**Procedure Document: Text Analysis Pipeline**

Requirements:

* Python 3.11.5 environment
* Necessary libraries: numpy, pandas, requests, BeautifulSoup, nltk

Procedure Steps:

1. **Import Required Libraries:**
   * Import necessary libraries such as numpy, pandas, requests, BeautifulSoup, re, and nltk. Ensure nltk data for 'punkt' and 'stopwords' is downloaded.
2. **Read Input Data:**
   * Read input data from a CSV file named "Input.csv" using pandas **read\_csv()** function. Assign it to DataFrame **df1**.
3. **Scrape Data from Websites:**
   * Iterate through each URL in the DataFrame **df1**.
   * Use the requests library to retrieve HTML content from the URL.
   * Parse the HTML content using BeautifulSoup.
   * Extract text content from specific **<div>** tags identified by class name.
   * Store the extracted text content in a new column named 'para' in the DataFrame **data\_new**.
4. **Save Scraped Data:**
   * Save the DataFrame **data\_new** with the newly added 'para' column to a CSV file named "data\_withPara.csv".
5. **Text Analysis:**
   * Read positive and negative words from external text files.
   * Read stopwords from multiple text files in a specified folder and concatenate them into a single list.
   * Load the scraped data from "data\_withPara.csv" into DataFrame **df**.
   * Define functions to count positive words, negative words, clean words, clean tokenized sentences, count complex words, and count personal pronouns.
   * Initialize new columns in DataFrame **df** for storing calculated scores and features.
6. **Compute Variables:**
   * Iterate through each row in DataFrame **df** and compute positive score, negative score, subjectivity score, polarity score, complex percentage, word count, and personal pronouns count using the defined functions.
   * Store the computed values in their respective columns in DataFrame **df**.
7. **Save Output Data:**
   * Save the DataFrame **df** with computed variables to a CSV file named "Output.csv".

Note:

* Make sure that the input CSV file "Input.csv", negative and positive word files, and stop word files are available in the working directory.
* The procedure assumes a specific structure of HTML content on the websites for text extraction. Adjustments may be needed if the structure varies.