Natural Language Processing
Computer Engineering Dept.
Dr. Sandra Wahid

Fourth year
Jan 15th, 2023

NLP Final Exam 2023

Question 1: Multiple Choice Questions (Only ONE correct choice, and no negatives):

1. The following statement is true about Skip-Gram:

a) Predicts target word from context

[✓ b) Predicts context from given target word]

[✓ c) Has one hidden layer that is fully-connected]

d) None of the above

2. Calculate cross-entropy loss of a single output of a model that predicts one of 4 classes uniformly randomly

a) -0.25 log(0.25)

b) -4 log (1/4)

[✓ c) -log (1/4)]

d) -4 log (4)

3. It is better to use **Macro** average when performance in all classes is equally important.

[✓ a) True]

b) False

4. The following statement is NOT an advantage of dense vectors to sparse vectors:

a) Easier to use as a feature in other Machine Learning models

[✓ b) Handles unknown words unlike long sparse vectors]

c) More general representation of rare words

d) Dense vectors are shorter than sparse vectors

5. The unknown words problem is solved by:

a) Laplace smoothing

b) interpolation

c) backoff

[✓ d) None of the above]

6. The CKY algorithm can parse a sentence under any context-free grammar.

a) True

[✓ b) False]

7. Encoding bigrams using a one-hot encoding, gives a vector representation of size ….

a) |V|

b) $|V^2|$

c) |2V|

[✓ d) $|V|^2$]

8. (0|1)*000 is a regular expression that matches …..

a) All even binary strings

b) All binary strings that are divisible by 4

[✓ c) All binary strings that are divisible by 8]

d) All binary strings ending with 00

9. Truncated backpropagation through time that can be used with RNNs ….

a) Limits forward but not backward

[✓ b) Limits backward but not forward]

c) Limits both forward and backward

d) Limits neither forward nor backward

10. The Markov chain representing a ~~b~~tri-gram language model is computed by:

a) $P(w_1)\prod_{i=2}^{n} p(w_i|w_{i-1}, w_{i-2})$

[✓ b) $P(w_1)P(w_2|w_1)\prod_{i=2}^{n-1} p(w_{i+1}|w_i, w_{i-1})$]

c) $P(w_1)P(w_2|w_1)\prod_{i=1}^{n} p(w_i|w_{i-1}, w_{i-2})$

d) $P(w_2|w_1)\prod_{i=2}^{n} p(w_i|w_{i-1}, w_{i-2})$

11. The Gazetteer that can be used in NER is …..

a) List of people names

b) List of animals names

[✓ c) List of places names]

d) None of the above

12. Attention mechanism allows the encoder to get information from all the hidden states of the decoder.

العكس

a) True

[✓ b) False]

~~**13.** Mean Reciprocal Rank (MRR) of a factoid Question Answering model decreases indicating an increase in performance.~~

~~a) True~~

~~b) False~~

مش في المنهج

لو حد لقاه يبعتلي اللوكيشن

14. Teacher Forcing method that omits redundancy by using a single set of embeddings at the input of the softmax layers.

a) True

[✓ b) False]

15. What is the CORRECT statement about Logistic Regression

[✓ a) Better to use when the features are strongly correlated]

b) Trains faster than Bayes's classifier

c) Considered a generative model

d) All of the above

Question 2:

A] CKY Problem

Non terminal:	Terminal:
S -> NP VP	NP -> b
VP -> V NP	V -> b
VP -> NP V	N -> a
VP -> VP VP	
NP -> N NP	

Required word: 'babbbb', is it generatable using above grammar rules? (Yes/No) Yes

b	a	b	b	b	b
NP, V	∅	VP From V NP	S	VP From VP VP	S
	N	NP From N NP	VP From NP V	S	VP From VP VP
		NP, V	VP From NP V , From V NP	S From NP VP	VP From VP VP
			NP, V	VP From NP V , From V NP	S From NP VP
				NP, V	VP From NP V , From V NP
					NP, V

B] Viterbi Problem

Given the tags (a1, a2, a3) and their initial probabilities (π), transition probabilities matrix (A), emission probabilities matrix (B), and given the sentence "b3b1b3b3b2".

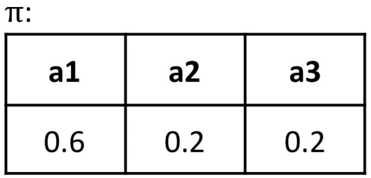
π:

a1	a2	a3
0.6	0.2	0.2

A (Transition Probabilities Matrix):

	a1	a2	a3
a1	0.8	0.1	0.1
a2	0.2	0.7	0.1
a3	0.1	0.3	0.6

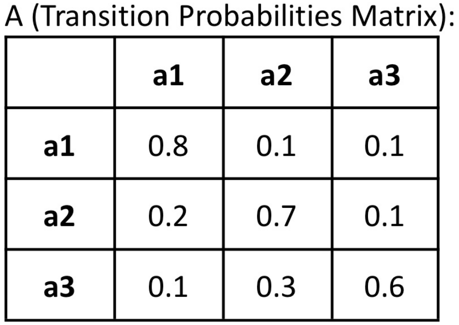
	b3	b1	b3	b3	b2
a_1	0.18	$0.7 \times \max(0.18 \times 0.8$ $, 0.16 \times 0.1)$ $= 0.1008$ back ← Pointer: a1	$0.3 \times \max(0.8 \times 0.1008$ $, 0.2 \times 0.048)$ $= 0.024192$ ← a1	$0.3 \times \max(0.8 \times 0.024192,$ $0.1 \times 8.064 \times 10^{-3})$ $= 5.80608 \times 10^{-3}$ ← a_1	0
a_2	0	$0.1 \times \max(0.18 \times 0.1$ $, 0.16 \times 0.3)$ $= 0.0048$ back ← Pointer: a3	0	0	$0.9 \times \max(5.80608 \times 10^{-3} \times 0.1$ $, 3.87072 \times 10^{-3} \times 0.3)$ ← a_3 $1. \_ \times 10^{-3}$
a_3	0.16	0	$0.8 \times \max(0.1 \times 0.1008$ $, 0.1 \times 0.048)$ ← a1 $= 8.064 \times 10^{-3}$	$0.8 \times \max(0.1 \times 0.024192$ $, 0.6 \times 8.064 \times 10^{-3})$ $= 3.87072 \times 10^{-3}$ ← a_3	$0.2 \times \max(0.1 \times 5.80608 \times 10^{-3}$ $, 0.6 \times 3.87072 \times 10^{-3})$ ← a_3 $= 4.64 \times 10^{-4}$



π:

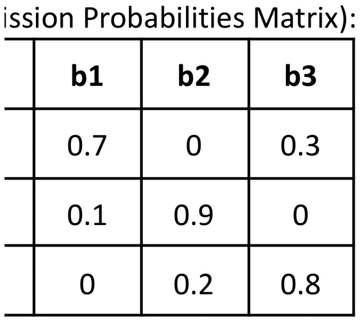
a1	a2	a3
0.6	0.2	0.2

$\boxed{a_1 \quad a_1 \quad a_3 \quad a_3 \quad a_2}$



A (Transition Probabilities Matrix):

	a1	a2	a3
a1	0.8	0.1	0.1
a2	0.2	0.7	0.1
a3	0.1	0.3	0.6



ission Probabilities Matrix):

	b1	b2	b3
	0.7	0	0.3
	0.1	0.9	0
	0	0.2	0.8

B (Emission Probabilities Matrix):

	b1	**b2**	**b3**
a1	0.7	0	0.3
a2	0.1	0.9	0
a3	0	0.2	0.8

Required to fill the Viterbi table and get the best tags for the given sentence.

Lecture 8, Slide 30

C] Give examples of Intrinsic and Extrinsic dense vectors performance evaluation.

D] Describe **very** briefly how Fasttext handles Unknown words.

by using subword models, representing each word as itself plus a bag of constituent n-grams, with special boundary symbols < and > added to each word. For example, with n = 3 the word where would be represented by the sequence <where> plus the character n-grams: <wh, whe, her, ere, re> Now, the vocabulary is the union of the subwords of all words (subwords refer to the words themselves + n-grams). Then a skip-gram/cbow embedding is learned for each subword. The vector of a word is the sum of the whole word vector in addition to the sum of its subwords' vectors. Unknown words can then be presented only by the sum of the constituent n-grams.

Question 3:

A]

An RNN with $U_{1\times1} = V_{1\times1} = W_{1\times1} = 1$, and linear activation & output functions $g(x) = x,\ f(h_i) = h_i$. What does it learn after training?

$h_t = h_{t-1} + X_t$ $\therefore$ Accumulator (Summer)

B] RNN Bottleneck Problem

You use an RNN to do sentiment analysis on hotel reviews. Each review receives a score of 0 (very negative) to 5 (very positive). The output depends on the last hidden layer only. The RNN was given this input: "The vacation was ruined, the food was bad, the rooms were bad. But surprisingly the staff seemed very happy". The model outputs "very positive". Why did this misclassification occur?

The misclassification may have occurred because the RNN relies on the last hidden layer, which might not capture the overall sentiment of the entire review effectively. In this case, the last part of the input, carries a positive sentiment, which influences the final hidden layer. Since the model focuses on this positive information at the end, it will incorrectly classify the overall sentiment as "very positive" despite the negative context in the earlier part of the review.

C]

A CBOW word2vec model learns the word vector matrix W_1, and a co-occurrence count based method also learns another word vector matrix W_2.

1. Are W_1, W_2 identical? (Yes/No) No
2. If the CBOW model was retrained, would the new vector matrix be identical to the old one W_1? (Yes/No) No لو عملت retraining ل neural network مش هتطلع نفس الحاجة
3. If the co-occurrence model was retrained, would the new vector matrix be identical to the old one W_2? (Yes/No) Yes, because we just count occurrences of words

D]

O O O O

Given the following sentence "NLP exam is held at faculty of engineering". Apply **BIOES** tagging. Entities of interest are "COURSE", "ORGANIZATION", and "PERSON".

S-COURSE B-ORG I-ORG E-ORG

E]

Given a random model and a test set containing a sentence of 30 characters drawn randomly from the English alphabet. (Note that the English alphabet contains 26 characters) Compute the Perplexity of this test sentence.

$$\sqrt[30]{\frac{1}{\left(\frac{1}{26}\right)^{30}}} = \sqrt[30]{26^{30}} = 26$$

Question 4:

A] Tweets Confusion Matrix Problem (Binary Classification)

	Golden "RUMOR"	Golden "NOT RUMOR"
System "RUMOR"	100	50
System "NOT RUMOR"	200	150

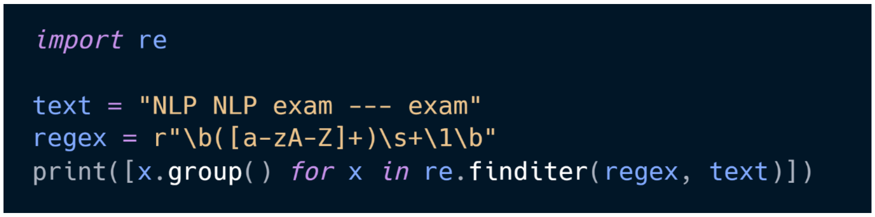
i. Calculate Precision, Recall, F1 $2/3$, $1/3$, $4/9$

ii. If it is very harmful to predict that tweets that are true "RUMOR" as "NOT RUMOR", is the above system the right choice? (Yes/No) No because cost of false negatives is high

iii. What is the reason for your answer to the question above? but recall of the model is low

B] Three Regex Problems. Write the output of the following code snippets:

i.

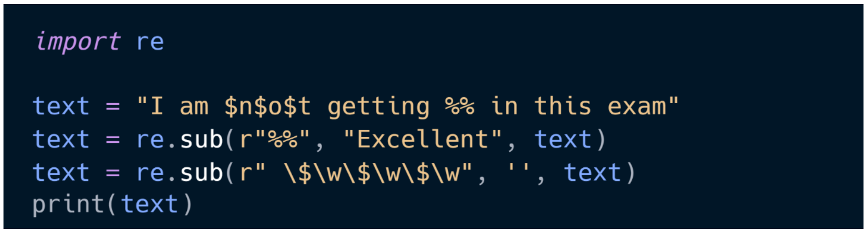


```
import re

text = "NLP NLP exam --- exam"
regex = r"\b([a-zA-Z]+)\s+\1\b"
print([x.group() for x in re.finditer(regex, text)])
```

NLP NLP

ii.

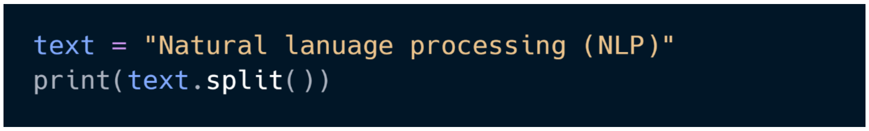


```
import re

text = "I am $n$o$t getting %% in this exam"
text = re.sub(r"%%", "Excellent", text)
text = re.sub(r" \$\w\$\w\$\w", '', text)
print(text)
```

I am getting Excellent in this exam

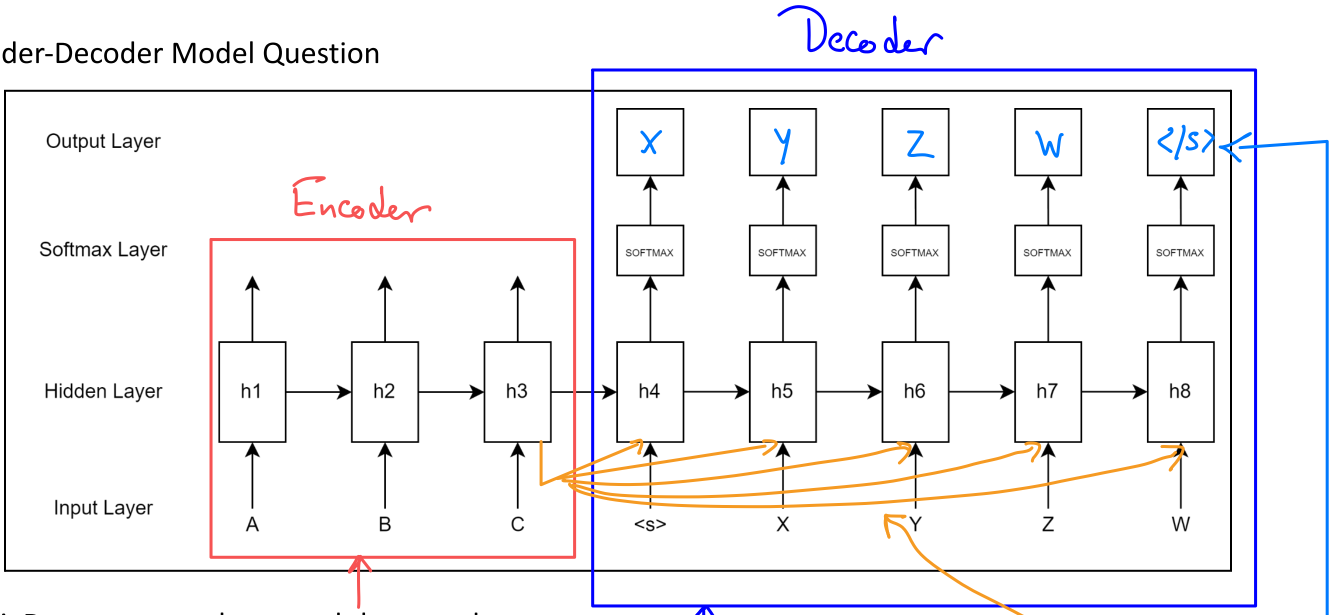
iii.



```
text = "Natural lanuage processing (NLP)"
print(text.split())
```

Natural, language, Processing, (NLP)

C] Encoder-Decoder Model Question



i. Draw a rectangle around the encoder part.

ii. Draw a rectangle around the decoder part.

iii. What is the hidden state vector in the encoder that represents the context vector? (h3)

iv. Complete the output layer of the decoder in the above diagram.

v. Draw a potential solution for the problem of the influence of the context vector being decreased as the decoder generates output. Lecture 12, slide 11

vi. There is still a problem, the context vector is a bottleneck, which can be fixed using (____) attention

vii. Mention the **name** of one score that calculates the relevance of each encoder's hidden state to the decoder. (dot Product)