Hisham Ijaz

20i-0951

MLOps Assignment 2

**Data Preprocessing and DVC Setup Documentation**

Steps in Data Preprocessing

* Text Rewriting: Regular expressions can be used to eliminate special characters. One whitespace should be used in place of several.
* Lowercase: Make sure that all text is lowercase for uniformity.
* It is optional to tokenize: For additional analysis, divide the text into tokens (words or subwords).
* Removal of stopwords (optional): Eliminate frequently used stopwords (such as "the," "is," and") if they don't add any useful information.
* Optional lemmatization and stemming:To normalize text, reduce words to their most basic form.

**DVC Setup**

Use pip install dvc to install DVC. Use dvc init to initialize a DVC repository in your project directory.

Setup:

* Set up your DVC repository to track data versions by using a storage location (Google Drive).
* Versioning Data: To begin tracking changes, use dvc add to add your data files to DVC.
* To commit modifications and generate fresh data versions, use dvc commit.
* Use DVC Push to push updates to your distant storage.
* Versioning of Metadata: Make a metadata file (such as metadata.yaml) to record further details about every data version. For every version, add pertinent information to the metadata file, such as the timestamp, author, description, etc. To guarantee consistency, use DVC to commit the metadata file together with the data.

**Overview of Workflow**

Preprocessing text data taken from Dawn.com and BBC.com articles is part of the workflow.

Text cleaning, lowercasing, and optional procedures like tokenization, stopword removal, and lemmatization are examples of data preprocessing procedures. DVC is used to maintain version control and repeatability by tracking versions of the metadata and preprocessed data.

**Challenges:**

1. Handling Dependencies:Confirming that the environment has all necessary libraries installed and accessible. Resolving problems with compatibility between various library versions.
2. Setup and Configuration: Establishing DVC and setting up storage spaces for data versioning.
3. regulating rights and authentication to access remote storage.
4. Management of Metadata: Creating a clear and standardized metadata structure to record pertinent information for every data version.Keeping consistency in the DVC workflow by integrating metadata versioning.
5. Integration of Workflow: Combining DVC commands with data pretreatment procedures to achieve smooth version control. Ensuring that updates to the metadata are reflected in changes to the data preparation code and vice versa.