Module 1 Introduction to NLP

1. Minimum Edit Distance technique is not used for Spelling Correction tasks in NLP: False
2. Consider a scenario where our training corpus contains, say the words slow, fast, faster, but not slower, then if the word slower appears in our test corpus, our system will not know what to do with it. Which of the following tokenization will help in tackling such problems? Byte Pair Encoding
3. Which of the following tasks are commonly applied as part of any text normalization process? Normalizing word formats, Segmenting sentences, Tokenizing words
4. The general task of segmenting the given text into words is called as. Tokenization
5. Which of the following are used for tokenization of documents? Unigram Language Modeling, WordPiece, Byte-Pair Encoding
6. Each character or syllable generally represents a single unit of meaning and it is called as? Morpheme
7. The relationship between the number of word types |V| and number of tokens N is called Herdan's Law or Heaps' Law ? (Types are the number of distinct words in a corpus; if the set of words in the vocabulary is V, the number of types is the word token vocabulary size |V|. Tokens are the total number N of running words.) True
8. Which of the following algorithms are used to measure similarity between two strings? Minimum Edit Distance Algorithm, Viterbi Algorithm
9. The Porter Algorithm is best used for? Stemming

Module 2 Linear Text Classification

1. Is Sentiment Analysis of a given text or document is part of text categorization? True
2. Which of the following is/are examples of sentiment analysis? To Identify whether the given tweet of the citizen is useful for a politician in coming elections or not, Predicting whether a given review of a product in an e-commerce website is useful or not
3. Which of the following best describes about Generative Classifiers? They build a model of how a class could generate some input data i,e given an observation, they return the class most likely to have generated the observation
4. Which of the following are the examples of Discriminative Classifiers? Support Vector Machines, Logistic Regression
5. To which of the following metrics, The BootStrap Test can be applied? Recall, BLEU, Predcision
6. Which of the following are common non parametric tests used for NLP? BootStrap Test, Approximate Randomization
7. Naive Bayes Algorithm assumes that the probabilities of features(words) for a given class, are conditionally dependent on each other. False

8. Regularization technique is used to avoid which of the following problems? Overfitting
9. Which of the following function is used by Multinomial logistic regression to compute probabilities? Softmax function
10. Logistic regression is a supervised machine learning classifier that extracts real-valued features from the input, multiplies each by a weight, sums them, and passes the sum through a sigmoid function to generate a probability. True

Module 3 Non Linear Text Classification

1. Which of the following algorithms share same mathematics as with Neural Networks especially for classification? Logistic Regression
2. While working with neural networks, it is more common to make use of rich hand derived features instead of Automatic derivation of features? False
3. Which of the following are the right tool for solving large scale problems that offer sufficient data to learn features automatically? Deep Neural Nets
4. Which of the following are Non Linear Functions? ReLu, Sigmoid Function, Tanh
5. Which of the following nonlinear function is the simplest activation function and perhaps the most commonly used? ReLu
6. Which of the following is the problem of the gradients that are almost 0 which cause the error signal to get smaller and smaller until it is too small to be used for training? Vanishing Gradient Problem
7. Perceptron is a very simple neural unit that has a binary output and does have a non-linear activation function. False
8. Neural language models use a neural network as a probabilistic classifier, to compute the probability of the next word given the previous n words. True
9. In a fully-connected, feedforward network, each unit in layer i is connected to each unit in layer i + 1, and there are cycles in the architecture. False
10. Drop Out is one of the important regularization technique and is used to solve which of the following problem in Neural Networks? Overfitting

Module 4 Applications

1. Which of following is the locus of word meaning ;definitions and meaning relations are defined at the level of the word sense rather than word forms. Word Induce, Word Sense
2. WordNet is a large database of lexical relations for English, and WordNets exist for a variety of languages. True
3. Which of the following is the task of determining the correct sense of a word in context? Word sense induction, Word Sense Disambiguation
4. Which of the following is a knowledge-based Word Sense Disambiguation algorithm which chooses the sense whose dictionary definition shares the most words with the target word’s neighborhood? Lesk algorithm

5. Which of the following are the relations between word senses? Hypernymy, Synonym, Antonym
6. A confusion matrix is a table for visualizing how an algorithm performs with respect to the human gold labels, using two dimensions (system output and gold labels), and each cell labeling a set of possible outcomes. True
7. Which of the following metric will not perform well for unbalanced dataset? Accuracy
8. Which of the following measures the percentage of the items that the system detected (i.e., the system labeled as positive) that are in fact positive (i.e., are positive according to the human gold labels) Precision
9. Which of the following measures the percentage of items actually present in the input that were correctly identified by the system. Recall
10. There are many ways to define a single metric that incorporates aspects of both precision and recall. The simplest of these combinations is F-measure

Module 5 Unsupervised and semi supervised learning

1. In case of Agglomerative clustering algorithms, hierarchy of clusters are built from top down. False
2. Following is the table of pairwise Euclidean distances between 5 data points

	a	**b**	**c**	**d**	**e**
a	0				
b	9	0			
c	3	7	0		
d	6	5	9	0	
e	11	10	2	8	0

Assuming we are using single-linkage (min distance) for hierarchical clustering. The order in which the clusters will be formed is: Points b and d are merged first to form cluster {b, d}.Then point a is incorporated into cluster {b, d}. Then points c and e are merged into {c, e}. Finally, clusters {b, d, a} and {c, e} are merged.

3. Word sense induction from unlabelled data can be efficiently learned using clustering techniques. The number of senses/clusters, k, is usually assumed for Expectation-Maximization or k-Means clustering. Which of the following methods can be used to estimate an optimal value for k. Markov Chain Monte Carlo, Held-out log likelihood, The elbow method, Akaike Information Criterion
4. Match steps of Expectation-Maximization framework for discovering k topics in an archive of 100 documents.

   Step 1: Introduce laten variable z to write P(x, z); and decide k
   Step 2: Perform E-step to calculate expected counts; P(z_i | x)

Step 3: Calculate M-step to update parameters phi and mu of likelihood P(x | z; phi) and priors P(z, mu)
Step 4: Repeat until converged

5. Match steps of k-Means clustering algorithm for discovering k topics in an archive of 100 documents.

   Step 1: Vectorize the text content for each document to form a matrix X.
   Step 2: Assume value of K and randomly select K document vectors from X as centroids
   Step 3: Calculate distance from each document to each centroid
   Step 4: Assign each document vector to the nearest centroid.
   Step 5: Recalculate centroid based on assignments
   Step 6: Repeat until converged

Module 6 Meaning of Words and Distributional Semantics

1. The Point-wise Mutual Information is negative when a word and context occur together less often than if they were independent. True
2. Latent semantic analysis is most effective when the count matrix is transformed before decomposing the terms matrix into orthonormal eigen vectors. One such transformation is Sigular Value Decomposition
3. Learning algorithms like perceptron and conditional random fields often perform better with discrete feature vectors. Distributional representations of words can be discretized using clustering techniques such as Brown clustering, Hierarchical clustering
4. The theory that makes it possible to acquire meaningful representations from unlabeled data is referred to as the distributional hypothesis
5. What does the word2vec classifier attempt to predict? Whether or not a given word is likley to show up near a context word.
6. Pointwise mutual information (PMI) captures the degree of association between pairs of words and contexts
7. Which of the following distributional representation methods compute the local context as an average of embedding for words in the immediate neighborhood. Neural embedding (Continuous Bag Words)
8. Which of the following assumptions does skip-gram make? All context words are independent.
9. While creating Neural word embeddings (Eg. Word2Vec), if we are using the Skip Gram model we use the distributed representation of input word to predict the context. True
10. While creating Neural word embeddings (Eg. Word2Vec), if we are using the Continuous Bag Of Words (CBOW) model we use the distributed representation of surrounding(context) words to predict the middle word. True

Module 7 Language Models: N-Gram and RNNs

1. Language models offer a way to assign a probability to a sentence or other sequence of words, and to predict a word from preceding words. True
2. Which of the following is/are different ways of performing smoothing? Add K Smoothing, Kneser-Ney smoothing, Laplace Smoothing, Stupid Backoff
3. Language modeling is not usually an application in itself but a typically used as a component of a larger systems. Evaluating whether the language model improves performance on the application task, such as machine translation or speech recognition, is referred to as Extrinsic Evaluation
4. The perplexity of a language model on a test set is the inverse probability of the test set, normalized by the square of number of words False
5. The perplexity of two language models can also be comparable if they use different vocabularies. False
6. To calculate probability estimates of sequence of tokens, if all sequences are of length M in a corpus of vocabulary V, how many sequences will be calculated and stored by the n-Gram model. $\boldsymbol{V^n}$
7. In simple Recurrent Neural Networks sequences are processed one element at a time, with the output of each neural unit at time t based both on the current input at t and the hidden layer from time t − 1. True
8. Relative frequency of observing a sequence in the data gives estimate of the probability of the sequence. This estimate is unbiased if we have all possible written/spoken sequences in the corpus. This approach to calculate probabilities of sequences is data hungry. To calculate unbiased estimates, if all sequences are of length M in a corpus of vocabulary V, how many sequences will be needed to calculate the relative frequency estimates. $\boldsymbol{V^M}$
9. Which of the following are common language-based applications for RNNs and transformers? Sequence labeling like part-of-speech tagging, Probabilistic language modeling, Auto-regressive generation using a trained language model
10. Gorillas always like to groom their friends. To train an n-Gram model to handle the relationship between their and Gorillas, it is necessary to model 6-grams to calculate p(friends | Gorillas always like to groom their). For a vocabulary of 1000 words, how many events should be accounted for to calculate p(friends | Gorillas always like to groom their) $\boldsymbol{1000^6}$
11. Which of the following techniques reserve some probability mass from the observed data, and redistribute this probability mass equally to the unseen words. Discounting
12. Which of the following technique/s combine multiple n-Gram models to estimate the probability of a word given n previous words. Interpolation, Backoff