**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]**
   2. 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]**
2. 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]**

1. Write a for loop to print out the characters of a string, one per line. **[1 Mark]**
2. 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]**
3. Create a variable words containing a list of words. Experiment with words.sort() and sorted(words). What is the difference? [2 Marks]

**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]**
2. 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]**