3.1 tfidf_minibatchkmeans

October 19, 2017

0.1 TF-IDF MiniBatchKMeans

```
In [1]: import pandas as pd
        import numpy as np
        from sklearn.cluster import MiniBatchKMeans, KMeans
        from sklearn.manifold import TSNE
        import matplotlib.pyplot as plt
        from sklearn.feature_extraction.text import TfidfVectorizer
        from sklearn import metrics
        from sklearn.cluster import MiniBatchKMeans, KMeans
        import matplotlib.cm as cm
```

0.2 Read the data

The first step is to read the data of the 500 tweets that we are going to use, this data is stored in a pandas DataFrame.

```
In [4]: #df = pd.read_pickle('my_df.pickle')
        df = pd.read_pickle('./pickles/500dataset.pickle')
        #df = pd.read_pickle('77k_df.pickle')
```

```
In [6]: df
```

```
Out[6]:                                                  text  \
        0    @sarahtavel What #MustHave #tech gadget can yo...
        1    Start-ups from every continent heading to #web...
        2    I'm at the #WebSummit2015 this week. On ali(at...
        3    @jalak What #MustHave #tech gadget can you not...
        4    #websummit is about to kickoff in #dublin! Wha...
        5    I'm at the #WebSummit this week. On ali(at)gos...
        6    Liz Halash: one question you need to ask yours...
        7    @fabricegrinda What #MustHave #tech gadget can...
        8    Know someone who needs help with their compute...
        9    We are excited &amp; ready for the #WebSummit2...
        10   @amynauiokas we have a better solution for tra...
        11   @pepi_post Thrilled to meet @beautyxapp team a...
        12   Are people more likely to act unethically on s...
        13   Sage are supporting #SageMentorHours at #WebSu...
        14   @mattturck What #MustHave #tech gadget can you...
```

```
15   Welcome All "10 Russian Startups Coming to #we...
16   Have questions about setting up your #startup ...
17   @JoshGad Just watched the entire @TheInternshi...
18   @WebSummitHQ what is the speaker lineup for th...
19   @benwilson_ml Maybe You join also #WebSummit20...
20   @Bandrew What #MustHave #tech gadget can you n...
21   @clairescatwalk hi Claire I did see you at the...
22   Trying out airpor registration for the #WebSum...
23   The @WebSummitHQ is here! Follow us this year ...
24   Can see some improvements happening to the @bl...
25   Don't fall into the 'mentoring gap' -  read up...
26   Quote #websummit and receive 10% off your bill...
27   Something to read for the poor souls whose fli...
28   At Birmingham airport, heading for #WebSummit2...
29   @WebSummitHQ are food and drinks included in t...
..                                                ...
470  Cool Monday - off we go to #websummit in Dubli...
471  Play #huntly at #WebSummit @vodtalker @TomMitc...
472  Tilkee's team is on his way to Dublin ! #WebSu...
473                            @WebSummitHQ thank you
474                      boycott the fucking websummit
475  Looking forward to the #WebSummit it always sh...
476  We're in Dublin and ready for #WebSummit. If y...
477  Let's see what all the fuss is about. Looking ...
478  @RebrandlyBuzz We can't wait for #websummit! H...
479  Well hello @imaginarycloud @NataliaTerlecka @N...
480  Looking to talk to VC's at the #websummit inte...
481  Heading to #WebSummit15? Dont miss @vmatarran...
482  Touch down. @WebSummitHQ here we come @Phoodst...
483  Connect with #ArabNetME team as they take on #...
484  What the fog... Delayed on my way to #websummi...
485  @iamdanram @spiritradioire a friend of mine is...
486  Big business just voted against @EndaKennyTD &...
487  Approaching #dublin #websummit with Diumbo's t...
488  The @audiireland stand is coming together nice...
489  Chk #30DaysStartup  https://t.co/ScgIOF2NvL L...
490  THIS CAN'T BE TRUE! 'Taoiseach Enda Kenny send...
491  The first #websummit Tag Cloud.. https://t.co/...
492  @usmanjaved84 Come &amp; See @iGotForms #disru...
493  made it to @WebSummitHQ frm SFO w/ @Boombotix ...
494  WOW!!: 24HRs till #Websummit !! - Be sure 2 st...
495  @darraghdoyle @NewstalkFM Great, email sent! C...
496  Heading to #websummit and caught by media. We ...
497  @IrishTimes Come &amp; See @iGotForms #disrupt...
498  @AdrienChl Come &amp; See @iGotForms #disrupt ...
499  The Pope thinks the internet can be a force fo...

                                       text_processed  \
```

```
0       [gadget, travel, without, ?, stop, stand, wed]
1    [start-ups, every, continent, heading, including]
2    [i'm, week, ., ali, (, ), goss, (, dot, ), ie,...
3       [gadget, travel, without, ?, stop, stand, wed]
4                        [kickoff, looking, forward, ?]
5    [i'm, week, ., ali, (, ), goss, (, dot, ), ie,...
6    [liz, halash, :, one, question, need, ask, wou...
7       [gadget, travel, without, ?, stop, stand, wed]
8    [know, someone, needs, help, computer, science...
9    [excited, &, ready, today, ., weather, stoppin...
10   [better, solution, tracking, bugs, sending, vi...
11   [thrilled, meet, team, plz, share, stand, numb...
12   [people, likely, act, unethically, social, med...
13                [sage, supporting, sage, mentors, &]
14      [gadget, travel, without, ?, stop, stand, wed]
15   [welcome, ", russian, startups, coming, (, /, ...
16   [questions, setting, usa, ?, come, talk, team,...
17   [watched, entire, movie, final, scene, @, h, ...
18   [speaker, lineup, future, ireland, event, toni...
19           [maybe, join, also, ?, would, nice, chat]
20      [gadget, travel, without, ?, stop, stand, wed]
21                              [hi, claire, see, thank]
22   [trying, airpor, registration, lines, movibg, ...
23   [!, follow, us, year, we'll, giving, away, pri...
24   [see, improvements, happening, website, buildi...
25   [don't, fall, ', mentoring, gap, ', -, read, m...
26   [quote, receive, %, bill, complimentary, glass...
27   [something, read, poor, souls, whose, flight, ...
28   [birmingham, airport, ,, heading, (, fog, perm...
29          [food, drinks, included, ticket, price, ?]
..                                                  ...
470                    [cool, monday, -, go, dublin, !]
471                [play, win, min, talk, investors, !]
472                    [tilkee's, team, way, dublin, !]
473                                             [thank]
474                                  [boycott, fucking]
475  [looking, forward, always, shows, something, n...
476  [we're, dublin, ready, you're, dublin, downloa...
477  [let's, see, fuss, ., looking, forward, tomorr...
478  [can't, wait, here's, places, grab, bite, eat,...
479                    [well, hello, see, hours, craic]
480  [looking, talk, vc, ', interested, mobile, pay...
481  [heading, , miss, ,, santander, , head, grou...
482                                       [touch, ., come]
483                               [connect, team, take]
484                              [fog, ., delayed, way]
485  [friend, mine, year, ., still, running, pace, ...
486                         [big, business, voted, &, -]
```

```
487                      [approaching, diumbo's, team]
488                  [stand, coming, together, nicely]
489  [chk, , looking, talk, vc, ', interested, mob...
490  [', true, !, ', taoiseach, enda, kenny, sends,...
491                              [first, tag, cloud, .]
492  [come, &, see, forms, stand, &, could, win, ap...
493  [made, frm, sfo, w, /, &, want, go, home, mand...
494  [wow, !, :, hr, till, !, -, sure, stop, tues, ...
495     [great, ,, email, sent, !, cheers, darragh, .]
496        [heading, caught, media, ., want, go, back]
497  [come, &, see, forms, stand, &, could, win, ap...
498  [come, &, see, forms, stand, &, could, win, ap...
499  [pope, thinks, internet, force, good, -, think...

                                                vector
0    {'@': [-0.0704075, -0.0324585, 0.0529671, 0.06...
1    {'S': [-5.27091, 2.70473, -0.157593, 4.45988, ...
2    {'I': [0.0332858, -0.00211889, 0.0470522, 0.08...
3    {'@': [-0.0704075, -0.0324585, 0.0529671, 0.06...
4    {'#': [-1.99362, 0.557125, 0.229587, 1.53814, ...
5    {'I': [0.0332858, -0.00211889, 0.0470522, 0.08...
6    {'L': [-0.0862547, 0.0434853, 0.0529697, 0.047...
7    {'@': [-0.0704075, -0.0324585, 0.0529671, 0.06...
8    {'K': [-0.0409234, 0.0100177, 0.0456978, 0.110...
9    {'W': [-2.20337, 1.15484, -0.0837178, 1.83615,...
10   {'@': [-0.0704075, -0.0324585, 0.0529671, 0.06...
11   {'@': [-1.51883, 1.20023, -0.149934, 1.34665, ...
12   {'A': [-0.0442411, 0.0892976, 0.0495418, -0.01...
13   {'S': [-5.05913, 2.50813, 0.327637, 3.41255, -...
14   {'@': [-0.0704075, -0.0324585, 0.0529671, 0.06...
15   {'W': [-0.103236, -0.00874031, -0.025626, 0.05...
16   {'H': [-0.0974271, 0.00968351, 0.00799207, -0...
17   {'@': [-4.12516, 2.6672, 0.109932, 3.14893, -1...
18   {'@': [-0.0704075, -0.0324585, 0.0529671, 0.06...
19   {'@': [-0.0704075, -0.0324585, 0.0529671, 0.06...
20   {'@': [-0.0704075, -0.0324585, 0.0529671, 0.06...
21   {'@': [-0.0704075, -0.0324585, 0.0529671, 0.06...
22   {'T': [-0.0968081, 0.058163, 0.0107348, 0.0030...
23   {'T': [-0.0968081, 0.058163, 0.0107348, 0.0030...
24   {'C': [-0.0568754, -0.029026, 0.0464809, 0.047...
25   {'D': [-0.0581674, 0.0520085, 0.0708274, 0.014...
26   {'Q': [0.0296337, 0.0358552, 0.020629, 0.08382...
27   {'S': [-0.00980741, 0.0246739, 0.0388746, 0.01...
28   {'A': [-0.0442411, 0.0892976, 0.0495418, -0.01...
29   {'@': [-0.0704075, -0.0324585, 0.0529671, 0.06...
..                                                 ...
470  {'C': [-0.0568754, -0.029026, 0.0464809, 0.047...
471  {'P': [-6.17016, 3.79473, 0.15815, 4.86407, -2...
```

```
472  {'T': [-2.82877, 1.41027, 0.196507, 2.56148, -...
473  {'@': [-0.0704075, -0.0324585, 0.0529671, 0.06...
474  {'b': [-1.47326, 0.59906, -0.0112458, 1.14946,...
475  {'L': [-0.0862547, 0.0434853, 0.0529697, 0.047...
476  {'W': [-1.37132, 0.705387, 0.223342, 1.17413, ...
477  {'L': [-2.24839, 1.56547, -0.13066, 1.05811, -...
478  {'@': [-0.0704075, -0.0324585, 0.0529671, 0.06...
479  {'W': [-0.103236, -0.00874031, -0.025626, 0.05...
480  {'L': [-0.0862547, 0.0434853, 0.0529697, 0.047...
481  {'H': [-2.84192, 1.30551, 0.00290061, 2.25281,...
482  {'T': [-0.0968081, 0.058163, 0.0107348, 0.0030...
483  {'C': [-0.0568754, -0.029026, 0.0464809, 0.047...
484  {'W': [-0.103236, -0.00874031, -0.025626, 0.05...
485  {'@': [-6.10975, 3.18388, 0.113092, 5.31015, -...
486  {'B': [-0.0381987, -0.0201446, 0.0672753, 0.02...
487  {'A': [-0.0442411, 0.0892976, 0.0495418, -0.01...
488  {'T': [-0.0968081, 0.058163, 0.0107348, 0.0030...
489  {'C': [-2.74508, 1.27311, 0.4357, 2.20195, -1...
490  {'T': [-0.786056, 0.546372, 0.436922, 0.179107...
491  {'T': [-1.62455, 0.848571, -0.0730899, 0.86711...
492  {'@': [-0.903234, 0.542758, 0.196501, 1.07581,...
493  {'m': [-6.68876, 3.74138, -0.367255, 5.3879, -...
494  {'W': [-0.819099, 0.252387, -0.0837383, 0.3442...
495  {'@': [-0.851929, 0.583582, 0.19988, 0.518904,...
496  {'H': [-0.0974271, 0.00968351, 0.00799207, -0...
497  {'@': [-1.40238, 0.813444, 0.139089, 1.19777, ...
498  {'@': [-1.07709, 0.770572, 0.0888662, 1.01179,...
499  {'T': [-0.0968081, 0.058163, 0.0107348, 0.0030...

[500 rows x 3 columns]
```

```
In [8]: tweet_texts_processed = [str.join(" ", x) for x in df['text_processed']] # list of pre

        # We use the TfidfVectorizer from sklearn which is a feature extraction technique whic
        # a vector representation of the text.
        vectorizer = TfidfVectorizer(min_df=4, max_features = 5000)
        vz = vectorizer.fit_transform(tweet_texts_processed)
```

```
In [13]: vz.tocsr()
```

```
Out[13]: <500x202 sparse matrix of type '<class 'numpy.float64'>'
                 with 2019 stored elements in Compressed Sparse Row format>
```

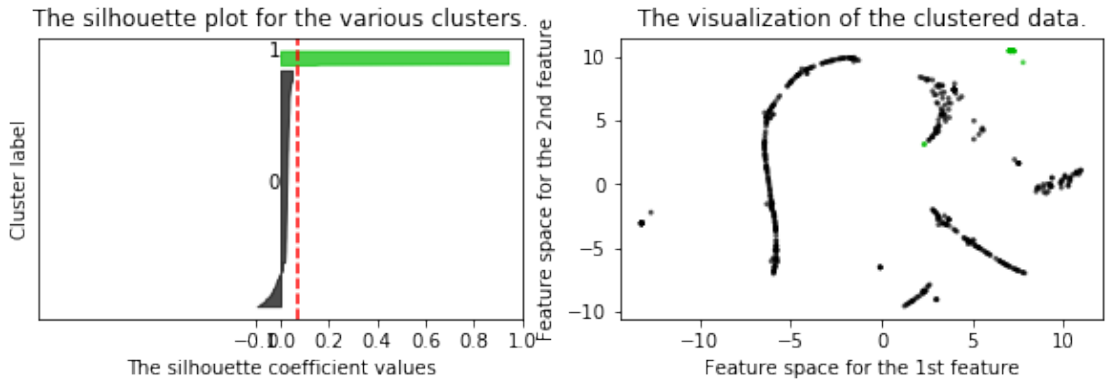
```
In [14]: kmeans = MiniBatchKMeans(n_clusters=3, init='k-means++', n_init=1,
                                  init_size=1000, batch_size=1000, verbose=False, max_iter=100
         kmeans_clusters = kmeans.fit_predict(vz)
         kmeans_distances = kmeans.fit_transform(vz)
```

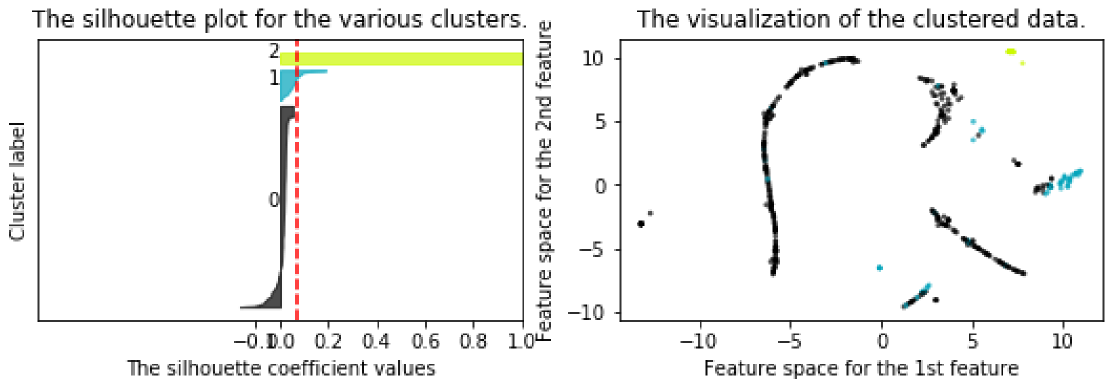
```
In [15]: kmeans_distances.shape
```

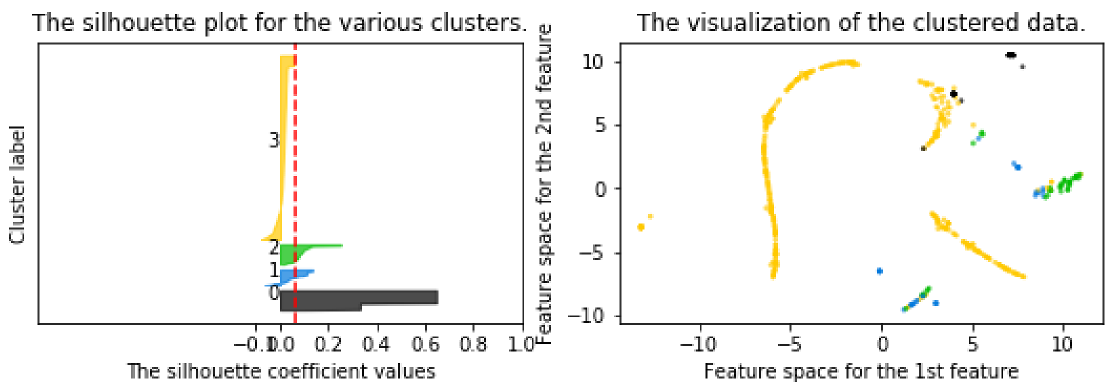
```
Out[15]: (500, 3)

In [16]: X_tsne = TSNE(learning_rate=200, perplexity=50, random_state=10).fit_transform(kmeans_
         X_tsne.shape

Out[16]: (500, 2)

In [17]: # We define diferent number of clusters

         range_n_clusters = [2, 3, 4, 6, 8, 12]

In [18]: for n_clusters in range_n_clusters:

             fig, (ax1, ax2) = plt.subplots(1, 2)
             fig.set_size_inches(10, 3)
             ax1.set_xlim([-1, 1]) # rango de silhouette de -1 a 1

             # Generate a cluster using the Mini Batch K-Means algorithm for diferent numbers
             kmeans = MiniBatchKMeans(n_clusters = n_clusters, random_state=10)
             kmeans_clusters = kmeans.fit_predict(vz)

             # Apply the dimensionality reduction to obatin the 2D representation of the point
             X_tsne = TSNE(learning_rate=200, perplexity=50, random_state=10).fit_transform(kme

             # Calculate the silhouette value of the cluster
             silhouette_avg = metrics.silhouette_score(vz, kmeans_clusters)

             print("For n_clusters =", n_clusters, "The average silhouette_score is :", silhoue

             sample_silhouette_values = metrics.silhouette_samples(vz, kmeans_clusters)

             y_lower = 10
             for i in range(n_clusters):
                 # Aggregate the silhouette scores for samples belonging to
                 # cluster i, and sort them
                 ith_cluster_silhouette_values = sample_silhouette_values[kmeans_clusters == i]

                 ith_cluster_silhouette_values.sort()

                 size_cluster_i = ith_cluster_silhouette_values.shape[0]
                 y_upper = y_lower + size_cluster_i

                 color = cm.spectral(float(i) / n_clusters)
                 ax1.fill_betweenx(np.arange(y_lower, y_upper), 0, ith_cluster_silhouette_value

                 # Label the silhouette plots with their cluster numbers at the middle
                 ax1.text(-0.05, y_lower + 0.5 * size_cluster_i, str(i))

                 # Compute the new y_lower for next plot
```

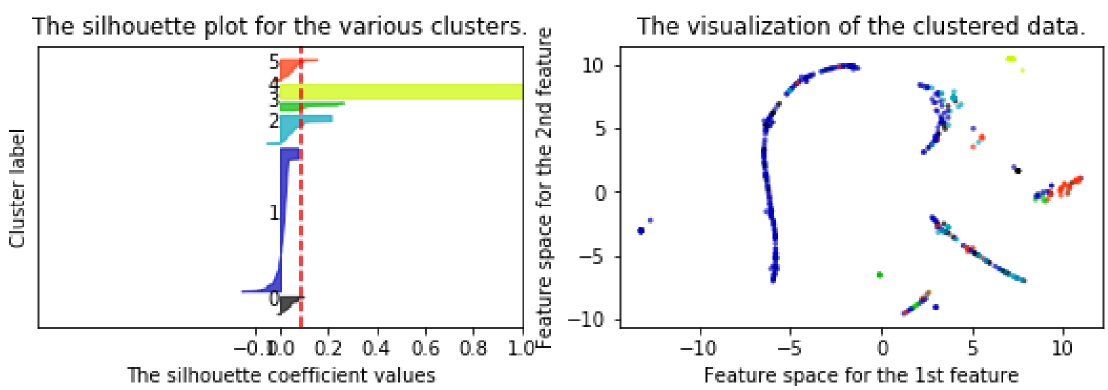
```
            y_lower = y_upper + 10  # 10 for the 0 samples
        ax1.set_title("The silhouette plot for the various clusters.")
        ax1.set_xlabel("The silhouette coefficient values")
        ax1.set_ylabel("Cluster label")

        # The vertical line for average silhouette score of all the values
        ax1.axvline(x=silhouette_avg, color="red", linestyle="--")

        ax1.set_yticks([])  # Clear the yaxis labels / ticks
        ax1.set_xticks([-0.1, 0, 0.2, 0.4, 0.6, 0.8, 1])

        # 2nd Plot showing the actual clusters formed
        colors = cm.spectral(kmeans_clusters.astype(float) / n_clusters)
        ax2.scatter(X_tsne[:, 0], X_tsne[:, 1], marker='.', s=30, lw=0, alpha=0.7, c=colo

        ax2.set_title("The visualization of the clustered data.")
        ax2.set_xlabel("Feature space for the 1st feature")
        ax2.set_ylabel("Feature space for the 2nd feature")

        ax2.figure.subplots_adjust(bottom=0.2)
        ax2.figure.savefig('./figures/tfidf_500_cluster%d_%f'% (n_clusters, silhouette_av

    plt.show()
```

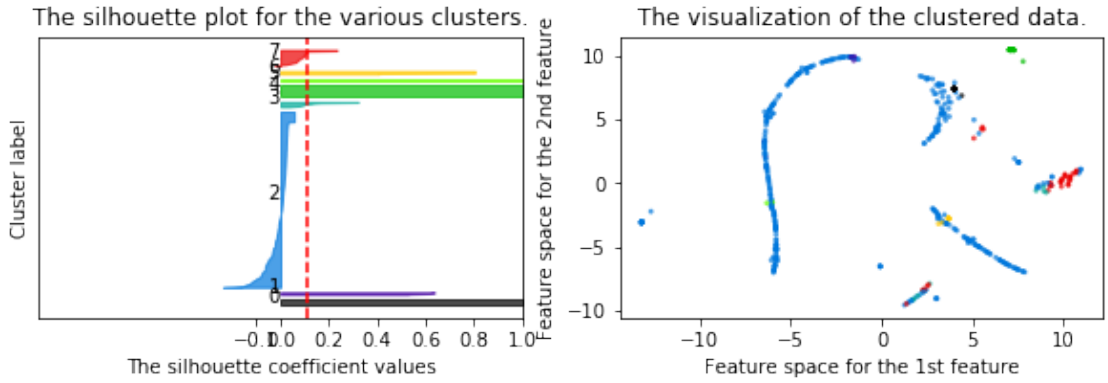
```
For n_clusters = 2 The average silhouette_score is : 0.0677904917774
For n_clusters = 3 The average silhouette_score is : 0.0698147438029
For n_clusters = 4 The average silhouette_score is : 0.0658527609612
For n_clusters = 6 The average silhouette_score is : 0.0834111961797
For n_clusters = 8 The average silhouette_score is : 0.112094563449
For n_clusters = 12 The average silhouette_score is : 0.105279187109
```

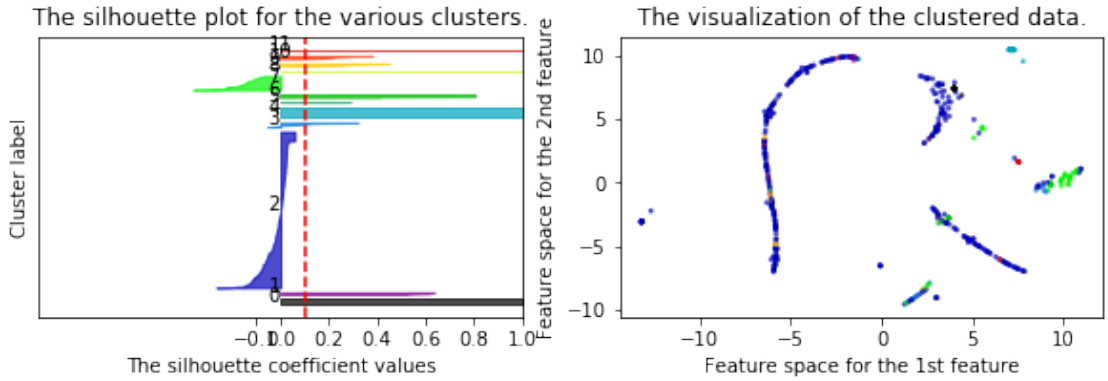



The silhouette plot for the various clusters.
Cluster label
The silhouette coefficient values
−0.1 0.0 0.2 0.4 0.6 0.8 1.0
The visualization of the clustered data.
Feature space for the 2nd feature
Feature space for the 1st feature
−10 −5 0 5 10


The silhouette plot for the various clusters.
Cluster label
The silhouette coefficient values
−0.1 0.0 0.2 0.4 0.6 0.8 1.0
The visualization of the clustered data.
Feature space for the 2nd feature
Feature space for the 1st feature
−10 −5 0 5 10


The silhouette plot for the various clusters.
Cluster label
The silhouette coefficient values
−0.1 0.0 0.2 0.4 0.6 0.8 1.0
The visualization of the clustered data.
Feature space for the 2nd feature
Feature space for the 1st feature
−10 −5 0 5 10


The silhouette plot for the various clusters.
Cluster label
The silhouette coefficient values
The visualization of the clustered data.
Feature space for the 2nd feature
Feature space for the 1st feature


The silhouette plot for the various clusters.
Cluster label
The silhouette coefficient values
The visualization of the clustered data.
Feature space for the 2nd feature
Feature space for the 1st feature

In [ ]:

In [ ]: