Iran University of Science & Technology

School of Computer Engineering

Assignment #1

Natural language processing

BY:

DR. Behrouz Minaei, Fall 2024

Teaching Assistants:

Mohammad Mosavi

Mahmood Kalantari

Due: 1403/08/06

Contents

Notes

1. Submit the answers in a complete PDF file and the code for the questions in the .ipynb format (including the notebook cell outputs) in a compressed file named HW1_StudentID.zip by the specified deadline.
2. A total of 72 hours of delay in submitting the answers is allowed across all projects. After that, for each additional day of delay, 10% of the score will be deducted.
3. If a student submits the project earlier than the deadline and achieves 75% of the score, up to 24 hours will be added to their allowable delay time.
4. The maximum delay for submitting each assignment is 4 days, and after 4 days, submission will not be accepted.
5. It is important to note that the explanation of the code and the obtained results must be included in the PDF file. Code without a report will result in a score deduction.
6. The evaluation of the assignment will be based on the correctness of the solution and the completeness and accuracy of the report.
7. Assignments must be completed individually, and group work on assignments is not allowed.
8. Please allocate sufficient time for the assignment and avoid leaving it until the last days.
9. You can ask your questions in the relevant group.

good luck.

Problem 1

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. **(10 points)**

Problem 2

Answer the following questions about regex (regular expressions), providing explanations for your answers. **(20 points)**

i. Write a regex pattern that matches a string starting with "a", followed by any number of any characters, and ending with "z". How would you modify it to ensure the string must contain at least one number between "a" and "z"?

ii. Explain the following regex pattern and state whether the string 'jhg4jhbhj291' is accepted by it. Provide your reasoning.

$$\text{\^{}}([\text{\^{}}\mathbf{0-9}]*\backslash\boldsymbol{d})\{\mathbf{3}\}[\text{\^{}}\mathbf{0-9}]*\$$$

Problem 3

Consider the following sentence:

S = “He eats an apple”

You are given the following conditional probabilities from a trained model:

- P(He)=0.1
- P(eats|He)=0.4
- P(an|He, eats)=0.3
- P(apple|He, eats, an)=0.5

Using the Chain Rule, calculate the probability of S. **(15 points)**

Problem 4

What is tokenization? In Chinese, since there are no spaces between words, how is tokenization performed? **(10 points)**

Problem 5

Apply four common preprocessing techniques (Lemmatization, Case Folding, Normalization, and Stemming) to a given text. You will perform these tasks step by step and observe how each technique transforms the text. **(15 points)**

S = "Running runners ran quickly to achieve their goals. Cats' paws are cleaner than dogs'."

Problem 6

Refer to the Notebook RegEx.ipynb. Review and run the Notebook, then complete the exercise. **(30 points)**