

MACHINE LEARNING STUDY 2020 WINTER

## TEAM A : STUDY PROJECT PROGRESS #2

---

Baek.S.W / Bae.S.M / Lee.J.W /Kwon.J.W

2020 . Jan . 15

# #1 Dataset

Lukious / Aladin-crawler

Code Issues Pull requests Actions Projects Wiki Security Insights Settings

Web Crawler for korea online book store aladin 알라딘 책표지 및 이름 크롤러

Manage topics

11 commits 1 branch 0 packages 0 releases 1 contributor GPL-3.0

Branch: master New pull request Create new file Upload files Find file Clone or download

Lukious Update README.md Latest commit 78c072b 2 days ago

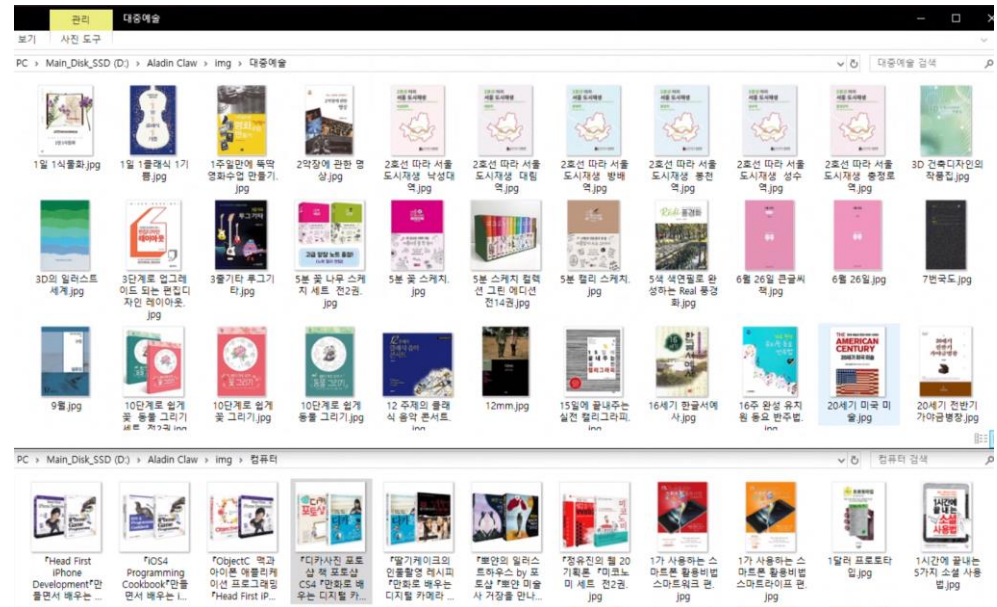
images	Add example images	4 days ago
.gitignore	Initial commit	6 days ago
AladinBookCoverCrawler.py	Typo removed	6 days ago
LICENSE	Initial commit	6 days ago
README.md	Update README.md	2 days ago
preprocessing.py	Add simple PreProcessor:	2 days ago

README.md

## Requirements

- beautifulsoup
- Numpy
- re [https://www.graphviz.org/]

```
pip install re
# Or conda install re
```



Make Korean Book Cover Dataset By Crawling Aladin

## #1 Dataset

---

가정요리 뷰티	-> 8,668 images	} Total 87,000 Images
과학	-> 15,507 images	
대중예술	-> 9,851 images	
소설/시	-> 6,378 images	
여행	-> 6,953 images	
유아	-> 10,948 images	
인문학	-> 7,598 images	
종교	-> 3,712 images	
참고서	-> 4,366 images	
컴퓨터	-> 13,026 images	

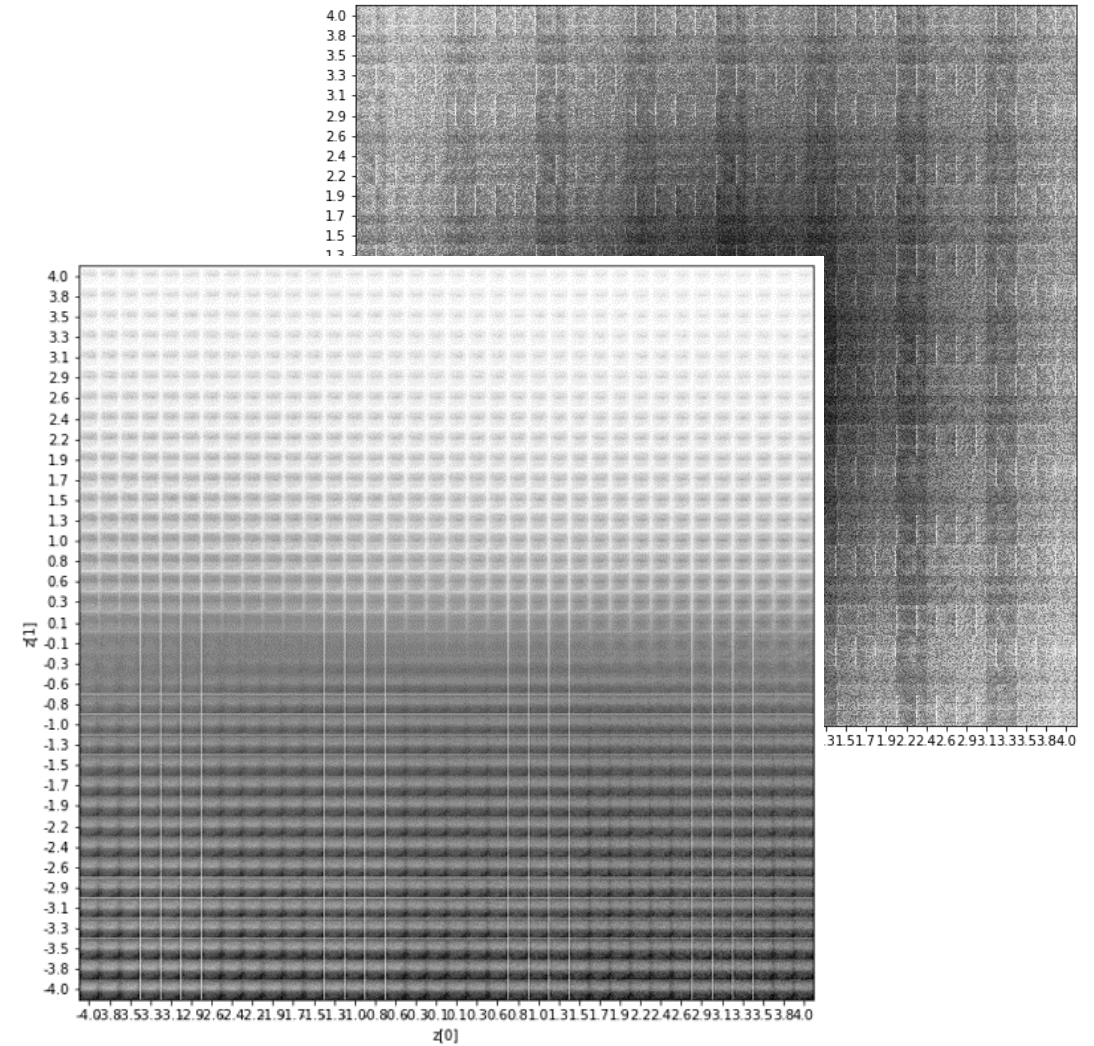
Make Korean Book Cover Dataset By Crawling Aladin

## #2 Tries

### Try VAE Method

```
Epoch 1/50  
15507/15507 [=====] - 11s 737us/step - loss: 39.5617  
Epoch 2/50  
15507/15507 [=====] - 10s 637us/step - loss: 117.8342  
Epoch 3/50  
15507/15507 [=====] - 10s 630us/step - loss: 18518.2015  
Epoch 4/50  
15507/15507 [=====] - 10s 629us/step - loss: 559.7961  
Epoch 5/50  
15507/15507 [=====] - 10s 633us/step - loss: 533.0516  
Epoch 6/50  
15507/15507 [=====] - 10s 675us/step - loss: 508.5692  
Epoch 7/50  
15507/15507 [=====] - 10s 628us/step - loss: 482.6156  
Epoch 8/50  
6144/15507 [=====>.....] - ETA: 6s - loss: 464.4500
```

- Avg Loss are 100~14000 [Useless]
- Gonna Try with other Preprocessing Method



### Preprocessing

Set Preprocessing size / Data Cleaning

Book Cover Data Preprocessing [Delete Name]

- Tesseract + Denoising Method + Manually processing
- OpenCV Segmentation Method







### NPL Model Construction

Considering use BERT LISA ...

- Adapt KoNLPy on Korean Words
- Adapt Attention Algorithm

## #2 Schedule

---

	Week1	Week2	Week3	Week4	Week5	Week6	Week7	Week8	Week9
Topic Select Project Sketch									
Data collection									
Data preprocessing									
Word Classifier Design									
GAN Design									
Model Validation									
Model Testing								