UNIVERISTY OF CAPE COAST

COLLEGE OF HUMANITIES AND LEGAL STUDIES

SCHOOL OF ECONOMICS

DEPARTMENT OF DATA SCIENCE AND ECONMOC POLICY

MSc. DATA MANAGAMENT AND ANAYSIS (SANDWICH)

DATA CURATION AND MANAGEMENT PLANS

DMA 821S

TERM PAPER 2024/2025

FREDERICK TETTEH

SE/DMD/23/0006

QUESTION 1

Effective data curation and management require both metadata and data preprocessing. Preprocessing converts raw data into a format that is appropriate for analysis, while metadata gives data meaning and organisation. Together, meta data and data preprocessing have the potential to greatly improve the precision and efficiency of data-driven operations. Meta data refers to data about data. It provides useful information about the data such as the type of data (numeric, categorical, text), the original source of the data, accuracy of the data (accuracy, completeness and consistency) as well as the format (Comma Seperated Version [CSV], Extensible Markup Language [XML]). Data preprocessing on the other hand entails processes of converting unprocessed data into an organized, clean and usable data. These processes inlcude data cleaning (handling missing values, outliers, and inconsistencies), data normalization (scaling data to a specific range), feature engineering (creating new features from existing ones) and data transformation (converting data into a different format).

Metadata and data preprocessing can provide a useful means of ensuring the efficiency of data curationa and management by enhancing informed decision making. This is seen where metadata provides insight into the data's characteristics, hence guiding data pre processing decisions. Again metadata and data preprocessing cana work together to provide efficient data cleaning where having the knowledge of data types helps to identify appropriate cleaning methods. Furthermore, understanding the data distribuation elps select the appropriate normalization techniques. Data curation and managament can be also be made efficient through metadata and data preprocessing by ensuring data quality assessment where in the case when metadata alone could not reveal data quality issues, data preprocessing will make it possible. Moroever, metadata and data preprocessing can work together to esnure data lineage tracking

such that preprocessing steps can be recorded in metadata, enabling traceability and reproducibility.

Real world example of how the use of metadata and data preprocessing can enahnce effucency of data curation and management can be seen in customer churn prediction. Metadata in this case will include the customer demographics, purchase history and tenure.

The data preprocessing will inlvove handling missing values, normalizing numerical features, creating new features (e.g., customer lifetime value). The beneift of metadata and data precprocessing together in this case is that, a model can be built to accurately predict customer churn, leading to targeted retention efforts.

QUESTION 2a

Open data sources are freely accessible, usable, and redistributable datasets that are made available to the public. These sources are frequently produced by commercial businesses, non-governmental organisations, government agencies, and research facilities. There are many uses for open data, including business, education, research, and community involvement. Two global open data sources identified are Kaggle and Data.gov.

Kaggle is a widely used platform for data scientists and machine learning devotees which offers a vast repository of public datasets. The following steps guides to accessing data from kaggle.

Create a Kaggle acount;

Visit https://www.kaggle.com/

Create and account by signing up for free using email or google or Github credentials

Search for datasets;

Use the search bar at the top of the page to search for specific keywords related to your data needs or use the "Datasets" section on the Kaggle pltaform to select from the curated list of popular datasets.

Explore dataset details;

Click on dataset to view details such as description, data fields, file formats as well as ratings and reviews from other users.

Downloading the dataset;

Click on the downloading button after identifying the dataset of interest to downlaod and save the data on the computer. Certain terms and conditions might show up to be agreed on with some datsets before downloading is continued.

Data.gov is an official open data portal of the United States government. It serves as a hub for a vast collection of datasets across various disciplines, from economics and healthcare to environment and transportation.

Open preferred web browser (e.g., Chrome, Firefox, Safari).

Enter URL: https://data.gov/ in the address bar at the top of the browser window and press enter to get into the webesite.

Once on the webesite, there are two way to access data from it; these are basic and advanced methods.

For the basic access method;

Search by Keyword or Topic:

Using the search bar, enter relevant keywords or topics related to the data you need in the search bar at the top of the homepage.

Using the filter tab, filter by category (e.g., Education, Health), agency (e.g., Department of Agriculture), format (e.g., CSV, JSON), and keywords within the dataset description.

Browse categories by clicking categories of interest to view decsription of selected data categories.

Browsing organization categories. Each government agency has a dedicated page. Navigate to the "Organizations" section and select a specific agency to see a list of all datasets published by that agency.

Explore datasets

Once you find a relevant dataset, click on it to access its detail page. This page provides crucial information like: Description: A summary of the data content. Format: Available formats for download (e.g., CSV, JSON, Excel).

Downlaod dataset

Download Links for downloading the data in each available format.

Choose the most convenient format for your needs from the available download links.

After clicking the download link, the data file will be saved to the computer's designated downloads folder.

Adanced method;

This is done using the Application Programming Interface (API)

The API allows automated download and manipulation of data within programs. This is useful for integrating data into applications or conducting large-scale analysis. However, this method requires users to have some programming skills and an API key to access the API.

Question 2b

ADVANTAGES OF OPEN DATA SOURCES

1. Open data can promote collaboration among researchers, policymakers, and citizens.
2. Open data sources provides cost effective means of conducting resrearch as it provides data for analysis instead of collecting data from the scratch.
3. Open data sources foster creativity and innovation by creating new avenues for entrepreneurship, research and development.

4. Open data can increase transparency and accountability in government and other organizations as data is made availabale publicly and available to public scruitny and criticism.
5. Open data promote accessibiity as data is freely available to anyone, regardless of their background or resources. This liberalises access to information and knowledge.

CHALLENGES OF OPEN DATA

1. Data from open data sources may not be clean, validated nor standardized making the quality of the data vary widely.
2. Open data may not be complete or comprehensive, and it may conatain missing values which might would have been useful for analysis.
3. Data from different sources may not be consistent or comparable, thus making it difficult to integrate and analyze.
4. Even though open data soruces are made public and and openly accessible, it may contain sensitive informations that users may have to protected.

QUESTION 3a

A crucial phase in the data warehousing process is data preprocessing. In order for raw data to be efficiently stored and analysed in a data warehouse, it must be transformed into a clear, consistent, and useable format. For the data kept in the warehouse to be accurate, dependable, and valuable, this procedure is necessary. Data preprocessing inlcude data cleaning which deals with handling missing values, getting rid of duplicates, and fixing discrepancies. It also include data intergration which enables integrating of data from many sources. Again data preprocessing inlcude data transformation which is the process of changing the format or organisation of data. Furthermore, data standardization is data preprocessing technqiue which

ensures uniformity in data forms and standards. Another data preprocesing technique is data which is the process of scaling data to a predetermined range.

The fundamental importance of data preprocessing in data warehousing include;

1. Improvement of data consistency through data standardization and data normalization by ensuring data is data is formatted consistently across different sources can facilitate integration and analysis as well as improving data comparability and reduce storage requirements.
2. Data security is also one importance of data preprocessing which is made oossible through data masking where sensitive data is protected by replacing and obscuring certain values.
3. Data preprocessing also improves data data quality through identifying and addressing missing data points to prevent errors and enhance analysis and reporting. Identifying and correcting outliers can also improve data accuracy and prevent skewed results.
4. Performance Optimization is another importance of data preprocessing in data warehousing. This is achived through data compression where reducing the physical size of data can improve storage efficiency and query performance. By building indexes on frequently accessed columns, data preparation further enhances data optimisation by speeding up query execution.

QUESTION 3b.

The practice of "data piling," where large quantities of raw data are accumulated without proper preprocessing techniques, can lead to significant challenges in data analysis, decision-making, and overall data management efficiency.

Step-by-step Advocacy Plan for an organization focusing on "data piling" without proper preprocessing techniques.

1. Research and Analysis:

i. Conduct a comprehensive literature review to identify existing research and case studies on the impact of data piling.

ii. Gather data on the prevalence of data piling practices in various industries and organizations.

iii. Analyze the costs associated with data piling, including wasted resources, delayed decision-making, and potential negative outcomes.

2. Develop Advocacy Materials:

i. Create a compelling narrative that highlights the negative consequences of data piling and the benefits of effective data preprocessing.

ii. Develop educational materials such as whitepapers, infographics, and presentations to explain the importance of data preprocessing.

iii. Develop a toolkit with resources and best practices for organizations to implement effective data preprocessing strategies.

3. Engage with Stakeholders:

i. Identify key stakeholders: Identify relevant stakeholders, including policymakers, industry leaders, data professionals, and researchers.

ii. Build relationships: Establish relationships with these stakeholders through meetings, conferences, and online platforms.

iii. Provide education and training: Offer workshops and training sessions on data preprocessing best practices.

4. Advocate for Policy Changes:

i. Lobby for government policies that support data quality and open data initiatives.

ii. Collaborate with industry associations to develop standards and guidelines for data preprocessing.

iii. Advocate for investment in data infrastructure and tools to support effective data management.

5. Measure and Evaluate:

i. Track progress towards achieving advocacy goals.

ii. Evaluate the impact of advocacy efforts on organizational practices and policies.

iii. Make adjustments to the advocacy plan as needed based on feedback and results.

QUESTION 4a

Humans' primary means of expression and communication is language, which emerges in early childhood and changes throughout life (Pinker & Morey, 2014; Hauser, Chomsky, Fitch, 2002).

However, without the aid of strong artificial intelligence (AI) algorithms, machines lack the innate capacity to comprehend and communicate in human language. A long-standing scientific problem has been to make machines capable of reading, writing, and communicating like humans (Turing, 2016).

From a technical standpoint, one of the main strategies for improving machine language intelligence is language modelling (LM). Generally speaking, LM seeks to forecast the probability of future (or absent) tokens by modelling the generative likelihood of word sequences.

The Langauge Models (LMs) can be divided into four major revolution stages;

Stage 1: Statistical Language Models (SLM)

Statistical learning approaches that emerged in the 1990s form the basis for the development of SLMs. The fundamental concept is to construct the word prediction model utilising the Markov assumption, which for example makes it possible to predict the next word based on

the most recent context. SLMs with a predetermined context length n are referred to as n-gram language models with exapmles being bigram and trigram language models.

SLMs have been used extensively to improve task performance in natural language processing (NLP) and information retrieval (IR). However, they frequently suffer from the peril of dimensionality, which makes it challenging to estimate high-order language models effectively because doing so requires estimating an exponential number of transition probabilities. To address the issue of data sparsity, specifically created smoothing techniques like Good-Turing estimation and back-off estimation have been established (Bahl, Brown, De Souza, & Mercer, 1989; Liu & Croft, 2005; Thede, & Harper, 1999; Zhai, 2008)

Stage 2: Neural Language Models (NLM)

Neural network-based NLMs describe the likelihood of word sequences. The study made a noteworthy contribution by introducing the idea of distributed representation of words and developing a word prediction function that was dependent on the distributed word vectors, or aggregated context characteristics. A generic neural network technique was created by expanding the concept of learning efficient features for text data in order to provide a cohesive, end-to-end solution for a variety of NLP jobs. Additionally, word2vec was put out to construct a reduced shallow neural network for distributed word representation learning, which proved to be highly successful in a range of natural language processing applications. These research have had a significant influence on the area of natural language processing (NLP) by introducing the use of language models for representation learning (beyond word sequence modelling).

Stage 3: Pre-trained Language Models (PLM)

Instead of learning fixed word representations, ELMo was an early attempt to capture context-aware word representations by pre-training a bidirectional LSTM (biLSTM) network and then

fine-tuning the biLSTM network based on certain downstream tasks. Additionally, BERT was suggested by pre-training bidirectional language models with specifically built pre-training tasks on large-scale unlabelled corpora, based on the highly parallelizable Transformer architecture with self-attention mechanisms. As general purpose semantic features, these pre-trained context-aware word representations are highly effective and have significantly improved the performance of NLP tasks. The "pre-training and fine-tuning" learning paradigm was established by the numerous follow-up studies that were sparked by this study. Numerous research on PLMs have been produced in accordance with this paradigm, proposing either enhanced pre-training procedures or alternative designs (such as GPT-2 and BART). This paradigm frequently necessitates adjusting the PLM to accommodate various downstream requirements.

Stage 4: Large Language Models (LLM)

Researchers discover that improving model capacity on downstream tasks (i.e., according to the scaling law) is frequently the result of scaling PLM (e.g., scaling model size or data size). By training an ever-larger PLM (such as the 540B-parameter PaLM and the 175B-parameter GPT-3), several research have investigated the performance limit. These large-sized PLMs exhibit distinct behaviours from smaller PLMs (such as 330M-parameter BERT and 1.5B-parameter GPT-2) and exhibit unexpected abilities in solving a series of complex tasks, despite the fact that scaling is primarily done in model size (with similar architectures and pre-training tasks). For instance, through in-context learning, GPT-3 is able to tackle few-shot problems but GPT-2 is unable to perform effectively. One outstanding application of LLMs is ChatGPT2, which transforms the GPT series' LLMs for discussion and offers a great capacity for human-to-human communication.

QUESTION 4b.

Natural language processing (NLP) has been transformed by pre-trained language models (PLMs). Large volumes of text data are used to train these models, which enable them to comprehend and produce human language in previously unthinkable ways. They have a big influence on data curation and management plans.

1. Enhanced Data Quality: PLMs may be utilised to detect and fix flaws and inconsistencies in text data, improving data cleaning methods in terms of data quality. PLMs may also be helpful in improving the quality of data by enriching it with context and meaning, which increases the data's value for analysis and decision-making.
2. Automated Data Curation: PLMs are able to automate data curation through text summerization where large volumes of text material may be automatically summarised using PLMs, facilitating study and comprehension. Again PLMs can aid in data annotation where PLMs can help with entity labelling and sentiment classification.
3. Improved Data Accesibility: PLMs may make data easier for non-technical people to access and query by enabling natural language interfaces.
4. Improved data understanding: PLMs are able to improve semantic understanding as they have the ability to understand the context and meaning of text data, allowing for more precise data search, classification, and categorisation. PLMs also imoprove data understanding by aiding information extraction in that PLMs can from unstructured text data, PLM may extract essential information such identified entities, relationships, and attitudes. In addition, PLMs can provide personalized recommendations based on user preferences and behavior.

REFERENCES

Bahl, L. R., Brown, P. F., De Souza, P. V., & Mercer, R. L. (1989). A tree-based statistical language model for natural language speech recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *37*(7), 1001-1008.

Hauser, M. D., Chomsky, N., Fitch W. T. (2002). "The faculty of language: what is it, who has it, and how did it evolve?" science, vol. 298, no. 5598, pp. 1569–1579.

Liu, X., & Croft, W. B. (2005). Statistical language modeling for information retrieval. *Annu. Rev. Inf. Sci. Technol.*, *39*(1), 1-31.

Pinker, S., & Morey, A. (2014). The language instinct: How the mind creates language (unabridged edition). *Brilliance Audio*.

Thede, S. M., & Harper, M. (1999, June). A second-order hidden Markov model for part-of-speech tagging. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics* (pp. 175-182).

Turing, I. B. A. (2016). Computing machinery and intelligence-AM Turing. *Mind*, *59*(236), 433.

Zhai, C. (2008). Statistical language models for information retrieval a critical review. *Foundations and Trends® in Information Retrieval*, *2*(3), 137-213.