Iran University of Science & Technology

School of Computer Engineering

Assignment #2

Natural language processing

BY:

DR. Behrouz Minaei, Fall 2024

Teaching Assistants:

Mahdi Khoursha

Mohammad Mosavi

Due: 1403/08/21

Contents

Notes 3
Problem 1 4
Problem 2 4
Problem 3 4
Problem 4 4
Problem 5 5

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

Explain the potential risks and limitations associated with using N-grams in natural language processing. Be sure to address issues such as data sparsity, scalability with larger N values, and context limitations. (10 points)

Problem 2

Provide an intuitive explanation of perplexity and discuss its limitations in evaluating model performance. (10 points)

Problem 3

a. Explain the purpose of perplexity smoothing and why it is beneficial in language modeling. (10 points)
b. Discuss the advantages and disadvantages of different approaches to smoothing perplexity, highlighting their impact on model performance and complexity. (10 points)

Problem 4

Consider the following sentences and answer the questions based on them. (40 points)

Sentences:

I. Iran is advancing rapidly in artificial intelligence.
II. Artificial intelligence research is thriving in Iran.
III. Artificial intelligence is common in Iran.
IV. Iran is focused on artificial intelligence.

Questions:

a. Construct a bi-gram table for this corpus, extracting all bi-grams and calculating the frequency of each.
b. Calculate P(intelligence | artificial) and P(artificial | intelligence). Explain the difference between them.
c. Estimate P(S_1) using this bi-gram model, where S_1 is the first sentence.
d. Calculate the perplexity of S_1 using this bi-gram model.
e. Calculate the smoothed perplexity of S_1 using Laplace smoothing and discuss the differences observed.
f. Given the input "Iran is," use your bi-gram model to predict the next word. Then, consider the input "Iran was" and discuss how to handle this situation.

Problem 5

Open the word2vec.ipynb notebook. In this notebook, you'll implement skip-gram word2vec and gain familiarity with the model. Complete the sections marked with Your code here. (30 points)

Note: For reference, the expected cell outputs are provided.