**Documentation of Data Preprocessing steps and DVC setup**

**Data Pre-processing Steps:**

1. **Extracting Links:**

* The python script utilizes the extract\_links function to extract links from the landing pages of dawn.com and BBC.com.
* It utilizes the requests library to get the webpage HTML content and BeautifulSoup for parsing the HTML and extracting links.
* Each link is appended to a list.

1. **Extracting Article Information:**

* The extract\_article\_info function is employed to extract titles and descriptions from articles displayed on the homepages of dawn.com and BBC.com.
* Similar to the link extraction, it utilizes requests and BeautifulSoup to parse the HTML and extract relevant information.
* The titles and descriptions are stored in a list of dictionaries, where each dictionary represents an article.

1. **Data Transformation:**

* The preprocess\_text function preprocesses the text data.
* It removes HTML tags, special characters, and converts the text to lowercase to ensure consistency and cleanliness for further analysis.
* Processed articles are stored in lists of dictionaries (dawn\_articles\_processed and bbc\_articles\_processed).

1. **Storing Processed Data:**

* The processed data is written to separate text files for Dawn and BBC articles (dawn\_processed\_data.txt and bbc\_processed\_data.txt).
* Each file contains titles and descriptions of the articles, with each article separated by new lines

**DVC (Data Version Control) Setup:**

1. **Installation and Setup:**
   * Installed DVC using (pip install dvc) command.
   * Initialized current directry with dvc init.
2. **Tracking Data:**

* Place the processed data files (dawn\_processed\_data.txt and bbc\_processed\_data.txt) under DVC control using dvc add.
* This command creates corresponding .dvc files, which are small text files that contain metadata about the tracked data files.

1. **Committing Changes**:

* After adding files, committed them to local repository (git add ., git commit -m "Added processed data files").

1. **Pushing to Github repository:**

* Pushed the changes to GitHub repository (git push origin master).
* Additionally, used dvc push to push the tracked data files along with their metadata to DVC remote storage ( Google Drive).
* Ensured proper authentication and permissions for accessing the remote storage.

1. **Versioning Meta-data:**

* DVC automatically tracks metadata changes and associates them with the respective data files.
* Each dvc push or commit triggers metadata versioning, maintaining a history of changes.

1. **Workflow Maintainance:**

* DVC-tracked data files must be regularly updated and metadata as the preprocessing pipeline or data changes.
* Any significant changes or updates in the project's README file must be documented for transparency and reproducibility.