Final Project

5/28/2025

100 Points Possible

Attempt 1

In Progress

NEXT UP: Submit Assignment

Add Comment

Unlimited Attempts Allowed

4/29/2025

Details

In March of 2023, **Goldman Sachs published a report** (https://www.aei.org/articles/why-goldman-sachs-thinks-generative-ai-could-have-a-huge-impact-on-economic-growth-and-productivity/) , indicating that ~25% of the tasks in US and Europe can be automated using AI. However, as you can see in **this visualization** (https://substackcdn.com/image/fetch/f_auto,q_auto:good,fl_progressive:steep/https%3A%2F%2Fsubstack-post-media.s3.amazonaws.com%2Fpublic%2Fimages%2F20fefbbb-4dbb-41fe-bd6f-5f7e856abd52_1169x600.jpeg) , not all industries will be affected equally. According to the report, certain jobs, like office tasks, legal, architecture, and social sciences have a potential for 30%+ automation, while positions like construction, installation, and building maintenance are going to be largely unaffected.

You can also find supporting evidence in the **Facebook Research paper** (https://ai.facebook.com/blog/robots-learning-video-simulation-artificial-visual-cortex-vc-1/?utm_source=linkedin&utm_medium=organic_social&utm_campaign=research&utm_content=video) , which highlights Moravec's Paradox. This thesis posits that the hardest problems in AI involve sensorimotor skills rather than abstract thought or reasoning. Notably, these findings coincide with predictions made by Goldman Sachs.

For this final project, I have prepared a collection of ~200K news articles on our favorite topics, data science, machine learning, and artificial intelligence. Your task is to **identify what industries are going to be most impacted by AI over the next several years, based on the information/insights you can extract from this text corpus.**

Your goal is to provide **actionable recommendations** on what can be done with AI to automate the jobs, improve employee productivity, and generally make AI adoption successful. Please pay attention to the introduction of novel technologies and algorithms, such as AI for image generation and Conversational AI, as they represent the entire paradigm shift in adoption of AI technologies and data science in general.

You can access the data by running the following command in a Jupyter notebook:

```
import pandas as pd
df_news_final_project = pd.read_parquet('https://storage.googleapis.com/msca-bdp-data-
```

```
open/news_final_project/news_final_project.parquet', engine='pyarrow')
df_news_final_project.shape
```

Consider the following steps as your brainstorm how you approach this:

- Become one with the data: read, inspect, and and profile it to understand what you're working with
- Sanitize the data of web crawl remnants
- Discard irrelevant articles
- Detect major topics and draw connections to different industries / jobs
- Identify top industries / job that experienced successful or unsuccessful AI integration (think sentiment analysis)
  - Suggest why certain jobs are more likely to be impacted by AI
  - Plot a timeline to illustrate how sentiment is changing over time
- Identify technologies and AI solutions that might be affecting the employment landscape
  - Plot a timeline to illustrate the introduction of some of these technologies
- Demonstrate what companies, academic institutions, and government entities can do to accelerate the development of these transformative capabilities
- Leverage appropriate NLP techniques to identify organizations, people, and locations. Then, apply sentiment analysis.
  - What types of companies are planning to invest in these technologies today or near future (success stories)?
  - What types of applications cannot by transformed by AI, based on the state of technology (failures)?
  - Can you provide text extracts / summarizations that support the answers to the above questions?

Some additional guidance:

- Don't start writing code until you've created a plan for yourself for how you're going to approach this project (what order will you do tasks in, what techniques apply to each task, how the results of the different components can be used together)
- For long-running code, test for bugs on samples of data before applying to the entire dataset
- Default sentiment will likely be wrong from any out-of-the-box package and will require a custom approach
  - Use sentiment lexicons if helpful
  - Custom models (transformer-based or not) should be trained on data that is somewhat similar to your project data
  - You can either find open source labeled data online, or get GPT3.5 (or GPT 4 if your mortgage is paid off and you're set for retirement) to label some for you. NOTE: you may not use any other models to label your data.
- You may use any combination of techniques to identify key topics:
  - Topic modeling
  - Custom classification model
  - Zero-shot modeling using transformer models
- The deliverable

- Your PowerPoint should remain under 30 slides. This is more than enough room to communicate your insights. The average is around 20 slides.
- Submit your PowerPoint as .pptx or .pdf as is**, without zipping it up.**
- If you have multiple notebooks, zip those up.
- The presentation should be clear, logical, and well organized. Use proper grammar and run spell check. Repeated typos, sloppy errors, and egregious formatting issues will be penalized.
- It is your responsibility to make sure the submitted document looks as you intend (for example, sometimes converting ppts to pdfs will shift visuals, etc)
- The slides should be self-sufficient and there should be no need to go to the notebooks for clarification.
- Your audience is your grader: someone that has an understanding of proper data science techniques and also understands communicating the results to a business. Having said this, provide enough technical information to describe methodology approaches but not so granular that it distracts from the point of the presentation.
- The slides should clearly answer all the questions and the answers should be supported by plots / tables / visuals produced in the notebooks, based on the actual data.
- Plots and visuals should be quality. No fuzzy plots, untitled plots, unreadable/overlapping labels.
- Any statements you make should be supported by the data, not outside knowledge or experience.

Grading Rubric:

Rubric	**Points**
Executive Summary w/ meaningful insights	20
Article clean-up and filtering	15
Topic detection	20
Entity extraction (people, organizations, technologies)	10
Topic-level sentiment analysis (customized)	15
Entity-level sentiment analysis	10
Visualization of sentiment analysis over time	10
Total	**100**

Choose a submission type

Upload

More

Duncan Calvert *(He/Him/His)* **(https://canvas.uchicago.edu/courses/62682/users/177889)**

AUTHOR | TA

Posted May 12 11:48pm | Last edited May 12 11:55pm

Final Project Tips and Grading Rubric

Hi Class,

Welcome to Week 8! I'm writing to share tips on the final as well as to provide some guidance on the grading rubric.

Speeding Up Your Analysis

For the final, you have been provided with a fairly large data set of ~200k news articles. Due to its size, many of the techniques that you've covered this quarter will take a long time to run. To mitigate this, you can use multiprocessing and clustered computing to speed up your processing time. Here are a few recommendations on options:

Name	Description
Macbook Pro M1-M3	If you already have a new Mac or a gaming PC with a GPU, you may be able to get through the final without using cloud computing as you can use the onboard cores to do multiprocessing.
Google Colab	Google provides 2 cores for free (for a limited time) when you switch your runtime to a T4. This likely will not be enough power for your final, but is a good place to start with a subset of data
Google Colab Pro	Colab Pro allows you to switch your runtime to a A100 GPU, offering additional cores for a limited amount of credits

Google Colab Pro+	Google Colab Pro+ allows you to use all available runtimes and additional compute credits
GCP Vertex AI Workbench	GCP is an enterprise solution to clustered computing. For the purposes of these assignments, you should use GCP's Vertex AI Workbench service to spin up a notebook with multiple cores. • Be sure to turn your notebook off when you are not using it to avoid excess charges. • I would also recommend setting up billing alerts and billing limits to avoid any unwanted large charges. Alternatives to GCP are Azure and AWS. I recommend using the one that you are most familiar/comfortable with.
Lambda Cloud	• Lambda Cloud is a small cloud provider that specializes in high GPU/CPU workloads for data science. • They often have cheaper per unit compute costs than GCP/Azure and are worth investigating if you are focused on minimizing cost.

Final Project - General Tips

- FinOps
  - I'd recommend regularly checkpointing your data after completing different processes (I.e. cleaning, topic modeling, etc.) to ensure you don't have to run processing steps multiple times
  - Start with a manageable sample of data on your local machine and only run the superset on the cloud once your code is fully debugged
  - Ensure you stop your cloud machines when not in use
  - Ensure you set up billing alerts
  - Ensure you set up billing budgets and auto shutoffs
- Data
  - The data that you've been provided is messy. Read, inspect, and profile it to understand what you're working with
  - Don't try to get to 100% clean data, instead, clean the majority of data issues and then move on to analysis

Final Project - Grading Rubric Guidance

The below is in addition to the guidance provided on the final assignment page in Canvas. If you have any questions regarding the rubric or potential approaches, please reach out to me via email or office hours.

- General:
  - Create a PowerPoint deck that is high quality enough that you would feel comfortable presenting it to an executive leader at your company
    - Make sure your visualizations are clear, well-formatted, have labeled axes, legends, and support the main point that you are trying to make on your slide
    - Don't include code screenshots as a general rule in PowerPoint, sadly executives don't care about how beautiful your code is
    - Use slide numbers, titles, and other general best practice slide formatting techniques
  - Defend your arguments - don't just say "I did X approach", explain your rationale for why it was the best approach to take
- Article clean-up and filtering:
  - Describe the process that you used to clean and filter your data.
    - Ordinal graphical visualizations are a good way to show this (i.e. DAGs, Filter diagrams, etc.)
  - If you decide to discard large amounts of data, ensure that you explain how you were confident that the discards were irrelevant articles
    - For example, random sampling 10,000 articles is not a valid approach unless you can explain how you know that 10,000 is the right sample size.
- Executive Summary w/ Meaningful Insights:
  - This slide should do the following:
    - Answer the central question of your analysis (i.e. "identify what industries are going to be most impacted by AI over the next several years, based on the information/insights you can extract from this text corpus.")
    - Provide **actionable recommendations** on what can be done with AI to automate jobs, improve employee productivity, and generally make AI adoption successful
    - Provide enough of a summary that if someone were only to read this single slide, they would still understand the main findings of your presentation
    - Do all of the above in a clean, visually appealing format that is easy for an executive to digest quickly.
  - I recommend looking up examples of executive summary slides online to get a sense of the format
- Topic detection
  - Use a combination of different topic modeling algorithms to identify topics in your data set to support your analysis
  - Be sure to provide human interpretable topic names and analysis
  - Provide commentary on if there are any outliers, trends, etc.
- Entity extraction (people, organizations, technologies):
  - Identify the top people, orgs, and technologies related to the central question of the analysis
  - Visualize the occurrences of entities
  - Provide commentary on if there are any outliers, trends, etc.

    - Describe why you used a specific modeling approach and why it's the right one
- Topic-level sentiment analysis (customized):
    - Visualize the topic-level sentiment
    - Provide commentary on if there are any outliers, trends, etc.
    - Describe why you used a specific modeling approach and why you're confident that it's the right one
- Entity-level sentiment analysis:
    - Visualize the sentiment for the most important entities you extracted earlier
    - Provide commentary on if there are any outliers, trends, etc.
- Visualization of sentiment analysis over time
    - Provide a visualization of sentiment over time.
    - Discuss any major turning points, trends, outliers, etc.
    - Did the introduction of any technologies sharply impact sentiment?
    - Discuss why you aggregated to a specific time interval.

Good luck with the analysis and please reach out if you have any questions!

Best,

Duncan

Reply