CE49X
Introduction to Data Science for Civil Engineering

Final Project

Civil Engineering & AI Integration: Analyzing Industry Trends through News & Media

Boğaziçi University

Fall 2025

Dr. Eyuphan Koc

December 4, 2025

Contents

1 Project Overview **2**
1.1 Project Approach 2
1.2 Learning Objectives 2

2 Team Formation **2**

3 Task 1: Data Collection (Corpus Creation) **2**
3.1 Objective 2
3.2 Data Sources 3
3.2.1 Suggested Sources 3
3.3 Requirements 3
3.4 Deliverables for Task 1 3

4 Task 2: Text Preprocessing & NLP **3**
4.1 Objective 3
4.2 Requirements 4
4.2.1 Preprocessing Pipeline 4
4.2.2 Feature Extraction 4
4.3 Deliverables for Task 2 4

5 Task 3: Categorization & Trend Analysis **4**
5.1 Objective 4
5.2 Requirements 4
5.2.1 Dictionary-Based Classification 4
5.2.2 Analysis Logic 5
5.3 Deliverables for Task 3 5

6 Task 4: Visualization & Insights **5**
6.1 Objective 5
6.2 Requirements 5
6.3 Deliverables for Task 4 5

7 Technical Requirements **6**
7.1 Programming Languages and Tools 6

8 Final Deliverables **6**
8.1 Code Repository 6
8.2 Final Report (PDF) 6
8.3 Final Presentation 6

1 Project Overview

This final project focuses on **Natural Language Processing (NLP)** and **Trend Analysis** to investigate the intersection of Artificial Intelligence (AI) and Civil Engineering. Students will scrape, process, and analyze a large corpus of news articles, blog posts, and industry reports to determine which sub-disciplines of Civil Engineering (e.g., Structural, Geotechnical, Transportation) are most actively adopting AI technologies and for what purposes.

1.1 Project Approach

This project utilizes a data-driven text mining approach:

1. **Data Collection:** Building a dataset of relevant articles from engineering news outlets and tech media.
2. **NLP Pipeline:** Preprocessing text, extracting keywords, and performing entity recognition.
3. **Trend Analysis:** Quantifying the relationship between specific Civil Engineering domains and AI applications (e.g., "Computer Vision in Construction Safety").

1.2 Learning Objectives

By completing this project, students will:

- Implement web scraping and API-based data collection pipelines.
- Apply Natural Language Processing (NLP) techniques to unstructured text data.
- Perform Topic Modeling (e.g., LDA) to discover hidden themes.
- Visualize text data using word clouds, network graphs, and frequency heatmaps.
- Gain industry insights into the digitization of Civil Engineering.
- Communicate technical findings through data storytelling.
- Collaborate effectively in a team environment.

2 Team Formation

- **Group Size:** 1–2 students per group
- **Collaboration:** All team members must contribute equally to the project
- **Deliverables:** Each group submits one final report and code repository

3 Task 1: Data Collection (Corpus Creation)

Points: 30/100

3.1 Objective

Build a substantial dataset of textual content (news articles, press releases, technical blog posts) related to Civil Engineering and Artificial Intelligence.

3.2 Data Sources

Students should target sources that cover construction technology ("ConTech"), smart cities, and general engineering news.

3.2.1 Suggested Sources

- **Industry News Portals:** ENR (Engineering News-Record), Civil + Structural Engineer Media, Construction Dive, BIMplus.
- **Tech News (Filtered):** TechCrunch, Wired, VentureBeat (searching for specific keywords like "construction", "infrastructure", "concrete").
- **Aggregators:** Google News (via scraping or library), NewsAPI.
- **Professional Blogs:** Company blogs (e.g., Autodesk, Bentley Systems, Trimble).

3.3 Requirements

- Collect **minimum 100 unique articles/documents**.
- Keywords for search/scraping should include combinations of:
  - *Civil Engineering terms:* "Construction", "Structural", "Geotechnical", "Transportation", "Infrastructure", "Concrete", "Bridge", "Tunnel".
  - *AI terms:* "Artificial Intelligence", "Machine Learning", "Computer Vision", "Generative AI", "Neural Networks", "Robotics", "Automation".
- Each entry must include:
  - Title
  - Publication Date
  - Source/Publisher
  - Full Text Content (or detailed abstract)
  - URL
- Store data in a structured format (CSV, JSON, or SQLite database).

3.4 Deliverables for Task 1

- Web scraping/API scripts.
- Raw dataset file(s).
- A "Data Description" document listing sources and search queries used.

4 Task 2: Text Preprocessing & NLP

Points: 25/100

4.1 Objective

Clean and prepare the raw text data for analysis, extracting meaningful features and tokens.

4.2 Requirements

4.2.1 Preprocessing Pipeline

Implement a standard NLP cleaning pipeline:

- **Tokenization:** Splitting text into words/sentences.
- **Normalization:** Lowercasing, removing punctuation and special characters.
- **Stopword Removal:** Removing common English words (and, the, is) and domain-specific noise (e.g., "subscribe", "click here").
- **Lemmatization/Stemming:** Reducing words to their root form (e.g., "building" → "build").

4.2.2 Feature Extraction

- **N-grams:** Identify common 2-word and 3-word phrases (e.g., "predictive maintenance", "smart city").
- **TF-IDF:** Calculate Term Frequency-Inverse Document Frequency scores to find unique/important words for each document.

4.3 Deliverables for Task 2

- Preprocessing script/notebook.
- Cleaned version of the dataset.
- Report on "Top 20 most frequent words" (excluding stopwords) and "Top 20 bi-grams".

5 Task 3: Categorization & Trend Analysis

Points: 30/100

5.1 Objective

Classify the articles to answer the core question: *Which Civil Engineering area is using AI the most?*

5.2 Requirements

5.2.1 Dictionary-Based Classification

Define keywords for major Civil Engineering sub-disciplines and AI technologies.

Civil Engineering Areas:

- **Structural:** Analysis, design, health monitoring, materials.
- **Geotechnical:** Soil, foundations, tunnels, excavation.
- **Transportation:** Traffic, roads, autonomous vehicles, logistics.
- **Construction Management:** Scheduling, safety, cost estimation, site monitoring.
- **Environmental Engineering:** Sustainability, waste management, green building.

AI Technologies:

- **Computer Vision:** Image recognition, drone inspection, safety monitoring.
- **Predictive Analytics:** Risk assessment, maintenance prediction.
- **Generative Design:** Optimization, parametric modeling.
- **Robotics/Automation:** Brick-laying robots, autonomous machinery.

5.2.2 Analysis Logic

- **Tagging:** Write a script to tag each article with one or more Civil Engineering Areas and AI Technologies based on keyword presence.
- **Co-occurrence Matrix:** Calculate how often specific CE areas appear with specific AI technologies.
- **Temporal Trends:** If data allows, show how mentions of specific combinations (e.g., "Generative Design in Structure") have changed over time.

5.3 Deliverables for Task 3

- Tagging/Classification script.
- Analysis results showing counts/percentages for each category.
- A "Heatmap" visualization (Civil Engineering Area vs. AI Technology).

6 Task 4: Visualization & Insights

Points: 15/100

6.1 Objective

Synthesize the findings into clear, compelling visualizations and a written conclusion.

6.2 Requirements

- **Bar Charts:** Number of articles per Civil Engineering Area.
- **Network Graph:** Visualize relationships between terms (e.g., linking "Concrete" to "3D Printing" and "Sustainability").
- **Word Clouds:** Generate separate word clouds for each major sub-discipline (e.g., what are the top words in "Transportation + AI" articles vs. "Structural + AI"?).
- **Final Conclusion:** Based on the data, rank the Civil Engineering areas by their "AI Maturity" or "AI Interest" level.

6.3 Deliverables for Task 4

- All visualization code and image files.
- Interpretation of the results in the final report.

7 Technical Requirements

7.1 Programming Languages and Tools

- **Primary Language:** Python 3.8+
- **Required Libraries:**
  - Web Scraping: `requests`, `BeautifulSoup`, `selenium` (if needed).
  - NLP: `nltk`, `spacy`, `textblob`, or `gensim`.
  - Data Manipulation: `pandas`, `numpy`.
  - Visualization: `matplotlib`, `seaborn`, `networkx`, `wordcloud`.
- **Environment:** Jupyter Notebooks are highly recommended for this iterative exploratory analysis.

8 Final Deliverables

8.1 Code Repository

- Well-organized GitHub repository.
- `requirements.txt` file.
- `README.md` explaining how to run the scraping and analysis scripts.

8.2 Final Report (PDF)

Length: 10–15 pages.

Structure:

1. **Title Page**
2. **Executive Summary:** The main answer to "Which area uses AI most?"
3. **Methodology:** How you scraped and cleaned the text.
4. **Quantitative Results:** Counts, frequencies, and statistics.
5. **Qualitative Insights:** Deep dive into specific trends (e.g., "Why is Computer Vision dominating Construction Safety?").
6. **Visualizations:** Heatmaps, graphs, clouds.
7. **Conclusion & Future Outlook.**
8. **References.**

8.3 Final Presentation

Length: 10-15 minutes

- 5-7 minutes Q&A
- 1-2 minutes summary
- 1-2 minutes future outlook