**AI Documentation**

1. **Project Breakdown:**

**You said:** I have attached my group project instruction with this chat breakdown what we have to do

**Chat GPT:** Thank you for sharing the full instructions for **Mini-Project 3**. Here's a **clear and actionable breakdown** of what you and your team need to do:

**✅ Mini-Project 3 Checklist**

**1. Getting Started**

* **Fork the repository**: One team member forks [this GitHub repo](https://github.com/openiti/fASDH25-portfolio3).
* **Add collaborators** to the fork on GitHub.
* **Each member clones** the forked repo locally.
* **Create a Trello board**, invite instructors (Mathew, Peter, Masoumeh).
* Use the **data folder** (contains CSVs and original articles).

**📁 Project Structure Guidelines**

* Use folders like this:

bash

CopyEdit

/data

/scripts

/exploration

/presentation

/visualisations

/docs (contains report, AI documentation, etc.)

**🔬 Tasks Per Team Member**

Each team member must:

* Work with **a different dataset** (lengths, n-grams, TF-IDF, topic modelling).
* Produce **2 scripts**:
  + **Exploration script** (e.g., filtering, initial plots, trying out ideas)
  + **Presentation script** (clean final visualisation with comments)
* Produce at least **one polished Plotly visualisation**
* Write a **1-page report section** explaining:
  + What dataset they used
  + What the method is (e.g. topic modelling, n-grams)
  + How their visualisation reveals insights
  + Why their visualisation method was appropriate
  + Critically reflect on limitations

**📊 Suggestions for Each Dataset**

**📏 Article Lengths**

* Show change over time (year/month)
* Visualise: bar chart, line plot

**🔤 N-grams (1, 2, 3)**

* Show popular terms by month/year
* Compare frequency of terms across time
* Visualise: heatmap, bar chart, line plot

**📈 TF-IDF Similarity**

* Cluster articles by similarity
* Show articles getting closer/further over time
* Visualise: heatmap, scatter plot (with PCA)

**📚 Topic Modelling**

* Show prevalence of topics over time
* Identify emerging/disappearing topics
* Visualise: stacked area chart, line graph

**📁 Bonus Option (10 Points)**

* Build your own dataset from the **article texts**
* Example ideas:
  + Sentiment analysis by month/year
  + Named Entity Recognition (NER)
  + Custom topic modelling (LDA)
* Include the script used to generate this dataset
* Use this new dataset for one member’s analysis

**🧾 Final Report (Word Doc, Submitted via Turnitin)**

Structure:

1. **Introduction (½ – 1 page)**
   * Overview of the corpus and the key argument/theme your findings support
2. **Documentation (1–2 pages)**
   * Explain project steps and assign team member responsibilities
   * Explain what each script does (but not line-by-line; that's for code comments)
3. **Data Set Sections (1 page per member, per dataset)**
   * Description of method (e.g., what is TF-IDF or topic modelling?)
   * Description of visualisation
   * Analysis of what is revealed
   * Include screenshot(s) of visualisation
4. **Conclusion (½ – 1 page)**
   * Bring all findings together
   * Restate and support your main argument

**📌 GitHub Submission**

* Push everything to GitHub fork (frequent commits, good messages)
* Include:
  + All scripts (exploration & presentation)
  + Any new datasets created
  + All visualisations
  + AI usage documentation
* Before submission:
  + Remove /data folder (if you didn't produce it)
  + Zip project folder
  + Submit zipped folder on Moodle
  + Submit report on Turnitin

**🤖 AI Usage Policy**

Allowed:

* Ask for code snippets
* Ask for help fixing bugs or understanding methods

Not allowed:

* Let AI write full scripts
* Use unexplained code

**Track AI use in**:  
/docs/AI\_documentation\_<your\_name>.docx  
Each entry should include:

1. Prompt
2. AI response
3. Your reflection on use

If you'd like help brainstorming visualisation ideas for your specific dataset (e.g., n-grams or topic modelling), or drafting your report section, just ask!

1. **Code error issue:**

**You said:** Traceback (most recent call last): File "C:/Users/DELL/Desktop/topic modeling portfolio 3.py", line 9, in <module> df = pd.read\_csv("dataframes/topic-model/topic-model.csv") File "C:\Users\DELL\AppData\Local\Programs\Python\Python313\Lib\site-packages\pandas\io\parsers\readers.py", line 1026, in read\_csv return \_read(filepath\_or\_buffer, kwds) File "C:\Users\DELL\AppData\Local\Programs\Python\Python313\Lib\site-packages\pandas\io\parsers\readers.py", line 620, in \_read parser = TextFileReader(filepath\_or\_buffer, \*\*kwds) File "C:\Users\DELL\AppData\Local\Programs\Python\Python313\Lib\site-packages\pandas\io\parsers\readers.py", line 1620, in \_\_init\_\_ self.\_engine = self.\_make\_engine(f, self.engine) File "C:\Users\DELL\AppData\Local\Programs\Python\Python313\Lib\site-packages\pandas\io\parsers\readers.py", line 1880, in \_make\_engine self.handles = get\_handle( File "C:\Users\DELL\AppData\Local\Programs\Python\Python313\Lib\site-packages\pandas\io\common.py", line 873, in get\_handle handle = open( FileNotFoundError: [Errno 2] No such file or directory: 'dataframes/topic-model/topic-model.csv' when i run above code this error is occuring.

**ChatGPT said:**

The error you're seeing:

vbnet

CopyEdit

FileNotFoundError: [Errno 2] No such file or directory: 'dataframes/topic-model/topic-model.csv'

means that Python cannot find the file at the path you provided:  
dataframes/topic-model/topic-model.csv

**Here's how to fix it:**

**✅ Check these things:**

1. **Correct File Path**  
   Make sure the file topic-model.csv actually exists on your computer at:

swift

CopyEdit

C:/Users/DELL/Desktop/dataframes/topic-model/topic-model.csv

If it's directly on your Desktop, change your code to:

python

CopyEdit

df = pd.read\_csv("C:/Users/DELL/Desktop/topic-model.csv")

1. **Use an Absolute Path** (recommended)  
   Try changing your code to:

python

CopyEdit

df = pd.read\_csv("C:/Users/DELL/Desktop/dataframes/topic-model/topic-model.csv")

1. **Check Spelling and Case**  
   Ensure the folder names and file name are exactly correct — even a small typo will cause this error.
2. **File Extension**  
   Confirm the file really is .csv and not accidentally .csv.txt or something else. Enable file extensions in File Explorer if needed.

Let me know if you'd like help verifying the correct path — you can even drag and drop the file here and I’ll give you the exact path to use in your code.

1. **Os module problem**

**You said:** Traceback (most recent call last): File "C:/Users/DELL/Downloads/FASDH25-portfolio3/scripts/Final\_Script (topic modeling).py", line 25, in <module> os.makedirs(output\_dir, exist\_ok=True) NameError: name 'os' is not defined. Did you forget to import 'os'? .... i have added the save codes but it is showing this error

**ChatGPT said:**

Yes — the error is because the os module is used but **was not imported**.

**✅ Fix:**

At the top of your script, add:

python

CopyEdit

import os

So your imports at the top should now look like this:

python

CopyEdit

import pandas as pd

import plotly.express as px

import os # <-- Add this line

Once you add that, the error will disappear and the script will successfully save all four figures in your output/ folder. Let me know if you’d like to also save them as PNGs or auto-open them in the browser.

1. **Stop words:**

**You said:** I used NLTK to remove stop words from the topic keyword columns. Can you help me explain why this step was important in simple terms?

**Chat GPT:** Removing stop words helps clean the data by eliminating common but meaningless words like "the," "is," and "and" that do not help identify real topics. This makes topic modeling more accurate and ensures that visualizations reflect meaningful content.

1. **Exporting Data and Graphs**

For visualization, ChatGPT helped me automate:

* Saving each filtered dataset as a CSV
* Exporting each graph as an interactive HTML file
* Organizing outputs in folders by topic or graph type

1. **Topic label problem while doing top5 key words**

**You said:** Traceback (most recent call last):

File "C:/Users/DELL/Downloads/FASDH25-portfolio3/scripts/Final\_Script (topic modeling).py", line 54, in <module>

fig3 = px.bar(top5\_keywords\_df, x="Frequency", y="Keyword", color="Keyword", facet\_col="Topic\_Label", facet\_col\_wrap=2, orientation="h", height=800, title="Top 5 Keywords per Topic (Horizontal View)")

File "C:\Users\DELL\AppData\Local\Programs\Python\Python313\Lib\site-packages\plotly\express\\_chart\_types.py", line 381, in bar

return make\_figure(

File "C:\Users\DELL\AppData\Local\Programs\Python\Python313\Lib\site-packages\plotly\express\\_core.py", line 2735, in make\_figure

fig = init\_figure(

File "C:\Users\DELL\AppData\Local\Programs\Python\Python313\Lib\site-packages\plotly\express\\_core.py", line 2888, in init\_figure

\_spacing\_error\_translator(e, "Vertical", "facet\_row\_spacing")

File "C:\Users\DELL\AppData\Local\Programs\Python\Python313\Lib\site-packages\plotly\express\\_core.py", line 2867, in \_spacing\_error\_translator

raise e

File "C:\Users\DELL\AppData\Local\Programs\Python\Python313\Lib\site-packages\plotly\express\\_core.py", line 2871, in init\_figure

fig = make\_subplots(

File "C:\Users\DELL\AppData\Local\Programs\Python\Python313\Lib\site-packages\plotly\\_subplots.py", line 565, in make\_subplots

\_check\_hv\_spacing(rows, vertical\_spacing, "Vertical", "rows", "rows")

File "C:\Users\DELL\AppData\Local\Programs\Python\Python313\Lib\site-packages\plotly\\_subplots.py", line 538, in \_check\_hv\_spacing

raise ValueError(

ValueError: Vertical spacing cannot be greater than (1 / (rows - 1)) = 0.026316.

The resulting plot would have 39 rows (rows=39).

Use the facet\_row\_spacing argument to adjust this spacing.

**Chat GPT:** Thanks for sharing the full traceback — this error occurs because **you are faceting by Topic\_Label**, and there are **too many unique values** (39 of them). Plotly Express tries to create a subplot for each Topic\_Label, which exceeds vertical space limits.

**🔍 Why this happens:**

In this part of your code:

fig4 = px.bar(top5\_keywords\_df, x="Frequency", y="Keyword", color="Keyword",

facet\_col="Topic\_Label", facet\_col\_wrap=2, orientation="h",

height=800, title="Top 5 Keywords per Topic (Horizontal View)")

You're faceting horizontally using facet\_col="Topic\_Label", which will try to create **39 subplots** across two columns — this causes too many rows and hits a spacing limit.

**✅ Recommended Fix Options**

**✅ Option 1: Only plot top 5 topics (not all 39)**

Filter top5\_keywords\_df to only include the top 5 topics — same way you did earlier:

top5\_keywords\_df = top5\_keywords\_df[top5\_keywords\_df["Topic\_Label"].isin(df\_top\_topics["Topic\_Label"].unique())]

Place this line **just before** creating fig4.

So full fix would be:

# Filter keywords to top 5 topics only (same as used in fig3)

top5\_keywords\_df = top5\_keywords\_df[top5\_keywords\_df["Topic\_Label"].isin(df\_top\_topics["Topic\_Label"].unique())]

fig4 = px.bar(top5\_keywords\_df, x="Frequency", y="Keyword", color="Keyword",

facet\_col="Topic\_Label", facet\_col\_wrap=2, orientation="h",

height=800, title="Top 5 Keywords per Topic (Horizontal View)")

fig4.show()

This will reduce the number of subplots and eliminate the spacing error.

**✅ Option 2: Don’t facet at all — show all in one chart**

Just plot a grouped bar chart:

fig4 = px.bar(top5\_keywords\_df, x="Frequency", y="Keyword", color="Topic\_Label",

orientation="h", height=800, title="Top 5 Keywords per Topic (Grouped View)")

fig4.show()

Let me know which you'd prefer (top 5 only vs all together) — I can help optimize the layout or colors too.