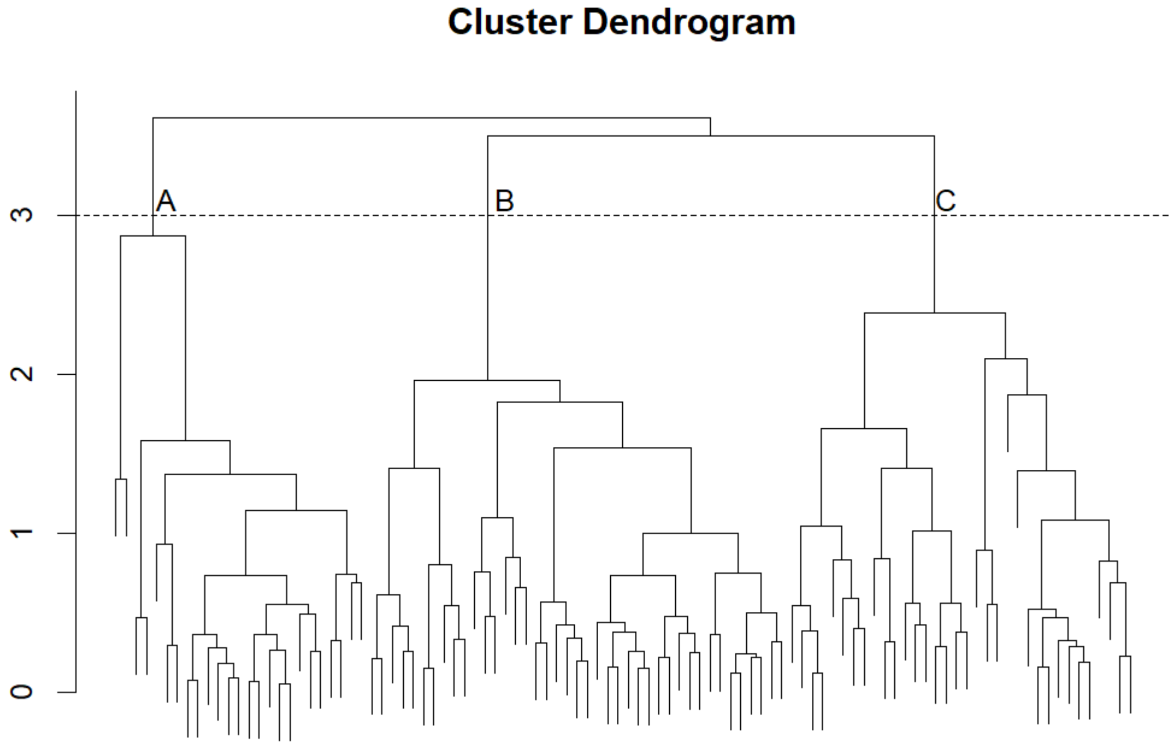
Explain in your own words how much of the variance is explained by a PCA retaining p components in a dataset containing p variables?

Rubric answer:

- 100%, because the number of components is equal to the number of variables.

You have performed hierarchical clustering on a dataset with Euclidean distance and the average linkage method. You obtain the following dendrogram, which you decide to cut at height = 3 (horizontal dashed line).



You label the clustered observations belonging to the leftmost branch A, the middle branch B and the rightmost branch C. Which clusters are, on average, most similar, as measured by the Euclidian distance?

Rubric answer:

- Clusters B and C are most similar. In practice, we could compute the average distances, but the figure shows that the hierarchical clustering method as a next step merges the clusters B and C, before merging with A as the final step.

You have a corpus of 20 000 tweets about a governmental election in the Netherlands, with a total number of 50 000 unique words. Your goal is to classify sentiment orientation (positive, negative) of each tweet. The first step is to create a vector representation of your text data using one of the two methods:

- Bag-of-Words (e.g., tf-idf)
- Word Embedding (e.g., word2vec)

Which method would be better in terms of memory allocation and capturing words relations? Explain your reasoning.

Rubric answer:

- Correctly stated word embedding (e.g., word2vec); mentioning at least two of the following terms: memory allocation, dense representation, distributional hypothesis, capturing words relations and neighborhood.

Text preprocessing is an important step for text mining. In your own words explain the purpose of text preprocessing.

Rubric answer:

One of these

- Text preprocessing transforms text into a more digestible form so that machine learning algorithms can perform better.
- To preprocess your text simply means to bring your text into a form that is analyzable for your task.
- Text preprocessing is an approach for cleaning and noise removal of text data.