

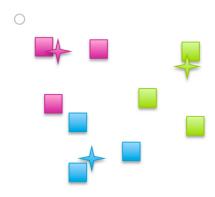
Clustering and Similarity

TOTAL POINTS 6

1. A country, called <i>Simpleland</i> , has a language with a small vocabulary of just "the", "on", "and", "go", "round", "paire "bus", and "wheels", for a word count vector with indices ordered as the words appear above, what is the word count vector for a document that simply says "the wheels on the bus go round and round." Please enter the vector of counts as follows: if the counts were ["the"-1", "on"-3", "and"-2", "go"-1", "round"-2", "bus"-1", "wheels"-1"], enter 1321211. 211211 2. In Simpleland a reader is enjoying a document with a representation: [1 3 2 1 2 1 1]. Which of the following articles would you recommend to this reader next? [[1 2 1 2 1 2 1 1]		
2. In Simpleland, a reader is enjoying a document with a representation: [1 3 2 1 2 1 1]. Which of the following articles would you recommend to this reader next? [7021001] [1700201] [1000712] [0200711] 3. A corpus in Simpleland has 99 articles. If you pick one article and perform 1-nearest neighbor search to find the closest article to this query article, how many times must you compute the similarity between two articles? 9. 38 98*2 = 195 98/2 = 49 (189)*2 99 4. For the TF-IDF representation, does the relative importance of words in a document depend on the base of 1 paint the logarithm used? For example, take the words "bus" and "wheels" in a particular document. Is the ratio between the TF-IDF values for "bus" and "wheels" different when computed using log base 2 versus log base 10? Ves No 5. Which of the following statements are true? (Check all that apply): 1 paint bedsels is a supervised learning problem. Dividing emails into two groups based on the text of each email is a supervised learning problem. Which of the following pictures represents the best k-means solution? (Squares represent observations, 1 point).	1.	
2. In Simpleland, a reader is enjoying a document with a representation: [1 3 2 1 2 1 1]. Which of the following articles would you recommend to this reader next? [7021001] [100201] [10020712] [0200711] 3. A corpus in Simpleland has 99 articles. If you pick one article and perform 1-nearest neighbor search to find the closest article to this query article, how many times must you compute the similarity between two articles? 9 8		
following articles would you recommend to this reader next? [7021001] [1700201] [1000712] [0200711] 3. A corpus in <i>Simpleland</i> has 99 articles. If you pick one article and perform 1-nearest neighbor search to find the closest article to this query article, how many times must you compute the similarity between two articles? 9 88 98*2 = 196 98*2 = 49 (98)*2 99 4. For the TF-IDF representation, does the relative importance of words in a document depend on the base off point the logarithm used? For example, take the words "bus" and "wheels" in a particular document. Is the ratio between the TF-IDF values for "bus" and "wheels" different when computed using log base 2 versus log base 10? Yes No No No No Which of the following statements are true? (Check all that apply): 1 point 2 Which of the following statements are true? (and a many statements are proposed learning problem. 1 point 2 If we are performing clustering, we typically assume we either do not have or do not use class labels in training the model.		2111211
 (1700201) [1000712] [10200711] 3. A corpus in Simpleland has 99 articles. If you pick one article and perform 1-nearest neighbor search to 1 point find the closest article to this query article, how many times must you compute the similarity between two articles? 98 98*2 = 196 98/2 = 49 (98)*2 99 4. For the TF-IDF representation, does the relative importance of words in a document depend on the base of 1 point the logarithm used? For example, take the words "bus" and "wheels" in a particular document. Is the ratio between the TF-IDF values for "bus" and "wheels" different when computed using log base 2 versus log base 10? Yes No 5. Which of the following statements are true? (Check all that apply): 1 point labels is a supervised learning problem. Dividing emails into two groups based on the text of each email is a supervised learning problem. If we are performing clustering, we typically assume we either do not have or do not use class labels in training the model. 6. Which of the following pictures represents the best k-means solution? (Squares represent observations, 1 point) 	2.	
[1000712] [1000712] [1000711] 3. A corpus in Simpleland has 99 articles. If you pick one article and perform 1-nearest neighbor search to find the closest article to this query article, how many times must you compute the similarity between two articles? 9. 98 98 982 = 196 98/2 = 49 (198)-2 99 4. For the TF-IDF representation, does the relative importance of words in a document depend on the base of 1 point the logarithm used? For example, take the words "bus" and "wheels" in a particular document. Is the ratio between the TF-IDF values for "bus" and "wheels" different when computed using log base 2 versus log base 10? Yes No 5. Which of the following statements are true? (Check all that apply): 1 point 1 po		O [7021001]
. A corpus in <i>Simpleland</i> has 99 articles. If you pick one article and perform 1-nearest neighbor search to 1 point find the closest article to this query article, how many times must you compute the similarity between two articles? 9 98 982 = 196 98/2 = 49 (98)*2 99 4. For the TF-IDF representation, does the relative importance of words in a document depend on the base of 1 point the logarithm used? For example, take the words "bus" and "wheels" in a particular document. Is the ratio between the TF-IDF values for "bus" and "wheels" different when computed using log base 2 versus log base 10? Yes No 1 point Deciding whether an email is <i>spam</i> or <i>not spam</i> using the text of the email and some <i>spam I not spam</i> labels is a supervised learning problem. If we are performing clustering, we typically assume we either do not have or do not use class labels in training the model.		[1700201]
3. A corpus in <i>Simpleland</i> has 99 articles. If you pick one article and perform 1-nearest neighbor search to 1 point find the closest article to this query article, how many times must you compute the similarity between two articles? 9 98 98*2 = 196 98/2 = 49 (98)*2 99 4. For the TF-IDF representation, does the relative importance of words in a document depend on the base of 1 point the logarithm used? For example, take the words "bus" and "wheels" in a particular document. Is the ratio between the TF-IDF values for "bus" and "wheels" different when computed using log base 2 versus log base 10? Yes No Which of the following statements are true? (<i>Check all that apply</i>): 1 point Dividing emails into two groups based on the text of each email is a supervised learning problem. If we are performing clustering, we typically assume we either do not have or do not use class labels in training the model.		○ [1000712]
find the closest article to this query article, how many times must you compute the similarity between two articles? 9 98 98*2 = 196 98/2 = 49 (98)*2 99 4. For the TF-IDF representation, does the relative importance of words in a document depend on the base of 1 point the logarithm used? For example, take the words "bus" and "wheels" in a particular document. Is the ratio between the TF-IDF values for "bus" and "wheels" different when computed using log base 2 versus log base 10? Yes No No Deciding whether an email is spam or not spam using the text of the email and some spam / not spam labels is a supervised learning problem. Dividing emails into two groups based on the text of each email is a supervised learning problem. If we are performing clustering, we typically assume we either do not have or do not use class labels in training the model.		O [0200711]
 98*2 = 196 98/2 = 49 (98)*2 99 4. For the TF-IDF representation, does the relative importance of words in a document depend on the base of 1 point the logarithm used? For example, take the words "bus" and "wheels" in a particular document. Is the ratio between the TF-IDF values for "bus" and "wheels" different when computed using log base 2 versus log base 10? Yes No S. Which of the following statements are true? (Check all that apply): 1 point Deciding whether an email is spam or not spam using the text of the email and some spam / not spam labels is a supervised learning problem. Dividing emails into two groups based on the text of each email is a supervised learning problem. If we are performing clustering, we typically assume we either do not have or do not use class labels in training the model. 6. Which of the following pictures represents the best k-means solution? (Squares represent observations, 1 point) 	3.	
 98/2 = 49 (98)^2 99 4. For the TF-IDF representation, does the relative importance of words in a document depend on the base of 1 point the logarithm used? For example, take the words "bus" and "wheels" in a particular document. Is the ratio between the TF-IDF values for "bus" and "wheels" different when computed using log base 2 versus log base 10? Yes No No 5. Which of the following statements are true? (Check all that apply): 1 point Deciding whether an email is spam or not spam using the text of the email and some spam I not spam labels is a supervised learning problem. Dividing emails into two groups based on the text of each email is a supervised learning problem. If we are performing clustering, we typically assume we either do not have or do not use class labels in training the model. 6. Which of the following pictures represents the best k-means solution? (Squares represent observations, 1 point) 		98
 (98)*2 99 4. For the TF-IDF representation, does the relative importance of words in a document depend on the base of 1 point the logarithm used? For example, take the words "bus" and "wheels" in a particular document. Is the ratio between the TF-IDF values for "bus" and "wheels" different when computed using log base 2 versus log base 10? Yes No 5. Which of the following statements are true? (Check all that apply): 1 point Deciding whether an email is spam or not spam using the text of the email and some spam / not spam labels is a supervised learning problem. Dividing emails into two groups based on the text of each email is a supervised learning problem. If we are performing clustering, we typically assume we either do not have or do not use class labels in training the model. 6. Which of the following pictures represents the best k-means solution? (Squares represent observations, 1 point) 		98*2 = 196
 99 4. For the TF-IDF representation, does the relative importance of words in a document depend on the base of 1 point the logarithm used? For example, take the words "bus" and "wheels" in a particular document. Is the ratio between the TF-IDF values for "bus" and "wheels" different when computed using log base 2 versus log base 10? Yes No No No No Deciding whether an email is spam or not spam using the text of the email and some spam / not spam labels is a supervised learning problem. Dividing emails into two groups based on the text of each email is a supervised learning problem. If we are performing clustering, we typically assume we either do not have or do not use class labels in training the model. 6. Which of the following pictures represents the best k-means solution? (Squares represent observations, 1 point) 		98/2 = 49
 4. For the TF-IDF representation, does the relative importance of words in a document depend on the base of 1 point the logarithm used? For example, take the words "bus" and "wheels" in a particular document. Is the ratio between the TF-IDF values for "bus" and "wheels" different when computed using log base 2 versus log base 10? Yes No 5. Which of the following statements are true? (Check all that apply): 1 point Deciding whether an email is spam or not spam using the text of the email and some spam / not spam labels is a supervised learning problem. Dividing emails into two groups based on the text of each email is a supervised learning problem. If we are performing clustering, we typically assume we either do not have or do not use class labels in training the model. 6. Which of the following pictures represents the best k-means solution? (Squares represent observations, 1 point) 		O (98)^2
the logarithm used? For example, take the words "bus" and "wheels" in a particular document. Is the ratio between the TF-IDF values for "bus" and "wheels" different when computed using log base 2 versus log base 10? Yes No No No Deciding whether an email is spam or not spam using the text of the email and some spam / not spam labels is a supervised learning problem. Dividing emails into two groups based on the text of each email is a supervised learning problem. If we are performing clustering, we typically assume we either do not have or do not use class labels in training the model.		99
the logarithm used? For example, take the words "bus" and "wheels" in a particular document. Is the ratio between the TF-IDF values for "bus" and "wheels" different when computed using log base 2 versus log base 10? Yes No No No Deciding whether an email is spam or not spam using the text of the email and some spam / not spam labels is a supervised learning problem. Dividing emails into two groups based on the text of each email is a supervised learning problem. If we are performing clustering, we typically assume we either do not have or do not use class labels in training the model.		
 No Which of the following statements are true? (Check all that apply): Deciding whether an email is spam or not spam using the text of the email and some spam / not spam labels is a supervised learning problem. Dividing emails into two groups based on the text of each email is a supervised learning problem. If we are performing clustering, we typically assume we either do not have or do not use class labels in training the model. 6. Which of the following pictures represents the best k-means solution? (Squares represent observations, 1 point) 	4.	the logarithm used? For example, take the words "bus" and "wheels" in a particular document. Is the ratio between the TF-IDF values for "bus" and "wheels" different when computed using log base 2 versus log
5. Which of the following statements are true? (Check all that apply): □ Deciding whether an email is spam or not spam using the text of the email and some spam / not spam labels is a supervised learning problem. □ Dividing emails into two groups based on the text of each email is a supervised learning problem. □ If we are performing clustering, we typically assume we either do not have or do not use class labels in training the model. 6. Which of the following pictures represents the best k-means solution? (Squares represent observations, 1 point)		○ Yes
 ✓ Deciding whether an email is <i>spam</i> or <i>not spam</i> using the text of the email and some <i>spam I not spam</i> labels is a supervised learning problem. ☐ Dividing emails into two groups based on the text of each email is a supervised learning problem. ✓ If we are performing clustering, we typically assume we either do not have or do not use class labels in training the model. 6. Which of the following pictures represents the <i>best</i> k-means solution? (<i>Squares represent observations</i>, 1 point 		● No
labels is a supervised learning problem. □ Dividing emails into two groups based on the text of each email is a supervised learning problem. ☑ If we are performing clustering, we typically assume we either do not have or do not use class labels in training the model. 6. Which of the following pictures represents the best k-means solution? (Squares represent observations, 1 point	5.	Which of the following statements are true? (<i>Check all that apply</i>):
 ✓ If we are performing clustering, we typically assume we either do not have or do not use class labels in training the model. 6. Which of the following pictures represents the best k-means solution? (Squares represent observations, 1 point) 		
training the model. 6. Which of the following pictures represents the <i>best</i> k-means solution? (<i>Squares represent observations</i> , 1 point		Dividing emails into two groups based on the text of each email is a supervised learning problem.
	6.	







I, S. M. Sabiul Hajjaj, understand that submitting work that isn't my own may result in permanent failure of this course or deactivation of my Coursera account.

3 P P

Learn more about Coursera's Honor Code

Save

Submit