Part 1: Binary Clustering

I ran K Means on the messages, making sure to stem them with the Tf-idf Vectorizer. I ran KMeans with k=2 because I know that the messages are either going to be spam or not (binary classifier).

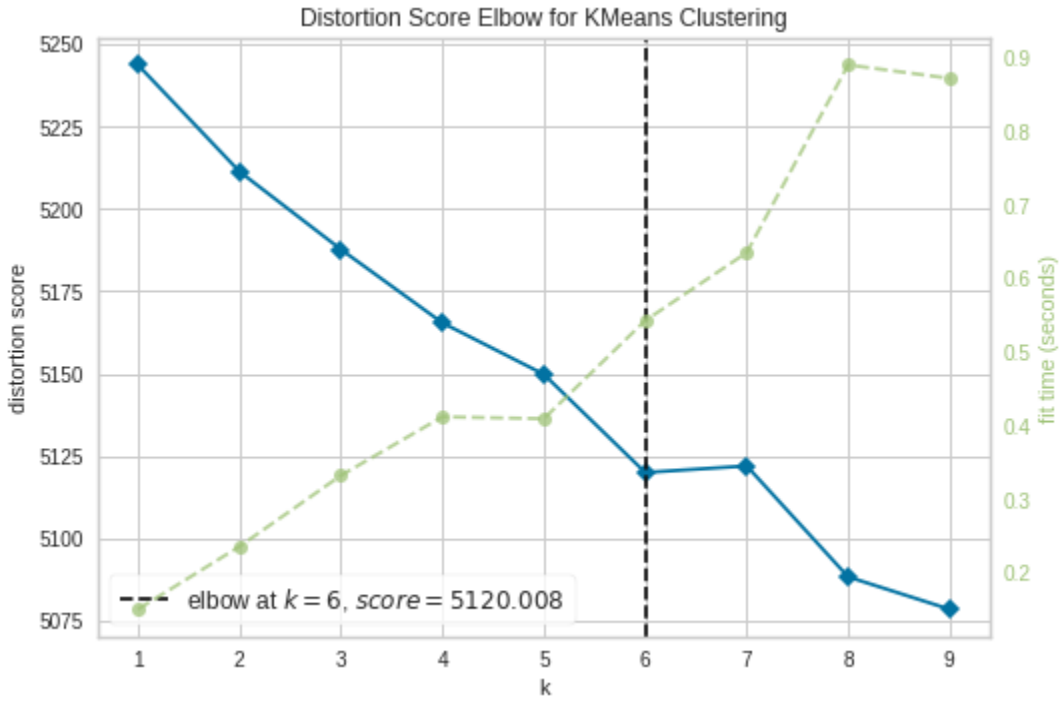
Accuracy: 0.17564006024096385
Precision: 0.13515115589804386
Recall: 0.9970845481049563
F1 Score: 0.23803723681921

Extra Credit:

I ran PCA with 3 principal components, making sure to preprocess the messages the same way as with K means (stemming and using the TFIDF vectorizer). Then I used KNN with k = 1 to classify the points.

Accuracy: 1.0
Precision: 1.0
Recall: 1.0
F1-Score: 1.0

Part 2: Clustering Quality



I used the elbow test to find out that k = 6 was the most optimal. Therefore, I ran KMeans with k = 6 on the stemmed and Tf-idf vectorizer messages.
Then I found how much spam each cluster contained:
Cluster 0 is 5.7763515181436675% spam
Cluster 1 is 0.0% spam
Cluster 2 is 0.3058103975535168% spam
Cluster 3 is 0.0% spam
Cluster 4 is 0.49261083743842365% spam
Cluster 5 is 92.42105263157895% spam

Clusters 0-4 contained very little spam and cluster 5 was mostly spam.

a.

Cluster 0:

Cluster size: 4051

5.8% spam

Messages:

1. U dun say so early hor... U c already then say...
2. Nah I don't think he goes to usf he lives around here though
3. Even my brother is not like to speak with me. They treat me like aids patent.
4. As per your request 'Melle Melle (Oru Minnaminunginte Nurungu Vettam)' has been set as your callertune for all Callers. Press *9 to copy your friends Callertune
5. I've been searching for the right words to thank you for this breather. I promise i wont take your help for granted and will fulfil my promise. You have been wonderful and a blessing at all times.

Cluster 1:

Cluster size: 59

0% spam

Messages:

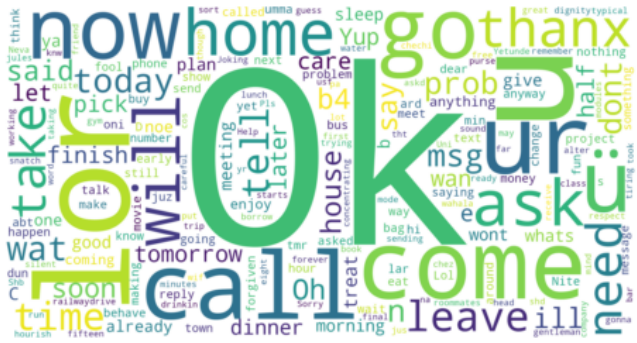
1. Sorry in meeting I'll call you later
2. Sorry I can't help you on this.
3. Sorry I'll call later
4. Sorry i've not gone to that place. I.ll do so tomorrow. Really sorry.
5. I'm in a meeting call me later a

Cluster 2:

Cluster size: 327
0.3% spam
Messages:

1. I'm going to try for 2 months ha ha only joking
2. So ü pay first lar... Then when is da stock comin...
3. Fair enough anything going on?
4. Ok lar i double check wif da hair dresser already he said wun cut v short. He said will cut until i look nice.
5. I am going to sao mu today. Will be done only at 12

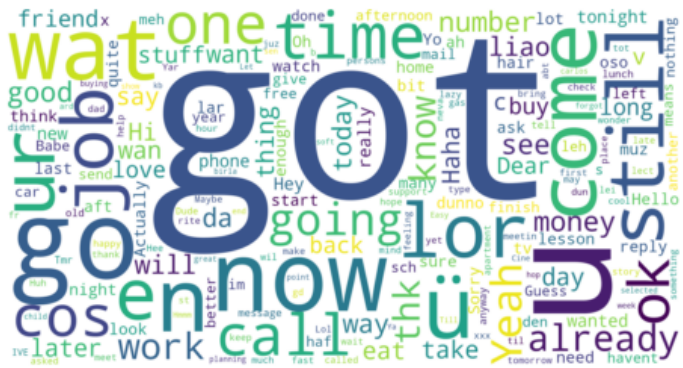
Cluster 3:



Cluster size: 186
0% spam
Messages:

1. Ok lar... Joking wif u oni...
2. Sorry my roommates took forever it ok if I come by now?
3. Ok... Ur typical reply...
4. ok. I am a gentleman and will treat you with dignity and respect.
5. Ok no prob. Take ur time.

Cluster 4:



Cluster size: 203

0.5% spam

Messages:

1. Go until jurong point crazy.. Available only in bugis n great world la e buffet... Cine there got amore wat...
2. Yeah he got in at 2 and was v apologetic. n had fallen out and she was actin like spoilt child and he got caught up in that. Till 2! But we won't go there! Not doing too badly cheers. You?
3. Got c... I lazy to type... I forgot ü in lect... I saw a pouch but like not v nice...
4. Sindu got job in birla soft ..
5. Let me know when you've got the money so carlos can make the call

Cluster 5:

Cluster size: 475
92.4% spam
Messages:

1. Free entry in 2 a wkly comp to win FA Cup final tkts 21st May 2005. Text FA to 87121 to receive entry question(std txt rate)T&C's apply 08452810075over18's
2. WINNER!! As a valued network customer you have been selected to receivea £900 prize reward! To claim call 09061701461. Claim code KL341. Valid 12 hours only.
3. SIX chances to win CASH! From 100 to 20000 pounds txt> CSH11 and send to 87575. Cost 150p/day 6days 16+ TsandCs apply Reply HL 4 info
4. URGENT! You have won a 1 week FREE membership in our £100000 Prize Jackpot! Txt the word: CLAIM to No: 81010 T&C www.dbuk.net LCCLTD POBOX 4403LDNW1A7RW18
5. 07732584351 - Rodger Burns - MSG = We tried to call you re your reply to our sms for a free nokia mobile + free camcorder. Please call now 08000930705 for delivery tomorrow

b.

Cluster 0 which is about 6% spam had “England v Macedonia - dont miss the goals/team news. Txt ur national team to 87077 eg ENGLAND to 87077 Try:WALES SCOTLAND 4txt/ú1.20 POBOXox36504W45WQ 16+” which is spam. It also had “SMS. ac Sptv: The New Jersey Devils and the Detroit Red Wings play Ice Hockey. Correct or Incorrect? End? Reply END SPTV” which is also spam.

Cluster 5 which was about 92% spam had “Turns out my friends are staying for the whole show and won't be back til ~### so feel free to go ahead and smoke that $### worth” which is not spam. It also had “Ela kano.il download come wen ur free..” which is also not spam.

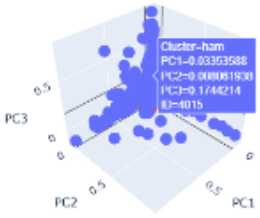
c.

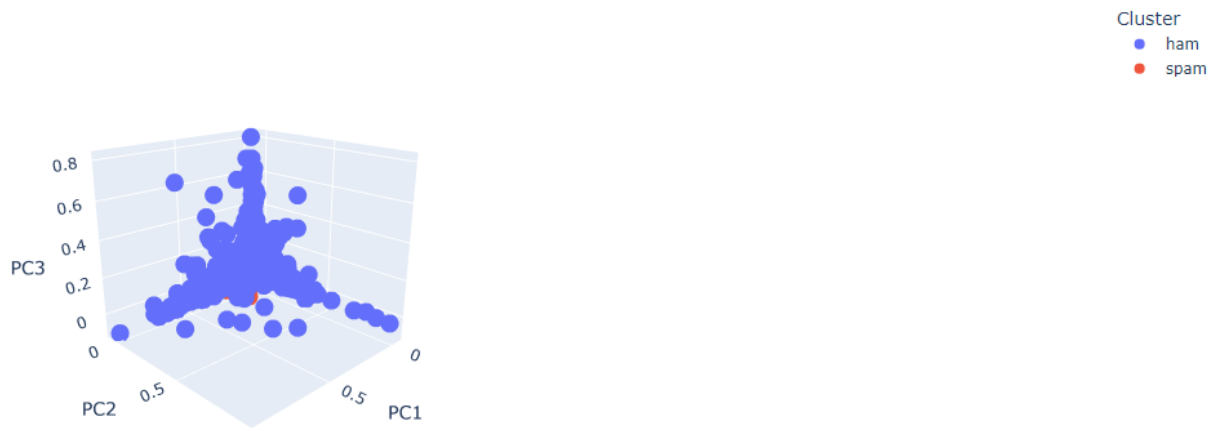
New texts:

1. “are u ok?”
    a. Predicted to fall into cluster 3 because it uses “ok” and “u” which were some of the biggest words in the word cloud
    b. Closest cluster according to KMeans: 3
2. “FREE cash! call or txt 800-444-free now to claim it”
    a. Predicted to fall into cluster 5 because it uses “FREE”, “cash”, “call”, “txt”, “now” and “claim” which were some of the largest words in the word cloud
    b. Closest cluster according to KMeans: 5
3. “sorry, cant talk, ill call you later”
    a. Predicted to fall into cluster 1 because it uses “sorry”, “call”, and “later” which were some of the largest words in the word cloud
    b. Closest cluster according to KMeans: 1
4. “wat job u got?”
    a. Predicted to fall into cluster 4 because it uses “wat”, “job”, “u” and “got” which were some of the largest words in the word cloud
    b. Closest cluster according to KMeans: 4

Part 3: PCA

a.



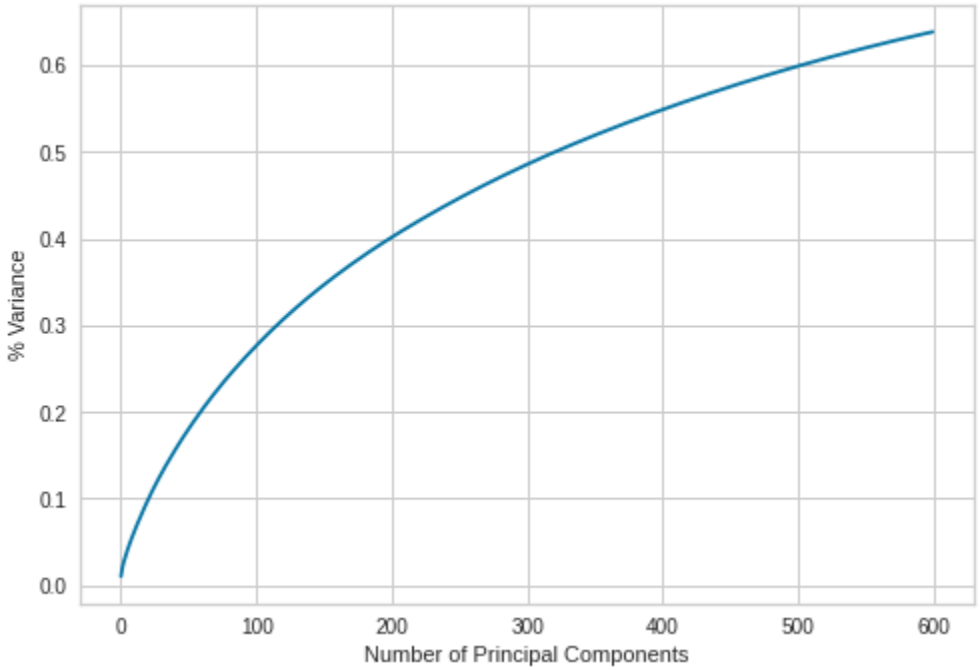


b.

1. “accordingly. I repeat just text the word ok on your mobile phone and send”
    a. It is probably far away from spam because it does not really have most of the words that often appear in spam, especially “FREE”
2. “How come it takes so little time for a child who is afraid of the dark to become a teenager who wants to stay out all night?”
    a. This one has pretty much no words in common with the other spam messages, which is why it was so far away from the other spam messages.

c.

There are 5301 features. You need at least 322 principal components to capture 50% of the variance in the data.



Part 4: Dimensionality Reduction and Classification

First Classifier:

I used KNN with the vectorized features and k = 1.

Accuracy: 0.9423844837421563
Precision: 1.0
Recall: 0.5627705627705628
F1-Score: 0.7202216066481995
CPU times: user 4.23 s, sys: 835 ms, total: 5.06 s
Wall time: 4.84 s

Second Classifier: PCA

I chose the number of principal components by measuring the accuracy every 100 components
With 1 PCs the accuracy is 0.8545350827153452
With 101 PCs the accuracy is 0.9606389047347405
With 201 PCs the accuracy is 0.9606389047347405
With 301 PCs the accuracy is 0.9560752994865944
With 401 PCs the accuracy is 0.9520821448944666
With 501 PCs the accuracy is 0.9509412435824302
With 601 PCs the accuracy is 0.9509412435824302
With 701 PCs the accuracy is 0.9509412435824302
With 801 PCs the accuracy is 0.9509412435824302
With 901 PCs the accuracy is 0.9509412435824302

I narrowed it down to being between 51 to 201:
With 51 PCs the accuracy is 0.9520821448944666
With 76 PCs the accuracy is 0.9577866514546491
With 101 PCs the accuracy is 0.9600684540787222
With 126 PCs the accuracy is 0.9646320593268682
With 151 PCs the accuracy is 0.9606389047347405
With 176 PCs the accuracy is 0.9600684540787222
With 201 PCs the accuracy is 0.9617798060467769

I chose k = 126 for the number of components because it had the highest accuracy.

So I used PCA with 126 principal components on the vectorized features and KNN with k = 1.

Accuracy: 0.970907016543069
Precision: 0.8884120171673819
Recall: 0.8922413793103449
F1-Score: 0.8903225806451613
CPU times: user 57.9 s, sys: 6.23 s, total: 1min 4s
Wall time: 55.5 s

Part 5: Spam Classification

I decided to go with a PCA + KNN classifier based on my findings from part 4. When comparing F-1 score, PCA + KNN with k=1 outperformed KNN with k =1 slightly.

Key:

S = Stemmed

SW = Stopwords

N = Normalized

C = Countvectorizer

Preprocessing	# of PCs	Accuracy	Precision	Recall	F1-Score
S + TDIDF + SW	26	0.966	0.883	0.868	0.876
S +TDIDF +SW + N	26	0.968	0.862	0.875	0.868
S + TDIDF	26	0.965	0.892	0.868	0.880
S + TDIDF + N	26	0.962	0.840	0.874	0.857
TDIDF + SW	51	0.970	0.893	0.860	0.876
TDIDF + SW + N	26	0.963	0.817	0.901	0.857
TDIDF	26	0.958	0.829	0.865	0.846
TDIDF + N	101	0.965	0.965	0.837	0.857
C	51	0.961	0.913	0.790	0.847
C + SW	51	0.962	0.884	0.797	0.838
C + S + SW	51	0.959	0.940	0.763	0.842
C + N	176	0.959	0.866	0.808	0.836
C + SW + N	126	0.950	0.806	0.809	0.807
C + S + SW + N	76	0.957	0.819	0.819	0.832

The best classifier is to use PCA + KNN with k=1 with a TFIDFvectorizer and stemming because it had the best F1-score. I made sure to focus on the F1 score because the dataset is unbalanced, so accuracy does not accurately describe the performance.