Assumptions :

- While removing the punctuations the letters are separated. For example u.s.a. will become u s a
- The distance measure is Euclidean Distance(RSS)

Bag of Words Model using K means :

a).Pre-Processing :

- Tokenization : Longer strings are splits into tokens. In this the text is tokenized into sentences and then sentences are tokenized into words. During tokenization the punctuations are also removed.
- After tokenizing the documents the number is converted to words(like 100 is converted to one hundred).If size is greater than each digit of number is converted to word.(like 100 is converted to one zero zero)
- Case-Folding(Lower Case) : Converting all words to lower case to make searching easy.
- Stop 54Words Removal : Stop words are removed.
- Stemming : It is process of eliminating affixes from the word to get the stem word. Generally it makes word running to run.

b).Methodology :

- Using Bag of Words:
  1. First step is to load the files from directory and reading the files using csopen() function.
  2. Once the file is read the all pre-processing is done like tokenization, conversion of number to words and stemming etc in train data.
  3. For each term in document we calculate term frequency to make Vocabulary that is number of unique terms.
  4. Now we create document vector for each document and it stores the term along with its count in that document and it is called our bag of words.
  5. Now once the bag of words created then we will perform K-means algorithm.
  6. First we initialize k points, called means, randomly.
  7. We categorize each item to its closest mean and we update the mean's coordinates, which are the averages of the items categorized in that mean so far.
  8. We repeat the process for a given number of iterations and at the end, we have our clusters.
  9. Now we have information about class and clusters then we can report error in terms of purity and ARI.

c) Test-Cases :

Number of Iterations	RSS(Using Bag of Words)
1	123209.78452706679
2	119858.42573451044
3	118920.23829753797
4	118692.63812001544
5	118556.50648281173
6	118546.94096735153

7	118502.76898131671
8	118423.1869258269
9	118332.50951127932
10	117982.57906314383
11	117672.56240731604
12	117545.60468030543
13	117491.34594651117
14	117377.06696809552
15	117221.25496067587
16	117096.37262126054
17	117094.67519271927
18	117089.67382450382

Error Measure and Cluster Distribution :

K_Means	Purity	Adjusted_Rand_Index	Cluster_Formation(After last iteration)
Bag of words	0.2654	0.3367	[586,1234,137,1589,1454]