생기는 문제
1. 데이터를 분석 결과를 고려하여 검토한 것
2. MainWordList가 제대로 분석되지 않고 누락되어 분석이 진행되고 있는 것
3. 통계를 분석하여 분석 결과에 대해
4. 데이터 분석 결과에 대해
5. 데이터 분석 결과에 대해
6. 데이터 분석 결과에 대해
7. 데이터 분석 결과에 대해
8. 데이터 분석 결과에 대해
9. 데이터 분석 결과에 대해
10. 데이터 분석 결과에 대해

농맞춤_데이터 분석 매뉴얼(20.03.06)

* 추가적인 시간이 소요되는 경우

#1 Input 정보 정리 과정(0.5h)

분석	분석 방법
분석 목적	분석 목적
분석 대상	분석 대상
분석 방법	분석 방법
분석 결과	분석 결과

* 분석 목적: 분석 목적
* 분석 대상: 분석 대상
* 분석 방법: 분석 방법
* 분석 결과: 분석 결과

#2 데이터 준비 과정(1.5-2h)

과정	분석 방법
데이터 수집	데이터 수집
데이터 정리	데이터 정리
데이터 분석	데이터 분석
데이터 시각화	데이터 시각화

* 데이터 수집: 데이터 수집
* 데이터 정리: 데이터 정리
* 데이터 분석: 데이터 분석
* 데이터 시각화: 데이터 시각화

* 데이터 분석: 데이터 분석
* 데이터 시각화: 데이터 시각화
* 데이터 분석: 데이터 분석
* 데이터 시각화: 데이터 시각화

#3 데이터 분석 과정(1.5h)

과정	분석 방법
데이터 수집	데이터 수집
데이터 정리	데이터 정리
데이터 분석	데이터 분석
데이터 시각화	데이터 시각화

* 데이터 수집: 데이터 수집
* 데이터 정리: 데이터 정리
* 데이터 분석: 데이터 분석
* 데이터 시각화: 데이터 시각화

08.EC21R&C internship

Project-1

Projects

농맞춤_aT해외시장맞춤조사 프로젝트(2020.01~2020.04)

발주: 한국농수산물유통공사

Objective

중소기업의 해외 진출을 위해 관련 상품의 국가별 트렌드 조사, 보고서 작성

Business Role

기존의 분석 방식에서 개선된 새로운 모델 도입

R, Python 모델링 코드 작성 및 효율화를 위한 기존 코드 개선

Modeling

- TF/DF 빈도 분석, N-gram 빈도 분석
- LDA 토픽 모델링
- 키워드 상관 분석 / 연관분석 / 네트워크 분석

농맞춤_데이터 분석 매뉴얼(20.03.06)

x 주가적인 시간이 소요되는 경우

#1 Input 정보 정리 과정(0.5h)

과정	제출 사항
분석 목적 및 현황명 파악	소싱 제품 사항 / 조사 대상 제품 분석
크롤링 데이터 소스인 사이트 정리	필요에 따라 소스인 URL
데이터 정리	필요에 따라 데이터 구조를 소스인 데이터 정리

※ 소스인 사이트에 가입된 데이터 개수와 최종적으로 크롤링 된 데이터 개수 차이가 크면 웹마스터에게 문의 후 피드백 받기

#2 데이터 준비 과정(1.5-2h)

과정	제출 사항
분석데이터 정리	Excel 프로그램 이용
소스인 URL 확인	①노이즈 여부 ②특정 데이터로 된 소스인인지의 여부 확인
노이즈 제거	① 데이터 수가 적은 경우(150개 이하), 노이즈의 영향력이 크므로 노이즈 제거의 더 집중 ② 온라인 스톱워드 제거 - Stopword 제거, 일관성 있는 데이터는 제거 ③ URL, 이메일, 빈도분석 데이터 - 특수+대문자 제거, 빈도 분석 - < > 앞뒤의 < > 제거, 빈도 분석 - < > 앞뒤의 < > 제거, 빈도 분석

※소스인이 데이터를 수집하지 않은 경우 + 노이즈가 너무 많은 경우

①소스인 수집 담당자에게 피드백

②노이즈에 해당하는 데이터 손수 제거

※데이터가 아닌 언어의 경우 노이즈 판별이 되는 어려움

-이런 경우 검색어 포함 여부만 확인하여 일차적으로 노이즈 데이터 제거

#3 데이터 분석 과정(1.5h)

과정	제출 사항
온라인 스톱워드	(제출물) TF-DF 빈도 분석 / TF-IDF 빈도 분석
키워드 상관	(제출물) TF-DF 빈도 분석 / TF-IDF 빈도 분석
연관분석	(제출물) TF-DF 빈도 분석 / TF-IDF 빈도 분석
네트워크	(제출물) TF-DF 빈도 분석 / TF-IDF 빈도 분석
인용	(제출물) TF-DF 빈도 분석 / TF-IDF 빈도 분석



Project-2

Projects

2020 미래예측 화장품 트렌드 분석 (2020.04~2020.05)

발주: 대한화장품산업연구원

Objective

각 국가의 화장품 시장 데이터를 이용해 중요한 이슈 인사이트 도출, 보고서 작성

Business Role

Excel을 통해 추가적으로 데이터 셋을 만들던 과정을 코드를 이용해 압축하여 효율화

기존 코드의 오류 수정 및 추가적인 모델링을 위한 코드 작성

Modeling

- CTM, DTM 토픽 모델링
- 미래신호탐지기법
- 상관분석

20.04.10 미래예측 코드 작성 현황

델리스 작성

· 기존 미래예측 분석 과정

Parameter	2019				2017				2018			
	1	2	3	4	1	2	3	4	1	2	3	4
매출액												
영업이익												
영업외이익												
이익총계												
영업비용												
영업외비용												
비용총계												
이익률												
비용률												
이익률												
비용률												

Parameter	2019				2017				2018			
	1	2	3	4	1	2	3	4	1	2	3	4
매출액												
영업이익												
영업외이익												
이익총계												
영업비용												
영업외비용												
비용총계												
이익률												
비용률												
이익률												
비용률												

→ Input 분석하고자 하는 Data / M_ulti_word_set 대입(/ 미래예측 대입이 되는 키워드 대입)
 → Output 결과를 키워드 TFI(키워드 통상 함수) / DF(분할 키워드 통상 함수)

Parameter	2019				2017				2018			
	1	2	3	4	1	2	3	4	1	2	3	4
매출액												
영업이익												
영업외이익												
이익총계												
영업비용												
영업외비용												
비용총계												
이익률												
비용률												
이익률												
비용률												

Parameter	2019				2017				2018			
	1	2	3	4	1	2	3	4	1	2	3	4
매출액												
영업이익												
영업외이익												
이익총계												
영업비용												
영업외비용												
비용총계												
이익률												
비용률												
이익률												
비용률												

→ 추가적인 Excel 상에서의 작업 과정
 DQ1/DQV 대입(1 셋 우선 입력하여 도출
 평균근비도 / 평균준가를 우선 입력하여 도출
 미래전도 결과표를 위한 X축, Y축 우선 입력하여 도출

→ 과정에서 생기는 문제
 1. 결과를 분석 개수를 고려하지 않았던 점
 2. M_ulti_word가 제대로 인식되지 않고 누락되어 분석이 진행되고 있었던 점
 3. 평균준가들에서 불모가 0인 경우
 → 대입(1) 속의 수치를 모두 고려하였던 점 + 의미 있는 증가를 감지할 수 있는 부분 발생





Project-4

Projects

CSF 중국전문가포럼 프로젝트 (2020.02~2020.03)

발주: 대외경제정책연구원

Objective

중국의 경제 뉴스 데이터를 이용하여 카테고리별 핵심 뉴스 선정하여 전달

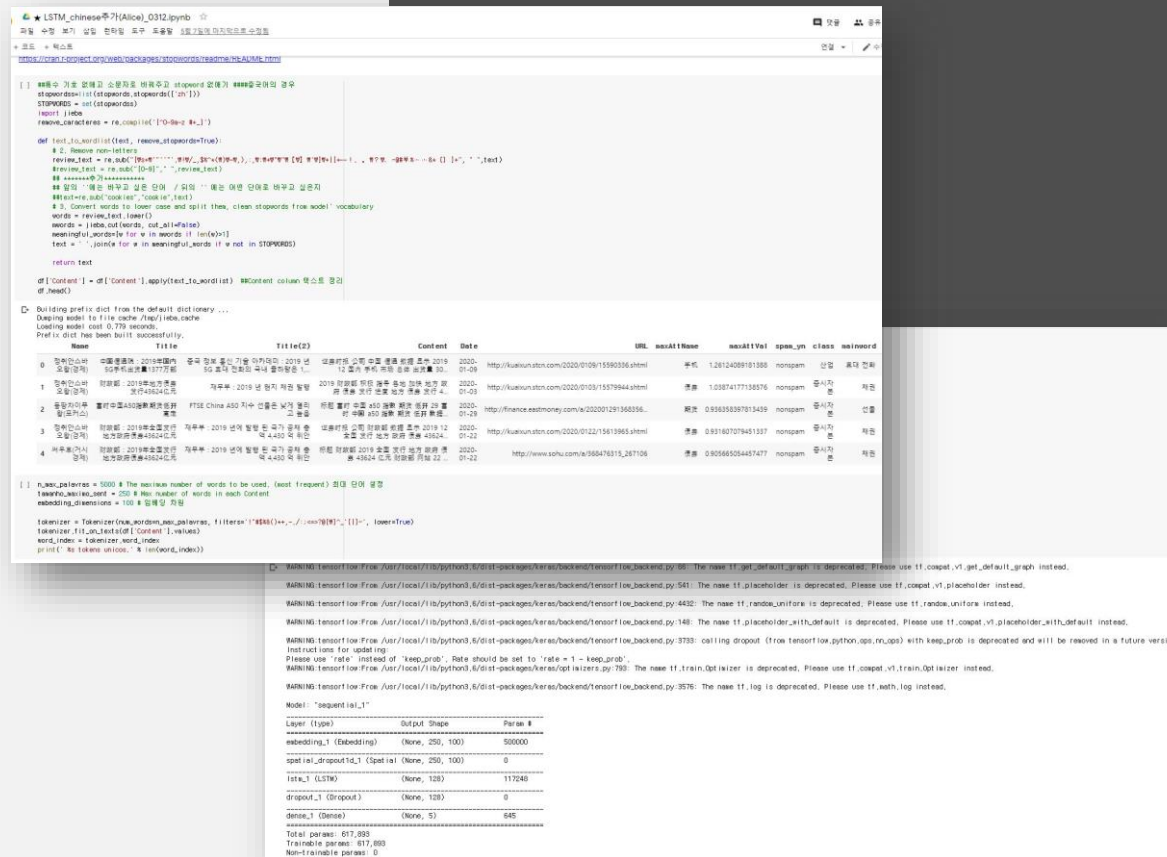
Business Role

중국어 텍스트 정제 처리 코드 작성

LSTM 모델 예측력 개선을 위한 파라미터 조정 및 코드 개선

Modeling

- LSTM 딥러닝 모델



09. ECG 심전도 데이터를 활용한 LHV 예측 모델

성균관대학교 '딥러닝 강의' 기말 팀 프로젝트
2020.11 ~ 2020.12

Why

좌심실 비대증(LHV)은 심혈관 질환 발생 예측의 중요한 전조 질환
심전도 검사의 주된 목적 중 하나
심장 초음파 검사로 나아가기 이전, 일차 의료에서 행해질 수 있는 심전도 검사로부터
LHV를 진단한다면 한 발 더 빠른 의학적 조치를 가능하게 함.

Objective

심전도 데이터를 입력 받아 좌심실 비대의 유무를
높은 성능으로 판별하는 모델 구축

Data

ECG 심전도 데이터(200 obs.)

문제 정의 좌심실 비대증 : 증상 발견의 중요성

좌심실 비대와 심전도



더 정확한 '심장 초음파검사'



BUT 높은 검사비용
일차 의료에서의 어려움



Github URL:

https://github.com/JiYeon9610/Deep-Learning_final-project

- 데이터 구조 파악 및 label 별 특성 파악을 위한 다양한 시각화
- 세부적인 전처리 과정을 거쳐 압축된 데이터를 input으로 DNN Framework 활용
- 주기적인 파동을 지닌 신호 데이터의 특성을 고려하여 1D-CNN Framework 활용
- 과적합 방지를 위해 Dropout, EarlyStopping, Regularizer 등 활용
- 200개의 적은 데이터를 최대한 활용하기 위해 k-fold cross validation 활용

적절한 파라미터 조정을 통해 1D-CNN 최종 모델 선정
Accuracy: 95.51

다양한 시각화를 통해 신호 데이터의 특성을 파악
다양한 input data 형태를 활용하여 걸맞는 모델에 각각 적용
적은 양의 데이터를 최대한 활용하며 동시에 과적합을 방지하기 위한 노력을 기울임

GlobalAveragePooling

10. 데이터 기반 코로나19 예측 공모전

Deep C-raker팀 우수상
2020.09 ~ 2020.10

Why

코로나19 시국의 장기화로 확진자 추이 예측 모델의 필요성 대두
코로나19의 확산에 주요 영향력을 행사하는 요인을 분석

Objective

추석 연휴 기간의 일별 확진자 수 예측
코로나19 일별 확진자 수 예측 모델 구축

Data

기본 제공 데이터: 경기도 감염병 관리지원단 기본 제공 데이터

수집 데이터:

- 코로나 관련 뉴스 기사 개수 데이터(Python 크롤링)
- 코로나 관련 키워드 검색어 추이 데이터(네이버 검색 추이)
- 날씨 데이터 수집(기상청 기상자료개방포털)



Github URL:

https://github.com/JiYeon9610/Covid19_prediction2020



Thank you



2,508,790	976,819	869,870	121,000	421,045	179,984	600,144	279,991	96,429	244,353	70,380	848,579	237,689	583,589	45,000	465,275	182,790	278,981	4,507,284
ENTER	Product A01	Product A02	Product A03	Product A04	Product A05	Product A06	Product A07	Product A08	Product A09	Product A10	Product A11	Product A12	Product A13	Product A14	Product A15	Product A16	Product A17	Product A18

	2015	2016	%Growth
Product A01	108,287	91,938	-15%
Product A02	91,938	125,819	+33%
Product A03	125,819	278,981	+122%
Product A04	278,981	11,827	-96%

	2015	2016
Product A01	8,714	39,912
Product A02	107,812	108,287
Product A03	89,918	91,938
Product A04	123,939	125,819