This is a document with answers to all questions asked after Quizzes

Quiz 2:

Which of these statements about the flat data model are correct? Check all that apply.	**In the flat data model, data can be represented in plain-text files**
	It is primarily used to represent complex relationships between data points
	To separate individual fields in the flat data model, we have to use delimiters such as comma or tab
	For large data, it gets more difficult to update samples stored in the flat data model than in the relational model

1. For large data, it gets more difficult to update samples stored in the flat data model than in the relational model

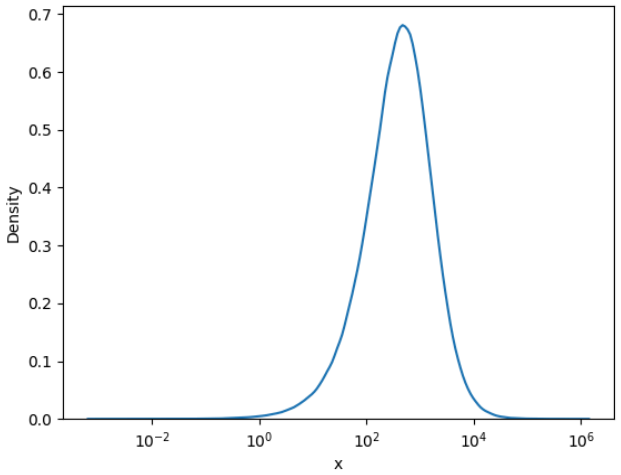
Some of you were unsure why this is a correct option. We will illustrate this with one simple example (and this is not the only one):

Imagine a situation where you have data about customers and orders that they have made. In a relational model, you decide to represent this with two tables, one consisting of information about customers (such as their id, name, surname, telephone number, etc.) and the other about the orders (the id of the customer who made the order, id of the order, type of order, quantity, etc.). In a flat model, you decide to represent each order as one record which will contain both the info about the customer and the order. Now imagine a person changes their telephone number. In a relational model you will only have to update one field in the customers table, while in the flat model, you will have to iterate through all of the records to update this information.

This is a document with answers to all questions asked after Quizzes

Quiz 3:

Given this kernel density estimation plot of the data, check the correct statement. Check all that apply.	**The data could come from a heavy-tailed distribution**
	The data could come from a Gaussian distribution
	The data is two-dimensional
	None of the above



Some of you were wondering why this cannot be a Gaussian curve. If you look closely at the x-axis, it is presented in logarithmic scale which means that this plot looks totally different on a linear scale. If you recall this slide (29) from the lecture, it is actually the heavy-tailed distribution that looks like this when it is presented on a logarithmic scale.

You have survey data from 100 people indicating their most preferred color out of red, blue, green, yellow, and purple. You want to visualize the results as counts/frequencies for each color. Which type of plot(s) is (are) suitable in this situation? Check all that apply.	Histograms
	Line chart
	Scatter plot
	Bar chart

Thank you for the feedback on this question. It is correct that histograms and bar plots may appear visually similar for categorical data. However, an important difference is that histograms are specifically used for **numerical data**, dividing a continuous axis into **bins** and showing frequencies within each bin range. \cite{https://en.wikipedia.org/wiki/Histogram}

Bar charts on the other hand are better suited for categorical data, with each bar corresponding directly to a discrete category. The gaps between bars help emphasize that the categories do not have an inherent ordering or continuity along the axis.

This is a document with answers to all questions asked after Quizzes

Quiz 4:

You have a method for classifying data points into 3 classes, A, B, and C. Your dataset is imbalanced: A appears in 70% of the cases, B in 25% of cases, and C in 5% of cases. You evaluate the performance of the method by calculating the fraction of data points classified correctly, for each class. Which statements are true? Check all that apply.	**If we care about overall performance, micro-average is a better metric than macro-average**
	If we care about performance per each class, macro-average is a better metric than micro-average
	If we care about overall performance, macro-average is a better metric than micro-average
	If we care about performance per each class, micro-average is a better metric than macro-average

Some of you didn't understand why micro-average is more insightful when we care about the overall performance, while macro-average is better if we care about performance of each class. This is correct because micro-average puts the same weight to each sample, regardless of the class it belongs to. The consequence of this is that, for instance, if we fail to correctly place samples from the less frequent classes, but do well for the more common ones, the overall performance would still be high, as most of the samples are correctly classified. On the other hand, macro-average puts the same weight on each class, resulting in lower performance in the same scenario, as one for one of the classes, our method is not working well.

Which of the following statements are true? Assume you are working with non-negative data. Check all that apply.	For skewed distribution, median is smaller than mean
	For heavy-tailed distribution, median can be smaller than mean
	For skewed distribution, median can be smaller, bigger or equal to the mean
	For symmetrical distributions, median is always the same as mean

There were some questions about the two last options. We agree that for categorical data it is possible to have symmetrical distribution for which median is not the same as mean. Similarly, for categorical data you can find examples where median is equal to mean. For continuous data, these are incorrect - for skewed distributions the median can be smaller or bigger only and the median is the same as the mean for symmetrical distributions. As this was a mistake from our side, **the points for this question will be altered** as follows: **(1)** checking any of the options 2, 3, and 4 will give you 100% **(2)** checking the first option will bring you -50%. This change won't be visible in Moodle.

This is a document with answers to all questions asked after Quizzes

You perform a hypothesis test of a null hypothesis H0 using Dataset 1 and obtain a p-value of 0.01, leading you to reject H0 at the 5% significance level. You then collect more data, acquiring a new dataset which we call Dataset 2. You perform the hypothesis test again, obtaining a p-value of 0.3, meaning you fail to reject H0 at the 5% significance level. What is the most valid conclusion based on these results?	The contradictory p-values mean H0 is equally likely to be true or false.
	The lower p-value indicates H0 is likely false, while the higher p-value indicates H0 is likely true.
	The first p-value implies that the alternative hypothesis is true. The second p-value implies that H0 is true.
	The p-values only suggest how likely each dataset was under H0. More evidence is needed to determine if H0 is true or false.

There was a question about the fact that you can never be 100% certain that H0 is true or false. Indeed in frequentist hypothesis testing we cannot prove definitively whether the null hypothesis is true or false. The p-values only provide evidence against the null, not evidence for it.
When we said "determine if H0 is true or false", we meant in a statistical sense - whether H0 is likely or unlikely to be the true state of nature based on the data, not metaphysical truth. You are correct that we can never have absolute 100% certainty about H0 being literally true or false. The language of "true" and "false" is shorthand for statistically likely or unlikely given the evidence.

Quiz 5:

Which of the following statements is correct? Check all that apply.	If we have negative predictors, we should not perform linear regression
	If a predictor has a positive linear regression coefficient, that means that an increase in the predictor is associated with an increase in the outcome
	Low R^2 score indicates that the predictors have no statistically significant correlation with the outcome
	Intercept represents the predicted value of the outcome when all predictor variables are set to zero

There were two types of questions about this quiz question:

(1) *If a predictor has a positive linear regression coefficient, that means that an increase in the predictor is associated with an increase in the outcome*: Some of you asked why this isn't true. This is incorrect because the p-value might be high, indicating that there isn't enough evidence to conclude that this coefficient is different from zero, and hence associated with no increase in the outcome.
(2) *Intercept represents the predicted value of the outcome when all predictor variables are set to zero*: Some of you were confused with this question because you were thinking about cases when intercept has no meaningful real-world value. While this can, indeed, be the case, the intercept is still the predicted value of the outcome when all the predictor variables are equal to zero. This might not be possible in the real world, and the value can be useless, but you can still plug these numbers into your regression equation.

This is a document with answers to all questions asked after Quizzes

Quiz 6:

When performing a randomized experiment (with a treatment group and a control group), which of the following statements is true?	Unobserved confounders may threaten the validity of your conclusions
	All participants have the same probability of being assigned to the treatment group
	Randomized experiments are usually harder to replicate than observational studies
	For every participant, the probability of being assigned to the treatment group is the same as the probability of being assigned to the control group

Some of you asked why the last option here is incorrect. While every participant has the same probability to be assigned to the treatment group, the treatment and control group don't necessarily have to be the same size in each case. As a consequence, these two probabilities might not be the same.

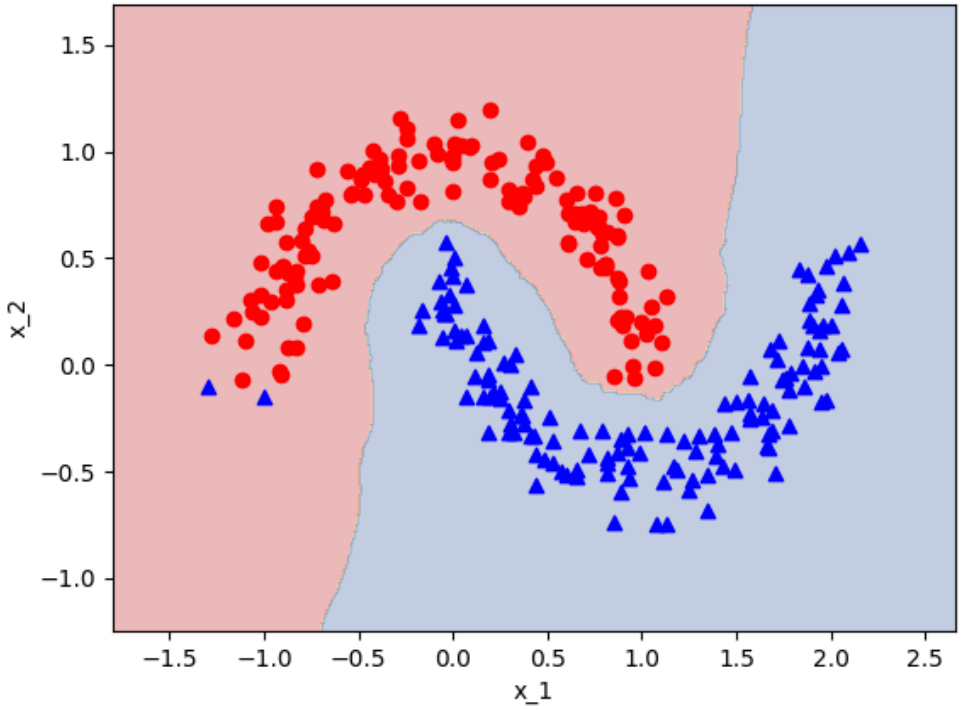
Which of the following statements about determining causality from observational data are true? Check all that apply.	Randomized experiments are usually cheaper to conduct than observational studies, as most modern data is found data
	Sensitivity analysis quantifies the potential impact of unmeasured confounding on study conclusions.
	Studies should match treated and control units on as many relevant covariates as possible to minimize confounding.
	Large sensitivity analysis parameter ($$\Gamma$$) means that two subjects with the same unobserved covariates have vastly different probabilities of getting the treatment.

We have made a mistake in this question. Indeed, you should match on *relevant* covariates. What we intended to show was that you shouldn't necessarily match on *every* covariate as this can in fact be harmful for your analysis. **The grading will be altered** as follows: **(1)** checking any of the options 2 or 3 will give you 100% of points **(2)** checking the first and the last option will give you -50% of the points. The change in the scores will be done on the final grading sheet and won't be visible through Moodle.

This is a document with answers to all questions asked after Quizzes

Quiz 7:

The plot shows data points from two classes (red circles and blue triangles) as well as the decisions made by some classifier: points on red background are classified as red, and points on blue background as blue. Which classifier could have been used to make these decisions (using x_1 and x_2 as features)? Check all that apply.	Logistic regression
	K nearest neighbours with K=10
	K nearest neighbours with K=2
	None of the above



There were some questions regarding this quiz question:

(1) *How can we know that this was generated using K nearest neighbours with K=10?* You cannot know for sure but the question doesn't ask you to be 100% sure. You are asked if it could have been the case, and there is nothing on the graph that would make you rule out this statement.

(2) *Why can't this be generated using logistic regression?* Logistic regression is a linear classifier. This means that you would expect the decision boundary (i.e. the line that separates the blue and red region) to be a line. Here you can clearly see that it is non-linear, and hence cannot come from a logistic regression classifier.

This is a document with answers to all questions asked after Quizzes

Which of the following statements about the bias-variance tradeoff are true? Select all that apply.	Increasing k in k-NN decreases bias but increases variance.
	Adding depth to a decision tree decreases bias but increases variance.
	Adding more relevant features to a logistic regression model decreases bias but increases variance.
	Increasing the number of trees in boosted decision trees decreases the bias.

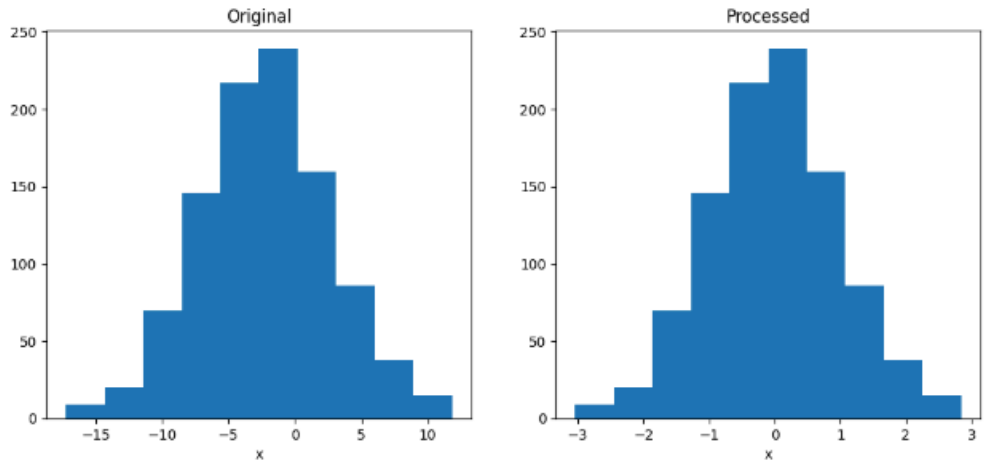
Two issues were raised with the statement "Adding more features to a logistic regression model decreases bias but increases variance":

- Adding constant features does not affect bias or variance.
- Adding irrelevant features may not increase variance.

The question aimed to test the conceptual understanding that, in general, adding more features increases a model's capacity and flexibility, allowing it to fit more complex functions and reducing bias. But this flexibility can lead to overfitting the training data, increasing variance. We made a mistake of not adding the word relevant to eliminate the corner cases. **The grading will be altered, everyone will get the point for this option**. The change in the scores will be done on the final grading sheet and won't be visible through Moodle.

Quiz 8:

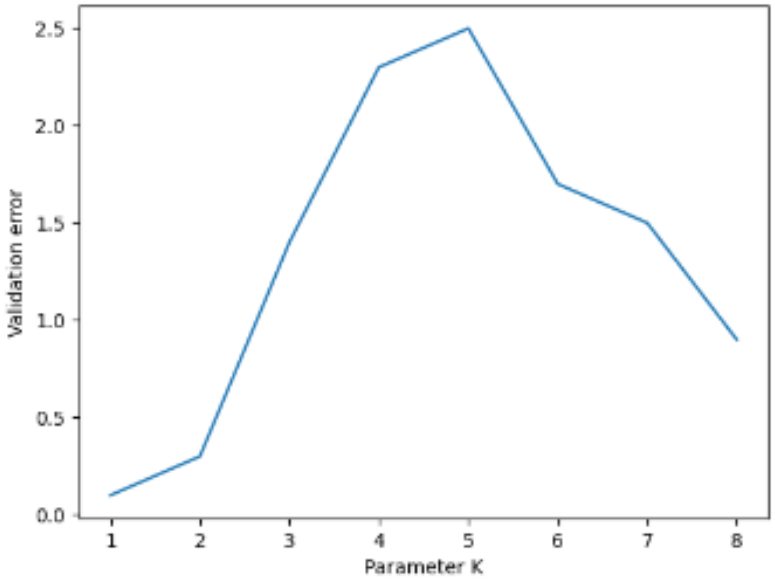
You have one-dimensional data whose distribution is shown in the left plot. After feature pre-processing, you obtain the data with the distribution shown in the right plot. Which feature transformation could have been performed?	**Standardization**
	Min-max scaling
	Logarithmic scaling
	None of the above



There were some questions asking why standardization is the correct option here. The question is asking which of these *could have been* performed. While for min-max and logarithmic scaling there are clear ways to conclude that they were not performed (the transformed data lies outside of (0,1) interval and hasn't changed the shape), there is nothing on the plot that eliminates the standardization out of the picture, as the result indeed looks like normal distribution.

This is a document with answers to all questions asked after Quizzes

You perform cross-validation to determine the best choice for a parameter K. You obtain the plot below. Which value of the parameter K should you choose?	**K=1**
	K=5
	K=4
	K=8



Some of you had questions why K=1 should be chosen as the value for the parameter K. Keep in mind that K isn't the number of folds, but a random parameter you are trying to choose using cross-validation. Otherwise the question would make little sense, as there would be no point in searching for the optimal number of folds for cross-validation using cross-validation. With that said, K=1 yields the lowest validation error and as such should be chosen.

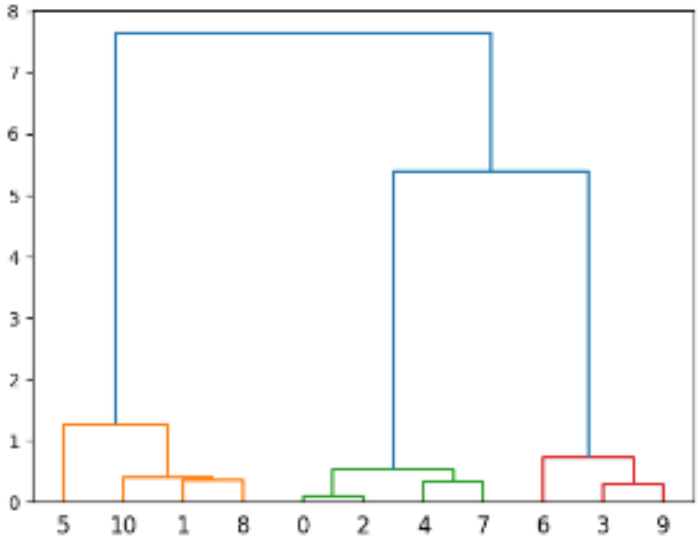
Which of these statements about the precision/recall curve are correct? Check all that apply.	The precision/recall curve is sensitive to the classification threshold
	The precision/recall curve for a random classifier corresponds to the identity line (y=x)
	The area under the precision/recall curve for a perfect classifier is 1
	The area under the precision/recall curve for a random classifier depends on number of positive and negative samples in the dataset

There were questions asking why the option *"The precision/recall curve is sensitive to the classification threshold"* is incorrect. Precision/recall curve by definition is generated by moving the threshold from the smallest to the biggest value. By changing the threshold you move along the precision/recall curve, but the curve doesn't change the shape. Because of that, it isn't sensitive to the classification threshold.

This is a document with answers to all questions asked after Quizzes

Quiz 9:

Which conclusions can be drawn from the given dendrogram? Check all that apply.	**The dataset from the analysis has 11 data points**
	The analysis could have been done by performing agglomerative clustering
	First two points that were grouped were point 0 and point 2
	Euclidean distance has certainly been used to determine distance between centroids



For this question, there were two options you had questions about:

(1) *The dataset from the analysis has 11 data points* - Some of you were mentioning that we cannot know when the clustering stops if we take the top down approach and hence cannot know the number of data points. Dendrogram is made by performing the whole procedure, regardless of the approach (top-down or bottom-up), so this is a correct option.

(2) *First two points that were grouped were point 0 and point 2* - There were some remarks that the question doesn't mention whether we performed agglomerative or divisive clustering, and because of that this option is not correct. It is true that this dendrogram could have been generated both ways. However, the question asks which **can** be drawn from the dendrogram. One valid option that fits this is that the dendrogram was generated in agglomerative fashion by first grouping the point 0 and point 2. This is not in contradiction with the divisive clustering (it is just one option), and it isn't incorrect either - these two points will always be grouped first.

Which of the following are advantages of the DBSCAN clustering algorithm over k-means clustering? Check all that apply.	**DBSCAN can detect clusters of arbitrary shapes, unlike k-means which assumes convex clusters.**
	DBSCAN does not require specifying the number of clusters k in advance.
	DBSCAN has a faster runtime than k-means on large datasets.
	DBSCAN clustering results do not depend on the initialization, unlike k-means.

This is a document with answers to all questions asked after Quizzes

Some of you had feedback on option 4. Here are two points that we wanted to clarify:

(1) Epsilon is a hyperparameter, **not** a parameter initialized randomly at the beginning of the algorithm – some other examples of such hyperparameters are K in k-means, the regularization parameter in regression, or the maximum depth of the tree in a decision tree.

(2) In the description of the DBSCAN algorithm on slide 48, there is no random sampling of points, or random initialization of clusters. Some of you argued that a point will be assigned to the cluster that was first initialized if it is density-reachable by two potential clusters. We want to emphasize that in k-means, if you run the algorithm 2 times you can get two different results according to your random initialization of cluster centers. However, in DBSCAN results are the same as there is nothing to initialize at random.

The scores for this problem remain unchanged.

Which statements about k-means are true? Check all that apply.	**The k-means problem is NP-hard but initializations can help avoid bad solutions found by the iterative algorithm.**
	K-means++ improves on random sampling by spreading out initial centers, choosing them more widespread than expected to be found by indpendent random sampling.
	Lloyd's algorithm is a greedy algorithm for solving k-means that is robust to centroid initialization.
	K-means clusters are convex shapes, which means it is robust to noisy data and outliers.

There was an argument that the k-means problem itself is not known to be NP-hard; however, finding the optimal solution is NP-hard. The distinction between the k-means problem and finding the optimal solution involves a minor colloquialism and we apologize if it made any confusion. However, we believe that in the context of this question there in little place for this confusion, as np-hardness is a property of a **problem, not a specific algorithm** (say, k-means with a specific initialization)

The scores for this problem remain unchanged.

This is a document with answers to all questions asked after Quizzes

Quiz 10:

Given the bag-of-words matrix below, which was obtained without any preprocessing of the text, what conclusions can you draw? Check all that apply.	**The corpus has 5 documents**
	All documents consist of at most 4 unique words
	All documents consist of at most 4 words
	The corpus has 4 documents

Bag-of-words matrix

2	3	0	0
1	0	2	0
2	0	0	0
0	1	2	1
3	1	1	2

There were a few remarks regarding the option “All documents consist of at most 4 unique words”. First one was talking about ngrams and the possible different length of the vocabulary in that case. However, bag-of-words works with unique words, not ngrams. There was no mention of such transformation. Second, the question was about words that are capitalized. These are, in the computational sense, unique words, so the statement still holds. Finally, some of you mentioned that vocabulary does not necessarily need to be made from all the words in the documents. While this can in practice be done, there was no mention of this in the question (or in the slides), so you should have assumed that vocabulary was generated from the documents and that they have at most 4 unique words.

Which of the following statements are true for character encodings: Check all that apply.	**Latin-1 always encodes each character with 1 byte**
	Standard ASCII encoding can be used when you are working with text containing only English alphabet
	UTF-8 encoding can encode only 256 different characters
	It is safe to read UTF-8 encoded data with ASCII

Why can't we read UTF-8 encoded data with ASCII? UTF-8 sometimes uses more bytes to encode one character and generally can encode more different characters than ASCII. Reading a text encoded with UTF-8 using ASCII might result in wrong decoding. Overall, it is always a good practice to encode and decode the text using the same method.

Which of the following transformations can be used to penalize the words that appear often in the corpus? Check all that apply.	**IDF**
	Stopword removal
	Row normalization
	Lemmatization

There were some remarks about stopword removal, and some of you didn’t think that this option should be marked as stopword removal can be chosen with different heuristics. This is true, the stopword removal can be done with different methods depending on the task. This is mentioned on slide 24 from the lectures. On the same slide, you can also see that taking frequent words is often a good heuristic, as it is often a good option. The question here asks what “**can be used**” for this

purpose, so marking stopword removal as one is indeed true, as one can use frequent words as heuristic. Another remark was talking about penalizing vs banning - removing a word is one way of penalizing its impact (in this case fully).

Which of the following statements about text tokenization are true? Check all that apply.	**Lemmatization maps words to normalized lexicon entries.**
	Stemming reduces sparsity but loses information such as the part of speech (i.e., whether a word is a verb, noun, adjective, etc.).
	Tokenization is more challenging for languages without explicit word delimiters like whitespace, such as Chinese, or those with compound words, like German.
	Lemmatization is most useful for languages with little morphology (i.e., where words aren't inflected in many different ways).

Some of you doubted the correctness of option one because it referred to "words" rather than "tokens" compared to the lecture slides. The objection was that lemmatization applies to tokens, not words. Some also argued that this option doesn’t encompass tokens such as punctuation marks, numbers, symbols, or tokens from a BPE tokenization.
However, this reflects a basic misunderstanding of what the linguistic process of lemmatization refers to.
Lemmatization is a method specifically for normalizing different inflected forms of **words** to a standard lexicon entry. Applying it to **non-word tokens** like punctuation does not make sense linguistically. Conflating lemmatization of word variants with processing of non-word tokens reflects the very misconception the question was trying to uncover.

This is a document with answers to all questions asked after Quizzes

Quiz 11:

Which of the following matrices U and S cannot be obtained by the singular value decomposition of an unknown 2x2 matrix? (Here matrices are represented as lists of rows.) Check all that apply.	**S = [[1, 1], [1, 1]]**
	U = [[1, -2], [2, 1]]
	U = [[1, -1/2], [0, sqrt(3)/2]]
	S = [[1, 0], [0, 2]]

Some of you had questions regarding the validity of the last option. As mentioned in slide 56, the values on the diagonal of matrix S are ordered by their decreasing value, following the convention. In the last option, these are not ordered by the convention, so this is not a correct option.

A document classification model achieves 99% training accuracy but only 60% test accuracy. Which methods could help address this overfitting? Check all that apply.	**Applying L2 regularization**
	Substituting the TF-IDF matrix with less sparse matrix representations obtained from dimensionality reduction techniques like LSA or LDA.
	Training on fewer documents
	Adding more words to the vocabulary

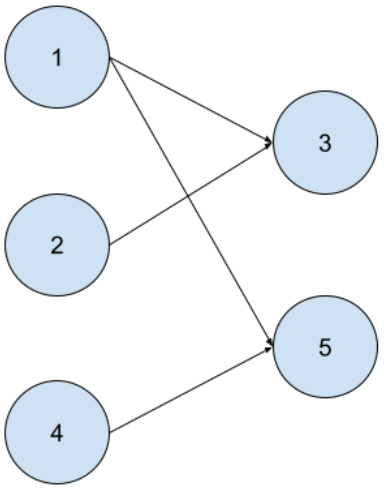
Option 3 is not correct as training on a smaller dataset would result in a simpler dataset for the model to memorize; hence, even more overfitting.

This is a document with answers to all questions asked after Quizzes

Quiz 12:

What can you say about the graph that is represented by the following set of edges: (1, 3) (2, 3) (1, 5) (4, 5) Check all that apply.	**This is a bipartite graph**
	This is a weighted graph
	The graph does not have self-loops
	This graph does not have multi-edges (multiple edges connecting the same nodes)

There was some confusion on how this graph looks as someone has claimed this is a half-empty bipartite graph, not seen during lectures. The graph represented with the set of edges looks like this:



As you can see, this is a bipartite graph.

Given the unweighted graph in the image, which statements are true? Check all that apply.	**Clustering coefficient of node 6 is higher than clustering coefficient of node 7.**
	Clustering coefficient of node 2 is 0
	Clustering coefficient of node 8 is 0.7
	Clustering coefficient of node 9 is 1

Some of you complained that 0.7 in option 3 is misleading as you rounded up the correct clustering coefficient, ⅔. First we wanted to say that the proximity of the wrong option to the correct one was not intentional. We didn't anticipate that this confusion would arise. Nevertheless, the phrase "Clustering coefficient of node 8 is 0.7" is wrong, and rounding up the correct answer, 2/3, when it is not explicitly mentioned in the problem, is not justifiable.

This is a document with answers to all questions asked after Quizzes

Which of the following statements about real-world networks and Erdös-Rényi random graphs are correct? Check all that apply.	**In an Erdös-Rényi random graph with n nodes, where edges exist with probability p independently from one another, the average centrality measure of a node is linearly proportional to p.**
	Short paths are easily discoverable in both real-world networks and Erdös-Rényi random graphs with decentralized algorithms.
	Short paths are greedily discoverable in real-world networks.
	Short paths typically exist in real-world networks, but not in Erdös-Rényi random graphs.

In Erdös-Rényi random graphs, the **average Clustering coefficient** and the **Degree centrality** are linearly proportional to P.
The centrality measure in option 1 is indeed vague and the score for this option would be given to everyone. Thank you for bringing this to our attention.