

Final Project Data Science

3/7/2024

Late **90.5/100 Points**

Attempt 1



Review Feedback
3/8/2024

Attempt 1 Score:
90.5/100



View Feedback

Anonymous Grading: **no**

Unlimited Attempts Allowed

2/14/2024 to 3/13/2024

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/) (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) (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 see 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) (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, the thesis that the hardest problems in AI involve sensorimotor skills, not abstract thought or reasoning, which coincide with Goldman Sachs predictions.

While both of these papers are very impressive, they also heavily influenced by the recent advances in Large Language Models (LLMs). For this final project I have prepared a collection of ~200K news articles (about 900 MB) articles on our favorite topics: Data Science, Machine Learning and Artificial Intelligence, and I want you to identify what industries and job lines are going to be most impacted by the AI over the next several years, based on the information you can disseminate from this text corpus.

The objective of your final project is to identify what types of tasks and jobs are most likely to see the biggest impact from AI by extracting **meaningful insights** from unstructured text. Your goal is to provide **actionable recommendations** on what can be done with AI to automate the jobs and / or improve employee productivity. 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 using one of the following methods:

- Download data by following this link from your browser: https://storage.googleapis.com/msca-bdp-data-open/news_final_project/news_final_project.parquet (https://storage.googleapis.com/msca-bdp-data-open/news_final_project/news_final_project.parquet)

- Use Pandas from anywhere (your laptop, Colab or any cloud) `df_news_final_project = pd.read_parquet('https://storage.googleapis.com/msca-bdp-data-open/news_final_project/news_final_project.parquet', engine='pyarrow')`
 - [NLP GCP 11.1 Final Project Starter.ipynb](https://canvas.uchicago.edu/courses/54288/files/10500307?wrap=1) (<https://canvas.uchicago.edu/courses/54288/files/10500307?wrap=1>)  https://canvas.uchicago.edu/courses/54288/files/10500307/download?download_frd=1 **Note:** *this is live data, so the layout and record counts in your dataframe will vary from the counts in the attached notebook*

To complete your assignment, I suggest considering the following steps:

- Clean-up the noise, by eliminating newlines, tabs, remnants of web crawls, and other irrelevant text
- Discard irrelevant articles
- Detect major topics
- Identify top candidates for AI integration - these can be related to any industry and yield positive or negative results (sentiment analysis).
 - Suggest why certain types of jobs are more likely to be impacted by AI
 - Plot a timeline to illustrate how the sentiment is changing over time
- Identify new 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 targeted sentiment
 - What types of companies (based on the lines of business) are planning to invest in these technologies today or near future (success stories)?
 - Create appropriate visualization to summarize your recommendations (i.e. word cloud chart or bubble chart)
 - What types of applications cannot currently be transformed by AI, based on today's state of technology (failures)?
 - Create appropriate visualization to summarize your recommendations (i.e. word cloud chart or bubble chart)

Additional guidance:







- Clean-up or sample data if you need to shorten processing times or reduce memory usage
- Default sentiment will likely be wrong from any software package and will require some tweaking
 - Keyword / dictionary approach
 - Data annotation and development of custom classifier
 - Building custom model on open-source data (i.e. Yelp)
 - Fine-tuning Transformer Pipeline
- You are encouraged to explore a combination several techniques to identify key topics:
 - Topic modeling (i.e. LDA using gensim or ktrain) or using BERTopic
 - Classification (hand-label several topics on a sample and then train classifier)
 - Clustering (cluster topics around pre-selected keywords or word vectors)

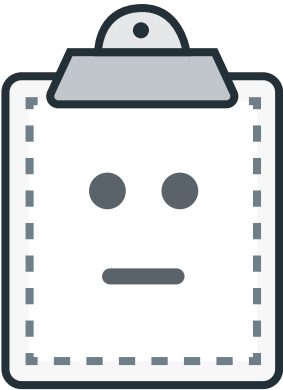
- Zero-shot (NLI) modeling
- Please ensure your PowerPoint presentation (in PPTX or PDF format) is submitted to the course module as-is (not zipped). Otherwise we are unable to use Canvas SpeedGrader.
- The presentation should look professional – not a collection of screenshots from your analytical software
- Roughly 8-12 pages is reasonable for this kind of project but there are no strict restrictions.
- On your slides you will want to provide:
 - Executive Summary
 - Methodology and source data overview
 - Actionable recommendations
 - Apply text summarization algorithms where possible to synthesize your insights
- Please submit your actual program codes (Jupyter notebooks) along with your PowerPoint
- The slides should be self-sufficient and after reading them, there should not be any need to read the notebook (we are still asking you to provide the notebooks as a proof or work though).
- The slides should clearly answer all the questions and the answers should be supported with the plots/tables/numbers produced in the notebook based on the actual data.
- The slides should contain the RIGHT amount of supporting material for each question, putting too much supporting material is as bad as putting too little: too much - you would not be able to keep the audience attention and your presentation would be a mess, too little - your statements would not look convincing.
- Everything should be clear, logical, well organized, as simple as possible. Use proper English grammar and run spell check.
- All the plots should be of production quality and easily readable. Fuzzy plots, untitled plots, unreadable labels, overlapping labels are unacceptable.
If you formatting somehow gets corrupted when you put your slides into Canvas (sometimes it happens), it is your responsibility to fix it. For example, try saving it in some other format like PDF, HTML.
- Any statements you make should be supported by data. Only recommendations or goals of the project sections can contain elements not directly supported by the data
- Please submit your actual program codes (i.e. Python Notebook) along with your PowerPoint – as a separate attachment
 - Your presentation should be targeted toward business audience and must not contain any code snippets
- You are welcome to use any software packages of your choice to complete the assignment

Grading Rubric:

| Rubric | Points |
|--|--------|
| Executive Summary with meaningful insights | 20 |
| Visualize actionable recommendations | 10 |
| Article clean-up and filtering | 10 |
| Topic detection | 20 |
| Sentiment analysis (explicitly customized) | 10 |
| Sentiment over time analysis and visualization | 10 |

| | |
|--|-----|
| Entity (organizations and people) identification | 10 |
| Targeted (entity) sentiment identification | 10 |
| Total Points | 100 |

| File Name | | Size | |
|---|---------------------------------------|---------|---|
|  | NLP_Final_Part1.ipynb | 221 KB |  |
|  | NLP_Final...tion.pptx | 3.97 MB |  |
|  | NLP_Final_Part2.ipynb | 10.3 MB |  |



Preview Unavailable
NLP_Final_Part1.ipynb

 [Download](#)

(https://canvas.uchicago.edu/files/10993331/download?download_frd=1&verifier=2UhWyPhAcpH3qj5dGss0Mg2jQcPEqr1vU2Y0sp4S)