1. When performing logistic regression on sentiment analysis, you represented each tweet as a vector of ones and zeros. However your model did not work well. Your training cost was reasonable, but your testing cost was just not acceptable. What could be a possible reason?

A. The vector representations are sparse and therefore it is much harder for your model to learn anything that could generalize well to the test set.
B. You probably need to increase your vocabulary size because it seems like you have very little features.
C. Logistic regression does not work for sentiment analysis, and therefore you should be looking at other models.
D. Sparse representations require a good amount of training time so you should train your model for longer

A

2. Which of the following are examples of text preprocessing?

A. Stemming, or the process of reducing a word to its word stem.
B. Lowercasing, which is the process of removing changing all capital letter to lower case.
C. Removing stopwords, punctuation, handles and URLs
D. Adding new words to make sure all the sentences make sense

ABC

3.

B

The sigmoid function is defined as h(x^(i) ,θ) = 1 / (1+e^(−θ^Tx^(i))) . Which of the following is true.

A. Large positive values of θ^Tx^(i) will make h(x^(i), θ) closer to 1 and large negative values of θ^Tx^(i) will make h(x^(i), θ) close to -1.
B. Large positive values of θ^Tx^(i) will make h(x^(i), θ) closer to 1 and large negative values of θ^Tx^(i) will make h(x^(i), θ) close to 0.
C. Small positive values of θ^Tx^(i) will make h(x^(i), θ) closer to 1 and large positive values of θ^Tx^(i) will make h(x^(i), θ) close to 0.
D. Small positive values of θ^Tx^(i) will make h(x^(i), θ) closer to 0 and large negative values of θ^Tx^(i) will make h(x^(i), θ) close to -1.

4. The cost function for logistic regression is defined as J(θ) = −1/m * ^{i=1}_m [y^(i)logh(x^(i), θ)+(1−-y^(i))log(1−h(x^(i), θ))]. Which of the following is true about the cost function above. Mark all the correct ones. AC

A. When y^(i) = 1, as h(x^(i), θ) goes close to 0, the cost function approaches .
B. When y^(i) = 1, as h(x^(i), θ) goes close to 0, the cost function approaches 0.
C. When y^(i) = 0, as h(x^(i), θ) goes close to 0, the cost function approaches 0.
D. When y^(i) = 0, as h(x^(i), θ) goes close to 0, the cost function approaches .

5. 0

For what value of θ^Tx in the sigmoid function does h(x^(i), θ)=0.5.

Enter answer here

6. Select all that apply. When performing logistic regression for sentiment analysis using the method taught in this week's lecture, you have to: ABD

A. Performing data processing.
B. Create a dictionary that maps the word and the class that word is found in to the number of times that word is found in the class.
C. Create a dictionary that maps the word and the class that word is found in to see if that word shows up in the class.
D. For each tweet, you have to create a positive feature with the sum of positive counts of each word in that tweet. You also have to create a negative feature with the sum of negative counts of each word in that tweet.

7. When training logistic regression, you have to perform the following operations in the desired order. B

A. Initialize parameters, get gradient, classify/predict, update, get loss, repeat
B. Initialize parameters, classify/predict, get gradient, update, get loss, repeat
C. Initialize parameters, get gradient, update, classify/predict, get loss, repeat
D. Initialize parameters, get gradient, update, get loss, classify/predict, repeat

8. Assuming we got the classification correct, where y^(i) = 1 for some specific example i. This means that h(x^(i), θ) > 0.5. Which of the following has to hold: D

A. Our prediction, h(x^(i), θ) for this specific training example is exactly equal to its corresponding label y^(i).
B. Our prediction, h(x^(i), θ) for this specific training example is less than 1−y^(i)).
C. Our prediction, h(x^(i), θ) for this specific training example is less than (1−h(x^(i), θ)).
D. Our prediction, h(x^(i), θ) for this specific training example is greater than (1−h(x^(i), θ)).

9. What is the purpose of gradient descent? Select all that apply. AC

A. Gradient descent allows us to learn the parameters θ in logistic regression as to minimize the loss function J.
B. Gradient descent allows us to learn the parameters θ in logistic regression as to maximize the loss function J.
C. Gradient descent, grad_theta allows us to update the parameters θ by computing θ = θ − α grad_theta
D. Gradient descent, grad_theta allows us to update the parameters θ by computing θ = θ + α grad_theta

10. What is a good metric that allows you to decide when to stop training/trying to get a good model? Select all that apply. AC

A. When your accuracy is good enough on the test set.

B. When your accuracy is good enough on the train set.
C. When you plot the cost versus (# of iterations) and you see that your the loss is converging (i.e. no longer changes as much).
D. When α, your step size is neither too small nor too large.

11. Assume that there are 2 happy people and 2 unhappy people in a room. Concretely, persons A and B are happy and persons B and C are unhappy. If you were to randomly pick a person from the room, what is the probability that the person is happy. A

A. 1/2
B. 1/4
C. 3/4
D. 0

12. Assume that there are 2 happy people and 2 unhappy people in a room. Concretely, persons A and B are happy and persons B and C are unhappy. If a friend showed you the part of the room where the two happy people are, what is the probability that you choose person B? A

A. 1/2
B. 1/4
C. 3/4
D. 1

13. From the equations presented below, express the probability of a tweet being positive given that it con- A

tains the word happy in terms of the probability of a tweet containing the word happy given that it is positive
P(Positive #"happy") = P(Positive)"happy") / P("happy")
P("happy" #Positive) = P("happy")Positive) / P(Positive)

A. P(Positive #happy) = P(happy #Positive) × P(Positive) / P(happy)
B. P(Positive #happy) = P("happy" #Positive) × P(happy) / P(Positive)
C. P(Positive Âhappy) = P(happy #Positive) × P(Positive) / P(happy)
D. P(Positive Âhappy) = P("happy" #Positive) × P(happy) / P(Positive)

14. Bayes rule is defined as — A

A. P(X¥) = P(Y¥) × P(X) / P(Y)
B. P(X¥) = P(Y¥) × P(Y) / P(X)
C. P(X¥) = P(X¥) × P(X) / P(Y)
D. P(X¥) = P(Y¥) × P(X) / P(Y¥)

15. Suppose that in your dataset, 25% of the positive tweets contain the word 'happy'. You also know that a total of 13% of the tweets in your dataset contain the word 'happy', and that 40% of the total number of tweets are positive. You observe the tweet: ''happy to learn NLP'. What is the probability that this tweet is positive? — 0.77

Enter answer here

16. The log likelihood for a certain word w_i is defined as: log(P(w_i #pos) / P(w_i #neg)). AC

A. Positive numbers imply that the word is positive.
B. Positive numbers imply that the word is negative.
C. Negative numbers imply that the word is negative.
D. Negative numbers imply that the word is positive.

17. The log likelihood mentioned in lecture, which is the log of the ratio between two probabilities is bounded between B

A. -1 and 1
B. − and
C. 0 and
D. 0 and 1

18. When implementing naive Bayes, in which order should the following steps be implemented. A

A. Get or annotate a dataset with positive and negative tweets
Preprocess the tweets: process_tweet(tweet) ž
Compute freq(w, class)
Get P(w | pos), P(w | neg)
Get λ(w)
Compute logprior = log(P(pos) / P(neg))

B. Get or annotate a dataset with positive and negative tweets
Preprocess the tweets: process_tweet(tweet) ž
Compute freq(w, class)

Get λ(w)
Get P(w | pos), P(w | neg)
Compute logprior = log(P(pos) / P(neg))

C. Get or annotate a dataset with positive and negative tweets
Compute freq(w, class)
Preprocess the tweets: process_tweet(tweet) ž
Get P(w | pos), P(w | neg)
Get λ(w)
Compute logprior = log(P(pos) / P(neg))

D. Get or annotate a dataset with positive and negative tweets
Compute freq(w, class)
Preprocess the tweets: process_tweet(tweet) ž
Compute logprior = log(P(pos) / P(neg)
Get P(w | pos), P(w | neg)
Get λ(w)

19. To predict using naive bayes, which of the following are required. A

A. X_val, Y_val, λ, logprior
B. X_val, Y_val, logprior
C. X_val, λ, logprior
D. Y_val, λ, logprior

20. Which of the following is NOT an application of naive Bayes? E

A. Sentiment Analysis
B. Author identification

C. Information retrieval
D. Word disambiguation
E. Numerical predictions

21. Given a corpus A, encoded as (1, 2, 3) and corpus B encoded as (4, 7, 2), What is the euclidean distance between the two documents? A

A. 5.91608
B. 35
C. 2.43
D. None of the above

22. Given the previous problem, a user now came up with a corpus C defined as (3, 1, 4) and you want to recommend a document that is similar to it. Would you recommend document A or document B? A

A. Document A
B. Document B

23. Which of the following is true about euclidean distance? AB

A. When comparing similarity between two corpuses, it does not work well when the documents are of different sizes.
B. It is the norm of the difference between two vectors.
C. It is a method that makes use of the angle between two vectors
D. It is the norm squared of the difference between two vectors.

24. What is the range of a cosine similarity score, namely s, in the case of information retrieval where the vectors are positive? C

A. -1 ≤ s ≤ 1
B. -∞ ≤ s ≤ ∞
C. 0 ≤ s ≤ 1
D. -1 ≤ s ≤ 0

25. The cosine similarity score of corpus A = (1, 0, −1) and corpus B = (2, 8, 1) is equal to ? A

A. 0.0851256530758748 6
B. 0
C. 1.251903
D. -0.3418283

26. We will define the following vectors, USA = (5, 6), Washington = (10, 5), Turkey = (3, 1), Ankara = (9, 1), Russian = (5, 5), and Japan = (4, 3). Using only the following vectors, Ankara is the capital of what country? D

A. Japan
B. Russia
C. Morocco
D. Turkey

27. Please select all that apply. PCA is AB

A. used to reduce the dimension of your data.
B. visualize word vectors

C. make predictions
D. label data

28. Please select all that apply. Which is correct about PCA? ABD

A. You can think of an eigenvector as an uncorrelated feature for your data.
B. The eigenvalues tell you the amount of information retained by each feature.
C. If working with features in different scales, you do not have to mean normalize.
D. Computing the covariance matrix is critical when performing PCA

29. In which order do you perform the following operations when computing PCA? A

A. mean normalize, get Σ the covariance matrix, perform SVD, then dot product the data, namely X, with a subset of the columns of U to get the reconstruction of your data.
B. mean normalize, perform SVD, get Σ the covariance matrix, then dot product the data, namely X, with a subset of the columns of U to get the reconstruction of your data.
C. get Σ the covariance matrix, perform SVD, then dot product the data, namely X, with a subset of the columns of U to get the reconstruction of your data, mean normalize.
D. get Σ the covariance matrix, mean normalize, perform SVD, then dot product the data, namely X, with a

subset of the columns of U to get the reconstruction of your data.

30. Vector space models allow us to — ABC

A. To represent words and documents as vectors.
B. build useful applications including and not limited to, information extraction, machine translation, and chatbots.
C. create representations that capture similar meaning.
D. build faster training algorithms

31. Assume that your objective is to minimize the transformation of X as similar to Y as possible, what would you optimize to get R? (XR≈Y) — A

A. Minimize the distance between XR and Y
B. Maximize the distance between XR and Y
C. Minimize the dot product between XR and Y
D. Maximize the dot product between XR and Y

32. When solving for R, which of the following is true? — C

A. Create a forloop, inside the forloop: (initialize R, compute the gradient, update the loss
B. Create a forloop, inside the forloop: (initialize R, update the loss, compute the gradient.
C. Initialize R, create a forloop, inside the forloop: (compute the gradient, update the loss)
D. Initialize R, compute the gradient, create a forloop, inside the forloop: (update the loss)

33. The Frobenius norm of A = (1, 3; 4, 5) is — 7.14

Enter answer here

34. Assume XR^m×n, RR^n×n, YR^m×n which of the following is the gradient of ‖XR−Y‖F^2? — A

A. 2/m × X^T × (XR-Y)
B. 2/m × X × (XR-Y)
C. 2/m × (XR-Y) × X
D. 2/m × (XR-Y) × X^T

35. Imagine that you are visiting a city in the US. If you search for friends that are living in the US, would you be able to determine the 2 closest of ALL your friends around the world? — B

A. Yes, because I am already in the country and that implies that my closest friends are also going to be in the same country.
B. No

36. What is the purpose of using a function to hash vectors into values? — AB

A. To speed up the time it takes when comparing similar vectors.
B. To not have to spend time comparing vectors with other vectors that are completely different.
C. To make the search for other similar vectors more accurate.
D. It helps us create vectors.

37. Given the following vectors, determine the true statements.
P: (1, 1)
V_1: (1, 1)
V_2: (2, 2)
V_3: (−1, −1)

A. P × V_1^T and P × V_2^T have the same sign.
B. P × V_1^T and P × V_2^T are equal in magnitude.
C. P × V_1^T and P × V_3^T have the same sign.

A

38. We define H to be the number of planes and h_i to be 1 or 0 depending on the sign of the dot product with plane i. Which of the following is the equation used to calculate the hash for several planes.

A. _i^H 2^i × h_i
B. _i^H 2^i × h_i^i
C. _i^H 2i × h_i
D. _i^H 2^h_i × i

A

39. How can you speed up the look up for similar documents.

A. PCA
B. Approximate Nearest Neighbors
C. K-Means
D. Locality sensitive hashing

BD

40. Hash tables are useful because

A. allow us to divide vector space to regions.
B. speed up look up

ABD

C. classify with higher accuracy
D. can always be reproduced

41. The Transition matrix A defined in lecture allows you to: C

A. Compute the probability of going from a word to another word.
B. Compute the probability of going from a part of speech tag to a word.
C. Compute the probability of going from a part of speech tag to another part of speech tag.
D. Compute the probability of going from a word to a part of speech tag.

42. The Emission matrix B defined in lecture allows you to: D

A. Compute the probability of going from a word to another word.
B. Compute the probability of going from a part of speech tag to another part of speech tag.
C. Compute the probability of going from a word to a part of speech tag.
D. Compute the probability of going from a part of speech tag to a word.

43. The column sum of the emission matrix has to be equal to 1. A

A. False.
B. True.

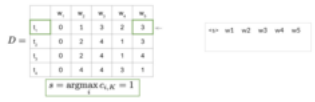
44. The row sum of the transition matrix has to be 1. A

A. True
B. False, it has to be the column sum.

45. Why is smoothing usually applied? Select all that apply. AC

A. Applying smoothing, for the majority of cases, allows us to decrease the probabilities in the transition and emission matrices and this allows us to have non zero probabilities.
B. Applying smoothing is a bad idea and we should not use it.
C. Applying smoothing, for the minority of cases, allows us to increase the probabilities in the transition and emission matrices and this allows us to have non zero probabilities.
D. Applying smoothing, for the majority of cases, allows us to increase the probabilities in the transition and emission matrices and this allows us to have non zero probabilities.

46. Given the following D matrix, what would be the sequence of tags for the words on the right? A



A. t_2, t_3, t_1, t_3, t_1
B. t_3, t_4, t_2, t_2, t_1
C. t_1, t_3, t_1, t_2, t_1
D. t_3, t_4, t_2, t_3, t_1

47. Previously, we have been multiplying the raw probabilities, but in reality we take the log of those proba- C

bilities. Why might that be the case?

A. The log probabilities help us with the inference as they bound the numbers between -1 and 1.
B. Because the log probabilities force the numbers to be between 0 and 1 and hence, we want to take a probability.
C. We take the log probabilities because probabilities are bounded between 0 and 1 and as a result, the numbers could be too small and will go towards 0.
D. The log probabilities should not be used because they introduce noise to our original computed scores.

48. Which of the following are useful for applications for parts of speech tagging? ABD

A. Named Entity Recognition
B. Coreference Resolution
C. Sentiment Analysis
D. Speech recognition

49. Corpus: "In every place of great resort the monster was the fashion. They sang of it in the cafes, ridiculed it in the papers, and represented it on the stage. " (Jules Verne, Twenty Thousand Leagues under the Sea)
In the context of our corpus, what is the probability of word "papers" following the phrase "it in the". D

A. P(papers|it in the) = 0
B. P(papers|it in the) =1
C. P(papers|it in the) = 2/3
D. P(papers|it in the) = 1/2

50. Given these conditional probabilities D
P(Mary)=0.1; P(likes)=0.2; P(cats)=0.3 . P(Mary|likes) =0.2; P(likes|Mary) =0.3; P(cats|likes)=0.1; P(likes|cats)=0.4
Approximate the probability of the following sentence with bigrams: "Mary likes cats"

A. P(Mary likes cats) =1
B. P(Mary likes cats) = 0
C. P(Mary likes cats) = 0.008
D. P(Mary likes cats) = 0.003

51. Given these conditional probabilities B
P(Mary)=0.1; P(likes)=0.2; P(cats)=0.3
P(Mary|<s>)=0.2; P(</s>|cats)=0.6
P(likes|Mary) =0.3; P(cats|likes)=0.1
Approximate the probability of the following sentence with bigrams: "<s> Mary likes cats </s>"

A. P(<s> Mary likes cats </s>) = 1
B. P(<s> Mary likes cats </s>) = 0.0036
C. P(<s> Mary likes cats </s>) = 0
D. P(<s> Mary likes cats </s>) = 0.003

52. Given the logarithm of these conditional probabilities: D
log(P(Mary|<s>))=-2; log(P(</s>|cats))=-1
log(P(likes|Mary)) =-10; log(P(cats|likes))=-100
Approximate the log probability of the following sentence with bigrams : "<s> Mary likes cats </s>"

A. log(P(<s> Mary likes cats </s>)) = 2000
B. log(P(<s> Mary likes cats </s>)) = 113

C. log(P(<s> Mary likes cats </s>)) = -112
D. log(P(<s> Mary likes cats </s>)) = -113

53. Given the logarithm of these conditional probabilities:
log(P(Mary|<s>))=-2; log(P(</s>|cats))=-1
log(P(likes|Mary)) =-10; log(P(cats|likes))=-100
Assuming our test set is W="<s> Mary likes cats </s>", what is the model's perplexity. C

A. log PP(W) = -113
B. log PP(W) = (-1/5)*(-113)
C. log PP(W) = (-1/4)*(-113)
D. log PP(W) = (-1/5)*113

54. Given the training corpus and minimum word frequency=2, how would the vocabulary for corpus preprocessed with <UNK> look like?
"<s> I am happy I am learning </s> <s> I am happy I can study </s>" D

A. V = (I,am,happy,learning,can,study)
B. V = (I,am,happy,learning,can,study,<UNK>)
C. V = (I,am,happy,I,am)
D. V = (I,am,happy)

55. Corpus: "I am happy I am learning"
In the context of our corpus, what is the estimated probability of word "can" following the word "I" using the bigram model and add-k-smoothing where k=3. C

A. P(can|I) = 0
B. P(can|I) =1

C. P(can|I) = 3/(2+3*4)
D. P(can|I) = 3/(3*4)

56. Which of the following are applications of n-gram language models? ABCD

A. Speech recognitions
B. Auto-complete
C. Auto-correct
D. Augmentative communication
E. Sentiment Analysis

57. The higher the perplexity score the more our corpus will make sense. B

A. True
B. False

58. The perplexity score increases as we increase the number of <UNK> tokens. B

A. False.
B. True.

59. Which one of the following word representations is most likely to correspond to a word embedding representation in a general-purpose vocabulary? In other words, which one is most likely to capture meaning and important information about the words? B

A. car -> 2
caravan -> 3

B. car -> (0.1 1)

caravan -> (-0.1 0.9)

C. car -> (0 1 0 0)
caravan -> (0 0 1 0)

D. car -> (1 0.1)
caravan -> (-1 -0.9)

60. Which one of the following statements is correct? D

A. To learn word embeddings you only need a vocabulary and an embedding method.
B. Learning word embeddings using a machine learning model is unsupervised learning as the input data set is not labelled.
C. The objective of a machine learning model that learns word embeddings is to predict word embeddings.
D. The meaning of the words, as carried by the word embeddings, depends on the embedding approach.

61. Which one of the following statements is false? C

A. word2vec-based models cannot create word embeddings for words they did not see in the corpus they were trained on.
B. ELMo may have different word embeddings for the word "stable" depending on the context.
C. You need to train a deep neural network to learn word embeddings.
D. You can use a pre-trained BERT model to learn word embeddings on a previously unseen corpus.

62. Consider the corpus "A robot may not injure a human being or, through inaction, allow a human being to come to harm." and assume you are preparing data to train a CBOW model. Ignoring punctuation, for a context half-size of 3, what are the context words of the center word "inaction"? B

A. "being inaction human"
B. "being or through allow a human"
C. "being or through inaction allow a human"
D. "through inaction allow"

63. Which one of the following statements is false? C

A. Given the corpus "I think therefore I am", the word "think" could be represented by the one-hot vector (1 0 0 0).
B. Consider the corpus "A robot may not injure a human being or, through inaction, allow a human being to come to harm." and assume you are preparing data to train a CBOW model. Ignoring punctuation, for a context size of 3, the context words of the center word "inaction" are: "a", "allow", "being", "human", "or", and "through"
C. The continuous bag-of-words model learns to predict context words given a center word.
D. Given the corpus "I think therefore I am", the word "you" cannot be represented.

64. You are designing a neural network for a CBOW model that will be trained on a corpus with a vocabulary of 8000 words. If you want it to learn 400-dimensional C

word embedding vectors, what should be the sizes of the input, hidden, and output layers?

A. 8000 (input layer), 8000 (hidden layer), 400 (output layer)
B. 400 (input layer), 400 (hidden layer), 8000 (output layer)
C. 8000 (input layer), 400 (hidden layer), 8000 (output layer)
D. 8000 (input layer), 400 (hidden layer), 400 (output layer)

65. If you are designing a neural network for a CBOW model that will be trained on a corpus of 8000 words, and if you want it to learn 400-dimensional word embedding vectors, what should be the size of W1, the weighting matrix between the input layer and hidden layer, if it is fed training examples in batches of 16 examples represented by a 8000 row by 16 column matrix?
Hint: if X is the input matrix, H the matrix for the hidden layer, and B1 the bias matrix, then H = ReLU(W1X + B1). A

A. 400 rows by 8000 columns
B. 16 rows by 8000 columns
C. 400 rows by 16 columns
D. 8000 rows by 16 columns

66. Given the input vector x below, a trained continuous bag-of-words model outputs the vector ŷ below. What is the word predicted by the model? C

A. Think
B. am
C. Therefore
D. I

67. The following weighting matrix W_1 has been learned after training a CBOW model. You are also given word-to-row mapping for the input column vectors. What is the word embedding vector for "ring"?

A. [-1.9; -1.18; 2.61; -1.7; -4.36; -4.54]
B. [-2.39; -1.3; -1.7; 1.75]
C. [4.56; -2.94; 2.61; -1.16]
D. [0; 0; 1; 0; 0; 0]

C

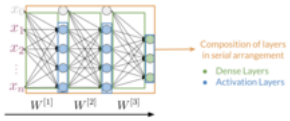
68. Select all that are correct.

A. You can perform intrinsic evaluation by using a clustering algorithm to group similar word embedding vectors, and determining if the clusters capture related words.
B. Extrinsic evaluation evaluates actual usefulness of embeddings, is time consuming and is more difficult to trouble shoot.
C. To evaluate word embeddings with intrinsic evaluation, you use the word embeddings to perform an external task, which is typically the real-world task that you initially needed the word embeddings for. Then, use the performance metric of this task as a proxy for the quality of the word embeddings.
D. To evaluate word embeddings with extrinsic evaluation, you use the word embeddings to perform an

ABD

external task, which is typically the real-world task that you initially needed the word embeddings for. Then, use the performance metric of this task as a proxy for the quality of the word embeddings.

69. How many layers does the following neural network have? C



A. 1
B. 2
C. 3
D. 4

70. Let us analyze the following class: D

What would be the output above?

A. 12
B. Null
C. 2
D. 14

71. The ReLU layer, is an activation layer that typically follows a dense fully connected layer, and transforms all values between 0 and 1 before sending them on to the next layer. B

A. True
B. False

72. The ReLU layer is an activation layer that typically follows a dense fully connected layer, and transforms any negative values to 0 before sending them on to B

the next layer.

A. False
B. True

73. For the embedding layer in your model, you'd have to learn a matrix of weights of what size? D

A. Equal to your vocabulary times the dimension of the number of layers
B. Equal to your vocabulary times the dimension of the number of classes
C. Equal to the dimension of the embedding times the first dimension of the matrix in the first layer.
D. Equal to your vocabulary times the dimension of the embedding

74. What would be the probability of a five word sequence using a penta-gram? B

A. P(w5#w4,w3,w2,w1) = count(w5,w4,w3,w2,w1) / count(w4,w3,w2,w1)
B. P(w5,w4,w3,w2,w1) = P(w1)×P(w2#w1)×P(w3#-w1,w2)×P(w4#w1,w2,w3)×P(w5#w1,w2,w3,w4)
C. P(w5,w4,w3,w2,w1) = P(w1)×P(w2)×P(w3)×P(w4)×P(w5)
D. P(w5,w4,w3,w2,w1) = P(w5#w4,w3,w2,w1)

75. The number of parameters in an RNN is the same regardless of the input's length. B

A. False
B. True.

76. Select all the examples that correspond to a "many to one" architecture. BD

A. An RNN which inputs a sentiment and generates a sentence.
B. An RNN which inputs a sentence and determines the sentiment.
C. An RNN which inputs a topic and generates a conversation about that topic.
D. An RNN which inputs a conversation and determines the topic.

77. What should be the size of matrix Wh, if h<t> had size 4x1 and x<t> 10x1? A
h<t> = g(Wh[h<t−1>,x<t>]+bh)

A. 4x14
B. 14x4
C. 4x4
D. 14x14

78. In the next equation, why is there a division by the number of time steps but not one for the number of classification categories? A
J = −1/T ^T_t=1 ^K_j=1 yj<t> log ŷj<t>

A. Because there is just one value in every vector y<t> different from zero.
B. Because the equation is wrong.
C. Because this equation is given for a single example.
D. Because for most classification tasks there are only two categories.

79. What problem, related to vanilla RNNs, do GRUs tackle? A

A. Loss of relevant information for long sequences of words.
B. Overfitting
C. High computational time for training and prediction.
D. Restricted flow of information from the past to the present.

80. Bidirectional RNNs are acyclic graphs, which means that the computations in one direction are independent from the ones in the other direction. A

A. True
B. False

81. Compared to Traditional Language models which of the following problems does an RNN help us with? AB

A. Helps us solve memory issues.
B. Helps us solve RAM issues.
C. They require almost no knowledge to use when compared to the traditional n-gram model.
D. They are much simpler to understand.

82. What type of RNN structure would you use when implementing machine translation? D

A. Many to one
B. One to many

C. One to one
D. Many to Many

83. In the scan() function the variable cur_value corresponds to the hidden state in an RNN. A

A. True
B. False

84. Identify the correct order of the gates that information flows through in an LSTM unit. B

A. Input gate, forget gate, output gate.
B. Forget gate, input gate, output gate.
C. Output gate, forget gate, input gate.
D. Forget gate, output gate, input gate

85. Which are some applications of LSTMs? ABCDE

A. Music composition
B. Image captioning
C. Next character prediction
D. Chatbots
E. Speech recognition

86. The tanh layer ensures the values in your network stay numerically stable, by squeezing all values between -1 and 1. This prevents any of the values from the current inputs from becoming so large that they make the other values insignificant. B

A. False
B. True

87. What type of architecture is a named entity recognition using? — A

A. Many to many
B. Many to one
C. One to many

88. Extract the named entities from the following sentence:
Younes, a Moroccan artificial intelligence engineer, travelled to France for a conference. — B

A. Younes, Moroccan, engineer.
B. Younes, Moroccan, France.
C. Younes, Moroccan, conference.
D. Younes, Moroccan engineer, France.

89. In a vectorized representation of your data, equal sequence length allows more efficient batch processing. — B

A. False
B. True.

90. Which built-in Python method would you use to iterate over your test set during the evaluation step? Assuming you are using a data generator. — A

A. next()
B. slice()
C. list()
D. enumerate()

91. — B

Why is it important to mask padded tokens when computing the loss?

A. We add the loss of the padded tokens independently.
B. Padded tokens are not part of the data and are just used to help us keep the same sequence length for more efficient batch processing. We should not include their loss.

92. In which of the following orders should we train an Named Entity Recognition with an LSTM? C

A. Create a tensor for each input and its corresponding number
Put them in a batch => 64, 128, 256, 512 ...
Run the output through a dense layer
Feed it into an LSTM unit
Predict using a log softmax over K classes

B. Create a tensor for each input and its corresponding number
Put them in a batch => 64, 128, 256, 512 ...
Run the output through a dense layer
Predict using a log softmax over K classes
Feed it into an LSTM unit

C. Create a tensor for each input and its corresponding number
Put them in a batch => 64, 128, 256, 512 ...
Feed it into an LSTM unit
Run the output through a dense layer
Predict using a log softmax over K classes

93. LSTMS solve vanishing/exploding gradient problems when compared to basic RNNs. A

A. True
B. False

94. Classification allows you to identify similarity between two things while siamese networks allow you to categorize things. B

A. True
B. False

95. Do the two subnetworks in a siamese network share the same parameters? B

A. No
B. Yes

96. When training a siamese network to identify duplicates, which pairs of questions from the following questions do you expect to have the highest cosine similarity ?
Is learning NLP useful for me to get a job? (ANCHOR)
What should I learn to get a job? (POSITIVE)
Where is the job? (NEGATIVE) A

A. Anchor, Positive
B. Anchor, Negative
C. Negative, Positive

97. In the triplet loss function below, will decreasing the hyperparameter alpha from 0.5 to 0.2 require more, A

or less, optimization during training ?
diff = s(A,N) − s(A,P)
L(A,P,N) = max(diff+α, 0)

A. Less
B. More.

98. The orange square below corresponds to the similarity score of question duplicates? B

A. True
B. False

99. What is the closest negative in this set of numbers assuming a duplicate pair similarity of 0.6? C
[-0.9,-0.4,0.4, 0.8]

A. -0.9
B. -0.4
C. 0.4
D. 0.8

100. In one shot learning, is any retraining required when new classes are added? For example, a new bank customer's signature. B

A. Yes
B. No

101. During training, you have to update the weights of each of the subnetworks independently. A

A. False.
B. True.

102. The mean negative is defined as the closest off-diagonal value to the diagonal in each row (excluding the diagonal). B

A. True
B. False

103. In what order are Siamese networks performed in lecture? A

A. Convert each input into an array of numbers
Feed arrays into your model
Compare ~~1, 2 u~~sing cosine similarity
Test against a threshold

B. Convert each input into an array of numbers
Feed arrays into your model
Run logistic regression classifier
Classify by using the probability

C. Convert each input into an array of numbers
Feed arrays into your model
Run soft-max classifier for all classes
Take the arg-max of the probabilities

D. Convert each input into an array of numbers
Feed arrays into your model
Compare ~~1, 2 u~~sing euclidean distance
Test against a threshold

104. Which of the following are bottlenecks when implementing seq2seq models? AB

A. You are trying to store variable length sequences in a fixed memory, for example, you are trying to store articles of different lengths in a fixed 100 dimensional vector.
B. There are vanishing/exploding gradient problems.
C. They require a lot of memory.
D. They are not that useful

105. What are some of the benefits of using attention? BD

A. The use of attention ends up giving you less accurate results.
B. It helps with the information bottleneck issue.
C. It is significantly slower to use attention and therefore it is not recommended to use it.
D. It allows you to focus on the parts that matter more.

106. What are the major components in the attention mechanism that are required? Select all that apply. ACDE

A. Softmax
B. Cosine similarity.
C. Queries: described in the lesson as the "ask" you are trying to match with the key.
D. Values: not really described in lecture, but you can think of them just like the keys for now. (Hint: you need this for attention).
E. Keys: desrcibed in the lesson as the object you are looking for.

107. Which sentinel is used in lecture to represent the end of sentence token in machine translation? B

A. 0
B. 1
C. infty
D. -infty

108. Teacher forcing uses the actual output from the training dataset at time step y(t) as input in the next time step X(t+1), instead of the output generated by your model. A

A. True.
B. False.

109. The BLEU score's range is as follows: A

A. The closer to 0, the worse it is, the closer to 1, the better it is.
B. The closer to 1, the worse it is, the closer to 0, the better it is.
C. The closer to -1, the worse it is, the closer to 1, the better it is.
D. The closer to −, the worse it is, the closer to , the better it is.

110. Bleu is defined as: A

A. (Sum of unique n-gram counts in the candidate)/ (total # of words in candidate).
B. (Sum of n-gram counts in the candidate)/ (total # of words in candidate).
C. (Sum of overlapping unigrams in model and reference)/(total # of words in reference)

D. (Sum of unique unigrams in model and reference)/(total # of words in reference)

111. What is the difference between precision and recall in A Rouge?

A. Precision is defined as:
(Sum of overlapping unigrams in model and reference)/(total # of words in model)
Recall is defined as:
(Sum of overlapping unigrams in model and reference)/(total # of words in reference)

B. Recall is defined as:
(Sum of overlapping unigrams in model and reference)/(total # of words in model)
Precision is defined as:
(Sum of overlapping unigrams in model and reference)/(total # of words in reference)

C. Recall is defined as:
(Sum of unigrams in model and reference)/(total # of words in model)
Precision is defined as:
(Sum of overlapping unigrams in model and reference)/(total # of words in reference)

D. Precision is defined as:
(Sum of overlapping bigrams in model and reference)/(total # of words in model)
Recall is defined as:
(Sum of overlapping bigrams in model and reference)/(total # of words in reference)

112. Greedy decoding — A

A. Allows you select the word with the highest probability at each time step.
B. Allows you randomly select the word according to its own probability in the softmax layer.
C. Selects multiple options for the best input based on conditional probability.
D. Makes use of the Minimum Bayes Risk method.

113. When implementing Minimum Bayes Risk method in decoding, let's say with 4 samples, you have to implement the following.
Calculate similarity score between sample 1 and sample 2
Calculate similarity score between sample 1 and sample 3
Calculate similarity score between sample 1 and sample 4
Average the score of the first 3 steps (Usually a weighted average)
Repeat until all samples have overall scores
Pick the golden one with the highest similarity score. — A

A. True
B. False

114. Select all the correct answers. — ABD

A. With transformers, the vanishing gradient problem isn't related with length of the sequences because we have access to all word positions at all times.

B. Transformers are able to take more advantage from parallel computing than other RNN architectures previously covered in the course.
C. Transformers are models that use both recurrent units and attention mechanisms.
D. Even RNN architectures like GRUs and LSTMs don't work as well as transformers for really long sequences.

115. Which of the following are applications of transformers? E

A. Text summarization.
B. Translation
C. Question Answering
D. Chatbots
E. All of the above.

116. What is one of the biggest techniques that the T5 model brings about? B

A. It's attention mechanism is far more superior than the one used in other models.
B. It makes use of transfer learning and the same model could be used for several applications. This implies that other tasks could be used to learn information that would benefit us on different tasks.
C. T5 model is very cheap to train from scratch.
D. It allows for interpretability.

117. When it comes to translating french to english using dot product attention: ACD

A. The queries are the english words and the keys and values are the french words.
B. A CPU is more than enough to train this type of model.
C. You find the distribution by multiplying the queries by the keys (you might need to scale), take the softmax and then multiply it by the values.
D. The intuition is that each query qi, picks most similar key kj. This allows the attention model to focus on the right words at each time step.

118. Which of the following corresponds to the causal (self) attention mechanism? B

A. One sentence (decoder) looks at another one (encoder)
B. In one sentence, words look at previous words (used for generation). They can not look ahead.
C. In one sentence, in this attention mechanism, words look at both previous and future words.
D. In causal attention, queries and keys come from different sentences and queries search among words before only

119. Let's explore multi-headed attention in this problem. Select all that apply. ABC

A. Each head learns a different linear transformations to represent words.
B. Those linear transformations are combined and run through a linear layer to give you the final representation of words.

C. Multi-Headed models attend to information from different representations at different positions
D. Multi-Headed attention allows you to capture less information than single headed attention.

120. Which of the following is true about about bi-directional attention? C

A. It only attends to words before.
B. It used an encoder and decodes it using a decoder.
C. It could attend to words before and after the target word.
D. It is less powerful than regular uni-directional attention.

121. Why is there a residual connection around each attention layer followed by a layer normalization step in the in the decoder network? A

A. To speed up the training, and significantly reduce the overall processing time.
B. To help with the interpretability.
C. To help with the parallel computing component during the training.
D. To break the symmetry in the back-prop.

122. The structure of the text input when implementing a summarization task is as follows: C

A. Article <EOS> separator, the summary, and another <EOS>
B. <SOS> Article <EOS> <SOS> the summary, <EOS>

C. ARTICLE TEXT <EOS> SUMMARY <EOS> <pad>
D. <SOS> Article, the summary, and <EOS>

123. In the lecture, the way summarization is generated is using: C

A. Next sentence prediction.
B. Next character generation.
C. Next word generation.
D. By extracting key sentences from the original article.

124. The English wikipedia is about 13 GB. The T5 model, that you will be working with is trained on the C4 corpus, which is how many GB? D

A. 26 GB
B. 130 GB
C. 256GB
D. 800 GB

125. Which of the following are true about pre-training in NLP? ABC

A. It allows you to get better results.
B. It speeds training.
C. It allows you to use information learned from a different task while working on a specific task.
D. It is not recommended because it takes a long time to pre-train a model.

126. What is fine-tuning in NLP? A

A. Fine tuning means taking existing weights of deeplearning model, and tweaking them a little bit to get a desired output, usually better results, on some specific task.
B. Fine tuning means taking existing weights from a deeplearning model, let's say word embeddings, and then using those weights in another model as they are without changing them.
C. Fine-tuning slows down your training.
D. Fine tuning allows you to better prepare your data for training.

127. Select all that apply for Masked Language Modeling. (MLM) ABC

A. The goal is to predict the masked token.
B. Choose 15% of the tokens at random: mask them 80% of the time, replace them with a random token 10% of the time, or keep as is 10% of the time.
C. The cross entropy loss over V classes is used when doing the prediction.
D. There could only be one masked span in a sentence.

128. What does the BERT objective consist of? B

A. It consists of a binary loss for next sentence prediction.
B. It consists of the sum of a binary loss used for next sentence prediction and a cross entropy loss over V tokens used for the masked language modeling.
C. It consists of a cross entropy loss over V tokens used for the masked language modeling.

D. It consists of a triplet loss similar to the one you have seen used for siamese networks.

129. Which of the following inputs could be used for the BERT model? D

A. Question/Answer
B. Article/Summary
C. Hypothesis/Premise
D. All of the above

130. How does the prefix language model attention work in the T5 model? C

A. It uses bidirectional attention for the X's and the Y's.
B. It uses an encoder decoder attention.
C. It uses bidirectional attention for the inputs (i.e. X's) and causal attention mapping the outputs (Y's) at time t, to all the previous X's and outputs before timestep t.
D. It only uses causal attention through out.

131. When training these latest NLP models, you end up training a model that can do many tasks. For example, you usually have data for sentiments, QA, chatbot, summarization, etc. The question now is how do you combine the datasets using temperature scaled mix-ing? B

A. You will sample in proportion to the size of each task's dataset.
B. You will adjust the "temperature" of the mix-

ing rates. This temperature parameter allows you to weight certain examples more than others. When T = 1, this approach is equivalent to examples-proportional mixing and as T increases the proportions become closer to equal mixing.
C. Each example in each batch is sampled uniformly at random from one of the datasets you train on.
D. You will just use the data for the specific task you are training on.

132. When doing fine-tuning, how do adapter layers work? A

A. It allows you to add a new layer and then you only fine-tune the new layer you added.
B. You freeze only the last layer, and then you gradually unfreeze each layer as you modify and fine-tune each layer starting from the end.
C. You freeze half the layers, and then you gradually unfreeze each layer as you modify and fine-tune one at a time.
D. You just take the pre-trained weights and start fine tuning on all of them in one go.

133. Which of the following is not evaluated using the GLUE benchmark? D

A. Similarity
B. Paraphrase
C. Question duplicates
D. Machine Translation

134. Which of the following are issues with transformers? AC

A. Attention on sequence of length L takes L^2 time and memory.
B. They help with the vanishing gradient problem.
C. N layers take N times as much memory.
D. They allow for parallel computing.

135. Why do we need to store activations somewhere when implementing the transformer network? B

A. We will need to keep track of all the activations we used so we can make predictions.
B. We need to save them to compute the back-propagation.
C. They are important for interpretability.
D. To allow us to debug our model incase it stops working.

136. Why do we use locality sensitive hashing when computing attention? AB

A. It allows us to not have to compare each query with each key. Instead we only compare the vectors that are found in the same bucket.
B. It is a faster way to compute attention.
C. It is more accurate when finding the most similar vectors than regular attention.
D. It is not worth using.

137. What is the point of using reversible layers? C

A. It allows you to have a symmetry in your model, and thus breaks it in the backprop.
B. It allows your model to capture dependencies that

you would not have been able to capture otherwise.
C. It allows you to reconstruct the the activations and as a result you do not have to save them.
D. It speeds up training.

138. Standard Transformer is defined as: A
y_a = x + Attention(x)
y_b = y_a + FF(y_a)
Reversible:
y_1 = x_1 + Attention(x_2)
y_2 = x_2 + FF(y_1)
To recompute x_1 from y_1 you can use the following:
x_1 = y_1 - Attention(x_2)
How would you recompute x_2?

A. x_2 = y_2 - FF(y_1)
B. x_2 = Attention(x_1) + FF(y_1)
C. x_2 = x_1 - FF(y_2)
D. x_2 = y_2 - Attention(x_1)

139. Select two main components that the reformer uses which makes it more efficient than the transformers. AB

A. Reversible layers
B. Locality sensitive hashing.
C. K-nearest neighbors
D. Skip connections.

140. What are the pros and cons of having more hashes when implementing LSH? B

A. The more hashes you have the less accurate your model is, but the faster it is.

B. The more hashes you have the more accurate your model is, but the slower it is.
C. The more hashes you have the faster you can train your model, and the more accurate it gets.
D. The more hashes you have the slower your model gets and the lower the accuracy becomes.

141. How many words can a reformer hold on a single 16GB GPU? C

A. 500,000
B. 200,000
C. 1 million
D. 50,000

142. In LSH, you want to attend to a bucket in a previous chunk because it covers the case with a hash bucket that is split over more than 1 chunk. B

A. False.
B. True.

143. One reason according to the lecture why the BLEU score for transformers is slightly better than the one where reversible layers are used is due to parameter tuning of the transformer network in the past 3 years. B

A. False
B. True.