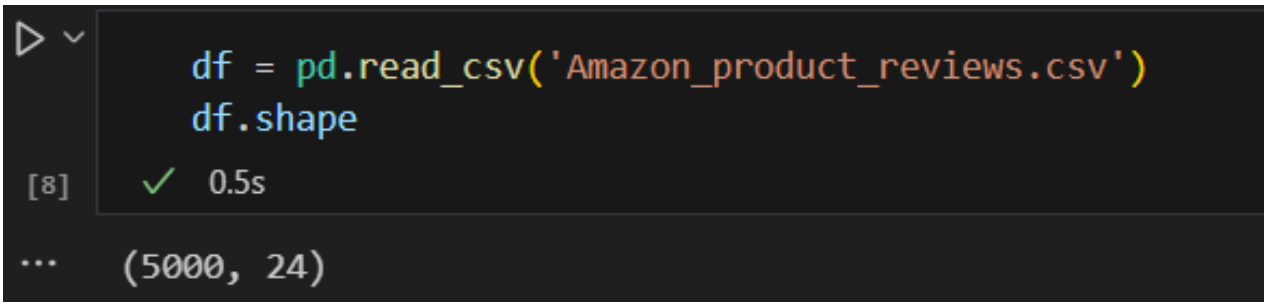
Task 21 - Capstone Project - NLP Applications

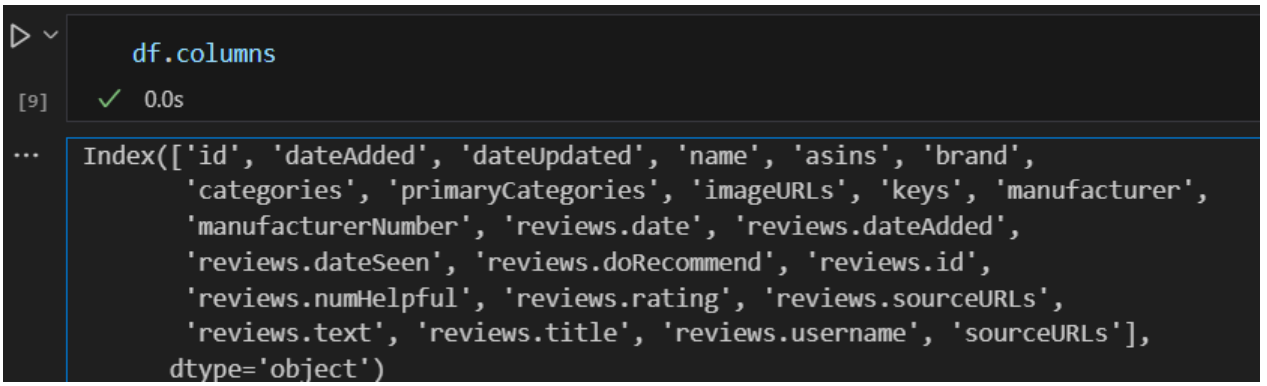
5. Write a brief report or summary in a PDF file:

5.1. A description of the dataset used:

This is a sample of a dataset with over 34,000 consumer reviews, the sample we use has 5000 reviews. The dataset includes basic product information like rating, review text, and more for each product. We will analyze only the text review column using NLP with Spacy Library.



5000 products (rows) and 24 columns.



These are our columns and the data we access for each product.

5.2. Details of the preprocessing steps:

```
df = pd.read_csv('Amazon_product_reviews.csv')
clean_data = df.dropna(subset=['reviews.text']) # Removes blank rows in our reviews.text column.
reviews_data = clean_data['reviews.text'] # Selects the column of reviews we want to analyse.
```

After reading our dataset we will address the only column we are studying, in this case, the "reviews_text" column. We will then drop any column that has information missing. In this case, the data frame will continue to be 5000 long as it has no missing data.

```
def remove_stop_words(review):
    """ Function to remove stop words from reviews """

    # Creates a list with all the words not included in the stop words from Spacy.
    filtered_review = [token for token in review if not token.is_stop]
    # Gets our list of filtered tokens back as a spaCy doc.
    filtered_doc = spacy.tokens.Doc(review.vocab, words=[token.text for token in filtered_review])
    return filtered_doc
```

After the data frame is ready to be worked on I will then use a function to remove all the stop words from a desired review I want to analyze so the analysis will be more accurate.

5.3. Evaluation of results:

```
Review nr: 14 : I use this every day on my commute. Great battery life, no backlight but very readable with normal lighting. I like the
built in dictionary. Email yourself pdf or mobi files for easy transfers.
Review nr: 15 : It does its job but I would buy one which the screen is brighter. There are times that it's difficult to read because sc
reen is not too bright
Polarity for review nr 14: 0.4611111111111111, positive feeling.
Polarity for review nr 15: 0.1000000000000003, neutral feeling.
Similarity between both: 0.7670457666535326
```

Running some test cases we can see that the similarity function of this model is working fine and giving us accurate responses.

We can see in this example that both test cases talk about lighting and so the similarity between reviews is high (0,76). Test case number 14 has a positive polarity (0,46) so I can conclude that the model works fine for this test case.

```
Review nr: 7 : I bought my Kindle about 2 months ago and the battery is already dead and will not charge
Review nr: 27 : I've wanted a kindle for a while and decided to get it when BB put it on sale. I am not disappointed.
Polarity for review nr 7: -0.2, negative feeling.
Polarity for review nr 27: -0.75, negative feeling.
Similarity between both: 0.3339228727867011
```

In this example, I looked for negative polarity looking for the first 2 reviews where the polarity would be negative using a simple while loop. When looking at this test case we can see that the module is accurate on test subject nr 7 but not on test subject nr 27. We can conclude that the model looked into the word "disappointed" and rated it negatively, however, what the human in the review tried to say is that "he is not disappointed" making it a bit faulty.

5.4. Insights into the models' strengths and limitations:

We can conclude that Spacy is an easy-to-apply model, fast, and can deal with big texts and data. It has different language packages and can do tagging, named entity recognition, and tokenization. We can also use it to take out the stop words with its library of stopped words and make our analysis more accurate.

I would say it lacks strength in sentiment analysis as we can see from the example of negative reviews above as the model might not be looking for expressions such as "not dissatisfied" or in general, double negatives.

In general, I think it's a good library and with the help of its pre-trained models, we can do big work.