CS5344 Course Project Handout

AY2024/2025 Semester 2

In this project, you will explore multimodal data analytics, where you will analyze and integrate various types of data (such as text, images, video, and audio) to derive meaningful insights. The project will involve data processing, integration, and the application of big data analytics technologies to handle and analyze multimodal datasets, as well as visualization to communicate findings effectively.

Requirements.

1. **Select a topic or use case.** Choose a topic or problem where multimodal data analysis is relevant, e.g., social media analysis, multimedia recommendation systems.

2. **Data collection.** Identify at least two types of data, e.g., text and images, or audio and video. This data can be sourced from publicly available datasets. Provide a detailed description of the dataset, including its source, size, and characteristics.

3. **Data preprocessing**. Clean, preprocess and transform the data into a format suitable for analysis. This may include text preprocessing (e.g., tokenization, stop word removal), image processing (e.g., resizing, normalization), audio/video preprocessing (e.g., audio-to-text transcription, frame extraction from videos). Document the steps with justifications.

4. **Data analysis.** Apply appropriate big data analytics techniques to analyse the multimodal data. Justify the choice of techniques based on your chosen dataset and objectives.

5. **Insights and Findings.** Interpret the results of the analysis and discuss insights.

6. **Data Visualization**. Use data visualization tools to present findings effectively. Include charts, graphs, and other visual aids.

7. **Tools and technology**. You may use a variety of tools and platforms, such as Apache Spark, NoSQL databases, Python libraries etc.

Deliverables.

1. **Project proposal.** Maximum one-page due 21 Feb 2025. Provide a brief description of the chosen topic or use case, planned data sources and preliminary thoughts on analysis.

2. **Presentation.** This will be scheduled on Week 13. Provide a 10 minute presentation summarizing your project.

3. **Final report**. Maximum six pages due 21 April 2025. Provide a final report with data analysis and visualization. Use the double-column ACM template, available at http://www.acm.org/publications/proceedings-template

Evaluation.

This is a team-based project. You will be graded on (a) relevance and complexity of topic, (b) appropriateness use and processing of multimodal data, (c) quality and depth of data analysis and insights derived, (d) effectiveness of visualizations and clarity of written report and oral presentation.