Mohammad Hossein Khojaste

Iran University of Science and Tech.
www.iust.ac.ir

Assignment 1 Problems

NLP : Fall 1400 : Dr. Minaei
Due Wednesday, Azar 3, 1400

Contents

Problem 1

In this section, you should explain some questions. Feel free to search for topics on the internet, but do not copy the answers. Once you find the answer, read and understand it and answer the question in your own words. No score is given for the answers copied directly from the internet. It is recommended to write your answers in Persian.

(a)

Ambiguity in human language is one of the reasons why NLP is hard. In the course, ambiguity was explained at four levels. Name and explain these four levels, and give an example for each in Persian.

(b)

In this question, we will focus on three tasks in NLP. For each of the following tasks, explain what the task is about. It should include only an introduction to the task. It is not necessary to write about different techniques used to solve the given task.

1. Text Summarization
2. Entity Linking
3. Machine Translation

Bonus Section: Read an article about one of the given tasks and write an explanation for it. Your chosen article must be in 2021 and accepted in one of the following two conferences:

1. ACL 2021 (https://2021.aclweb.org/program/accept/)
2. EMNLP 2021 (https://2021.emnlp.org/papers)

Your explanation about the chosen article must be at least two pages and include the following sections:

1. Introduction and idea: Briefly explain the idea of the selected article and the problem that the article wants to solve.
2. Algorithm: Explain in detail the proposed algorithm and model of the article.
3. Results: Write about the experiments performed by the authors and state the results of the experiments.

Please write the name of the selected article.

(c)

Explain the Maximum Matching Word Segmentation Algorithm and provide an example for it. Also, explain the application of the algorithm.

(d)

Explain Lemmatization and Stemming and state examples in Persian.
(20 + 15 pts)

Problem 2

In this section, you should answer questions about Regular Expressions. In each part, you must write a regex that can specify the desired text strings. You can use this website to check your written regex: https://www.regextester.com. On this website, you should check the multiline flag and write one string in each line.

Problem 2 continued on next page. . .

(a)

Present a regex that can specify all strings that start with a capital letter and end with the letter f. (There is no condition for the middle letters of the string. For example, the string Agh_23f is acceptable, and $aghf$ is not.

(b)

Present a regex that can specify all strings that have at least three numbers 4. (These numbers can be anywhere in the string, and anything can come between them. For example, 14450414 and m_444 are acceptable and 441 is not)

(c)

Present a regex that can specify all strings that start with an odd number and end with an even number. There should also be a lowercase letter in the middle of the strings. (Any character can be between odd and even numbers, but there must be a lowercase letter. For example, $1m8$ and 111_Aha2 are acceptable, and $1AS2$ is not. (15 pts)

Problem 3

In this question, you have to solve a question in Probabilistic Language Models. Consider the following table. Calculate the probability of the test data using the Bigram model based on the train data. Please write your calculations in detail. You should use Laplace Smoothing in your calculations. (20 pts)

text	data
$< s >$ Ali Daei footballer $< /s >$	train
$< s >$ Hadi Saei taekwondo athlete $< /s >$	train
$< s >$ Persepolis football club $< /s >$	train
$< s >$ Ali Daei Persepolis $< /s >$	train
$< s >$ football club $< /s >$	test
$< s >$ Hadi Saei Persepolis $< /s >$	test
$< s >$ Saei Daei taekwondo $< /s >$	test
$< s >$ footballer Persepolis club $< /s >$	test

Problem 4

In this section, you should implement codes in Python.

(a)

In this question, you should implement regex in Python. You should use The Python module ”re.” For more information about this module, you can use this link:
https : //www.tutorialspoint.com/python/python_reg_expressions.htm
For each of the following parts, implement a function in Python that can specify the desired text strings. Your function receives a string as input and returns a Boolean variable that indicates whether the given string corresponds to the implemented regex. Inside the function, you should write a regex using the ”re” module that can specify desired text strings. (15 pts)

1. Validating an Email address: In your first function, you should implement a regex that can verify an email address. An email address is in the following form

   username@domain.tld

Problem 4 [(a)] continued on next page...

username can only have English letters (uppercase and lowercase), numbers, underline, and dot

domain can only contain English letters (small and capital) and numbers

tld must have only three characters and should consist of only English letters (uppercase and lowercase).

2. Validating a phone number: In your second function, you should implement a regex that can verify a phone number. A phone number is valid if it has one of the following forms:

   - Start with 09 and have 11 characters

   - Start with +989 and have 13 characters

   - Start with 00989 and have 14 characters

(b)

In this question, you have to implement Python code to calculate the Bigram model. Suppose we have a small language consisting of the following Persian words:

آنها، کتاب‌ها، بوستان، درختان، فردا، اتصالات، می روم، رفته‌اند، گفته بودند، می‌گویید، می‌خورم، خوردند،
نوشتید

First, you should implement a function for Lemmatization and Stemming. This function receives a string or an array of strings and performed Lemmatization and Stemming on the string(s) and returns the result. The following figure shows an example of the input and output of this function:

آنها -> آن

نوشتید -> نوشت

For the second part of this question, you should implement a function that receives a string and calculate the probability of the given string based on the train data and using the Bigram model with Laplace Smoothing. Consider the following strings as the train data:

<s> کتاب‌ها، گفته بودند، نوشتید، فردا، اتصالات </s>
<s> کتاب‌ها، آن‌ها، درختان، می‌روم، می‌گویید </s>
<s> می‌خورم، می‌گویید، نوشتید، اتصالات</s>
<s>آن‌ها، درختان، رفته‌اند، کتاب‌ها، اتصالات</s>
<s>کتاب‌ها، رفته‌اند، می‌خورم، نوشتید</s>

Do not consider the comma. It is just for separating words. In your second function first, implement the train data as arrays, then call the Lemmatization function, and get the result for all words. For the test data, which is the input of your function, call the first function for Lemmatization and Stemming. Then using the Bigram model with Laplace Smoothing, calculate the probability of the test data. In this question, you can only use the NumPy library, and using any other library is not allowed. Use the following table as the test data:

Problem 4 [(b)] continued on next page. . .

<s> کتاب‌ها، می‌گویید، می‌خورم، بوستان، اتصالات </s>
<s> بوستان، رفته‌اند، کتاب‌ها، می‌گویید، می‌گویید </s>

(30 pts)

Notes

- Codes should be implemented in .ipynb format (notebooks)
- All Code cells should be executed before turning in the assignment (Make sure your outputs are there before you submit your assignment)
- Please explain the code and the results in the notebook
- If you have any questions, feel free to ask. You can ask your questions in the Telegram group.
- Please upload your assignments as a zipped folder with all necessary components. Upload your file in *HW1_NLP_YourStudentID_YourName.zip* format.