1. **Get 40 more CNN news articles online (try to find them in different categories, e.g., sports and finance). Record the category of each article.**

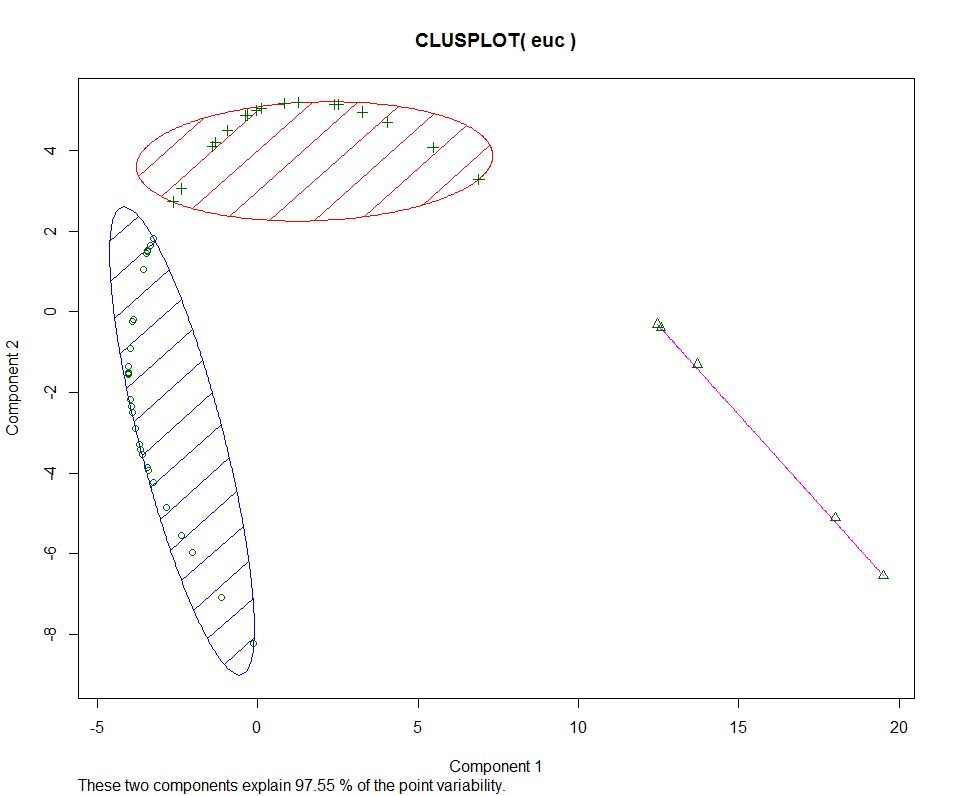
1. **Convert all 50 CNN news articles to data matrix (each row is an article and each column is a unique term)**

**.**

1. **Run K­means clustering with Euclidean, Cosine and Jarcard similarity. (Specify K a s the number of categories of your 50 CNN news articles)**

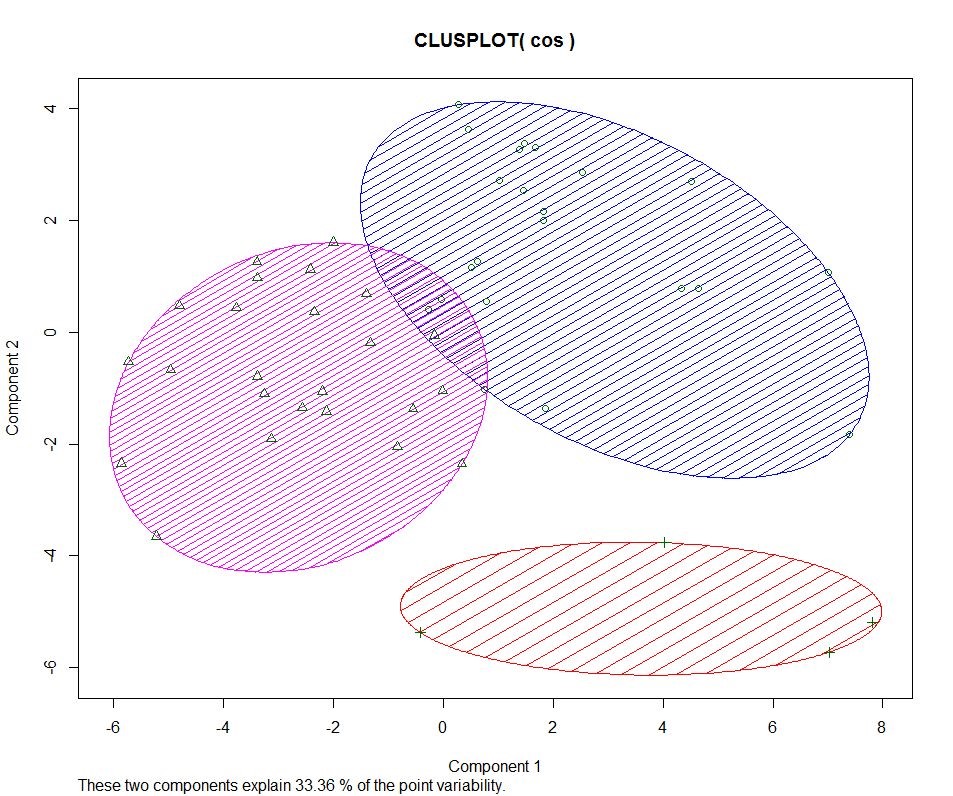
EUCLIDEAN:

­K­means clustering with 3 clusters of sizes 28, 5, 17



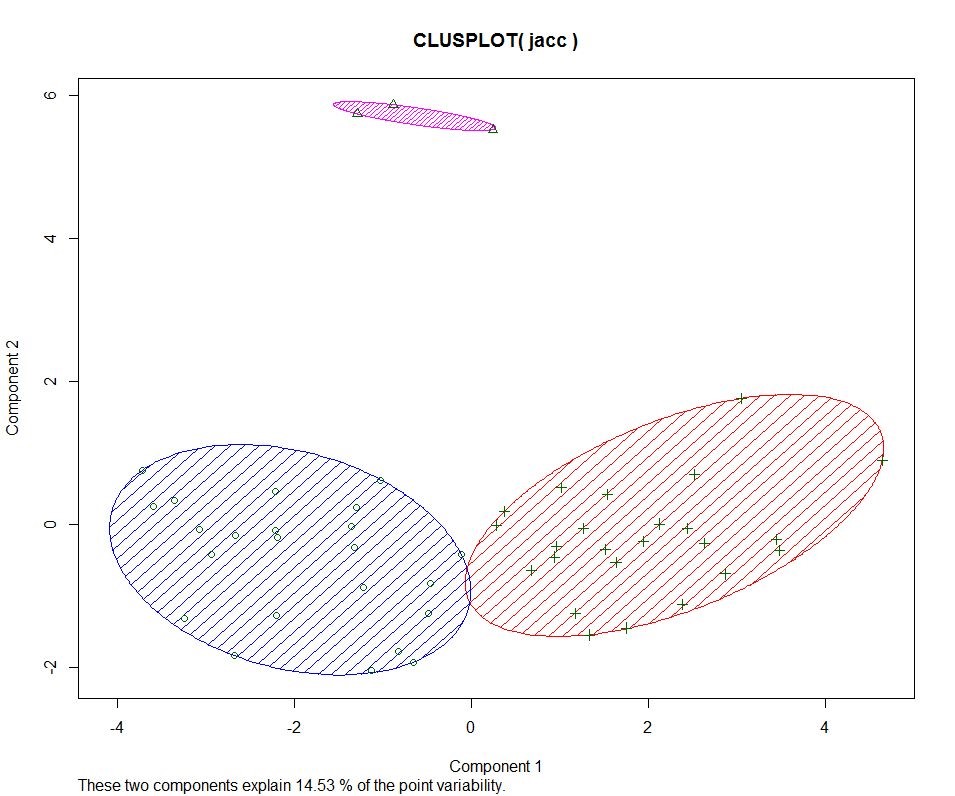
COSINE:

­​K­means clustering with 3 clusters of sizes 22, 24, 4



JACCARD:

­​K­means clustering with 3 clusters of sizes 3, 23, 24



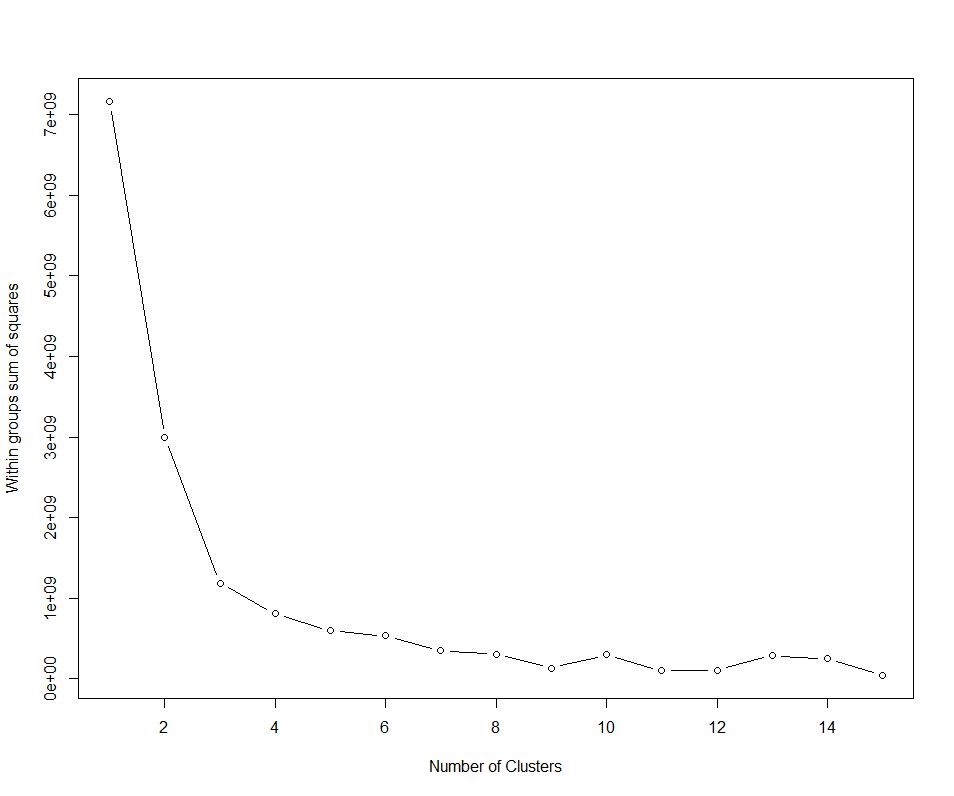
1. **Evaluate K­means clustering results with SSE**

EUCLIDEAN:

­​Within cluster sum of squares by cluster:

[1] 207983335 398248464 580462712

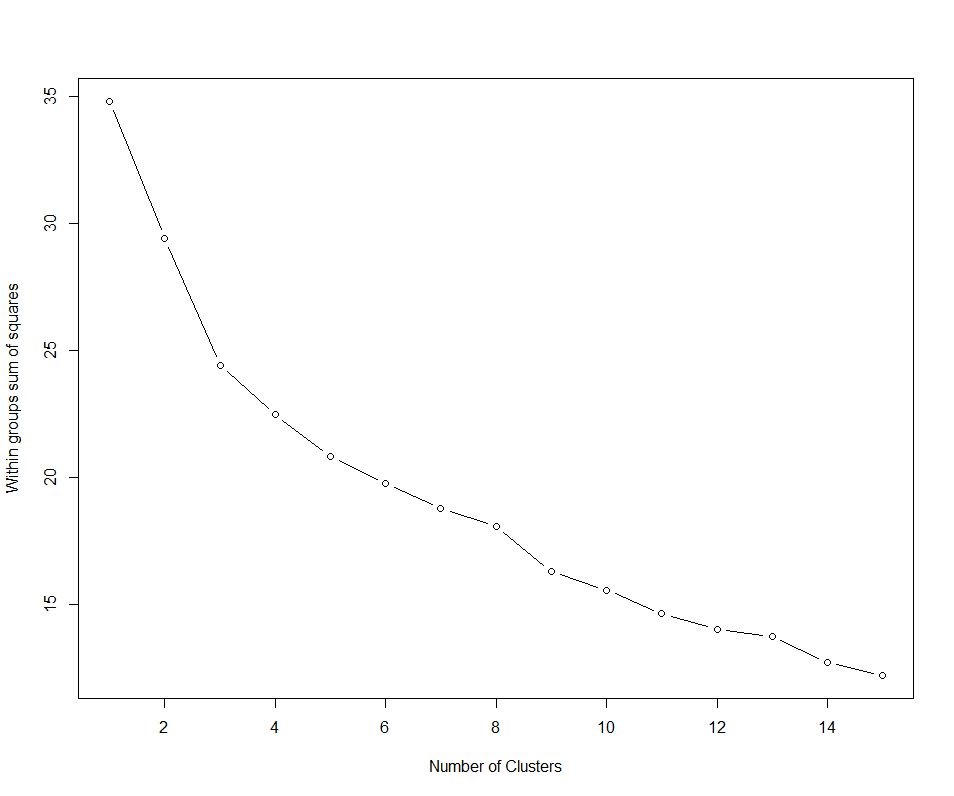
(between\_SS / total\_SS = 83.4 %)



COSINE:

­Within cluster sum of squares by cluster:

[1] 12.6744129 10.8218117 0.9139836 (between\_SS / total\_SS = 29.9 %)

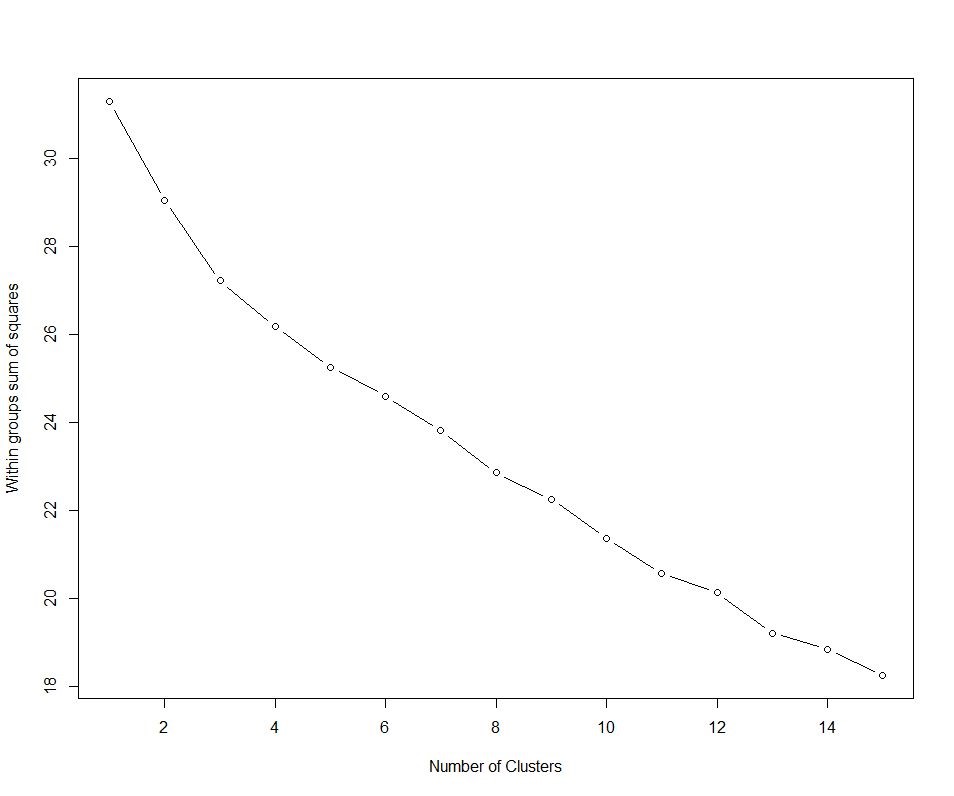


JACCARD:

­​Within cluster sum of squares by cluster:

[1] 0.5708887 14.2528880 12.3981739

(between\_SS / total\_SS = 13.0 %)



Looking at the results the Euclidean distance showed the best results. The sum of squares value was the highest, implying that the Euclidean K­means created more accurate cluster results than the other two.