Natural Language

MAP 1

2025

Student number: _____

Student name: _____

This MAP has two parts:

- The first part contains multiple-choice questions. There is only one correct answer for each question. Multiple-choice questions might have different scores.
- Write down your answer in the table bellow (**mandatory!**). The scoring only considers this table. If you make a mistake, cross out the wrong answer and write the new answer. If you leave two or more answers, the question will be considered wrong.
- The second part contains open question(s). Limit the size of responses to the available space. You can answer in Portuguese or English. Write legibly or your answer might not be evaluated.

During this evaluation there is no clarification of doubts. If you detect an error, mark it. If you are right, you will be given the full score for that question.

Good luck!

Table for Multiple-Choice Answers

Scoring: correct answer = 1, wrong answer = $-1/\#options$, no answer = 0.

Question	Chosen answer	Question	Chosen answer
1		6	
2		7	
3		8	
4		9	
5		10	

Professor corner (ignore this, please):

# Correct	# Wrong	# Not Answered	Open Questions

Formulas

$$\text{Precision: } P = \frac{TP}{TP + FP} \qquad \text{Recall: } R = \frac{TP}{TP + FN} \qquad F1 = \frac{2PR}{P + R}$$

$$\text{Jaccard}(s, t) = \frac{|s \cap t|}{|s \cup t|} \qquad \text{Dice}(s, t) = \frac{2|s \cap t|}{|s| + |t|}$$

TF-IDF:

$$\text{TF-IDF}(t, d, D) = TF(t, d) \times IDF(t, D)$$

$$TF(t, d) = \text{freq}(t, d) \qquad IDF(t, D) = \log \frac{|D|}{|\{d \in D : t \in d\}|}$$

Cosine Similarity:

$$\text{cosine\_similarity}(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \, \|\mathbf{B}\|}$$

Naïve Bayes:

$$\hat{c} = \arg\max_{c_i \in \mathcal{C}} P(c_i \mid x) \approx \arg\max_{c_i \in \mathcal{C}} P(c_i) \prod_{j=1}^{m} P(E_j \mid c_i)$$

Soundex:

1. Retain first letter.
2. Drop a, e, i, o, u, y, h, w.
3. Replace consonants:
    - b, f, p, v $\rightarrow$ 1
    - c, g, j, k, q, s, x, z $\rightarrow$ 2
    - d, t $\rightarrow$ 3
    - l $\rightarrow$ 4
    - m, n $\rightarrow$ 5
    - r $\rightarrow$ 6
4. Remove consecutive duplicates.
5. Pad with 0s if needed.

Multiple-Choice Questions (10 values)

Examples of possible multiple-choice questions:

1. Which of the following is an example of syntactic ambiguity?

   a) “John kissed his own wife”
   b) “The bank is on the river”
   c) “I saw the man on the hill with a telescope”
   d) “Step” as in “Your dog’s name is Step”

   Answer: c)

2. Which strings match the Regular Expression /(should\s)+I?\s(stay|go)\snow/ (/ are used as delimiters)

   a) should should I go now
   b) shouldIstaynow
   c) should I stay go
   d) shoud should go
   e) I stay now

   Answer: a)

3. Which of the following pairs of Regular Expressions are NOT equivalent?

   a) 1(01)* and (10)*1
   b) a* and (a|a+)
   c) x(xx)* and (xx)*x
   d) a+ and (a|a+)
   e) aa* and a+

   Answer: b)

4. You are evaluating a named entity recognizer that identifies ORGANIZATIONS against a reference. You obtain the following values: True Positive = 4, False Positives = 2, and False Negatives = 1. Which of the following sentences is ****NOT**** correct:

   a) Precision is 2/3
   b) The system missed one ORGANIZATION
   c) The system identified two entities that are not ORGANIZATIONS
   d) Recall is 4/6
   e) The system accurately identified 4 ORGANIZATIONS

   Answer: d)

5. Consider the Minimum Edit Distance (MED) between words “BIO” and “BIFE” and the following table:

		B	I	O
	0	1	2	3
B	1			
I	2		X	
F	3	Y		
E	4			

Which are the values of X and Y?

a) X is 1 and Y is 0

b) X is 0 and Y is 1

c) Both X and Y are 1

d) X is 0 and Y is 2

e) X is 1 and Y is 2

Answer: d)

Open Questions (10 values)

Write a short comment justifying the veracity/falsity of the following sentences or explain in up to five sentences a given concept. Convince us that you understand the concept. No formulas are needed to get the full score.

1. Question: Data hygiene refers to cleaning the data.

Possible answer: False. Data hygiene involves practices like ensuring training and test datasets are kept separate to maintain the integrity of model evaluation.

2. Explain two challenges associated with annotating toxic data?

Possible answer: Annotating toxic data is challenging due to the subjective nature of what is considered toxic, which can vary among individuals, and the potential emotional impact on annotators dealing with harmful content.

3. How does noise injection help in data augmentation?

Possible answer: Noise injection introduces intentional errors like typos or grammatical mistakes to simulate real-world data imperfections, enhancing the model's ability to handle diverse and noisy inputs.

4. What is the purpose of K-fold cross-validation?

Possible answer: K-fold cross-validation aims to evaluate model performance by splitting the dataset into K subsets, using each subset once as a test set while the rest serve as training data, to obtain a more reliable average performance estimate.