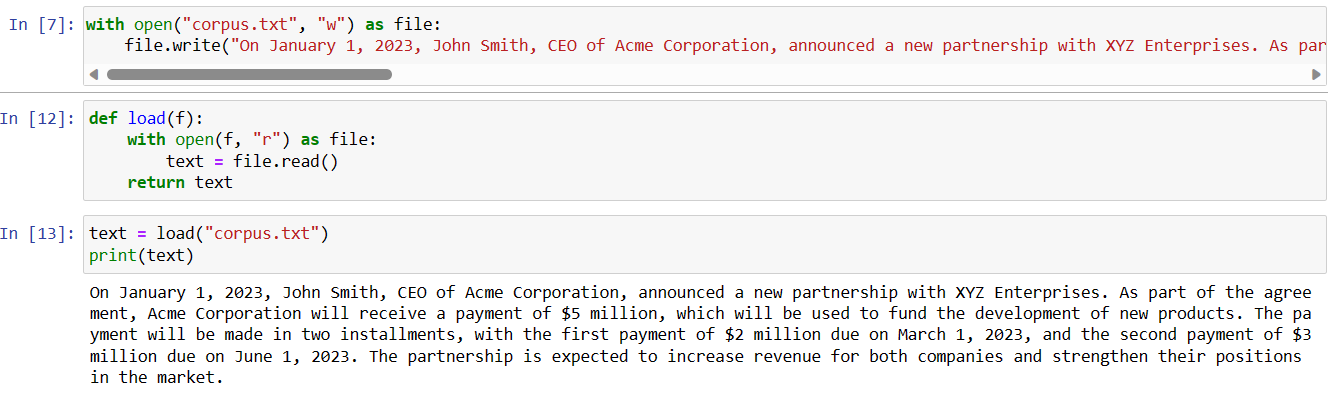
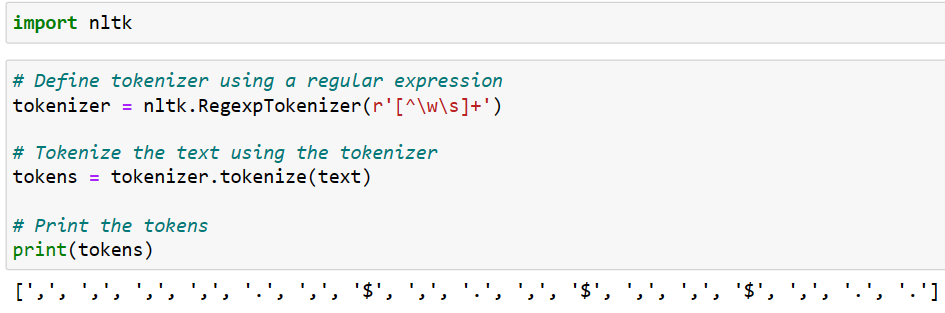
**Q1 \_\_\_\_\_ [9 Marks]**

**Part 01: Read and answer carefully the following questions:**

1. Save some text into a file corpus.txt. Define a function load(f) that reads from the file named in its sole argument, and returns a string containing the text of the file. **[1 Mark]**

****

* 1. Use nltk.regexp\_tokenize() to create a tokenizer that tokenizes the various kinds of punctuation in this text. Use one multi-line regular expression, with inline comments, using the verbose flag (?x). **[1 Mark]**

****

* 1. Use nltk.regexp\_tokenize() to create a tokenizer that tokenizes the following kinds of expression: monetary amounts; dates; names of people and organizations. **[1 Mark]**

Text

Description automatically generated

1. Rewrite the following loop as a list comprehension:

|  |  |  |
| --- | --- | --- |
| |  |  | | --- | --- | |  | **>>> sent = ['The', 'dog', 'gave', 'John', 'the', 'newspaper']**  **>>> result = []**  **>>> for word in sent:**  **... word\_len = (word, len(word))**  **... result.append(word\_len)**  **>>> result**  **[('The', 3), ('dog', 3), ('gave', 4), ('John', 4), ('the', 3), ('newspaper', 9)]** | |

* 1. Define a string raw containing a sentence of your own choosing. Now, split raw on some character other than space, such as 's'. **[1 Mark]**

Graphical user interface, text, application, email

Description automatically generated

1. Write a for loop to print out the characters of a string, one per line. **[1 Mark]**

Graphical user interface, text

Description automatically generated with medium confidence

1. What is the difference between calling split on a string with no argument or with ' ' as the argument, e.g. sent.split() versus sent.split(' ')? What happens when the string being split contains tab characters, consecutive space characters, or a sequence of tabs and spaces? (In IDLE you will need to use '\t' to enter a tab character.) **[2 Marks]**

Calling split() on a string with no argument is equivalent to calling it with ' ' (a space) as the argument.

When you call split() with no argument, it uses whitespace (spaces, tabs, and newlines) as the delimiter to split the string. For example:

Graphical user interface, text, application, Word

Description automatically generated

In this example, the string my\_string contains the text "Hello" followed by a tab character (\t) and then "world" followed by a newline character (\n). The split() method is called on the string with no argument, which causes the string to be split into a list of words based on the whitespace characters.

When you call split() with ' ' as the argument, it splits the string on spaces only, ignoring any other whitespace characters. For example:

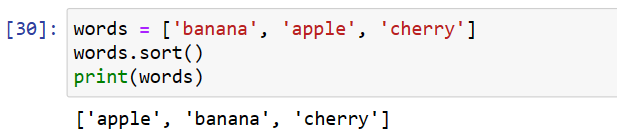
Graphical user interface, text, application, Word

Description automatically generated

In this example, the split() method is called on the string with ' ' as the argument, which causes the string to be split into a list containing only the original string, because there are no spaces in it. The tab and newline characters are not treated as delimiters in this case.

1. Create a variable words containing a list of words. Experiment with words.sort() and sorted(words). What is the difference? [2 Marks]

Both list.sort() and the built-in function sorted() can be used to sort a list. However, they differ in how they work and how they affect the original list.



sorted() is a built-in function that returns a new sorted list based on the input list. It doesn't modify the original list. The sorted list is returned as the result of the function call, so you need to assign it to a new variable to keep it. For example:

Graphical user interface, text, application

Description automatically generated

In summary, the main differences between list.sort() and sorted() are:

list.sort() sorts the list in-place, while sorted() returns a new sorted list. list.sort() doesn't return anything, while sorted() returns the sorted list as its result. list.sort() modifies the original list, while sorted() leaves the original list unchanged.

**Part 02: Solve the following problem [6 Marks]:**

1. Using any of the three classifiers described **Chapter** **6.** [**Learning to Classify Text**](https://www.nltk.org/book/ch06.html), and any features you can think of, build the best name gender classifier you can. Begin by splitting the Names Corpus into three subsets: 500 words for the test set, 500 words for the dev-test set, and the remaining 6900 words for the training set. Then, starting with the example name gender classifier, make incremental improvements. Use the dev-test set to check your progress. Once you are satisfied with your classifier, check its final performance on the test set. **[4 Marks]**



Text

Description automatically generated

Table

Description automatically generated

1. How does the performance on the test set compare to the performance on the dev-test set? Is this what you'd expect? **[2 Marks]**

The accuracy (performance) of the test set is higher than dev-test set.