## UiPath Process Documentation: LA Times News Scraper

**1. Process Name:** LA Times News Scraper

**2. Purpose:** This UiPath process automates the scraping of news articles from the Los Angeles Times website, extracts relevant data, filters articles based on a specified timeframe and date, and saves the results to an Excel file.

**3. Inputs:**

* **searchPhrase (String):** The keyword or phrase to search for on the LA Times website. This is user input.
* **selectedMonths (Integer):** The number of months back in time from the current date to retrieve news articles. This is user input.

**4. Outputs:**

* **Output.xlsx (Excel File):** An Excel spreadsheet containing the scraped news articles, including:
  + **articleTitle (String):** The title of the news article.
  + **articleDescription (String):** A brief description of the news article.
  + **articleDate (String):** The date the news article was published (or created, if a relative time is provided).
  + **imageURL (String):** The URL of the image associated with the article.
  + **searchPhraseCount (Integer):** The number of times the search phrase appears in the article title and description (case-insensitive).
  + **Contains money (Boolean):** Indicates whether the article title or description contains a money reference.
  + **Picture filename (String):** The filename of the downloaded image.

**5. Workflow Steps:**

**A. Input and Initialization:**

1. **Log Start:** Logs the start of the process.
2. **Search Phrase Input:** An "Input Dialog" activity prompts the user to enter the search phrase.
3. **Months Selection Input:** An "Input Dialog" activity prompts the user to select the number of months back to retrieve news.
4. **Switch:** A "Switch" activity processes the number of months input, converting it into an integer and assigning it to the selectedMonths variable. It provides various log messages based on the selected timeframe.
5. **Log Message:** Logs the start of the browser extraction process.

**B. Open Browser and Extract Data:**

1. **Open Browser:** An "Open Browser" activity opens the LA Times website.
2. **Wait for Element:** A "Wait for Element" activity waits for the search box to load, ensuring the page is ready.
3. **Go To URL:** A "Go To URL" activity navigates to the search results page using the searchPhrase variable.
4. **Extract Data:** An "Extract Data" activity scrapes the news article details into a data table (newsDT) and includes a "Next Link" to allow scraping of data on multiple pages. The extraction includes the following columns: articleTitle, articleDescription, articleDate, and imageURL.
5. **Add Columns:** "Add Data Column" activities add columns to the data table for searchPhraseCount, Contains money, and Picture filename.
6. **Copy DT:** An "Assign" activity copies the newsDT to the filteredDT variable.

**C. Filter Data by Timeframe and Date:**

1. **Log Message:** Logs the start of the filtering process.
2. **For Each Row:** A "For Each Row" activity loops through each row in the newsDT.
3. **Check Article Time Today:** An "If" activity checks if the article date contains "ago."
   * **If "ago" is found:** It extracts the time difference, calculates the actual date, and stores it in the rowDate variable.
   * **If "ago" is not found:** It attempts to parse the articleDate in "MMMM d, yyyy" format and if that fails, it tries "MMM. d, yyyy" format.
4. **Assign Filtered Date Value:** An "Assign" activity calculates the date rowFilteredDate that's selectedMonths months ago from the current date.
5. **If Condition Remove Row:** An "If" activity checks if the articleDate is older than rowFilteredDate and removes the row if it is.
6. **Check if DT is empty:** An "If" activity checks if the filteredDT is empty. If it is, the workflow throws an exception and logs an error message.

**D. Process Filtered Articles:**

1. **For Each Row:** A "For Each Row" activity loops through each row in the filteredDT.
2. **Assign Title/Description:** "Assign" activities extract the articleTitle and articleDescription from each row.
3. **Calculate Search Phrase Count:** An "Assign" activity calculates the number of occurrences of the searchPhrase in the article title and description (case-insensitive) and stores the count in searchPhraseCount.
4. **Check for Money:** An "Assign" activity checks if the title or description contains a money reference.
5. **Assign Image URL:** An "Assign" activity extracts the imageURL from the row.
6. **Split URL:** An "Assign" activity splits the imageURL to extract the filename.
7. **Assign Picture Filename:** An "Assign" activity assigns the extracted filename to the "Picture filename" column in the data table.
8. **Download File:** A "Download File" activity downloads the image using the imageUrl and saves it with the imageName in a designated folder. The "Download File" activity is wrapped in a "Try Catch" block to handle errors during the download process.

**E. Write to Excel:**

1. **Write Range:** A "Write Range" activity writes the data from the filteredDT to the "Output.xlsx" file.

**F. Log End:**

1. **Log Message:** Logs the end of the process.

**6. Error Handling:**

* **Try Catch Blocks:** "Try Catch" activities are used around the image download process to catch any errors that might occur during the download. Error messages are logged.
* **Empty Data Table:** The workflow checks if the filtered data table is empty. If it is, an exception is thrown to indicate that no news articles were found within the specified timeframe.

**7. Notes:**

* **XPath Selectors:** The XPath selectors used in the workflow might need to be adjusted if the LA Times website structure changes.
* **Date Format:** The workflow handles both full month names and abbreviated month names. This might need to be adjusted if the website's date format changes.