University of Central Missouri

Department of Computer Science & Cybersecurity

CS5760 Natural Language Processing

Fall 2025

Homework 2.

Student name:

Submission Requirements:

- Once finished your assignment push your source code to your repo (GitHub) and explain the work through the ReadMe file properly. Make sure you add your student info in the ReadMe file.
- Submit your GitHub link on the Bright Space.
- Comment your code appropriately ***IMPORTANT.***
- Any submission after provided deadline is considered as a late submission.

Q1. Bayes Rule Applied to Text (based on slide: Bayes' Rule for documents)
The PPT shows that classification is based on:

$$c_{MAP} = \arg\max_{c \in C} P(c)\, P(d \mid c)$$

Tasks:

1. Explain in your own words what each term means: P(c), P(d|c) and P(c|d).
2. Why can the denominator P(d) be ignored when comparing classes?

Q2. Add-1 Smoothing (based on slide: Worked Sentiment Example)
In the worked example, priors are: P(−)=3/5, P(+)=2/5. Vocabulary size = 20.

Tasks:

1. For the negative class, the total token count is 14. Compute the denominator for likelihood estimation using add-1 smoothing.
2. Compute P(predictable|−) if the word "predictable" occurs 2 times in the negative documents.
3. Compute P(fun|−) if "fun" never appeared in any negative documents.

Q3. Worked Example Document Classification (based on slide: Test document "predictable no fun")
Using the smoothed likelihoods and priors from Q2, compute the probability scores for the document *"predictable no fun"* under both the positive and negative classes.

Tasks:

1. Show each step of the multiplication.
2. Which class should the system assign to this document?

Q4. Harms of Classification (based on slide: Avoiding Harms in Classification)

Tasks:

1. Define **representational harm** and explain how the Kiritchenko & Mohammad (2018) study demonstrates this type of harm.
2. What is one risk of censorship in toxicity classification systems (based on Dixon et al. 2018, Oliva et al. 2021)?
3. Give one reason why classifiers may perform worse on African American English or Indian English, even though they are varieties of English.

Q5: Evaluation Metrics from a Multi-Class Confusion Matrix

The system classified 90 animals into Cat, Dog, or Rabbit. The results are shown below:

System \ Gold	**Cat**	**Dog**	**Rabbit**
Cat	5	10	5
Dog	15	20	10
Rabbit	0	15	10

Tasks:

1. Per-Class Metrics
   - Compute precision and recall for each class (Cat, Dog, Rabbit).
2. Macro vs. Micro Averaging
   - Compute the macro-averaged precision and recall.
   - Compute the micro-averaged precision and recall.
   - Briefly explain the difference in interpretation between macro and micro averaging.
3. Programming Implementation
   Write Python code that:

   1. Accepts the confusion matrix above as input.
   2. Computes per-class precision and recall.
   3. Computes macro-averaged and micro-averaged precision and recall.
   4. Prints all results clearly.

Q6. Bigram Probabilities and the Zero-Probability Problem

You are given the following bigram counts from a small training corpus:

Previous word	Next words (with counts)
<s>	I: 2 , deep: 1
I	love: 2
love	NLP: 1 , deep: 1
deep	learning: 2
learning	</s>: 1 , is: 1
NLP	</s>: 1
is	fun: 1
fun	</s>: 1
ate	lunch: 6 , dinner: 3 , a: 2 , the: 1

Tasks:

1. Bigram Sentence Probabilities
   Using maximum likelihood estimation (MLE):

$$P(w_i \mid w_{i-1}) = \frac{C(w_{i-1}, w_i)}{C(w_{i-1})}$$

   - Compute the probability of sentence S1: <s> I love NLP </s>.
   - Compute the probability of sentence S2: <s> I love deep learning </s>.
   - Which sentence is more probable under the bigram model?

2. Zero-Probability Problem
   Using the same table, compute:
   - P(noodle|ate) with MLE.
   - Explain why this probability creates problems when computing sentence probabilities or perplexity.

- Apply Laplace smoothing (Add-1) to recompute P(noodle|ate). Assume the vocabulary size is 10 and total count after “ate” is 12.

Q7. Backoff Model (based on “Activity: <s> I like cats … You like dogs” slide)

Training corpus:

<s> I like cats </s>
<s> I like dogs </s>
<s> You like cats </s>

Counts:

- I like = 2
- You like = 1
- like cats = 2
- like dogs = 1
- cats </s> = 2
- dogs </s> = 1

Tasks:

1. Compute P(cats|I,like).
2. Compute P(dogs|You,like) using trigram → bigram backoff.
3. Explain why backoff is necessary in this example.

Q8. Programming: Bigram Language Model Implementation (based on “Activity: I love NLP corpus” slide)

Tasks:

Write a Python program to:

1. Read the training corpus:

2. <s> I love NLP </s>
3. <s> I love deep learning </s>
4. <s> deep learning is fun </s>

5. Compute unigram and bigram counts.
6. Estimate bigram probabilities using MLE.
7. Implement a function that calculates the probability of any given sentence.
8. Test your function on both sentences:
    - <s> I love NLP </s>

- <s> I love deep learning </s>

9. Print which sentence the model prefers and why.