Tahrirchi's Test Task

Background information: Training an ML model requires a lot of data, and, unfortunately, Uzbek does not have a solid database of words/sentences/phrases that can be used for training. Your team decided to use the web scrapping technique to collect some data from various news websites.

Minimal requirements:

- Complete at least the first step
- Save your dataset(s) as CSV files
- Answer at least 1 theoretical question

Tools can include, but are not limited to:

- Python
- Pandas
- BeautifulSoup

Technical requirements:

Step1. Bronze layer. Data acquisition

- Find at least 3 articles that are written in Uzbek. You can use any news website like uznews, kunuz etc.
- Extract the text from the page. Usually, the most useful text is between tags.
- Parse it in a way that your dataset has the following structure

`source_url` - link through which you accessed the page
`access_datetime` - when you accessed the page
`content` - text that was in the article

Step 2. Silver layer. Basic transformations

The field `content` from the previous dataset contains textual data. Basically, it is just a group of sentences. At this stage, we need to split this text into words. So you need a new field

`word` - a group of characters that are surrounded by spaces
E.g. the sentence “Men seni kecha ko’rdim” has 4 words.

Now your dataset should have the following structure
`source_url` - link through which you accessed the page
`access_datetime` - when you accessed the page
`content` - text that was in the article
`word` - a group of characters that are surrounded by spaces

Note: Don’t worry about excessive data duplication for now.

Step 3. Let's gather some stats!
Take the dataset produced in step 2 and add a column indicating how often this word has occurred in the dataset.

Note: You should not remove any of the rows produced in step 2.

Bonus Part. Some questions for reasoning:

- How would you automate this process so that we can get new datasets every day?
- What file format would you use to store this data?
- How would you evaluate the quality of the collected data?

Expected results:

- The code used to perform the steps should be uploaded to GitHub. You can create a public repository and just share the link.
- Obtained datasets can be located in GitHub as well. If you have very large datasets, you can share them via telegram.
- Answers to theoretical questions should be included in the repository's ReadMe file or sent via telegram.

Note: Do not worry too much about the accuracy of the datasets. At this point, we are more interested in your problem-solving and practical coding skills.