✖

Clustering and Similarity

✓ 6/6 questions correct

Quiz passed!

Continue Course (/learn/ml-foundations/supplement/wNurw/download-the-ipython-notebook-used-in-this-lesson-to-follow-along)

Back to Week 4 (/learn/ml-foundations/home/week/4)

✓ 1.

A country, called *Simpleland*, has a language with a small vocabulary of just *“the”*, *“on”*, *“and”*, *“go”*, *“round”*, *“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.

✓ 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?

✓ 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?

✓ 4.

For the TF-IDF representation, does the relative importance of words in a document depend on the base of 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?

✓ 5.

Which of the following statements are **true?** (*Check all that apply*):

✓ 6.

Which of the following pictures represents the ***best*** k-means solution? (*Squares represent observations, plus signs are cluster centers, and colors indicate assignments of observations to cluster centers*.)
