DSCI 560 – Data Science Professional Practicum

Lab-2: Impact of Artificial Intelligence on Employment and Labor Markets

Team Details

Team Name: ILR

Member 1 – Lance Vijil Dsilva USC ID: 3824765644
Member 2 - Rafayel Mirijanyan– USC ID: 3487192016
Member 3 - Isabella Yoo – USC ID: 1305966908

Demo Video

Demo video link will be provided.

GitHub Repository

https://github.com/Lance-Dsilva/DSCI_560-Data-Science-Professional-Practicum-/tree/main/lab2

Objective

Artificial Intelligence (AI) is rapidly transforming labor markets across industries. While AI drives productivity and innovation, it also raises concerns about job displacement, skill gaps, and workforce restructuring. Understanding the real-world impact of AI on employment requires analyzing data from multiple sources, including structured datasets, reports, and online publications.

The objective of this project is to collect, extract, and organize employment related data from heterogeneous sources (CSV, PDF, and HTML) to support future analysis on how AI affects jobs, layoffs, and workforce trends.

Domain Selection & Team Decision

1. Individual Shortlisted Options

Each team member proposed a distinct domain based on data availability and potential use cases:

- **Lance (Healthcare/Medical):** Proposed focusing on clinical data and medical trends.
    - *Reasoning:* High impact, but potential challenges with data privacy (HIPAA) and sensitive information.

Dataset: Weekly Hospital Respiratory Data (HRD) Metrics by Jurisdiction, National Healthcare Safety Network (NHSN)

https://data.cdc.gov/Public-Health-Surveillance/Weekly-Hospital-Respiratory-Data-HRD-Metrics-by-Ju/ua7e-t2fy/data_preview

CDC RESP-NET dashboard (COVID / Flu / RSV hospitalization rates by age, state, season)
https://www.cdc.gov/fluview/surveillance/2026-week-01.html

CDC FluView weekly surveillance report (weekly updated page with key stats)

https://www.cdc.gov/fluview/surveillance/2026-week-01.html

Reuters: Flu cases rise across U.S. as holiday travel fuels spread

https://www.reuters.com/business/healthcare-pharmaceuticals/flu-cases-rise-across-us-holiday-travel-fuels-spread-2025-12-31/

Weekly Hospital Respiratory Data (HRD) Metrics by Jurisdiction, National Healthcare Safety Network (NHSN)

/eek ...	Geogr...	Numb...	Numb...	Numb...	Numb...	Numb...	Numb...	Numb...	Numb...	Numb...	Numb...	Numb...	Numb...	Numb...	Numb...	Total ...
020-08-08	AZ	8,086	5,671	133	8,407	6,453	381	2,211	989	182	1,813	975	149	2,166	13	2,179
020-08-15	AZ	8,008	5,500	206	8,122	6,089	370	1,941	960	183	1,605	933	149	1,534	10	1,544
020-08-22	AZ	9,397	6,612	500	8,255	6,253	385	2,378	1,186	184	1,661	923	157	1,208	11	1,219
020-08-29	AZ	10,547	7,801	507	8,178	6,194	394	2,201	1,444	181	1,351	872	155	944	8	951
020-09-05	AZ	10,222	7,477	508	7,872	5,999	403	2,292	1,366	183	1,425	820	160	712	8	719
020-09-12	AZ	10,682	7,967	518	8,095	6,239	331	2,194	1,474	195	1,270	845	142	436	10	446
020-09-19	AZ	10,719	7,961	512	8,399	6,477	423	2,330	1,461	187	1,407	836	163	409	5	415
020-09-26	AZ	10,836	7,902	711	8,405	6,499	395	2,255	1,383	185	1,337	807	155	406	7	413

- **Isabella (Project Management):** Suggested analyzing organizational workflows and efficiency.
    - *Reasoning:* Practical for corporate services, but data is often proprietary and difficult to find in public repositories.

CSV: Project Management DataSet Example, containing project name, status, priority, and task id in a structured data format. This dataset represents the state and progress of collaborative projects. It supports the AI assistant's ability to filter tasks or projects by their priority or pending status.

ASCII: Enron Email Dataset. Although Enron Email Dataset is distributed in CSV format, the actual email contents with subject and bodies are in ASCII text stored within the same column. Extracted text contents from the file will serve as examples of real-world, professional communication logs among team members. This dataset helps the AI assistant learn how to perform natural language processing and identify key information from the plain texts.

PDF: Guidelines for Project Proposals. Project guideline and description often described in a pdf file in an academic environment. It supports the AI assistant's ability to summarize key tasks and requirements in the project, generate possible milestones, or alarm users about deadlines.

- **Rafayel (AI & Workforce Dynamics):** Proposed datasets covering employment trends, tech layoffs, and the influence of automation.
    - *Reasoning:* Highly relevant to the current tech climate with a vast amount of publicly available structured and unstructured data.

Dataset: In further section

Describe what might be
missing in these existing chatbots. Discuss how your dataset might improve the overall performance and correctness.

2. Final Selection: Artificial Intelligence and Labor Markets

The team reached a consensus to focus on the **AI and Employment** domain. This decision was based on several strategic factors:

- **Data Diversity:** The domain allows for a robust pipeline that handles multiple formats:
    - **Structured:** Global layoff metrics (CSV).
    - **Semi-Structured:** Academic research papers and reports (PDF).
    - **Unstructured:** Real-time news articles and forum discussions (HTML).
- **Relevance to Final Project:** This domain provides the "Knowledge Base" necessary for the group's final AI agent. It enables the agent to answer complex student queries regarding job trends, skills demand, and market shifts.

Describe what might be missing in these existing chatbots. Discuss how your dataset might improve the overall performance and correctness

Many existing chatbots have limited usefulness because they rely on generic training data, lack strong context retention, and are not grounded in real-world, domain-specific information. As a result, they often produce vague or outdated answers that users may not trust.

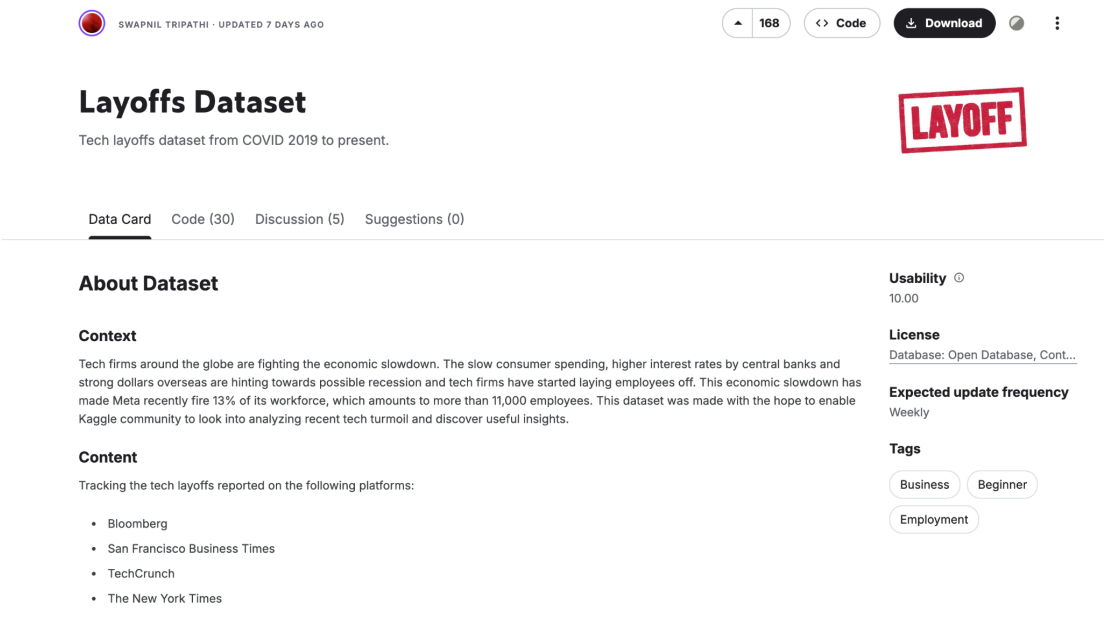
Our dataset improves chatbot performance by providing **structured, authoritative, and up-to-date data** focused on AI and employment. By grounding responses in real layoff statistics, industry trends, and expert reports, a chatbot built on this dataset can deliver more accurate, contextual, and reliable answers to student questions, improving both correctness and user confidence.

- **Technical Suitability:** The availability of APIs (Kaggle) and scrapeable research (MIT, Yale) makes it ideal for demonstrating data acquisition and cleaning techniques required for the lab.

Datasets

A) CSV Dataset – Tech Layoffs

Source: Kaggle
Link: https://www.kaggle.com/datasets/swaptr/layoffs-2022



This dataset contains global layoff information across companies, industries, locations, and time periods. It provides structured numerical data to analyze employment trends.

B) PDF Dataset – Future of Jobs Report 2025

Source: World Economic Forum

Link: https://www.weforum.org/publications/the-future-of-jobs-report-2025/

This report provides global insights on job creation, displacement, skill demand, and the impact of AI and automation on the workforce.

C) HTML Datasets – Web Articles

1. https://www.adpresearch.com/yes-ai-is-affecting-employment-heres-the-data/
2. https://news.crunchbase.com/startups/tech-layoffs/
3. https://mitsloan.mit.edu/ideas-made-to-matter/how-artificial-intelligence-impacts-us-labor-market
4. https://budgetlab.yale.edu/research/evaluating-impact-ai-labor-market-current-state-affairs

These sources provide qualitative insights and recent analysis on AI-driven employment trends.

Implementation

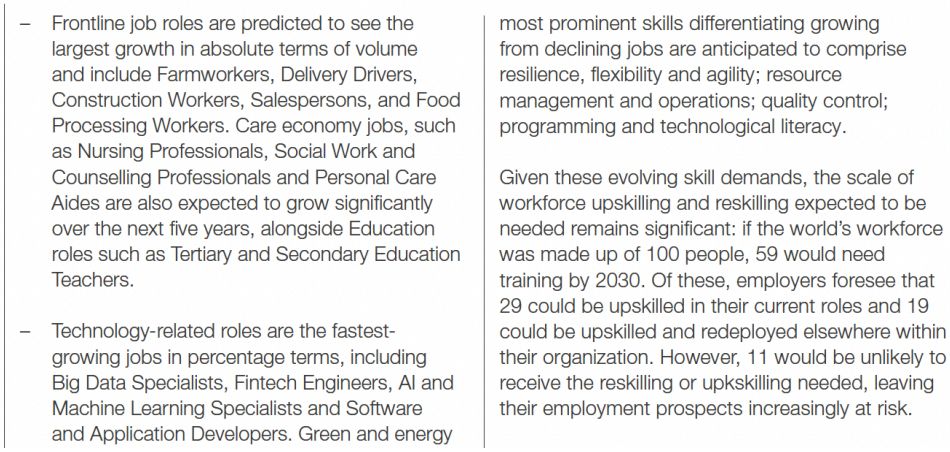
A) CSV Data Extraction

- Acquisition: Programmatically fetched the layoffs-2022 dataset using the kagglehub API.
- Localization: Utilized shutil to move the CSV from the system cache into a dedicated training_data/ directory for environment portability.
- Parsing: Loaded the dataset into Pandas and used .head(10) to sample and verify column alignment (Company, Industry, Date).

- Outlier Logic: Calculated the Z-score ($mean \pm 3 \cdot std$) on the total_laid_off column to identify and flag extreme layoff events that could skew statistical results.
- Filtering: Used select_dtypes(include=[np.number]) to isolate numerical features for correlation heatmaps and mathematical operations.

B) PDF Text Extraction (World Economic Forum Report)

To extract text from the *Future of Jobs Report 2025*, we implemented a **column-aware PDF extraction pipeline** using the pdfplumber library. Since the report follows a two-column layout, a standard line-by-line extraction would merge unrelated text across columns. To address this, the script reconstructs the logical reading order by leveraging word-level positional metadata.



– Frontline job roles are predicted to see the largest growth in absolute terms of volume and include Farmworkers, Delivery Drivers, Construction Workers, Salespersons, and Food Processing Workers. Care economy jobs, such as Nursing Professionals, Social Work and Counselling Professionals and Personal Care Aides are also expected to grow significantly over the next five years, alongside Education roles such as Tertiary and Secondary Education Teachers.

– Technology-related roles are the fastest-growing jobs in percentage terms, including Big Data Specialists, Fintech Engineers, AI and Machine Learning Specialists and Software and Application Developers. Green and energy

most prominent skills differentiating growing from declining jobs are anticipated to comprise resilience, flexibility and agility; resource management and operations; quality control; programming and technological literacy.

Given these evolving skill demands, the scale of workforce upskilling and reskilling expected to be needed remains significant: if the world's workforce was made up of 100 people, 59 would need training by 2030. Of these, employers foresee that 29 could be upskilled in their current roles and 19 could be upskilled and redeployed elsewhere within their organization. However, 11 would be unlikely to receive the reskilling or upkskilling needed, leaving their employment prospects increasingly at risk.

The extraction process groups words into lines based on vertical alignment and then separates content into left, right, and full-width sections using horizontal gap detection. This allows the script to correctly read full-width headings first, followed by left and right columns in sequence. Additional heuristics were applied to remove repeated headers and footers, merge hyphenated words split across lines, and preserve paragraph structure.

The extracted output is saved in two formats:

1. A **plain text file** containing human-readable page-by-page content
2. A **JSONL file** storing structured text along with page-level metadata

```
(base) rafayelmirijanyan@Rafayels-MacBook-Pro week 2 % python column_aware_pdf_e
xtraction.py WEF_Future_of_Jobs_Report_2025.pdf

Done.
- Text:   output.txt
- JSONL:  output_pages.jsonl
```

This approach ensures accurate reconstruction of multi-column PDF text and produces clean, logically ordered data suitable for downstream analysis.

B) Web data Text Extraction using Web scraping

- We used Selenium in headless mode to render JavaScript-heavy research sites (like MIT Sloan) that traditional scrapers cannot read, allowing the pipeline to capture data that only appears after a page fully loads.
- By disabling the AutomationControlled flag, the script hides the digital signature that identifies it as a bot; this allows the scraper to mimic a standard human browser and avoid being blocked by security filters.
- The implementation uses readability-lxml to act as a "Reader Mode" for the code, algorithmically stripping away non-essential elements like advertisements, sidebars, and navigation menus to isolate the core article text.
- We employed html2text to convert complex HTML into clean Markdown-style ASCII, which preserves document structure (like headers and lists) while removing redundant tags that would otherwise waste AI processing tokens.
- This multi-stage distillation process ensures that the final dataset is composed of high-density information, providing a "clean" and structured foundation for training an AI agent without the interference of website "fluff."

Pipeline Execution and Data Results

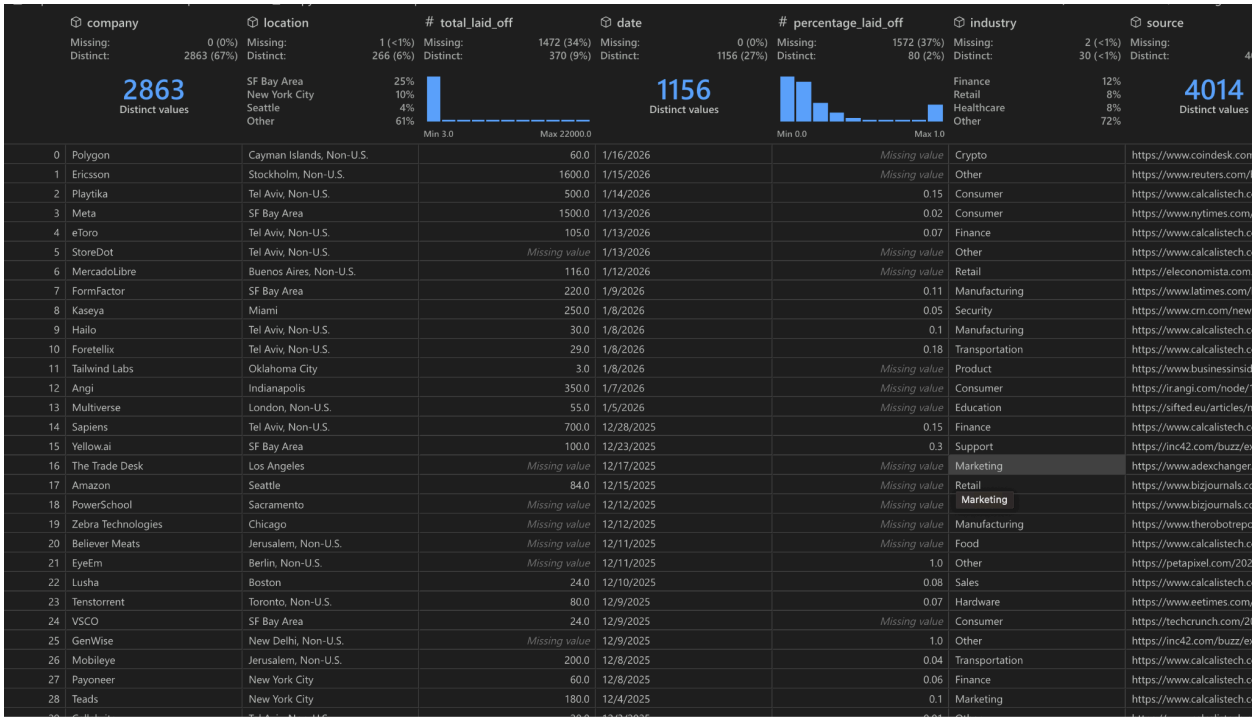
The script data_exploration.py was executed in a terminal environment to perform a complete end-to-end data acquisition and analysis workflow. The execution successfully processed structured CSV data, dynamic web content, and local PDF documentation.

```
(venv) (base) lancedsilva@Lances-MacBook-Air lab2 % python3 data_exploration.py
 STARTING A PIPELINE

[1/4] Downloading Kaggle Dataset...
 CSV saved to: training_data/layoffs.csv

[2/4] Scraping Dynamic Web Content...
🔍 Processing: https://www.adpresearch.com/yes-ai-is-affecting-employment-heres-the-data/
 Extracted: Yes_AI_is_affecting_employment_Here_s_the_data_ADP.txt
🔍 Processing: https://news.crunchbase.com/startups/tech-layoffs/
 Extracted: Tech_Layoffs_US_Companies_With_Job_Cuts_In_2024_An.txt
🔍 Processing: https://mitsloan.mit.edu/ideas-made-to-matter/how-artificial-intelligence-impacts-us-labor-market
 Extracted: How_artificial_intelligence_impacts_the_US_labor_m.txt
🔍 Processing: https://budgetlab.yale.edu/research/evaluating-impact-ai-labor-market-current-state-affairs
 Extracted: Evaluating_the_Impact_of_AI_on_the_Labor_Market_Cu.txt

[4/4]  Checking for PDFs...
 Reconstructing: WEF_Future_of_Jobs_Report_2025.pdf
PDF converted to text.
```

Data Acquisition: The pipeline initialized by downloading the layoffs.csv dataset from Kaggle and storing it in a local directory. It then utilized a headless Selenium browser to scrape four high-authority web sources, including MIT Sloan and Yale Budget Lab, converting them into clean text files.

PDF Reconstruction: A local PDF titled WEF_Future_of_Jobs_Report_2025.pdf was processed and successfully converted into a standardized text format for the training corpus.

```
[3/4]  Launching Deep Data Analysis...

---  CSV Data Preview (Top 10 Rows) ---
      company                 location  total_laid_off       date  percentage_laid_off  ...                                             source          stage funds_raised        country date_a
ded
0     Polygon  Cayman Islands, Non-U.S.            60.0  1/16/2026                  NaN  ...  https://www.coindesk.com/business/2026/01/16/p...         Unknown        451.0  Cayman Islands  1/16/
026
1     Ericsson       Stockholm, Non-U.S.          1600.0  1/15/2026                  NaN  ...  https://www.reuters.com/business/world-at-work...        Post-IPO        663.0         Sweden  1/16/
026
2     Playtika        Tel Aviv, Non-U.S.           500.0  1/14/2026                 0.15  ...  https://www.calcalistech.com/ctechnews/article...        Post-IPO          NaN         Israel  1/15/
026
3         Meta              SF Bay Area          1500.0  1/13/2026                 0.02  ...  https://www.nytimes.com/2026/01/12/technology/...        Post-IPO      26000.0  United States  1/14/
026
4        eToro        Tel Aviv, Non-U.S.           105.0  1/13/2026                 0.07  ...  https://www.calcalistech.com/ctechnews/article...         Unknown        322.0         Israel  1/14/
026
5     StoreDot        Tel Aviv, Non-U.S.             NaN  1/13/2026                  NaN  ...  https://www.calcalistech.com/ctechnews/article...        Series D        184.0         Israel  1/14/
026
6  MercadoLibre    Buenos Aires, Non-U.S.           116.0  1/12/2026                  NaN  ...  https://eleconomista.com.ar/economia/despidos-...        Post-IPO       5100.0         Brazil  1/14/
026
7    FormFactor              SF Bay Area           220.0   1/9/2026                 0.11  ...  https://www.latimes.com/business/story/2026-01...        Post-IPO         15.0  United States  1/13/
026
8       Kaseya                   Miami            250.0   1/8/2026                 0.05  ...  https://www.crn.com/news/channel-news/2026/kas...  Private Equity        547.0  United States  1/13/
026
9        Hailo        Tel Aviv, Non-U.S.            30.0   1/8/2026                 0.10  ...  https://www.calcalistech.com/ctechnews/article...        Series C        344.0         Israel  1/13/
026

[10 rows x 11 columns]
--------------------------------------

--- CSV Metrics ---
Records: 4268 | Features: 11
Extreme Layoff Events Detected: 37
```



	company	location	total_laid_off	date	percentage_laid_off	industry	source
	Missing: 0 (0%) Distinct: 2863 (67%)	Missing: 1 (<1%) Distinct: 266 (6%)	Missing: 1472 (34%) Distinct: 370 (9%)	Missing: 0 (0%) Distinct: 1156 (27%)	Missing: 1572 (37%) Distinct: 80 (2%)	Missing: 2 (<1%) Distinct: 30 (<1%)	Missing: Distinct: 4
	2863 Distinct values	SF Bay Area 25%, New York City 10%, Seattle 4%, Other 61%	Min 3.0 Max 22000.0	1156 Distinct values	Min 0.0 Max 1.0	Finance 12%, Retail 8%, Healthcare 8%, Other 72%	4014 Distinct values
0	Polygon	Cayman Islands, Non-U.S.	60.0	1/16/2026	*Missing value*	Crypto	https://www.coindesk.com
1	Ericsson	Stockholm, Non-U.S.	1600.0	1/15/2026	*Missing value*	Other	https://www.reuters.com/
2	Playtika	Tel Aviv, Non-U.S.	500.0	1/14/2026	0.15	Consumer	https://www.calcalistech.c
3	Meta	SF Bay Area	1500.0	1/13/2026	0.02	Consumer	https://www.nytimes.com/
4	eToro	Tel Aviv, Non-U.S.	105.0	1/13/2026	0.07	Finance	https://www.calcalistech.c
5	StoreDot	Tel Aviv, Non-U.S.	*Missing value*	1/13/2026	*Missing value*	Other	https://www.calcalistech.c
6	MercadoLibre	Buenos Aires, Non-U.S.	116.0	1/12/2026	*Missing value*	Retail	https://eleconomista.com
7	FormFactor	SF Bay Area	220.0	1/9/2026	0.11	Manufacturing	https://www.latimes.com/
8	Kaseya	Miami	250.0	1/8/2026	0.05	Security	https://www.crn.com/new
9	Hailo	Tel Aviv, Non-U.S.	30.0	1/8/2026	0.1	Manufacturing	https://www.calcalistech.c
10	Foretellix	Tel Aviv, Non-U.S.	29.0	1/8/2026	0.18	Transportation	https://www.calcalistech.c
11	Tailwind Labs	Oklahoma City	3.0	1/8/2026	*Missing value*	Product	https://www.businessinsid
12	Angi	Indianapolis	350.0	1/7/2026	*Missing value*	Consumer	https://ir.angi.com/node/
13	Multiverse	London, Non-U.S.	55.0	1/5/2026	*Missing value*	Education	https://sifted.eu/articles/
14	Sapiens	Tel Aviv, Non-U.S.	700.0	12/28/2025	0.15	Finance	https://www.calcalistech.c
15	Yellow.ai	SF Bay Area	100.0	12/23/2025	0.3	Support	https://inc42.com/buzz/ex
16	The Trade Desk	Los Angeles	*Missing value*	12/17/2025	*Missing value*	Marketing	https://www.adexchanger
17	Amazon	Seattle	84.0	12/15/2025	*Missing value*	Retail	https://www.bizjournals.c
18	PowerSchool	Sacramento	*Missing value*	12/12/2025	*Missing value*	Marketing	https://www.bizjournals.c
19	Zebra Technologies	Chicago	*Missing value*	12/12/2025	*Missing value*	Manufacturing	https://www.therobotrepo
20	Believer Meats	Jerusalem, Non-U.S.	*Missing value*	12/11/2025	*Missing value*	Food	https://www.calcalistech.c
21	EyeEm	Berlin, Non-U.S.	*Missing value*	12/11/2025	1.0	Other	https://petapixel.com/202
22	Lusha	Boston	24.0	12/10/2025	0.08	Sales	https://www.calcalistech.c
23	Tenstorrent	Toronto, Non-U.S.	80.0	12/9/2025	0.07	Hardware	https://www.eetimes.com/
24	VSCO	SF Bay Area	24.0	12/9/2025	*Missing value*	Consumer	https://techcrunch.com/2
25	GenWise	New Delhi, Non-U.S.	*Missing value*	12/9/2025	1.0	Other	https://inc42.com/buzz/e
26	Mobileye	Jerusalem, Non-U.S.	200.0	12/8/2025	0.04	Transportation	https://www.calcalistech.c
27	Payoneer	New York City	60.0	12/8/2025	0.06	Finance	https://www.calcalistech.c
28	Teads	New York City	180.0	12/4/2025	0.1	Marketing	https://www.calcalistech.c

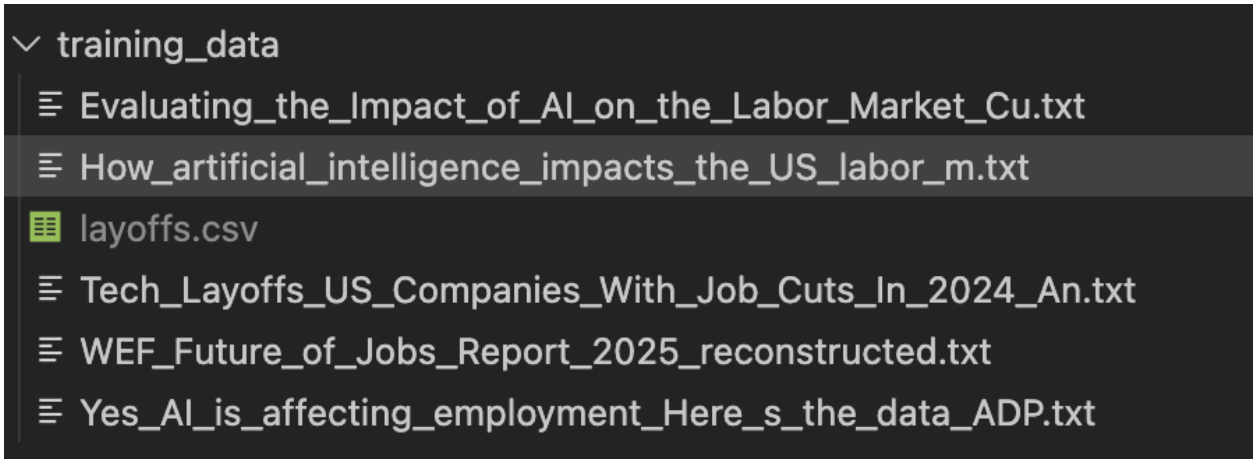
Structured Analysis (CSV): The system analyzed a dataset containing 4,268 records and 11 features. A preview of the top 10 rows revealed recent 2026 data from companies such as Meta, Ericsson, and Polygon. Statistical filtering identified 37 "Extreme Layoff Events" based on Z-score logic.

```
--- Text Corpus Linguistics ---
                                                file  word_count  lexical_richness  avg_sentence_complexity
How_artificial_intelligence_impacts_the_US_labor_m.txt        1418             0.374                    22.16
Tech_Layoffs_US_Companies_With_Job_Cuts_In_2024_An.txt        1711             0.359                    14.02
      WEF_Future_of_Jobs_Report_2025_reconstructed.txt      136395             0.033                    49.49
Evaluating_the_Impact_of_AI_on_the_Labor_Market_Cu.txt         385             0.538                    18.33
Yes_AI_is_affecting_employment_Here_s_the_data_ADP.txt         960             0.459                    16.55

Targeted Keyword Density:
- 'ai': 1147 occurrences
- 'intelligence': 46 occurrences
- 'automation': 72 occurrences
- 'layoffs': 48 occurrences
- 'jobs': 809 occurrences
- 'labor': 21 occurrences
- 'growth': 997 occurrences

 PIPELINE COMPLETE in 113.42s
```

Unstructured Analysis (NLP): A linguistic audit was performed across the text corpus. The reconstructed World Economic Forum report was the most significant contributor, with a word count of 136,395 and an average sentence complexity of 49.49.
Keyword Insights: Targeted density analysis showed a strong domain focus, with 'ai' (1,147 occurrences), 'growth' (997 occurrences), and 'jobs' (809 occurrences) emerging as the primary thematic drivers.

Data Storage

The script utilizes a single global configuration, DATA_DIR = "training_data", to house all project artifacts. This directory is automatically created at startup to ensure a consistent environment. It contains three primary types of files:



1. Structured Data (CSV)

- **`layoffs.csv`**: This is the primary dataset downloaded from Kaggle.
- **Purpose**: It serves as the "immutable" raw data source for the analysis engine to calculate layoff trends, industry distribution, and numerical correlations.

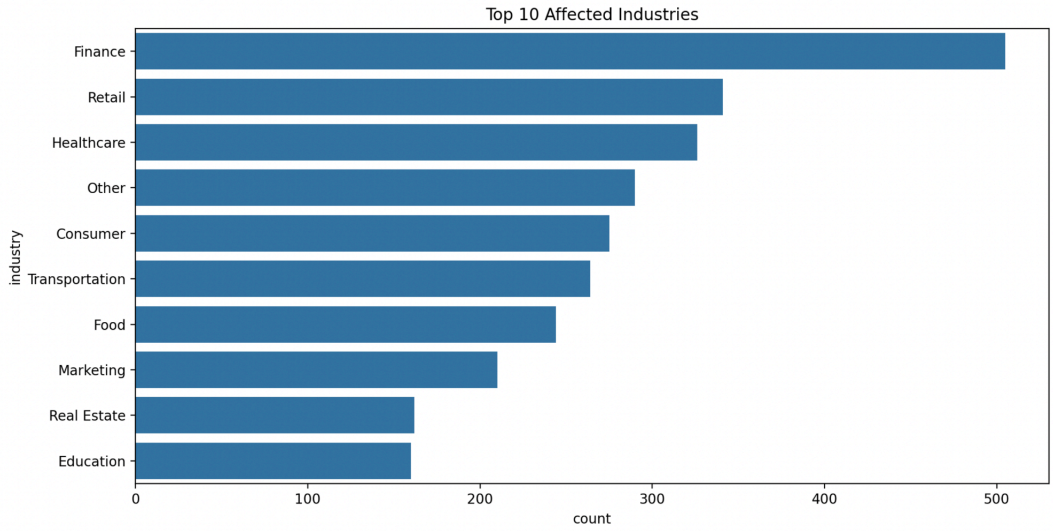
2. Unstructured Scraped Text (`.txt`)

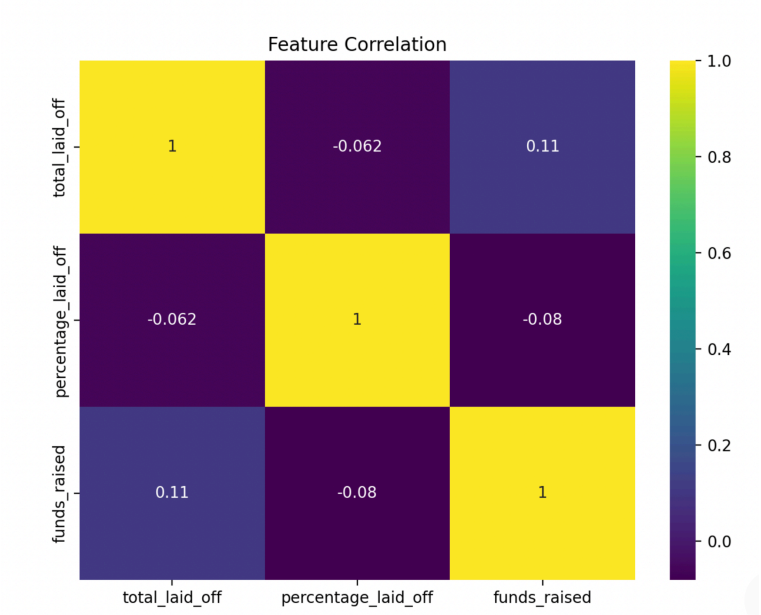
- **Article Files**: Each web article scraped by the Selenium engine is saved as a separate text file.
- **Naming Convention**: Filenames are generated using a "safe filename" logic sanitizing the article title by replacing non word characters with underscores and truncating them to 50 characters for OS compatibility.
- **Internal Content**: Each file contains the article title, the source URL, and the cleaned "main content" extracted via the readability engine.

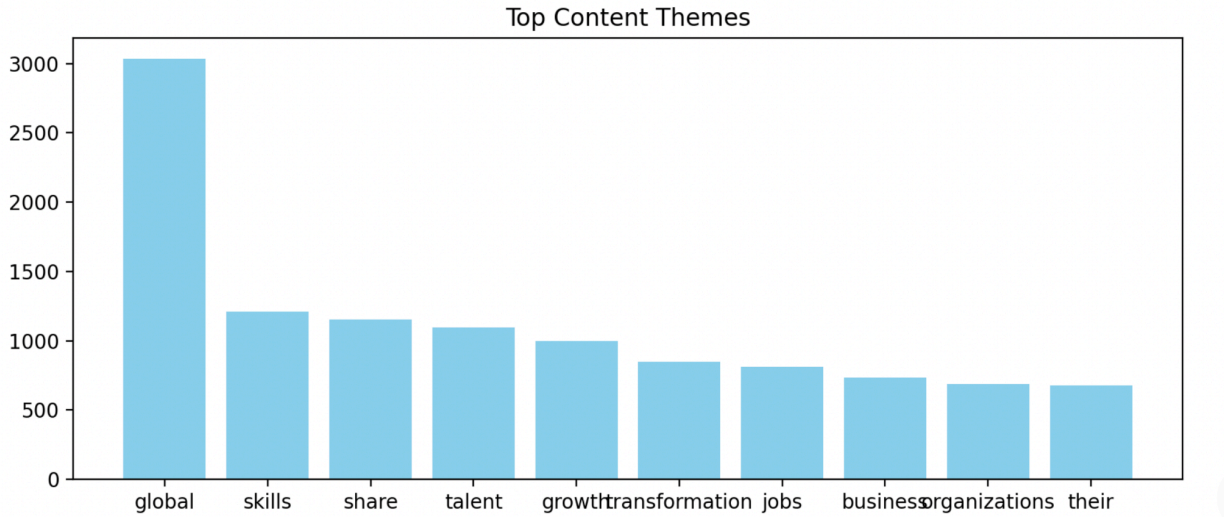
3. Reconstructed PDF Text (_reconstructed.txt)

- **PDF to Text**: Files ending in _reconstructed.txt are created during the PDF reconstruction stage.
- **Source**: These are generated from local .pdf files found in the working directory using pdfplumber to extract text page-by-page.

Data Analysis







The analysis highlights key patterns in both the textual and numerical data. The **Top Content Themes** visualization shows that terms related to *skills*, *talent*, *jobs*, and *transformation* dominate the text corpus, confirming a strong focus on workforce change driven by artificial intelligence. This aligns well with the project's goal of building an AI chatbot capable of answering student questions about AI and employment.

The **Top Affected Industries** chart ranks industries by layoff frequency, showing that sectors such as Finance, Retail, and Healthcare experience the highest number of

layoff events. This provides a clear overview of where workforce disruptions are most concentrated.

Finally, the **Feature Correlation Matrix** reveals weak correlations between total layoffs, percentage laid off, and funds raised, indicating that layoffs are not driven by a single financial factor. Together, these results demonstrate that the datasets capture both thematic context and quantitative trends, making them suitable for a question-answering AI agent focused on labor market insights.

Output

On execution, the scripts successfully extracted structured and unstructured data from all sources. Extracted data was stored locally and verified through console logs.

Individual Contributions

Member 1: CSV data extraction and exploration, Web scraping implementation
Member 2: PDF text extraction and data organization
Member 3: report creation, data exploration