1. Vector-Space Model
2. Create a dictionary of words present in the documents (exclude stop words, only extract important features);
   1. Select all terms;
   2. Convert it to a dimension in a vector space;
3. Build feature vectors for each documents using word frequency;
4. Tokenization, Counting and Normalization is called the **Bag of Words;**
5. Why k-mean clustering isn’t stable?
6. The results from Hierarchical clustering are not very delicate?
   1. How to make more precise clustering? (tf-idf?)
   2. How come tf-idf make the result even worse?
7. What are the ways to reduce dimensionality other than excluding stop-words?
8. How to visualize the data? All I get is the number of clusters and which documents are inside that cluster.

words that occur frequently within a document but not frequently within the corpus receive a higher weighting as these words are assumed to contain more meaning in relation to the document.

dist is defined as 1 - the cosine similarity of each document. Cosine similarity is measured against the tf-idf matrix and can be used to generate a measure of similarity between each document and the other documents in the corpus (each synopsis among the synopses). Subtracting it from 1 provides cosine distance which I will use for plotting on a euclidean (2-dimensional) plane.