Analysis of SOU similarities

In this problem you will use the classic vector space model from information retrieval to find similar SOU addresses.

a) Compute the tf-idf vectors for each SOU address. You should lower case all of the text, and remove punctuation.

b) Use a similarity measure to find the following:

- 50 most similar pairs of SOUs given by different Presidents
- 50 most similar pairs of SOUs given by the same President
- 25 most similar pairs of Presidents, averaging the cosine similarity over all pairs of their SOUs

c) Using this vector representation, cluster the speeches using k-means. (Experiment with different number of clusters, and display the clusters obtained. Comment on the clustering results, and whether or not the results are interpretable.