10 IDEAS FOR STARTING A GITHUB PORTFOLIO TO BREAK INTO DATA SCIENCE

MIGUEL FIERRO
https://miguelgfierro.com

v1.1

https://miguelgfierro.com

Welcome! I'm Miguel Fierro.

You might already know, but companies these days are more interested in practical knowledge than in theoretical knowledge.

That's why CVs are less and less relevant, your online presence is the best CV.

That's why creating a Data Science portfolio is the best way to get a job in Data Science.

Next, you'll find 10 ideas to build your portfolio.

Vamos!

1. HATE SPEECH DETECTION

Description:
Sentiment analysis of Tweets to identify whether the text is hate speech, offensive language, or neutral.

Dataset:
Kaggle Hate Speech and Offensive Language Dataset

Library:
Huggingface Transformers

Algorithm:
DeBERTa

2. FORECASTING OF COVID

Description:
Use time-series forecasting to predict the spread of COVID in different countries.

Dataset:
Kaggle COVID-19 dataset or Google COVID-19 Open Data

Library:
Meta AI Prophet

Algorithm:
Prophet

3. MOVIE RECOMMENDATION

Description:
Recommend a movie based on historical user behavior.

Dataset:
Movielens

Library:
Microsoft Recommenders

Algorithm:
SAR

4. CARTOON CLASSIFICATION

Description:
Identification of the Simpson characters in images

Dataset:
Kaggle The Simpsons Characters Data

Library:
Torchvision (PyTorch)

Algorithm:
ResNet18

5. FOOTBALL PLAYER TRACKING

Description:
Player tracking in football matches for obtaining match metrics.

Dataset:
Sports Videos in the Wild from Michigan University

Library:
Ultralytics YOLOv5

Algorithm:
YOLOv5

6. NEWS RECOMMENDATION

Description:
Recommendation of news articles mixing NLP and Recommendation Systems.

Dataset:
Microsoft MIND dataset

Library:
Microsoft Recommenders

Algorithm:
LSTUR

7. PAPER SUMMARIZATION

Description:
Summarization of papers

Dataset:
arXiv Summarization Dataset

Library:
Huggingface Transformers

Algorithm:
BART

8. AUTOPLAY SUPER MARIO

Description:
Use Reinforcement Learning to play Super Mario Bros.

Dataset:
Open AI gym Super Mario Bros

Library:
Ray

Algorithm:
PPO

9. EXPLANATION NOTEBOOK

Description:
Use a Jupyter notebook to explain logistic regression in detail.

Dataset:
Iris plant dataset from Scikit-learn

Library:
Vowpal Wabbit

Algorithm:
Logistic regression

10. CELEBRITY IDENTIFICATION

Description:
Using an API like the Face API of Microsoft Cognitive Services, identify the face of a celebrity.

Dataset:
Celebrity Face Recognition Dataset

Library:
Face API

Algorithm:
API

Your GitHub portfolio is your first step into an amazing Data Science career.

I would love to know your progress. Feel free to create a LinkedIn post about your project.

This is my LinkedIn:
https://www.linkedin.com/in/miguelgfierro/

Follow me and press the bell button to get tips on how to understand and apply AI.

Looking forward to hearing from you.

Best,
Miguel