THE UNIVERSITY OF ARIZONA   Great Learning POWER AHEAD

Applied NLP - Project - [Marks 10]

Note: Each Question carries 0.25 marks including all sections.

I. True or False

1. We can use the spacy library to pre-process documents in Spanish. _________True

2. In keras, a TimeDistributed layer is a wrapper that allows a Recurrent layer to return an output for every token in the sequence. _________False

3. When using bag-of-words, if we sort the vocabulary alphanumerically the resulting vectors will be the same as if we sort it randomly. _________False

4. To represent out-of-vocabulary words with one-hot encoding we can use a vector where all the values are zeros. _________True

5. We could represent all the 10,000 words of a vocabulary with word embeddings of 10 dimensions. _________True

6. You trained a logistic regression model to predict if a text contains misinformation or not. For an input article, the model returns a predicted value of 0.43, so you could classify the article as containing misinformation. _________False

7. If you were working on a binary text classification problem and all the examples in the training set were positive (equal to 1), the cross-entropy would be equal to

$$-\frac{1}{N}\sum_{i=1}^{N} y_i \cdot log(p(y_i))$$

. _________True

8. Training a bi-directional Recurrent Neural Network is more efficient than a unidirectional Recurrent Neural Network because it requires less parameters to capture long dependencies. _________False

9. In a LSTM, if the output of the forget gate is a vector with all zeros, the unit should forget all the information from the cell state. _________True

10. Training a Neural Network for a sequence-to-sequence problem, we usually need to pad/truncate the input sequences to have the same length, but we don't need to pad/truncate the output sequences. _________True

11. Contextual word embeddings can handle word polysemy. _________True

12. A Pre-trained Language Model that has been already fine-tuned on a specific task cannot be longer fine-tuned on a different one. ________False

13. A sparse self-attention can be used to process longer sequences because its computational requirements grow quadratically. ________False

14. Just like GPT, training chatGPT does not involve human supervision. ________True

15. A sigmoid function only accepts input values between 0 and 1. ________False

II. Fill in the Blank

1. The ________Adjudicator is in charge of resolving discrepancies among the annotations produced by the annotators.

2. ________Stop words are words that can be filtered out from textual data because they are so frequent that they provide little information.

3. You have the sentence *"Time flies when you're having fun."* tokenized by words and annotated with Part-of-Speech. To represent this annotation following the BIO schema, there should be ________0__tokens with the label O.

4. A n-gram language model that only attends to the previous word in the sequence, is called a ________Bigram__language model.

5. The dot product of 2 word-embeddings is 10. If their magnitudes were 5 and 4, their cosine similarity would be 0.50.

6. You are working on a text classification problem with 3 classes, and you have implemented a model with a softmax in the output layer. For a specific input, the model returns the following probabilities: 0.31, 0.14 and 0.55.

7. A character-based tokenization of the sentence *"Can't wait, it's almost vacation time."* would result in ________38 tokens.

   HINT: Do not include the double quotes.

8. The maximum value of BLEU's Brevity Penalty is 1______.

9. ________Self Attention allows a deep-learning model to selectively focus on certain parts of the input sequence based on the relevance of each token to the others.

10. Given the following confusion matrix for a sentiment analysis model:

		Prediction		
		positive	Negative	neutral
Truth	positive	37	5	0
	negative	3	17	8
	neutral	16	21	32

The macro-average f1 score is ________**0.634**.

HINT: A multi-class confusion matrix for N classes can be converted into N one-vs-all binary confusion matrices.

III. Multiple Choice

1. What AutoClass of the transformers library could be used to instantiate a pre-trained model for a sequence labeling task?

a. AutoModelForTokenClassification
b. AutoModelForMaskedLM
c. AutoModelForSequenceClassification
d. AutoModelForCausalLM

2. Which of the following pre-processing steps should always be taken?

a. None, it depends on the task.
b. Word tokenization
c. Sentence segmentation
d. Lemmatization

3. A word that is very frequent in a document but very infrequent in the rest of the documents in a corpus will have:

a. High tf and high idf.
b. High tf and low idf.
c. Low tf and high idf.
d. Low tf and low idf.

4. Given the co-occurrence probabilities ratio $p(w_1|w_2)/p(w_1|w_3) = 10$, GloVe will

learn embeddings for w_1, w_2 and w_3 such that:

a. w_1 and w_2 are closer together than w_1 and w3.
b. w_1 and w_3 are closer together than w_1 and w2.
c. w_2 and w_3 are close together but far apart from w1.
d. w_1, w_2 and w_3 are all close together.

5. The range of the output values of a relu function is:

a. [0, ∞)
b. (-∞, ∞)
c. [0, 1]
d. [0, 10]

6. A model for Named Entity Recognition is able to identify all the entities in a test set, however the model is only able to predict one token per entity. For example, for the named entity *"Frida Kahlo"*, the model only identifies *"Frida"*. Using a relaxed evaluation, the precision of the model would be:

a. 1
b. 0
c. 0.5
d. Depends on the total number of tokens per entity.

7. Both GTP and BERT are based on the transformer architecture, but they only use part of it:

a. GPT uses the decoder and BERT uses the encoder.
b. GPT uses the encoder and BERT uses the decoder.
c. Both use the encoder.
d. Both use the decoder.

8. Which of the following statements about the Embedding layer is not true?

a. It is a lookup table that maps words to their indices.
b. Its weights can be trained as parameters of a neural network.
c. It can be initialized with random vectors.
d. It can be a square matrix.

9. Which of the following NLP approaches is most suitable for sentence segmentation?

a. Sequence Labeling
b. Text Classification
c. Sequence-to-Sequence
d. Language Modeling

10. Which of the following corruptions of the tokenized input *"SGD is an optimizer. It*

learns from errors." would not be used for pre-training BART?

a. "SGD is an error. It learns from optimizer."

b. "It learns from errors. SGD is an optimizer. "

c. "from errors. SGD is an optimizer. It learns"

d. "SGD. an optimizer. It learns. errors."

IV. Short Answer

1. Why do we need to train Tfidfvectorizer of scikit-learn on the training data?

Training the TfidfVectorizer is crucial for learning the vocabulary and document frequency parameters from the training set leading to accurate feature representation. Fitting it only on the training data prevents data leakage between the training and test sets. This also ensures consistency in preprocessing steps, aiding in better generalization on the test data.

2. What information should be included in the annotation guidelines?

Annotation guidelines are crucial for understanding task requirements and producing high-quality annotated data. They should include task definition, annotation schema, guidelines for each label or category, annotation process, inter-annotator agreement (IAA), data examples, quality control, tool usage, and ethical considerations.

3. It is possible to make Beam Search behave as Greedy Search. How?

Beam Search and Greedy Search are used with neural networks for sequence generation such as text generation. Beam Search can mimic Greedy Search by setting the beam width to 1 adjusting it dynamically for top-k sampling or setting the temperature parameter to infinity.

4. The [CLS] token provides an aggregate representation of the sequence that is used to fine-tune BERT for text classification tasks. What is the [CLS] token used for during BERT pre-training?

During BERT pretraining, the [CLS] token is vital for the Masked Language Modeling (MLM) and Next Sentence Prediction (NSP) objectives. In MLM, it provides an aggregate representation to predict masked tokens while in NSP its embedding summarizes the entire input sequence for accurate predictions.

5. How can we help to distinguish training examples from different tasks when pre-training a multitask seq-to-seq model?

During pre-training a multitask sequence-to-sequence model distinguishing training examples from different tasks is crucial for effective learning. This can be achieved by using task-specific tokens, different input formats, task-specific loss

functions and pretraining on diverse tasks followed by fine-tuning on task-specific datasets.

Happy Learning 🙂