University of Westminster
School of Computer Science and Engineering

5DATA005W Data Engineering **Coursework 2** (2024/2025)	
Module Leader	Dr Habeeb Balogun
Weighting	50 %
Minimum mark required	30 %
Description and learning outcomes	The student needs to demonstrate their understanding of data engineering by implementing data pipeline combining multiple unstructured data sets (LO2). As part of this pipeline, they will need to analyse the data quality and suggest strategies to ensure high quality data (LO1), create and manage metadata (LO3). Finally, they need to store the data in a suitable format (LO4).
Due Date	**16th December 2024 at 1:00 pm**
Deliverables	1. Database documentation file (PDF file). 2. Python scripts used to process the data. 3. Image and text database files. Use the following format to name your files: 5DATA005W_StudentID_firstName_lastName.ext
Submission	Deliverables must be submitted via Blackboard
Feedback	Written feedback on the CW will be provided within 3 weeks after the submission. All marks will remain provisional until formally agreed by an Assessment Board.

5DATA005W DATA ENGINEERING COURSEWORK 2

DUE DATE: 16th December 2024 AT 1:00 PM.

Objective

This coursework aims to create two unstructured databases: the first should be suitable for a content-based image retrieval system, and the second should be ideal for training and evaluating sentiment analysis models.

1. Database for content-based image retrieval system

Content-Based Image Retrieval (CBIR) is a computer vision and image processing technology. It allows users to search and retrieve images from a database based on their visual content rather than relying on textual metadata. CBIR systems analyse images' visual features and characteristics to find similarities between query images and images in the database. CBIR's applications include image search engines, medical image analysis, fashion and e-commerce, art and cultural heritage, geospatial and remote sensing, and content management.

a) Image Collection: Gather various colour images from multiple sources, including photography, art, nature, products, etc. Ensure the dataset covers numerous subjects, styles, and characteristics. Alternatively, you can focus on a particular object type to create your database (i.e. animals, clothes, transport, medical images, etc.). The collection should consist of at least 50 images (colour). **[3marks]**

b) Image preprocessing: Rename the images systematically and transform them to have a uniform size, no larger than 500x500 pixels. For this purpose, you can rescale, rotate and denoise the images. Save the pre-processed images in a separate folder. **[10marks]**

c) Image Annotation: Manually or automatically annotate the images with metadata, including keywords and descriptions. This information will serve as the ground truth for evaluating the CBIR system. **[5marks]**

d) Image Feature Extraction: use the set of images created in (b) to extract relevant image features, such as mean and norm of the channel pixel intensities, texture descriptors, and shape features, to represent the visual content of each image. Create JSON files containing information on the extracted features. **[10marks]**

e) Database Design: Set up a NoSQL database to store the images and their associated metadata and features. **[10marks]**

2. Databases for sentiment analysis models

Sentiment analysis, also called opinion mining, is a natural language processing method that determines the sentiment or emotional tone expressed in a text, such as a sentence, paragraph, or document. The primary goal of sentiment analysis is to classify the sentiment conveyed in the text as positive, negative, neutral, or sometimes more fine-grained emotions. Some applications of sentimental analysis include customer feedback analysis, social media monitoring, market research, news and media analysis, political analysis, customer support and healthcare. Tasks:

a) Data collection: Gather a diverse collection of textual data from various sources, such as social media, product reviews, news articles, and customer feedback. Alternatively, you could focus on a specific type of text document (e.g. customer reviews, social media, political news, etc.). Ensure that the data reflects a broad range of topics and emotions. The collection must contain at least 50 textual documents. **[3marks]**

b) Data preprocessing: Implement TWO pre-processing techniques to prepare the text data for natural language processing. **[10marks]**

   For example,

   i. Text normalisation (lowercasing)
   ii. Tokenisation,
   iii. Removal of punctuation and stop words.

c) Text vectorisation: Convert the text data into TWO numerical representations suitable for machine learning models **[10marks]**

   For example,

   i. Bag of words
   ii. Term frequency-inverse document frequency
   iii. Word embeddings.

d) Metadata and labelling: Create a metadata document to explain the vectorisation process. Manually or automatically label the text data with sentiment labels (e.g., positive, negative, neutral) and include it in the metadata. **[5marks]**

e) Database Creation: Store the pre-processed text data and associated sentiment labels in a NoSQL database. **[10marks]**

3. Databases documentation

Write a report (2000-2500 words) detailing the databases you created. The report must include the following sections:

a) Objectives: Purpose of the databases **[2marks]**

b) Data Sources: Provide details on the sources from which the raw data was extracted. **[2marks]**

c) Methodology: For each database, describe the data processes that were used. Include images that show the steps in the processing methods you used. You might use, for instance, plots, processed images and screenshots. **[8marks]**

d) Storage and retrieval: Description of the NoSQL database used in each case. Demonstrate that the databases were implemented correctly by showing the results of some sample queries. **[4marks]**

e) Code: Submit a Python script (.py or .ipynb file) for ingesting, preprocessing and loading to the databases. This should be a different file and not part of the report. **[8 Marks]**

Coursework files for submission

a) Database documentation file (PDF file).

b) Python scripts used to process the data.

c) Image and text database files.

Marking Scheme

Section	Section Marks	Subsection	Subsection item description	Subsection item marks
Database for CBIRS	38	Image collection	Gather a collection of at least 50 colour images; give details on the source	3
		Image preprocessing	Preprocess and clean the images Size must be 500x500 pixels.	6
			Provide rationale on the chosen methods to acquire the required size and image quality	4
		Image annotation	Create metadata files of each image. The files must include keywords and description of the images.	5
		Feature extraction	At least two image features	6
			Al least one shape feature	2
			At least one texture feature	2
		Image Database	Produce a database to store images and the metadata	6
			Provide rationale on the chosen method of storage	4
Database for sentiment analysis models	38	Text collection	Gather a collection of at least 50 textual data;	3
		Text preprocessing	Preprocess and clean the textual data	8
			Provide rationale on the chosen methods to acquire the required size and image quality	2
		Feature extraction	Convert the text data into numerical representations suitable for machine learning.	10
		Metadata	Create a metadata file describing the dataset and the process to vectorize it;	3
			Include the textual data labels	2
		Image Database	Produce a database that includes the text data, the vectorization and the metadata	6
			Provide rationale on the chosen method of storage	4
Documentation	24	Objectives	Describe the purpose of the database	2
		Data source	Provide information on the sources give details on the type of images (animals, transport, people, etc) and text data (news, social media, customer review, etc)	2
		Methodology	Describe sequentially the steps followed to create the databases	4
			Include images on the intermediate step (e.g. grayscale, segmentate, and scaled images, etc)	4
		Storage and retrieval	Describe the software used to store the data and how the NoSQL database was generated. Demonstrate correct database implementation.	4
		Code	Python scripts used to process data/implementation	8
			Total	100