Big Data Engineer II

By Christan Versteeg, 500859503

Contents

[1. Summary/Abstract 2](#_Toc163105671)

[2. Introduction 2](#_Toc163105672)

[3. Background 2](#_Toc163105673)

[4. Methods 3](#_Toc163105674)

[4.1. Data Collection 3](#_Toc163105675)

[4.2. Data Preparation 4](#_Toc163105676)

[4.3. Sentiment Analysis Implementation 4](#_Toc163105677)

[5. Results 4](#_Toc163105678)

[6. Conclusion and Recommendations 4](#_Toc163105679)

[7. Bibliography 5](#_Toc163105680)

# Summary/Abstract

# Introduction

# Background

# Methods

First I started of by tackling what NoSQL program I would use. I decided to use MongoDB since it is suggested in the assignment, plus it has a clean UI called MongoDB Compass. I downloaded it on the website. (MongoDB, 2024)

A screenshot of a computer

Description automatically generated

I then created a Database and a collection for it, and imported the hotel reviews csv:

A screenshot of a computer

Description automatically generated

It imported roughly half a million entries:

A screenshot of a computer

Description automatically generated

After importing the csv into mongo I wanted to access it in the Python script, which I did like so:  
A screenshot of a computer

Description automatically generated

Which I did at the hand of PyMongo and Pandas documentation (MongoDB, 2024) (Pandas, 2023).

Interestingly, I figured out a limitation of the DB, for reasons I do not know yet, essentially, importing the csv allows for easy and quick display of ALL 500k reviews. However, if the DB tries to display even 100k reviews, it will given an overflow message saying it has exceeded usage of 200MB, (exceeded as 253MB specifically). I figured this out because I wanted a quick and easy local variant of my program, like so:

from pymongo import MongoClient

import pandas as pd

db\_driven = False

if db\_driven:

    client = MongoClient("localhost", 27017)

    db = client['Big']

    collection = db['Data']

    df = pd.DataFrame(list(collection))

else:

    csv\_file = r"C:\Users\pooti\Desktop\Big-Data-Engineer\App\Hotel\_Reviews.csv"

    df = pd.read\_csv(csv\_file)

Then I decided to work on a basic implementation of the dashboard. I wanted to figure out what dashboard library would be the best to utilize, so I compared it online. (Pavlovych, 2024). I ended up choosing Streamlit because of having experience with it before, and it being very quick to set up prototypes and easy to use, I do not need much flexibility.

The basic implementation of the dashboard ended up looking like this:

A screenshot of a hotel registration

Description automatically generated

A screen shot of a computer

Description automatically generated

Which due to the simplistic nature of streamlit, only required a few lines of code:  
A screen shot of a computer code

Description automatically generated

Which I wrote with the help of the streamlit documentation. (Streamlit, 2024).

After this, I wrote a very simple new script that stores the columns:

A screenshot of a computer program

Description automatically generated

This helps to later work with the columns, as well as currently filtering the only relevant columns in the dashboard:

A screenshot of a computer program

Description automatically generated

Which only changes the default view, they can still be added by the user:

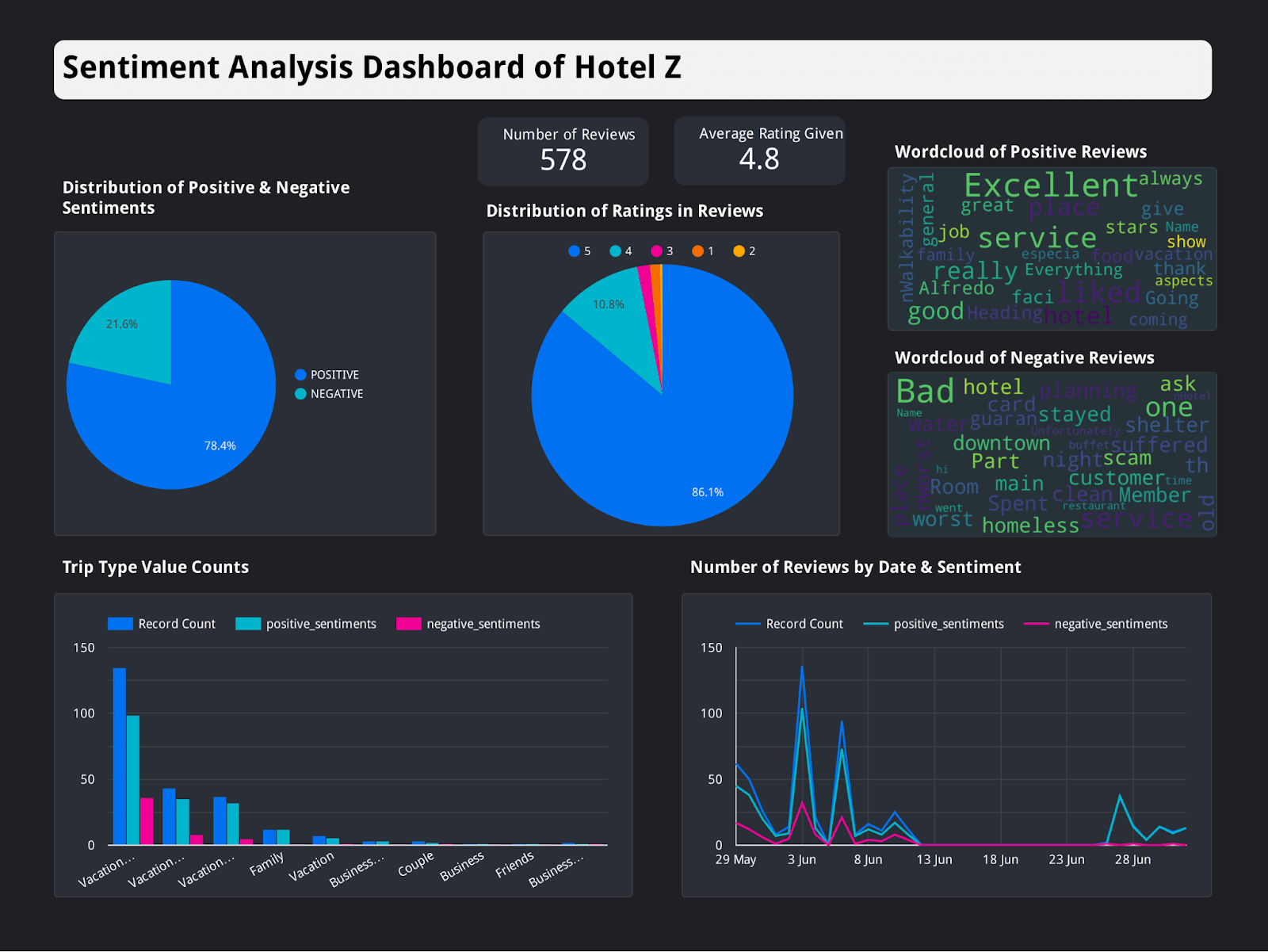
A screenshot of a phone

Description automatically generated

But it now gives a much cleaner table view, which you is visible without having to scroll horizontally:  
A screenshot of a computer

Description automatically generated

Next up I wanted to flash out my dashboard more, and took this source as a good reference for what I’d want my dashboard to roughly look like (Bagaskara, Güler, Rasyid, & Torcato, 2024):



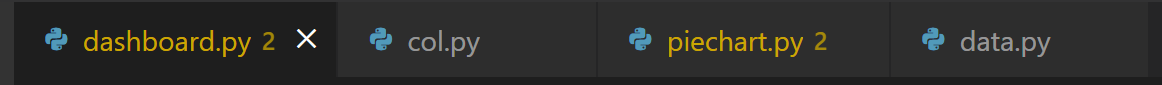
I needed a way to draw these graphs in python, which I chose matplotlib for. It’s the most popular visualization library in python. (Talaj, 2024).

With a basic implementation of the pie chart added into the program:

A green and red circle with white text

Description automatically generated

I decided to split up my classes:



Initially I tried to import the data file of dashboard into piechart, but this gave a circular dependency because piechart was being used in dashboard. So I created a data script to resolve this dependency, now every file that uses data will grab from the data file.

After this I created a new class for showing totals, like in the reference image. With some simple code I managed to get some a very simple layout:

A black background with white text

Description automatically generated

import col

import data

import streamlit as st

num\_positive = data.df[col.POSITIVE\_REVIEW].shape[0]

num\_negative = data.df[col.NEGATIVE\_REVIEW].shape[0]

total\_reviews = data.df.shape[0]

average\_rating = data.df[col.REVIEWER\_SCORE].mean()

def draw():

    st.markdown(

    f'<p>Number of Reviews</p><h1>{total\_reviews}</h1>',

    unsafe\_allow\_html=True

    )

    st.markdown(

    f'<p>Average Rating Given</p><h1>{average\_rating:.2f}</h1>',

    unsafe\_allow\_html=True

    )

This however didn’t look very good so I decided to use the build in columns feature of streamlit (Streamlit, 2024):

A screen shot of a number

Description automatically generated

def draw():

    col1, col2 = st.columns(2)

    with col1:

        st.markdown(

        f'<p>Number of Reviews</p><h1>{total\_reviews}</h1>',

        unsafe\_allow\_html=True

        )

    with col2:

        st.markdown(

        f'<p>Average Rating Given</p><h1>{average\_rating:.2f}</h1>',

        unsafe\_allow\_html=True

        )

Which already gave it a much cleaner design, and more clear too. However it could still use a bit of polish to stand apart, just like in the refence image. To do this, I use some simple CSS to define a box (Mmdn web docs, 2024) around my text:

A black square object with a black background

Description automatically generated

import col

import data

import streamlit as st

num\_positive = data.df[col.POSITIVE\_REVIEW].shape[0]

num\_negative = data.df[col.NEGATIVE\_REVIEW].shape[0]

total\_reviews = data.df.shape[0]

average\_rating = data.df[col.REVIEWER\_SCORE].mean()

def draw():

    st.html("""

    <style>

    .metric-box {

        background-color: #1e1e1e;

        color: white;

        padding: 10px;

        border-radius: 8px;

        text-align: center;

        box-shadow: 0px 2px 4px rgba(0, 0, 0, 0.2);

    }

    .metric-box h1 {

        margin: 0;

        font-size: 2em;

    }

    .metric-box p {

        margin: 0;

        font-size: 1.2em;

        font-weight: bold;

    }

    </style>

    """)

    col1, col2 = st.columns(2)

    with col1:

        st.html(f'<div class="metric-box"><p>Number of Reviews</p><h1>{total\_reviews}</h1></div>')

    with col2:

        st.html(f'<div class="metric-box"><p>Average Rating Given</p><h1>{average\_rating:.2f}</h1></div>')

After this I continued to work on the piechart class, because I wanted another piechart, just like the reference image, because having an overall score distribution on the dashboard presented as a piechart is a good metric for the user to see. I did this by first renaming the draw function in the class to the specific pie it draws, so draw\_positive\_negative, and then for the new piechart, gave it the name draw\_score\_distribution, so that both could be called in the dashboard.py class like so:

piechart.draw\_positive\_negative()

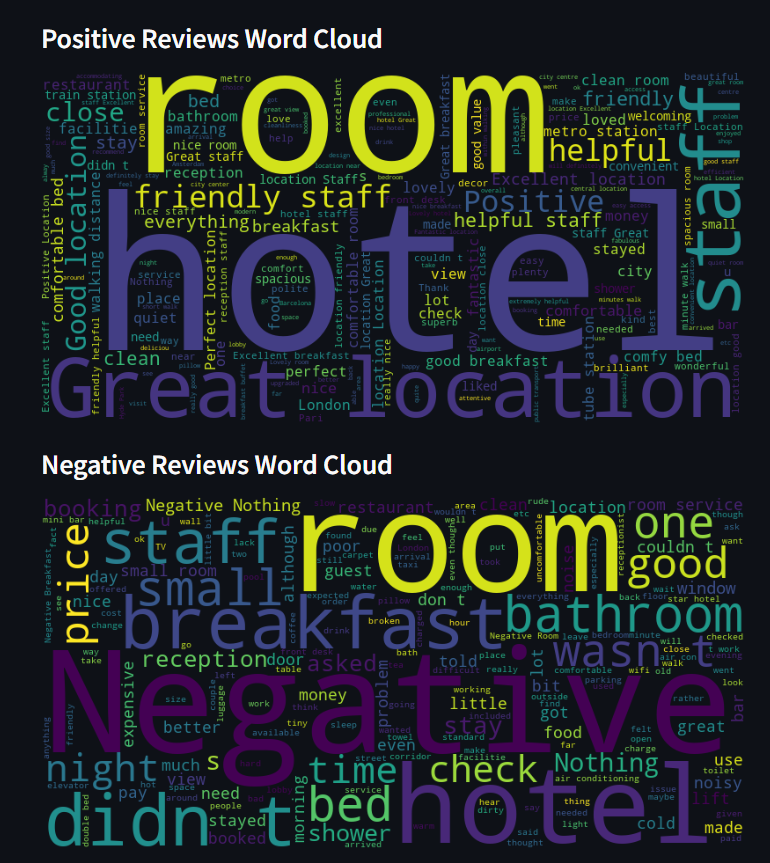
piechart.draw\_score\_distribution()

Other than this, the piechart makes use of histogram math (the graphs that are usually displayed as bars), to section it in each pie. Notably, every pie has been normalized to fit into 1 through 5, since the actual review score is from 1 to 10. (So 1 is 1-2, 2 is 3-4 and so on).

A pie chart with numbers and a number on it

Description automatically generated

For some playful visualization, I also decided to add a wordcloud, like the reference image. Just like the reference image I separated it into two separate wordclouds for negative and positive words. I used an article about wordcloud as reference, despite really liking the masks applied within the article, I don’t think it would be beneficial to add it to the dashboard, so I did not use that step. (Wang, 2022). There’s also some funkiness when it comes to the drawing of streamlit, which makes use of matplotlib, by doing this it will by default give my wordcloud a white background, as denoted in a source found (Mueller, 2020). After fixing the transparency issue, the wordcloud came out clean, looking like this:



Next up, I started working on the machine learning part of the assignment, for this I choice Pytorch and Keras to compare the results. And article written by Kaggle itself was used as basis to support the claims made here (Agrawal, 2021). Keras is said to be easier to use but more constrained, while Pytorch is more flexible but in general hard to use. However; it turns out Keras has several deprecated libraries related to it, such as Bleach (Draves, 2018). This makes the installation process quite tricky. Eventually I managed to fix it, which from what I could tell there were a couple distinct issues. First of all my attempt at trying to fix it with a package manager – Anaconda got me no further, so after that I tried to fix it within the local pip install again. One of the first issues I solved was that I accidentally named my script the same as the library, which then when you try to import your library, it will actually try to import your script (Tilwankar, 2015). Another one was trying to import it through Tensorflow, instead of Keras directly – there was conflicting information online, but eventually found a source that stated recent use that stated since Tensorflow 2.16 the import should detonate separately from tensorflow (Rakita, 2021). Because of dependencies the numpy version that Tensor/Keras needed was also lower than the default, the specific version 1.21.6, below the 2.0 release (Zhang, 2022). Later this was bumped up twice, ending up with 1.23.5, because specific packages required a higher version of it, this did not conflict with the other package requirements. Finally, the last issue was an obscure package dependency, for tensorflow.keras.preprocessing.text, for that you had to install keras-preprocessing, but that didn’t fix it. Eventually I found a obscure Stack Overflow post that said you to install tensorflow-text, which that ended up fixing it (Afnan, 2024). After fixing all of it up I used a simple example script (Awan, 2023) to test out if it was working correctly, and it worked like expected.

I rewrote the keras script to accommodate an actual ML sentiment analysis model on basis of an in depth article (Brownlee, Your First Deep Learning Project in Python with Keras Step-by-Step, 2022). I first wrote the foundational layers and then grouped it into functions. After that I did the same for PyTorch, the same author wrote an article about PyTorch as well (Brownlee, 2023), which I used as guidance to write the pytorch script, which I also did in the way of first overall structure and then sorting it into functions. After both scripts had their initial function I made a class that overarches both of them to hold common methods.

After finishing the commonalities, I wanted to have the the primitive form of the sentiment analysis of both keras and pytorch uploaded. Which I did by altering the data class, adding a new upload function to it:

def upload(data, collection\_name):

    records = data.to\_dict('records')

    for record in records:

        record.pop('\_id', None)

    client = MongoClient("localhost", 27017)

    db = client['Big']

    target\_coll = db[collection\_name]

    target\_coll.insert\_many(records)

    client.close()

Which was primitively implemented in the keras and pytorch class like so (Tim, 2019):

    p1 = predict\_sentiment("The hotel was fantastic!")

    p2 = predict\_sentiment("The room was dirty and the service was terrible.")

    df\_preds = pd.DataFrame({"Prediction": [p1, p2]})

    data.upload(df\_preds, "Keras")

prediction()

Which ended up looking like this in the database:

A screenshot of a computer

Description automatically generated

After having the initial primitive version of it working, I wanted the actual proper dataset in Mongo, so I could retrieve it from there for the dashboard. Which ended up looking like this:

A screenshot of a computer program

Description automatically generated

def prediction():

    test\_loss, test\_acc = model.evaluate(test\_padded, test\_labels, verbose=1)

    print(f"Test Loss: {test\_loss:.4f}")

    print(f"Test Accuracy: {test\_acc:.4f}")

    def predict\_sentiment(text):

        text\_cleaned = kerch.clean\_text(text)

        seq = tokenizer.texts\_to\_sequences([text\_cleaned])

        padded = pad\_sequences(seq, maxlen=kerch.MAX\_LEN, padding='post', truncating='post')

        prediction = model.predict(padded)[0][0]

        return prediction, kerch.POSITIVE if prediction > 0.5 else kerch.NEGATIVE

    data.sample = kerch.isolate\_sentiment\_columns(data.sample)

    for i, row in data.sample.iterrows():

        score, sentiment = predict\_sentiment(row['Review'])

        data.sample.at[i, col.REVIEWER\_SCORE] = score \* 10

        if sentiment == kerch.POSITIVE:

            data.sample.at[i, col.POSITIVE\_REVIEW] = row['Review']

        else:

            data.sample.at[i, col.NEGATIVE\_REVIEW] = row['Review']

    data.upload(data.sample, "Keras")

prediction()

A purple circle with text

Description automatically generated

After this I started looking into Dask, quickly I found that by using the Dask DataFrame, you can convert pandas to a Dask DataFrame, which allows you to execute operations in parallel (Dask, 2024). It’s only a small implementation, but if more data is loaded than the current limit it could save performance by saving on the regex cleaning of the text.

df\_dd = dd.from\_pandas(data.df, npartitions=4)

df\_dd[col.REVIEW] = df\_dd[col.REVIEW].fillna('')

df\_dd[col.REVIEW] = df\_dd[col.REVIEW].apply(

    kerch.clean\_text,

    meta=(col.REVIEW, 'object')

)

df = df\_dd.compute()

df = kerch.isolate\_sentiment\_columns(df)

Though ideally, Dask would also be used for the predictions so it can predict in parallel, making it much faster.

# Results

Pytorch appeared to be much more performant, up 3x as fast, taking only 25 seconds as opposed to the 75 seconds keras takes.

# Conclusion and Recommendations

# Bibliography

Afnan, S. (2024, November 8). *ModuleNotFoundError: No module named 'keras\_preprocessing'*. Opgehaald van Stack Overflow: https://stackoverflow.com/a/79170893

Agrawal, U. (2021). *Kaggle*. Opgehaald van Keras VS PyTorch: A perfect guide: https://www.kaggle.com/code/utcarshagrawal/keras-vs-pytorch-a-perfect-guide

Awan, A. R. (2023, Aug 24). *Youtube\_\_\_Sharings/computer vision/how to setup keras and tensorflow in vs code using python*. Opgehaald van GitHub: https://github.com/ahmedrohailawan/Youtube\_\_\_Sharings/blob/main/computer%20vision/how%20to%20setup%20keras%20and%20tensorflow%20in%20vs%20code%20using%20python/tensorflow%20and%20keras.py

Bagaskara, L., Güler, H., Rasyid, M., & Torcato, S. (2024, August 10). *Sentiment Analysis on Hotel Reviews*. Opgehaald van Data Hen: https://www.datahen.com/blog/sentiment-analysis-hotel-reviews/

Brownlee, J. (2022, August 16). *Your First Deep Learning Project in Python with Keras Step-by-Step*. Opgehaald van Machine Learning Mastery: https://machinelearningmastery.com/tutorial-first-neural-network-python-keras/

Brownlee, J. (2023, March 1). *PyTorch Tutorial: How to Develop Deep Learning Models with Python*. Opgehaald van Machine Learning Mastery: https://machinelearningmastery.com/pytorch-tutorial-develop-deep-learning-models/

Draves, S. (2018, January 26). *Dependency on old version of bleach (1.5)*. Opgehaald van GitHub: https://github.com/tensorflow/tensorflow/issues/16424#issuecomment-1229561997

Korab, P. (2023, May 14). *Fine-tuning VADER Classifier with Domain-specific Lexicons*. Opgehaald van Medium: https://pub.towardsai.net/fine-tuning-vader-classifier-with-domain-specific-lexicons-1b23f6882f2

Mmdn web docs. (2024). *box-sizing*. Opgehaald van Mozilla Developers: https://developer.mozilla.org/en-US/docs/Web/CSS/box-sizing

MongoDB. (2024, 8 26). *mongdb.com*. Opgehaald van Try MongoDB Community Edition: https://www.mongodb.com/try/download/community

MongoDB. (2024, June 24). *pymongo.readthedocs.io*. Opgehaald van PyMongo 4.8.0 documentation: https://pymongo.readthedocs.io/en/stable/tutorial.html

Mueller, A. (2020, July 17). *WordCloud results in a white background instead of a transparent one when when mode is “RGBA” and background\_color is None. #569*. Opgehaald van GitHub: https://github.com/amueller/word\_cloud/issues/569

Navlani, A. (2020, September 5). *Naive Bayes Classification using Scikit-learn*. Opgehaald van Medium: https://avinashnavlani.medium.com/naive-bayes-classification-using-scikit-learn-60bc5176f868

Pandas. (2023, April 24). *pandas.pydata.org*. Opgehaald van pandas.DataFrame: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.html

Pavlovych, A. (2024, July 25). *What Is the Best Python Dashboard Framework?* Opgehaald van Planeks: https://www.planeks.net/python-dashboard-development-framework/

Pius, A. (2023, November 22). *Using Python TextBlob for Text Classification*. Opgehaald van Medium: https://medium.com/chat-gpt-now-writes-all-my-articles/using-python-textblob-for-text-classification-7953014f54e6

Rakita, M. (2021, November 20). *PyCharm doesn't resolve anything under tensorflow.keras #53144*. Opgehaald van GitHub: https://github.com/tensorflow/tensorflow/issues/53144#issuecomment-2524058323

Streamlit. (2024, August 27). *Basic concepts of Streamlit*. Opgehaald van docs.streamlit.io: https://docs.streamlit.io/get-started/fundamentals/main-concepts

Streamlit. (2024). *st.columns*. Opgehaald van Streamlit docs: https://docs.streamlit.io/develop/api-reference/layout/st.columns

Talaj, T. (2024, February 6). *Most Commonly Used Python Libraries*. Opgehaald van Medium: https://medium.com/pythoneers/most-commonly-used-python-libraries-00a984919543#:~:text=Matplotlib%20is%20widely%20recognized%20as,and%20our%20data%20confirms%20this.

Tilwankar, T. (2015, September 15). *ImportError: No module named 'keras.models'; 'keras' is not a package #687*. Opgehaald van GitHub: https://github.com/keras-team/keras/issues/687#issuecomment-140506372

Wang, M. (2022, September 16). *WordClouds with Python*. Opgehaald van Medium: https://medium.com/@m3redithw/wordclouds-with-python-c287887acc8b

Zhang, M. (2022, May 9). *TypeError: Unable to convert function return value to a Python type! The signature was () -> handle*. Opgehaald van Stack Overflow: https://stackoverflow.com/a/72165870