CS57300: Homework 4

Hasini, Urala Liyanage Dona Gunasinghe

Note: I used 1 late day for this assignment.

Q1

(a)

I implemented the standard kmeans algorithm to be consistent with my implementation of spherical kmeans algorithm.
At first I compared the cluster score returned by kmeans implementation of scikit-learn library (with random initialization of centroids instead of initialization of centroids with KMeans++) and the cluster score returned by my implementation of standard kmeans for number of clusters = 50. As shown in Table 1, my implementation of standard kmeans resulted in a better cluster score since I ran standard kmeans with 10 random restarts and selected the solution with the best cluster score, in order to find a better local optimal solution.

Table 1: Comparing cluster scores returned by two implementation of standard KMeans for number of clusters = 50.

matrix	Cluster Score from Scikit KMeans	Cluster Score from my KMeans
W_p	82895.3737049	11370.4631867
W_{NP}	137950.792173	11323.7422572

Figure 1 and Figure 2 plot the cluster scores vs number of clusters for W_p and W_{NP} respectively. For standard kmeans, the cluster score is sum of squared euclidean distances from each data point to its cluster center.

Best K:
Although we can not see a clear knee/elbow point in the following graph, I see a slight knee point when the number of clusters is equal to 100 in both the graphs below, because the slope of the curve slightly gets decreased after 100. Therefore, I select 100 as the best K for kmeans.

Figure 1: Cluster Score vs K for Wp in kmeans.

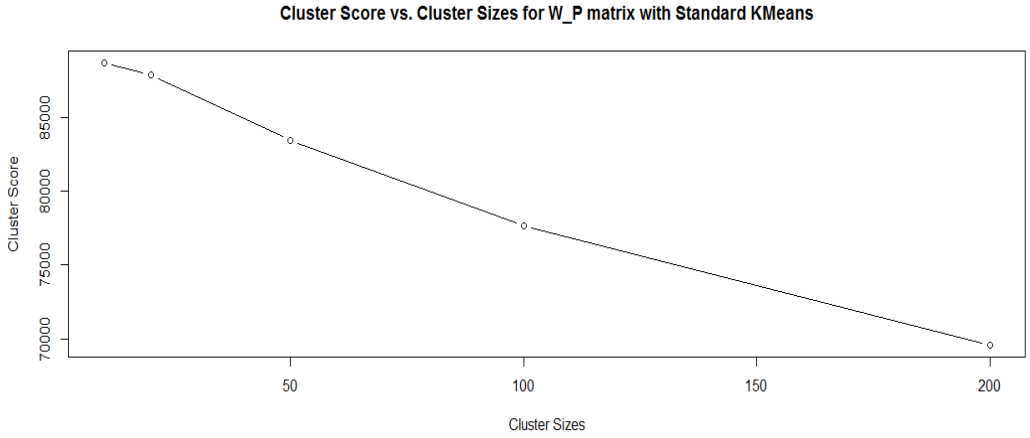
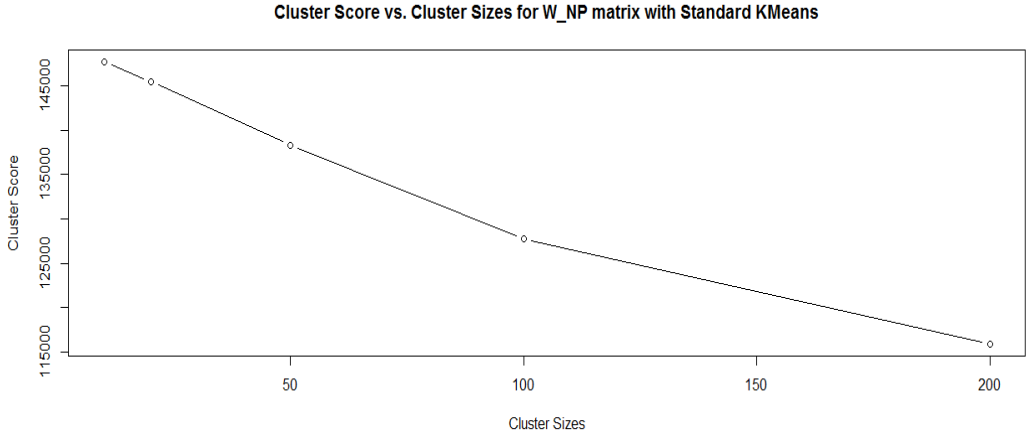


Figure 2: Cluster Score vs K for Wnp in kmeans.



(b) Interesting topics found by standard kmeans for the best K (=100):

Kmeans on W_p matrix:
- Following cluster of words is about a stay in hotel/inn:
"76.0": ["small", "once", "though", "madison", "couldnt", "such", "stay", "wonderful", "hotel", "floor", "bathroom", "inn"],

- Following two words of appreciation are clustered separate from other words - may be because they explains extreme positive cases:

"42.0": ["exceptional"],
"31.0": ["cooked", "perfectly"],

- Following cluster separates out kids/school:
"59.0": ["kids", "school"],

- Following cluster groups words relates to pets and warehouse/supplies:
33.0": ["dog", "pet", "supplies", "ryans", "warehouse"],

- Although following is a big cluster of words, it mainly contains lot of food item names of different kinds - we can find several words related to Mexican food too:
"90.0":["nothing", "both", "recommend", "done", "looked", "tell", "finally", "put", "sandwich", "review", "someone", "friends", "door", "drinks", "decided", "server", "wont", "maybe", "having", "top", "friend", "until", "those", "open", "least", "probably", "three", "ok", "reviews", "waitress", "however", "beer", "outside", "id", "doesnt", "each", "several", "burger", "couple", "plate", "does", "town", "brought", "atmosphere", "downtown", "disappointed", "absolutely", "fried", "eating", "coffee", "counter", "part", "anyone", "busy", "fact", "wife", "fries", "drive", "close", "whole", "keep", "name", "beef", "able", "reason", "second", "high", "fast", "live", "fish", "quite", "restaurants", "pick", "dish", "extra", "doing", "life", "bring", "cool", "bbq", "fantastic", "comes", "show", "sweet", "highly", "deal", "loved", "seriously", "yet", "red", "street", "making", "spot", "yelp", "tip", "theyre", "start", "especially", "soon", "dining", "pork", "mind", "often", "white", "friday", "wish", "looks", "light", "spicy", "glass", "ribs", "enjoyed", "although", "list", "between", "havent", "neighborhood", "options", "dishes", "remember", "tacos", "wow", "bowl", "hands", "warm", "sunday", "ice", "welcome", "burgers", "fan", "exactly", "authentic", "cafe", "style", "tea", "bite", "blue", "along", "patio", "bacon", "prepared", "seating", "french", "including", "dessert", "egg", "evening", "glad", "onions", "vegetarian", "la", "butter", "unique", "veggies", "veggie", "homemade", "curry", "montys", "brunch", "flavors", "chef", "peanut"]

In W_{NP} matrix:
- Following cluster separates out the word gym from the rest of words:
"74.0": ["gym"],

- Following cluster of words is from review(s) on health care organization:
"21.0": ["dr", "doctor", "surgery"],

- Following cluster also groups related words:
"52.0": ["salad", "dressing"],

- Following cluster of words perfectly separates out words related to a ladies salon:

"38.0": ["salon", "nail", "nails", "feet", "pedicure", "polish", "manicure"],

- Following cluster also groups related words:
"40.0": ["breakfast", "eggs"],

(c)

Figure 2 and Figure 3 plot the cluster scores vs number of clusters for W_p and W_{NP} respectively. For spherical kmeans, the cluster score is sum of cosine similarity between each data point to its cluster center.

Best K: In these graphs also we can not identify a clear knee point as best K. However, there is a noticeable decrease in the slope of the curves after k=50 in both the graphs. Therefore, I select k=50 as the best K for spherical kmeans.

Figure 3: Cluster Score vs K for Wp in spherical kmeans.

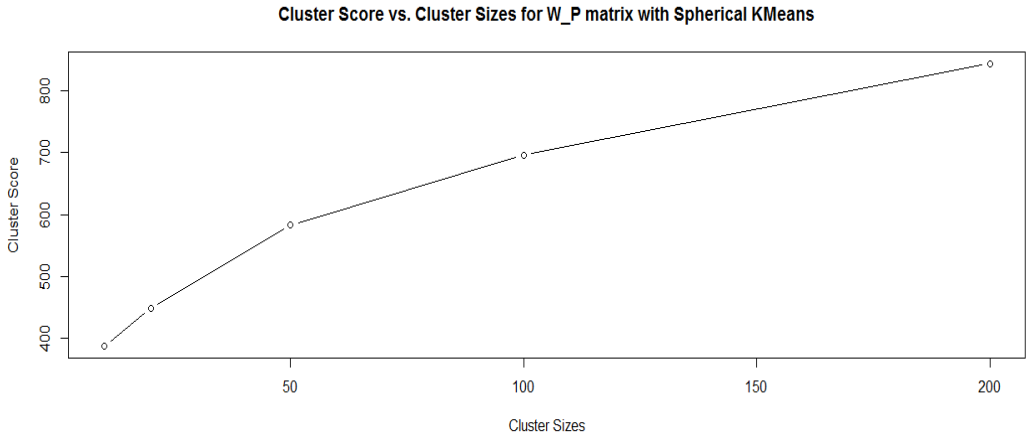
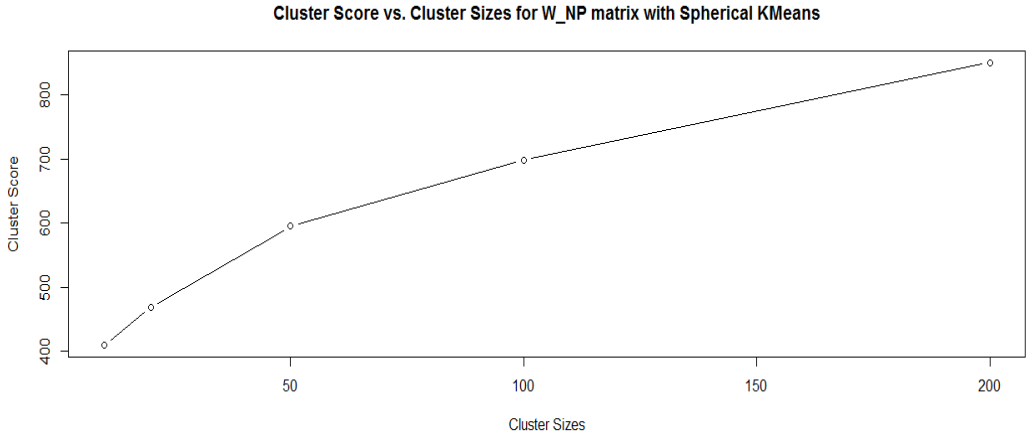


Figure 4: Cluster Score vs K for Wnp in spherical kmeans.



Interesting topics found in spherical kmeans:

- Following cluster of words is also about a health care organization:
48.0": ["care", "office", "high", "husband", "seen", "weeks", "please", "understand", "dr", "etc", "issue", "type", "questions", "returned", "true", "lack", "doctor", "positive", "multiple", "imagine", "expecting", "health", "smile", "patient", "agreed", "closer", "2nd", "schedule", "fall", "doctors", "deserve", "north", "missed", "respect", "7th", "medical", "lenses", "posted", "exam", "concerns", "clinic", "practice", "notch", "patients", "courteous", "news"],

- Following cluster of words contains words mostly related to payments:
"10.0": ["prices", "pay", "help", "bring", "anywhere", "amount", "goes", "25", "sell", "gotten", "dollars", "except", "az", "40", "arrive", "finding", "difference", "filling", "cat", "ryans", "internet", "shipping", "toys", "package", "minimum", "cobbler", "treats", "retail"],

- Following cluster of words is mainly from a review about buying a tv screen to an apartment:
"30.0": ["tv", "min", "airport", "older", "apartment", "300", "24", "stupid", "olive", "lounge", "screen", "acceptable", "150", "machines", "key", "payment", "wifi"],

- Following is a very interesting cluster with lot of words related to mexican food grouped together:
31.0": ["mexican", "green", "taco", "tacos", "authentic", "cashier", "corn", "bakery", "burrito", "la", "bag", "finish", "consistently", "mins", "enchi-

ladas", "soft", "stick", "onion", "enchilada", "grill", "prefer", "wtf", "poorly", "bites", "smoke", "de", "asada", "chile", "somewhat", "skip", "nachos", "heres", "carne", "bright", "ground", "shredded", "fire", "rositas", "tolteca", "dump", "guacamole", "bus", "rolled", "napkins", "yuck", "weak", "neighbors", "board", "closest", "mexico"],

- Following is also an interesting cluster with lot of words related to Thai food:
"37.0": ["served", "soup", "beef", "restaurants", "dish", "thai", "kitchen", "spicy", "says", "kept", "shrimp", "dishes", "tea", "somewhere", "rolls", "mediocre", "veggies", "appetizer", "portions", "wild", "veggie", "yum", "curry", "iced", "vegetables", "entrees", "crispy", "tender", "medium", "flavorful", "pad", "standard", "flavors", "managed", "dead", "thaiger", "appetizers", "salmon", "peanut", "tofu", "waitresses", "yellow", "generous", "mushrooms", "lemon", "cheesecake", "leaves", "milk", "repeat", "sounded", "pot", "raw", "basil", "sticky", "dine", "loves", "presentation", "cuisine", "coconut", "sampler", "wrapped"],

(d)

Differences in the results from the two algorithm:
- Standard kmean's score functions represent sum of squared distances which gets minimized as the number of clusters is increased, where as spherical kmean's score function is sum of cosine similarity which gets maximized as the number of clusters increased.
- Although we can not see a knee/elbow point in the graphs for both algorithm, which can be identified as best the k, we can select k=100 for standard kmeans and k=50 for spherical kmeans.
- When I inspect the words clusters created by standard kmeans and spherical kmeans for the above k's; in order to find interesting topics, I noticed that spherical kmeans has identified more interesting word clusters compared to standard kmeans.
- I have submitted the files containing word clusters identified by two algorithms for the above k values.

Q2.

(a)

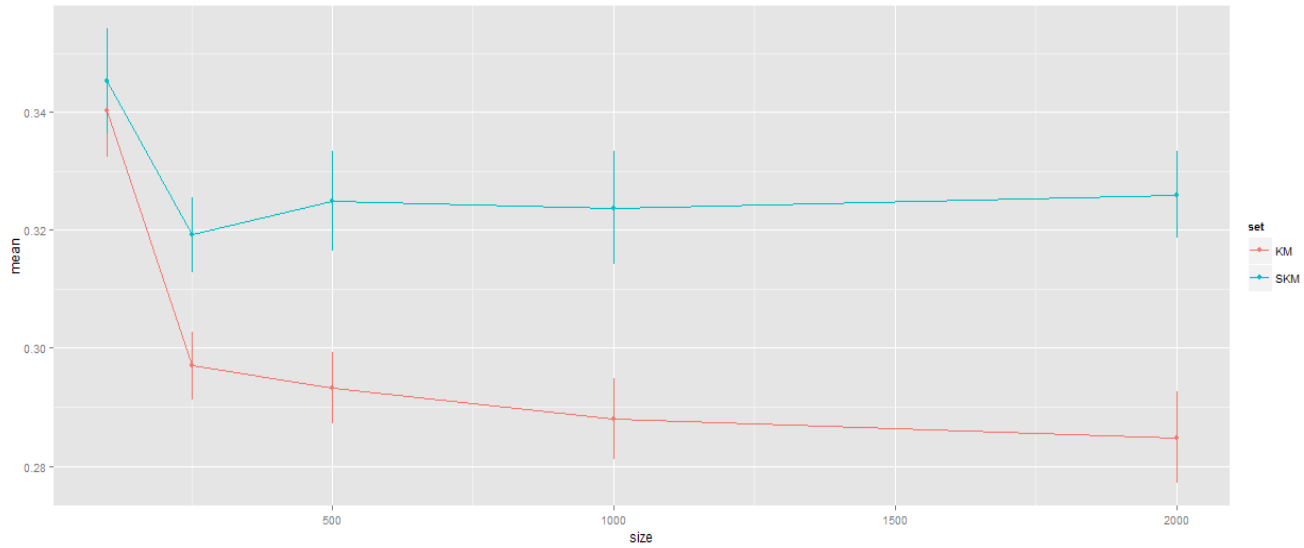
Figure 5 shows the numbers obtained in the experiment for question 2. Here 'avg' stands for the mean zero-one loss and 'stderr' stands standard error for prediction by NBC, whereas 'blavg' and 'blstderr' stands for average and standard error for baseline prediction.
As we can observe, both approach A (KM) and approach B (SKM) performs better than the baseline prediction where as approach A performs better than approach B.

Figure 5: Results obtained for experiment 2.

```
Results for KMeans-clustered words:
  {
  100: {'blavg': [0.50919999999999999], 'avg': [0.3402],             'blstderr': [0.0048265930012794779], 'stderr': [0.0078508598255222963]},
  250: {'blavg': [0.502],               'avg': [0.29699999999999999], 'blstderr': [0.0056000000000000043], 'stderr': [0.0057008771254956903]},
  500: {'blavg': [0.50719999999999998], 'avg': [0.29319999999999996], 'blstderr': [0.0051551915580315775], 'stderr': [0.0060279349697885762]},
 1000: {'blavg': [0.49880000000000002], 'avg': [0.28799999999999998], 'blstderr': [0.0056228106850577907], 'stderr': [0.006876045375068432 8]},
 2000: {'blavg': [0.50480000000000003], 'avg': [0.2848],             'blstderr': [0.0054273382057874421], 'stderr': [0.0077418344079423435]}})

Results for S-KMeans-clustered words:
  {
  100: {'blavg': [0.50919999999999999], 'avg': [0.34540000000000004], 'blstderr': [0.0048265930012794779], 'stderr': [0.0088546033225661777]},
  250: {'blavg': [0.502],               'avg': [0.31919999999999998], 'blstderr': [0.0056000000000000043], 'stderr': [0.006357357941786824]},
  500: {'blavg': [0.50719999999999998], 'avg': [0.32500000000000001], 'blstderr': [0.0051551915580315775], 'stderr': [0.0084958813551037765]},
 1000: {'blavg': [0.49880000000000002], 'avg': [0.32379999999999998], 'blstderr': [0.0056228106850577907], 'stderr': [0.0096434433684239561]},
 2000: {'blavg': [0.50480000000000003], 'avg': [0.32599999999999996], 'blstderr': [0.0054273382057874421], 'stderr': [0.0073756355658343084]}})
```

Figure 6: Learning curves for approach A and approach B: Mean zero-one loss vs training set size.



(b)

Null hypothesis: H_0 : There is no difference between the performance of approach A (KM) and approach B (SKM).
Alternative hypothesis: H_a : Approach A performs better than approach B.

(c)

As we can observe from figure 6, the learning curve of approach A reports lower mean zero-one loss than the learning curve of approach B, for all the training set sizes without overlapped error bars (which represents the variation of the mean), except for the training set size = 100. We can ignore that case since the training set size=100 is very small. Therefore, we can say that the observed results supports alternative hypothesis.

Q3.

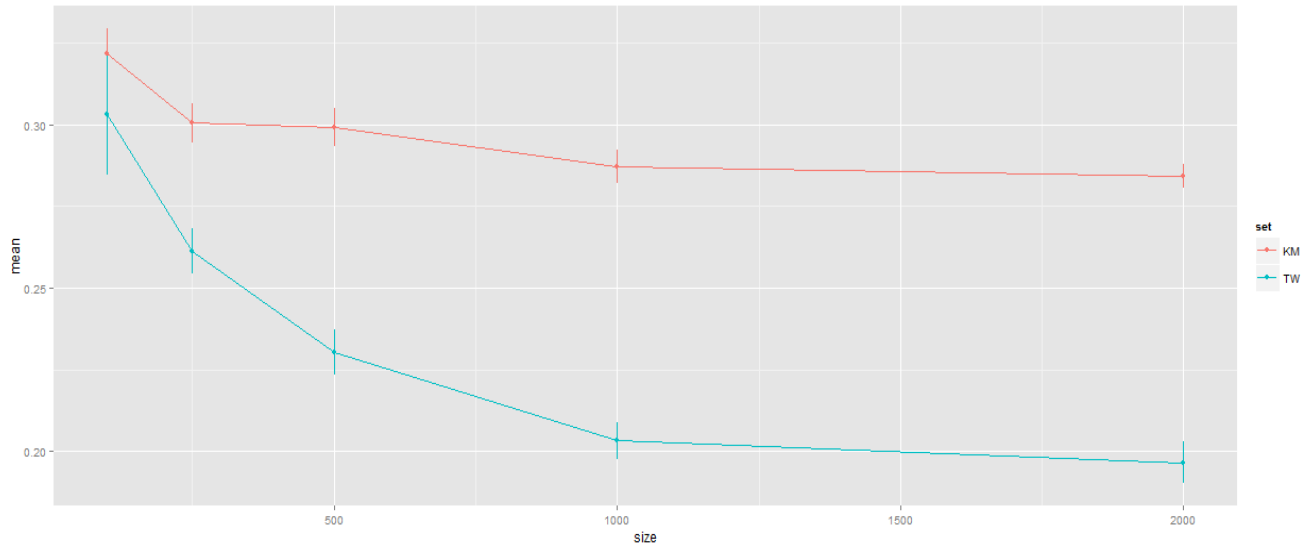
(a)

Since standard kmeans performed better in the experiments for question 2, I selected approach B to be the NBC model using 100 binary features from standard kmeans topics (word clusters).
Figure 7 shows the numbers obtained in the experiment for question 3. Here 'avg' stands for the mean zero-one loss and 'stderr' stands standard error for prediction by NBC, whereas 'blavg' and 'blstderr' stands for average and standard error for baseline prediction.
As we can observe, both approach A - NBC with binary features for top 2000 words (TW) and approach B - NBC with 100 binary features for topics selected by kmeans (KM) performs better than the baseline prediction where as approach A performs better than approach B.

Figure 7: Results obtained for experiment 3.

```
Results for KMeans-clustered words:
  {
   100: {'blavg': [0.49560000000000004], 'avg': [0.32179999999999997], 'blstderr': [0.005697718841782213], 'stderr': [0.007597104711664835]},
   250: {'blavg': [0.49519999999999997], 'avg': [0.30040000000000006], 'blstderr': [0.005665333176433673], 'stderr': [0.005877414397505079]},
   500: {'blavg': [0.5063999999999996], 'avg': [0.29920000000000002], 'blstderr': [0.005504906902028412], 'stderr': [0.005763332369384923]},
  1000: {'blavg': [0.5088000000000003], 'avg': [0.28700000000000003], 'blstderr': [0.005162944896084021], 'stderr': [0.005140038910358556]},
  2000: {'blavg': [0.5063999999999996], 'avg': [0.28419999999999995], 'blstderr': [0.005504906902028412], 'stderr': [0.003649109480407510]}})
Results for top words:
  {
   100: {'blavg': [0.49560000000000004], 'avg': [0.30320000000000003], 'blstderr': [0.005697718841782213], 'stderr': [0.018796169822599497]},
   250: {'blavg': [0.49519999999999997], 'avg': [0.26119999999999999], 'blstderr': [0.005665333176433673], 'stderr': [0.006986844781446259]},
   500: {'blavg': [0.5063999999999996], 'avg': [0.23020000000000002], 'blstderr': [0.005504906902028412], 'stderr': [0.006878662660721194]},
  1000: {'blavg': [0.5088000000000003], 'avg': [0.20319999999999999], 'blstderr': [0.005162944896084021], 'stderr': [0.005565608681896348]},
  2000: {'blavg': [0.5063999999999996], 'avg': [0.1966],            'blstderr': [0.005504906902028412], 'stderr': [0.006362703827776364]}})
```

Figure 8: Learning curves for approach A and approach B: Mean zero-one loss vs training set size.



(b)

Null hypothesis: H_0 : There is no difference between the performance of approach A (TW) and approach B (KM).
Alternative hypothesis: H_a : Approach A performs better than approach B.

(c)

As we can observe from Figure 8, the learning curve of approach A reports lower mean zero-one loss than the learning curve of approach B, for all the training set sizes without overlapped error bars (which represents the variation of the mean), except for the training set size = 100. We can ignore that case since the training set size=100 is very small. Therefore, we can say that the observed results supports alternative hypothesis.

Q4.

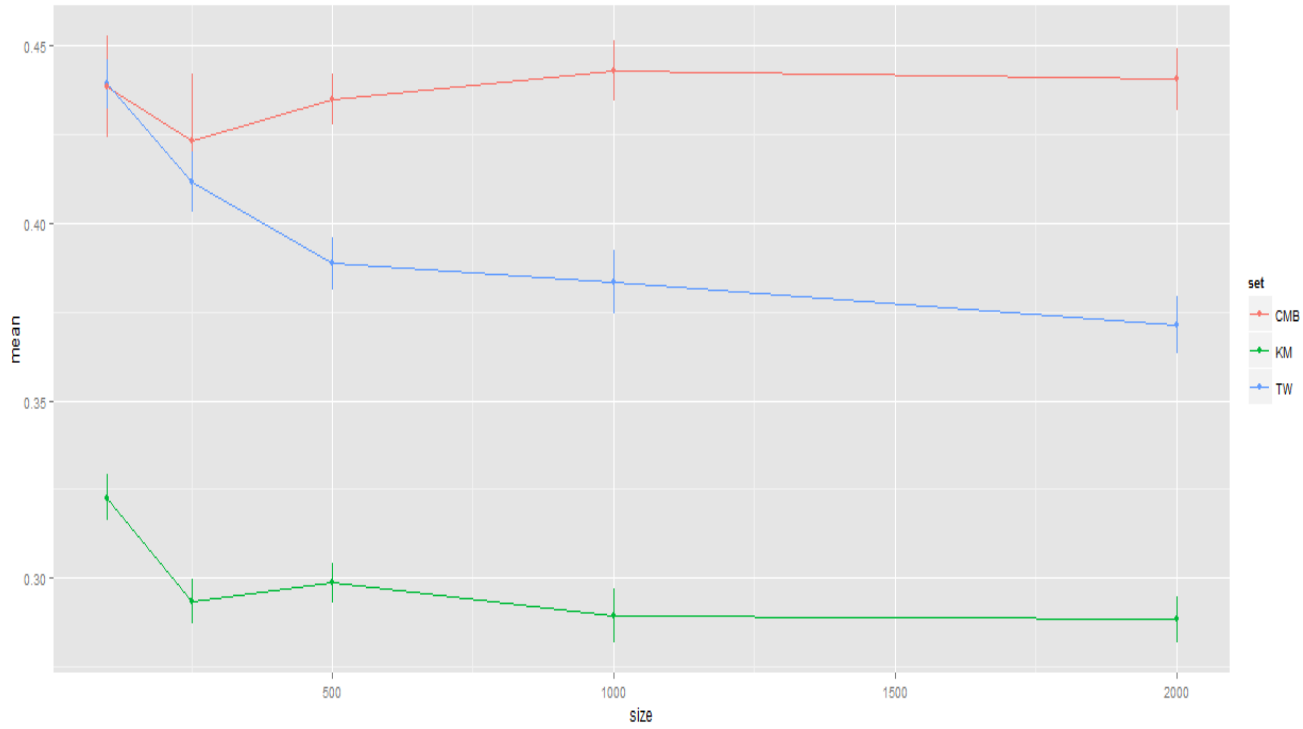
(a)

Figure 9 shows the numbers obtained in the experiment for question 4. Here 'avg' stands for the mean zero-one loss and 'stderr' stands standard error for prediction by NBC, whereas 'blavg' and 'blstderr' stands for average and standard error for baseline prediction.
As we can observe, all three approaches: approach A - NBC with binary features for randomly selected 100 top words (TW), approach B - NBC with 100 binary features for topics selected by kmeans (KM) and approach C - NBC with combined features from approach A and approach B performs better than the baseline prediction. Approach B (KM) performs the best and approach C performs the worst.

Figure 9: Results obtained for experiment 4.

```
Results for KMeans-clustered words:
  {
   100: {'blavg': [0.50480000000000003], 'avg': [0.32259999999999994], 'blstderr': [0.0075217019350676236], 'stderr': [0.00654858763398643]},
   250: {'blavg': [0.50079999999999991], 'avg': [0.29359999999999997], 'blstderr': [0.0076691590151723964], 'stderr': [0.006208381431581019 8]},
   500: {'blavg': [0.49000000000000005], 'avg': [0.29859999999999992], 'blstderr': [0.0069914233171794148], 'stderr': [0.005521231746630457 7]},
  1000: {'blavg': [0.49400000000000005], 'avg': [0.28939999999999999], 'blstderr': [0.0074350521181764453], 'stderr': [0.007626532632854853 7]},
  2000: {'blavg': [0.51639999999999997], 'avg': [0.28840000000000005], 'blstderr': [0.0056554398591091076], 'stderr': [0.006516440746297014 7]}})
Results for top words:
  {
   100: {'blavg': [0.50480000000000003], 'avg': [0.43920000000000003], 'blstderr': [0.0075217019350676236], 'stderr': [0.006963906949407064 9]},
   250: {'blavg': [0.50079999999999991], 'avg': [0.41160000000000008], 'blstderr': [0.0076691590151723964], 'stderr': [0.008356075633932473 5]},
   500: {'blavg': [0.49000000000000005], 'avg': [0.3886],              'blstderr': [0.0069914233171794148], 'stderr': [0.007294107210618721 4]},
  1000: {'blavg': [0.49400000000000005], 'avg': [0.38340000000000002], 'blstderr': [0.0074350521181764453], 'stderr': [0.009073257408450399 7]},
  2000: {'blavg': [0.51639999999999997], 'avg': [0.37140000000000001], 'blstderr': [0.0056554398591091076], 'stderr': [0.007960150752341315 2]}})
Results for combined words:
  {
   100: {'blavg': [0.50480000000000003], 'avg': [0.43840000000000001], 'blstderr': [0.0075217019350676236], 'stderr': [0.014404999132245721]},
   250: {'blavg': [0.50079999999999991], 'avg': [0.42300000000000004], 'blstderr': [0.0076691590151723964], 'stderr': [0.018976301009416986]},
   500: {'blavg': [0.49000000000000005], 'avg': [0.43499999999999994], 'blstderr': [0.0069914233171794148], 'stderr': [0.007206941098690899 8]},
  1000: {'blavg': [0.49400000000000005], 'avg': [0.44299999999999995], 'blstderr': [0.0074350521181764453], 'stderr': [0.008434453153583819]},
  2000: {'blavg': [0.51639999999999997], 'avg': [0.44060000000000005], 'blstderr': [0.0056554398591091076], 'stderr': [0.008704251834592103 4]}})
```

Figure 10: Learning curves for approach A, B and C: Mean zero-one loss vs training set size.



(b)

Null hypothesis: H_0 : There is no difference between the performance of approach A (TW) and approach B (KM).
Alternative hypothesis: H_a : Approach B performs better than approach A.

As we can observe from Figure 10, the learning curve of approach B (green) reports lower mean zero-one loss than the learning curve of approach A (blue), for all the training set sizes without overlapped error bars (which represents the variation of the mean. Therefore, we can say that the observed results supports alternative hypothesis.

(c)

Null hypothesis: H_0 : There is no difference between the performance of approach C (CMB) and approach B (KM).

Alternative hypothesis: H_a : Approach C performs worse than approach B.

As we can observe from Figure 10, the learning curve of approach C reports higher mean zero-one loss than the learning curve of approach B, for all the training set sizes without overlapped error bars (which represents the variation of the mean). Therefore, we can say that the observed results supports alternative hypothesis.

NOTE: I plotted the two/three learning curves in questions 2-4 in the same graph because at each training set sizes, I used the same training set and test set for experiments with all the approaches in a particular question. Please refer finalexperiments.py, inc10FCV.py in source code.