
CONTENT WORTH DEBUT ARTIST CLASSIFIER

흑인음악 전문 웹매거진의 신인 아티스트 분류기

패스트캠퍼스 데이터사이언스 스쿨 7기

서원영

INTRODUCTION

HIPHOPLE.com ??????????

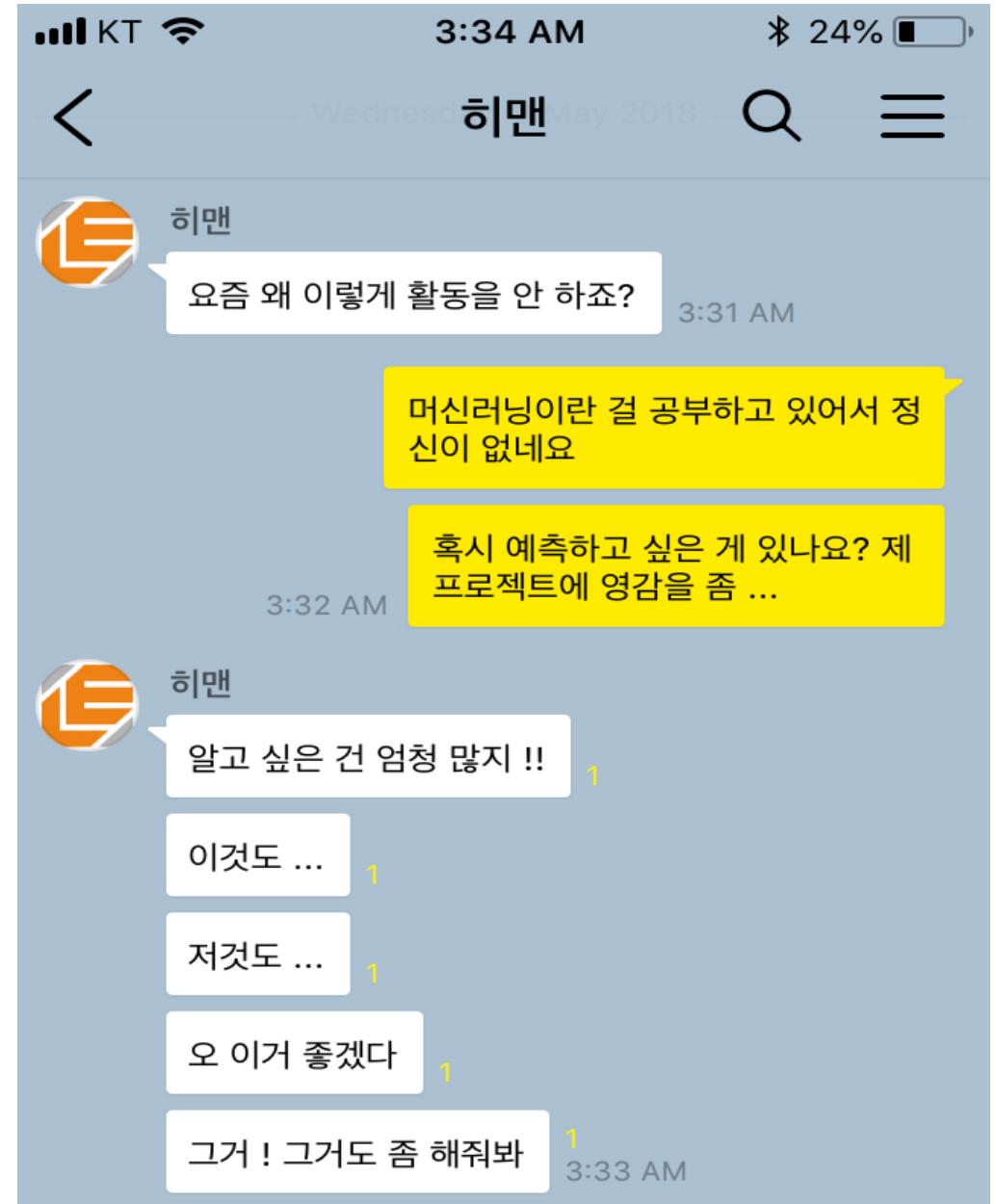
- 흑인음악 전문 웹매거진 & 컨텐츠 플랫폼
- 주요 컨텐츠
 - 번역 컨텐츠
 - 외국 힙합 / R&B 음악 가사 번역 (*)
 - 자막 뮤직비디오
 - 국외 뉴스
 - 매거진 컨텐츠
 - 기획 기사
 - 인터뷰
 - 신보 리뷰
 - 패션 / 라이프 스타일

The screenshot shows the homepage of HIPHOPLE.com. At the top, there's a navigation bar with links for NEWS, MAGAZINE, STYLE, THE LE, LE TV, SPECIAL, COMMUNITY, and SHOP. Below the navigation is a large banner featuring four men in hip-hop attire (hats, sunglasses) with the text "#RAP #LESSON #IN #SEOUL Part.3". Underneath the banner are two smaller sections: one for 'THEY. x DEAN' featuring a track from Dante's Creek (deantrbl Remix) and another for NF's 'PERCEPTION' album. The main content area has four columns: '국외 뉴스' (International News) showing Post Malone and J. Cole, '국내 뉴스' (Domestic News) showing a Monsta Truck, '카드 뉴스' (Card News) showing a man with a sign, and '공지사항' (Announcements) showing a sign for a 100,000 won discount. At the bottom, there are links for '인터뷰' (Interview), '기획 · 연재 기사' (Feature · Serial Article), '앨범 · 싱글 리뷰' (Album · Single Review), and '스타일' (Style).

프로젝트의 시작

- 히맨 (대표) 의 니즈 파악

- 아티스트 머천다이즈 판매 Buzz 분석
- 동남아 K-Hiphop 관련 Buzz 분석
- 대박 신인 아티스트 예측
- 각종 음악 차트 알림 메신저 등등 ...
- and many more ...





대박 신인 예측? (...에 선택과 집중!)

INTERVIEW

DESIIGNER

디자이너



발매일	2017/08/25
레이블	Independent
한줄평	내 영감의 원천은 오로지 당신.

(124.5.114.105) 조회수 3271 추천수 5 댓글 3

힙합 엘리의 컨텐츠 =

리뷰 + 인터뷰 + 기획기사 + SNS계정 + @

일약 스타덤에
차트 1위를 차
행하고, 머나먼
(Desiigner)가
통해 호평을 받
기를 나누고 있는
데

LE: 반갑다. 간
요! 때가 왔어.

LE: 먼저 당신의 대표곡에 대해서 이야기해보고 싶다. "Panda"를 발표한 이후 당신의 삶이 어떻게 달라졌는지 궁금하다.

축복 그 자체야. 모든 게 가장 좋은 쪽으로 바뀌었으니까. ("Panda")는 한 사람을 송두리째 바꿔놨어. 내가 진정한 남자가 될 수 있게 해준 노래야. 비즈니스에 집중하게 됐고, 가족을 돌볼 수 있게 됐어. 물론, 때로는 친구를 잊기도 했지만, 그만큼 얻기도 했지. 지금도 내 주변에 계속 함께 하는 이들이 있기도 하고, 말 그대로 온전한 한 사람이 되게 해준 거야. 그래서 "Panda"라는 곡에 정말 감사해. 판다 판다 판다

- 03. Hold Me Down
- 04. Neu Roses (Transgressor's Song)
- 05. Loose
- 06. We Find Love
- 07. Blessed
- 08. Take Me Away (Feat. Syd)
- 09. Transform (Feat. Charlotte Day Wilson)
- 10. Freudian

캐나다 흑인음악 씬이 계속해서 심상치 않다. 시작은 드레이크(Drake)와 OVO 사운드(OVO Sound), 위肯드(The Weeknd)와 XO였다. 이전까지의 움직임이 아예 없었던 건 아니겠지만, 대중적인 관심이 쏠린 건 그 두 뮤지션이 활약하기 시작했던 2010년대 초반부터였다. 그들은 인기를 얻음과 동시에 자신의 레이블에 같은 연고인 온타리오 주의 뮤지션들을 영입했다. 그때까지만 하더라도 얼터너티브 알앤비라는 음악적 공통분모에 의한 결집 정도로 보였다. 하지만 이후에도 캐나다에서는 이것저것 섞고, 공간감을 살리고, 톤 다운된 감성을 내뿜는 아티스트들이 꾸준히 등장



hiphople

...



Liked by shawna_le_, iandlals and 799 others

hiphople Childish Gambino, SNL에서 신곡 공개

- 올해 더 이상 컴백할 아티스트는 없다고 생각한 순간, 이번에는 Donald Glover가 Childish Gambino로서 신

의식의 흐름 ...

**대박 신인을
예측해보자 !!!**

정량적 요소를 토대로

신인 아티스트에 대한 컨텐츠를 제작하고

유저에게 배포할 가치가 있는지

머신러닝으로 **분류 예측**을 해보자

힙합엘리 매거진팀은

성공적인 데뷔를 한 신인들을 컨텐츠로 다룬다

신인에 대한 컨텐츠를 제작할 때는 :

- (i) 매거진팀 에디터들이 오랜 시간 축적한 도메인 지식을 토대로 컨텐츠 제작 여부를 판별한다
- (ii) 미리 많은 시간을 들여 리서치를 한다
- (iii) 이미 다른 매체에서 다뤄진 후, 뒤늦게 다룬다

리서치 시간을 줄이고

속도를 높일 수 있는 방법은 없을까 ?

프로젝트 진행 단계

-
1. 변수 정의
 2. 데이터 수집 & AWS MySQL 서버 저장
 3. 종속 변수 Labeling
 4. ORM Data Query
 5. Data 통합
 6. 모델 학습
 7. Feature Engineering (6 + 7 번 반복)
 8. 모델 교차 검증 & 최적화

변수 정의

종속 변수 (y :Target Variable)

Binary Classification

1 : 컨텐츠로 제작한다

0 : 컨텐츠로 제작하지 않는다

힙합엘리 에디터 4인

종속 변수 Labeling (1, 0) 작업 참여



독립 변수 (X : Independent Variables)

대상 : 2011 ~ 2018년 4월 18일 사이 발매된 미국 음반시장 데뷔 앨범

- 음악의 장르
- 정식 발매 전, 프로모션 싱글 음반 발매 수
- 대중 음악 매체 발행된 관련 기사의 수
- 대중 음악 매체에서 부여한 평균 평점
- 아티스트의 SNS 팔로워 수

데이터 수집

데이터 크롤링 → AWS EC2 MySQL DB 저장

크롤링 데이터	크롤링 대상	크롤링 라이브러리
연도 별 데뷔 앨범	www.wikipedia.com	Requests, BeautifulSoup
매체 평점	www.metacritic.com	Scrapy
	www.pitchfork.com	Requests
	www.albumoftheyear.com	Requests, BeautifulSoup
아티스트 별 SNS 팔로워	www.chartmetric.io	Selenium
대중 음악 매체 관련 기사 수	www.billboard.com	Requests, BeautifulSoup
	www.genius.com	Requests, BeautifulSoup
	www.thesource.com	Requests, BeautifulSoup
	www.xxlmag.com	Requests, BeautifulSoup

종속 변수 LABELING

종속 변수 Labeling

- 힙합엘이 매거진팀 에디터들의 도메인 지식 참고
- 구글 스프레드시트를 통한 Labeling 템플릿 공유

	C	D	E	F	G
	genre	Yes / NO			
	hiphop	0			
	hiphop	0			
	hiphop	0			
	hiphop	0			왼쪽의 데이터는 2011년~현재 사이 발매 된 데뷔 스튜디오 앨범의 목록입니다.
	hiphop	0			과거 힙합엘이에서 왼쪽의 앨범을 가지고 컨텐츠로 다루었던
	hiphop	1			이력을 기재해주시면 됩니다.
	hiphop	1			** 컨텐츠는 매거진팀 기획기사, 리뷰, SNS컨텐츠 등을 뜻합니다
	hiphop	0			** 국외 뉴스, 발매 뉴스 등은 포함하지 않습니다
Side Story	hiphop	1			** 너무 고민 마시고 술술 기재해주세요
	hiphop	0			
	hiphop	1			예시)
	hiphop	0			과거 컨텐츠로 다룬 적이 있다 ---> 1
	hiphop	0			기억이 안 나거나, 다른 적 없지만, 다룰 만 하다 ---> 1
	hiphop	0			컨텐츠를 만들지 않았다 / 다룰 필요가 없다 ---> 0
	hiphop	1			
	hiphop	1			예시)
J. Love II	hiphop	0			J. Cole - Sideline Story (hiphop) ---> 1
	hiphop	0			Sevyn Streeter - Girl Disrupted ---> 0
Ever Told	hiphop	0			DRAM - Baby DRAM ---> 1
	hiphop	1			
	hiphop	0			
	hiphop	1			

DATA 통합

Data size : 1083 Rows

- Label (종속변수 Target Variable)

Features : 15개

Album Info	Rating (평점)	Aritcle Counts	SNS
- 장르	- Album of the Year	- Billboard	- 트위터 팔로워
- 싱글 앨범 수	- Pitchfork - Metacritic	- Genius - The Source - XXL Magazine	- 인스타그램 팔로워 - 페이스북 페이지 좋아요 - 유튜브 채널 구독자 - 사운드 클라우드 팔로워 - 스포티파이 팔로워

모델 학습

분류 알고리즘 학습 : Scikit-Learn 라이브러리

Model	Classification Metrics (Class I)				Comment
	Precision	Recall	F1 Score	AUC	
KNN : K-Nearest Neighbours	0.79	0.43	0.55	0.83	
SGD : Steepest Gradient Decent	0.31	0.92	0.46	0.66	
Decision Tree	0.62	0.65	0.63	0.86	
Random Forest	0.74	0.67	0.70	0.91	★
Gradient Boosting	0.69	0.57	0.62	0.93	
XGBoost : Extra Gradient Boosting	0.78	0.64	0.70	0.93	

Random Forest Classifier 선택의 이유

Recall (Sensitivity) 가 매우 중요하다고 판단

- **Precision** 의 경우 False Positive (FP)이 미치는 영향

- 예측에 실패해서, 무시해도 될 신인인데, 굳이 컨텐츠를 생산한다.
- 기존과 비슷한 시간과 에너지 소비 → 비즈니스에 큰 영향을 끼치지 않음

$$Precision = \frac{TP}{TP + FP}$$

- **Recall (Sensitivity)** 의 경우 False Negative (FN)이 미치는 영향

$$Recall / Sensitivity = \frac{TP}{TP + FN}$$

- 예측에 실패해서 해당 신인 아티스트에 대해 신속하게 컨텐츠를 생산하지 않는다.

- 다른 매체에서 먼저 다룰 가능성
- 잠재적인 구독자 유입이 감소 할 가능성
- 부가적인 에이전시 계약, 머천다이즈 등의 기회를 놓쳐버릴 가능성 ...

FEATURE ENGINEERING

Feature Engineering 시도

Feature	유형		Performance Improvement
Genre	One-Hot Encoding	One-Hot Encoding을 통한 Dummy Variable 생성	○
Ratings	Null Value Imputation	대중 음악 매체의 평점, 평론, 주목을 받지 못한 아티스트의 경우 0점을 받은 것으로 처리	○
Buzz Rating SNS Followers	Bining	<ul style="list-style-type: none">- 실수 값의 범위를 토대로 구간 분할- 구간 별 Dummy Variable 생성	×
SNS Followers	Ratio	<ul style="list-style-type: none">- 각 SNS 플랫폼 / 전체 SNS 팔로워 수- 비율을 계산	×
SNS Followers	Re-Grouping	<ul style="list-style-type: none">- 개인 SNS (Twitter, Instagram)- 팬페이지 성향의 SNS (Facebook, Youtube)- 음악 기반 SNS (Soundcloud, Spotify)	×

모델 성능 개선 & 최적화

Recall 성능 향상

Imbalance Problem (데이터 불균형 문제)

- 종속 변수의 Class가 불균형하면 ...
 - Accuracy를 기반으로 예측 성능을 평가
 - Major Class를 예측하는 것이 상대적으로 쉽다
- **Under Sampling**을 통해 학습데이터 Class의 비율 50:50으로 조절
- Precision의 저하는 Trade-off → Recall 향상에 초점

Under Sampling : Imblearn 라이브러리

Model	Classification Metrics (Class I)				Comment
	Precision	Recall	F1 Score	AUC	
ClusterCentroids	0.45	0.95	0.61	0.86	
Random Under Sampler	0.67	0.88	0.89	0.94	★
CondensedNearestNeighbour	0.61	0.80	0.69	0.92	
AllKNN Metrics	0.54	0.89	0.67	0.92	
InstanceHardnessThreshold	0.60	0.89	0.72	0.92	
NearMiss	0.32	0.95	0.47	0.74	
NeighbourhoodCleaningRule	0.60	0.80	0.69	0.93	
OneSidedSelection	0.76	0.73	0.75	0.92	
TomekLinks Metrics	0.75	0.71	0.73	0.95	

모델 최적화 : Scikit-Learn GridSearchCV

하이퍼 파라미터 튜닝 부분 반영

```
1 %%time  
2 gs_result = gs.fit(X_resampled_rus, y_resampled_rus)
```

CPU times: user 1h 32min 3s, sys: 47.4 s, total: 1h 32min 50s
Wall time: 1h 37min 8s

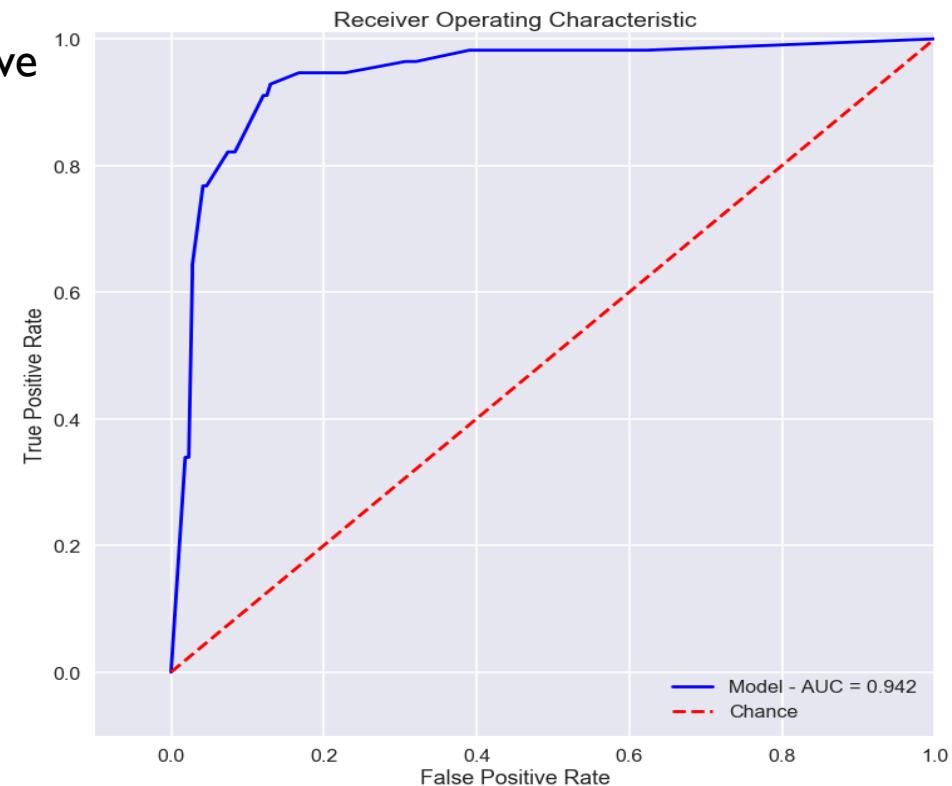
```
1 gs.best_params_
```

```
{'bootstrap': True,  
'class_weight': None,  
'criterion': 'entropy',  
'max_depth': 10,  
'max_features': 'auto',  
'max_leaf_nodes': None,  
'min_impurity_decrease': 0.0,  
'min_impurity_split': None,  
'min_samples_leaf': 4,  
'min_samples_split': 3,  
'min_weight_fraction_leaf': 0.0,  
'n_estimators': 10,  
'oob_score': False,  
'random_state': None,  
'verbose': 0,  
'warm_start': False}
```

Classification Report

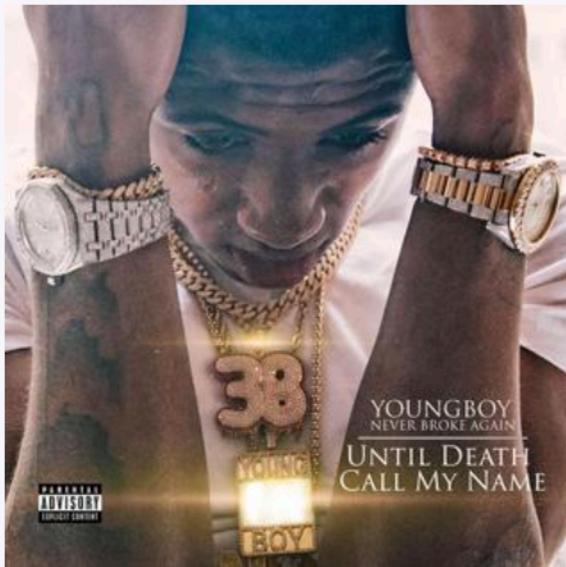
	Precision	Recall	F1 Score
Class 0	0.98	0.87	0.92
Class 1	0.65	0.93	0.76
Avg / Total	0.91	0.88	0.89

ROC Curve AUC



새로운 데이터 테스트 해보기

Until Death Call My Name



Studio album by **YoungBoy Never Broke Again**

Released April 27, 2018

Recorded 2017–2018

Genre Hip hop · trap

Label Never Broke Again · Atlantic

Producer 1040 · Ben Billions · Big Korey ·
Bighead · CashMoneyAP · DJ
Montay · DJ Swift · DMacTooBangin
· Drumma Boy · Dubba-AA · Judge ·
Kenny Beats · Mook · Schife
Karbeen · TM88 · Wheezy

Youngboy Never Broke Again – Until Death Call My Name

- 2018년 4월 27일 발매 (장르 : 힙합)

- 데뷔 스튜디오 앨범

* 5월 9일 현시점까지 힙합엘리에서 다뤄지지 않았음 *

그렇다면 ... 모델은

이 앨범을 어떻게 예측했을까?



Load Model

저장한 모델 불러오기

```
In [2]: 1 model = pickle.load(open("rfc_rus_gs.plk", "rb"))
```

Input New Data

```
In [3]:
```

```
1 data = {  
2     'single_count': [2],  
3     'genre_funk': [0],  
4     'genre_hiphop': [1],  
5     'genre_pop': [0],  
6     'genre_rnb': [0],  
7     'genre_soul': [0],  
8     'freq_billboard': [1],  
9     'freq_genius': [0],  
10    'freq_theSource': [0],  
11    'freq_xxl': [42],  
12    'rating': [69],  
13    'twitter': [912000],  
14    'instagram': [3900000],  
15    'facebook': [581939],  
16    'youtube': [2170000],  
17    'soundcloud': [384000],  
18    'spotify': [854874]  
19 }
```

아티스트, 앨범에
해당하는 데이터 입력

```
In [4]: 1 test = pd.DataFrame.from_dict(data)
```

Predict

Class 예측 결과 :

```
In [5]: 1 model.predict(test)
```

```
Out[5]: array([0])
```

0 → 컨텐츠로 다루지 않는다

Predict Probability

확률 예측 결과 :

```
In [6]: 1 model.predict_proba(test)[:, 1]
```

```
Out[6]: array([0.3])
```

20% → 매우 낮다 → 우선순위에서 밀려남

OH - YEAH



보완점 / 앞으로의 진행 방향

I. 기존 모델의 Precision까지 개선할 수 있는 다른 Imbalanced Problem Sampling 방법론을 적용

2. XGBoost 알고리즘을 적용과 최적의 Hyper Parameter 튜닝 방법 연구

3. 적은 데이터 양

- 향후 발매되는 데뷔앨범 데이터를 축적하여 데이터를 확보 → 모델 개선
- 다양한 매체의 기사 + 평점 +평론 데이터 등 추가 크롤링
- 평론 텍스트 자연어 분석

4. 기존의 Dash 기반 시각화 웹어플리케이션 개선

- 새로운 데뷔 아티스트 / 앨범 입력 시

- Feature Data 자동 크롤링 → 데이터셋 업데이트
- Pickle 파일로 저장한 모델 구동, 예측까지 구현

프로젝트에 대한 정보는...

아래의 링크를 참고해주세요 :D

Github Repository Source Code
<http://www.github.com/lucaseo>

Data Visualization Web Application
<http://www.luca.herokuapp.com>

그리고 저에 대해서는 ...

아래의 링크를 방문해주세요 :)

Github
<http://www.github.com/lucaseo>

LinkedIn
<http://www.linkedin.com/in/lucaseo>

Instagram
<http://www.instagram.com/seoluca>

Github Blog
<http://www.lucaseo.github.io>