**Clustering of text documents(tweets)**

Yiheng Pan†  
 Biomedical and Health Informatics  
 Case Western Reserve University  
 Cleveland, OH, USA  
 yxp273@case.edu

Wei Du  
 Computer and Data Science  
Case Western Reserve University  
 Cleveland,OH, USA  
 wxd163@case.edu

Minyang Tie  
 Computer and Data Science  
 Case Western Reserve University  
 Cleveland, OH, USA  
 mxt497@case.edu

**1 ABSTRACT**

This project develops three main clustering algorithms on tweets from health news accounts of new organizations. We define the feature matrix by Bag-Of-Words approach, and use two different distance measures. After clustering, for each method and distance combinations, we select two prominent clusters and identify their meanings.

**2 DATA**

The dataset is CNN health news from Twitter. It was collected from August 2011 to December 2014.

**3 METHOD**

3.1**Bag of Words**

Bag of Words (BOW) is a method to extract features from text documents. These features can be used for training machine learning algorithms. It creates a vocabulary of all the unique words occurring in all the documents in the training set. In simple terms, it’s a collection of words to represent a sentence with word count and mostly disregarding the order in which they appear.

3.2**Latent Dirichlet Allocation**

Latent Dirichlet Allocation (LDA) is one of the topic modeling approaches. LDA is a generative probabilistic model based on a three-level hierarchical Bayesian model. LDA assumes that documents contain latent topics and each topic can be represented by distribution across words.

3.3**Clustering Algorithm**

3.3.1**K-means Clustering**

To process the learning data, the K-means algorithm in data mining starts with the first group of randomly selected centroids, which are used as the beginning points for every cluster, and then performs iterative calculations to optimize the positions of the centroids.

3.3.2**Birch**

Balanced Iterative Reducing and Clustering using Hierarchies is suitable for large amounts of data and high k value. It can cluster the data by only one time of scan. It starts by clustruct the B+ tree, and check whether two data points belong to one Clustering Feature Tree by the threshold, then further decide whether to split a CF tree by branching factor.

3.3.3**Hierarchical Clustering**

Hierarchical clustering starts by treating each observation as a separate cluster. Then it repeatedly executes the following two steps: (1) identify the two clusters that are closest together, and (2) merge the two most similar clusters. This continues until all the clusters are merged together.

**4 RESULTS**

4.1**Dataset Feature Matrix**

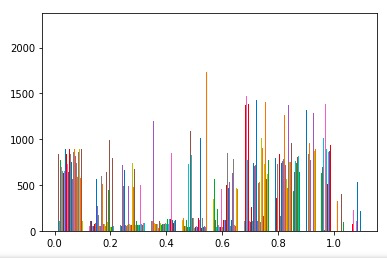
After feature extraction, we can see the summary result below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Dataset | #Doc | #Tokens | #Unique Tokens | #Avg/ Doc | Description |
| Cnn health | 4061 | 76295 | 13323 | 18.8 | medical tweets |

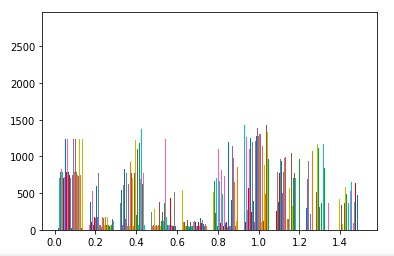
Table 1. Summary results of the feature matrix

4.2**Distance Distribution**

To access the distance distribution, we plot the histograms of each distance. We select Euclidean distance and Manhattan distance as similarity measurement.



Euclidean distribution



Manhattan distribution

Euclidean distance: the square roots of the sum of difference between the coordinates of the two points.

Manhattan distance: the sum of absolute differences of their Cartesian coordinates.

The reason we choose these two distance is that they are easy to interpret and calculate. They are good at time cost and future understanding.

4.3**Parameter Setting**

Each clustering algorithm has an individual parameter setting. We use the greedy search method and find optimal parameters with the highest silhouette score.

4.3.1 **K-means**

Euclidean distance: when topic size = 5, cluster size = 10, we have the silhouette score with 0.773.

Manhattan distance: when topic size = 5, cluster size = 10, we have the silhouette score with 0.751.

4.3.2 **Birch**

Euclidean distance: when topic size = 5, n\_cluster = 10, threshold = 0.5, we have the silhouette score 0.754, calinski\_harabasz\_score is 14121.98

Manhattan distance: when topic size = 5, cluster size = 10, threshold = 0.5 ,we have the silhouette score with 0.778, calinski\_harabasz\_score is 9814.266

The reason to choose threshold = 0.5 is because the samples are processed by LDA algorithm so the variance of the sample is not too large, so default 0.5 is token. The branching factor is chosen to default is also because the number of the samples are not too large.

4.3.3 **Hierarchical Clustering**

Euclidean distance: when topic size = 5, n\_cluster = 13, linkage = ward, we have the silhouette score with around 0.807.

Manhattan distance: when topic size = 5, n\_cluster = 13, linkage = average, we have the silhouette score with around 0.792.

4.4**Clustering Summary Result**

Kmeans-Euclidean: num\_cluster = 10, cluster distribution: 3: 922,1: 670, 6: 736, 2: 703, 7: 760, 4: 89, 9: 37, 5: 38, 8: 35, 0: 71, entropy: 11.7

Kmeans-Manhattan: num\_cluster = 10, cluster distribution: 0: 850, 3: 830, 1: 621, 4: 701, 7: 790, 9: 32, 8: 37, 5: 73, 2: 66, 6: 61, entropy: 11.52

Birch-Euclidean: num\_cluster = 10,cluster\_distribution: 3: 942, 8: 815, 7: 790, 4: 731, 0: 503, 1: 75, 2: 73, 5: 55, 6: 48, 9: 29 Entropy: 11.49

Birch-Manhattan: num\_cluster = 10, cluster\_distribution: 8: 930, 2: 802, 0: 746, 1: 665, 3: 605, 5: 79, 9: 68, 4: 59, 7: 57, 6: 50 Entropy:11.399

Hierarchical-Euclidean: num\_cluster = 13, cluster distribution: 2: 974, 7: 649, 11: 648, 3: 644, 9: 604, 0: 102, 1: 83, 6: 82, 10: 78, 5: 64, 4: 52, 8: 43, 12: 38, entropy: 11.7

Hierarchical-Manhattan: num\_cluster = 13, cluster distribution: 6: 802, 1: 790, 10: 782, 0: 724, 8: 691, 2: 49, 3: 46, 5: 43, 9: 40, 7: 37, 4: 35, 12: 21, 11: 1, entropy: 11.3

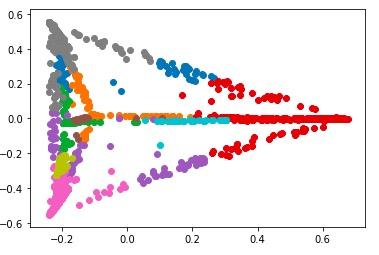
Purity between Birch Euclidean and Hierarchical Euclidean:0.000584

Purity between Birch Manhattan and Hierarchical Manhattan: 0.000698

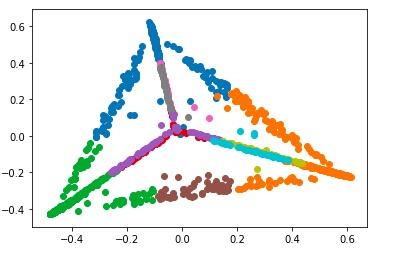
4.5**Visualization Analysis**

Principal component analysis(PCA) is a fast and flexible unsupervised method for dimensionality reduction in data. Its behavior is easiest to visualize by looking at a two-dimensional dataset.

4.5.1 **K-means**

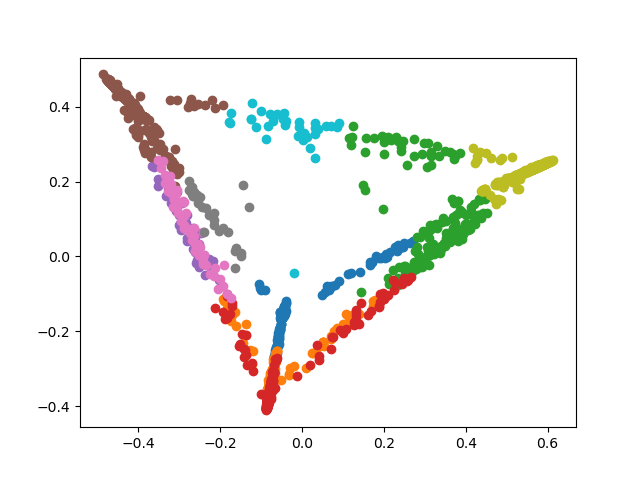
****

Euclidean distance

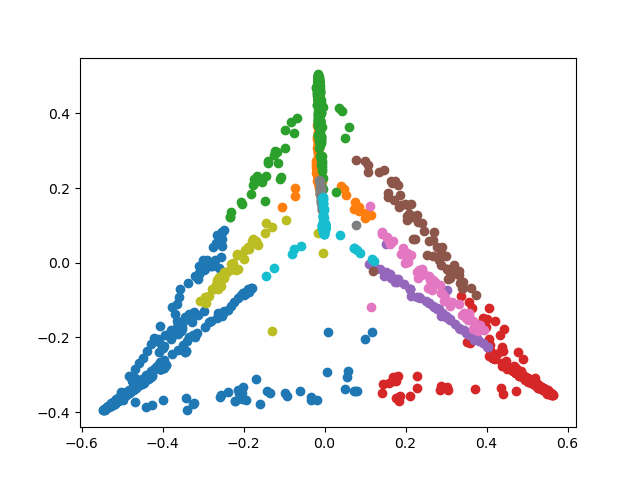
****

Manhattan distance

4.5.2 **Birch**

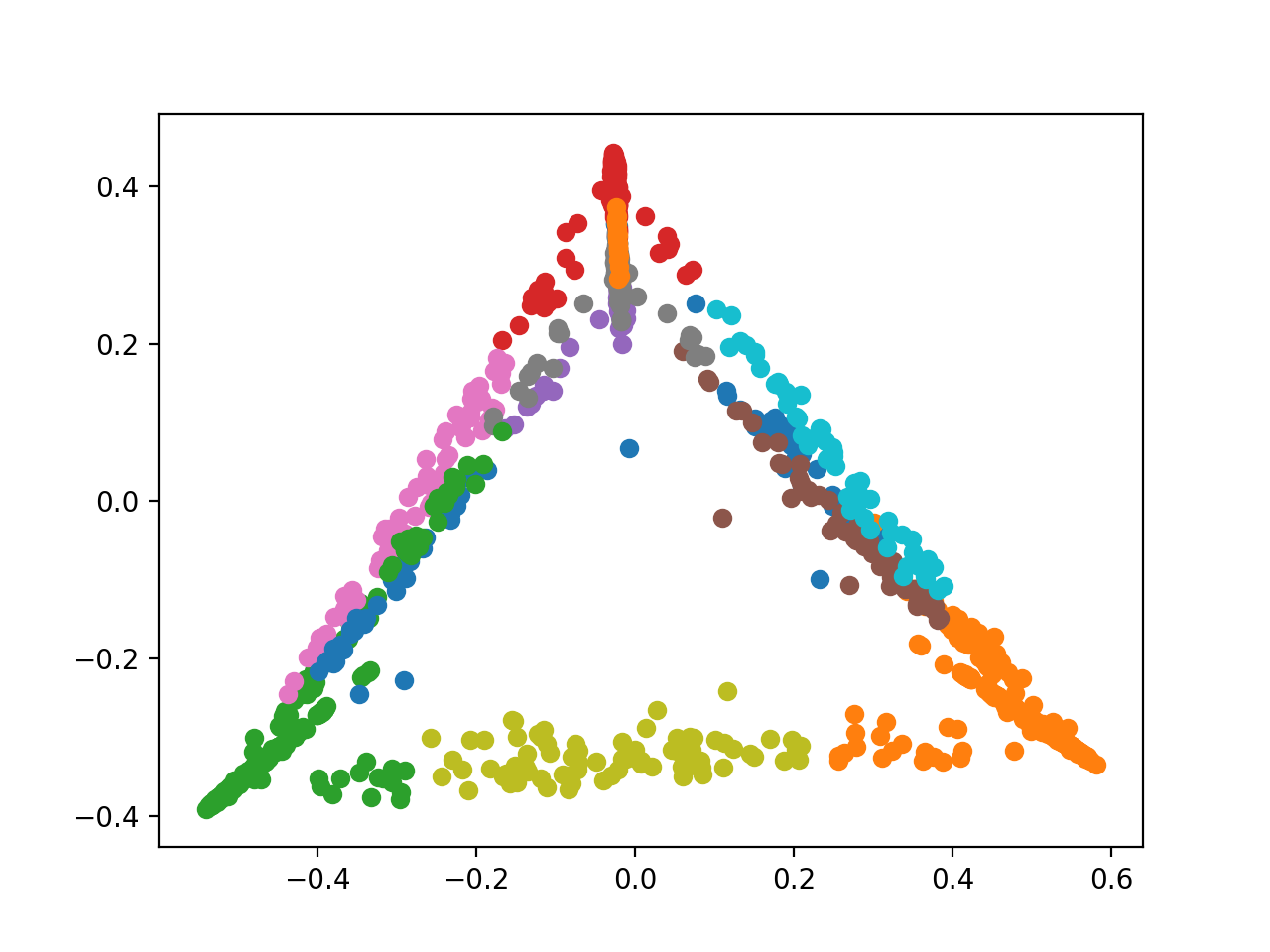
****

Euclidean distance

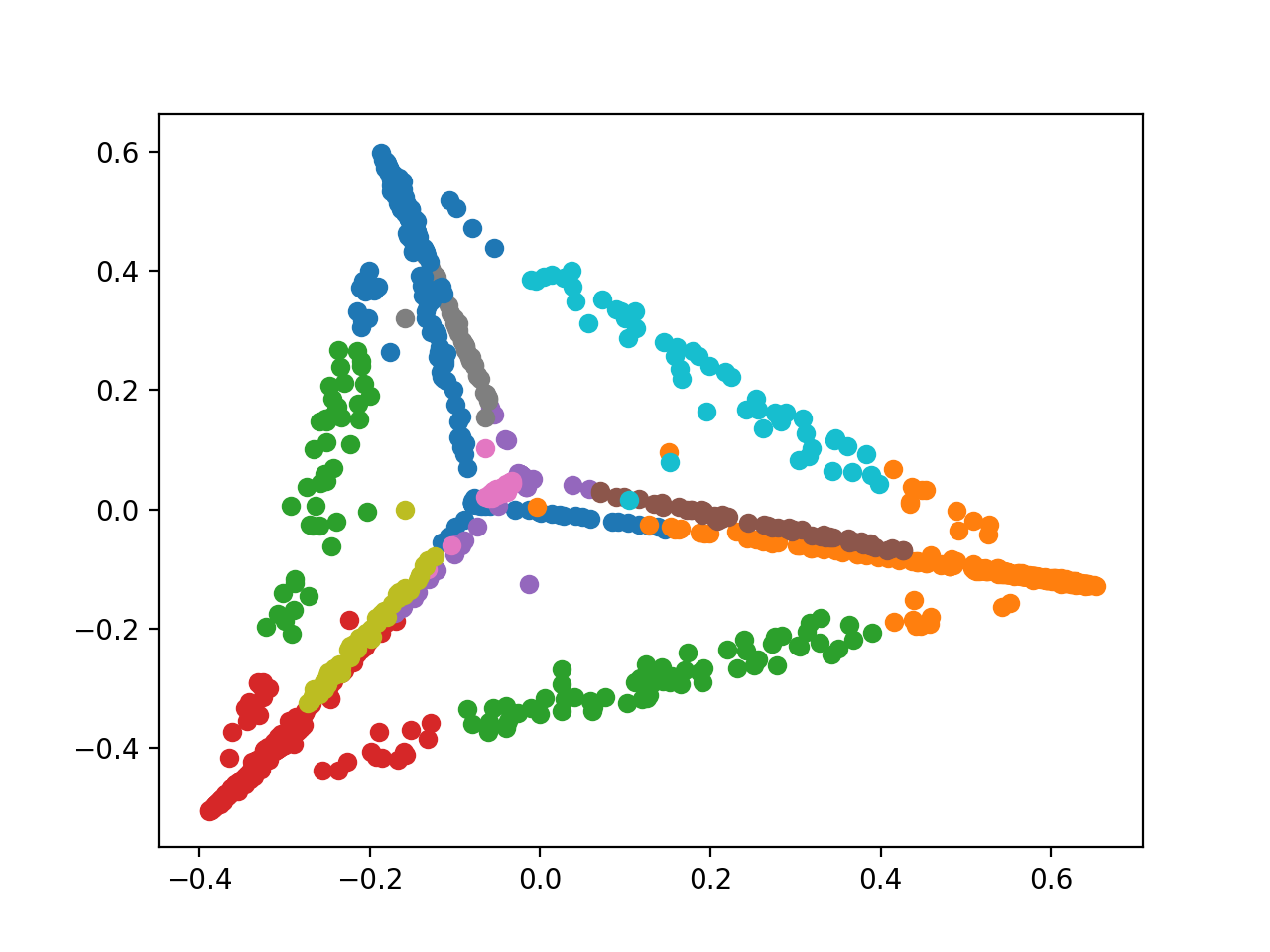


Manhattan distance

4.5.3 **Hierarchical Clustering**



Euclidean distance



Manhattan distance

4.6**Clustering Interpretation**

We look into the most prominent two clusters by average feature scores and corresponding top word features.

4.6.1 **K-means**

|  |  |
| --- | --- |
| T1: Euclidean best | ['you','via','rt','thing','ar','diet','tip','todai', 'ill','know','like','come','food','realli','less','remind','watch','want','readi', 'surround'] |
| T2:  Euclidean 2nd best | ['you','rt','130pm','help','est','todai','mai','twitter','chat','word','sai','we','join','us','think','make','allergi','expect','blood','thi'] |
| T3:  Manhattan best | ['know','thing','keep','brain','danger','viru','cancer','you','stress','fat','tip','lack','make','why','live','amp','we','prescript','home', 'dose'] |
| T4:  Manhattan 2nd best | ['doctor','we','medic','help','rt','ask','decad','share','mai','now','person','fight','nearli','thi','new','last','race','test','you', 'creat'] |

T1 is about food and diet’s association with illness. T2 is about allergic influence in blood. T3 is about brain cancer might be associated with stress and fat. T4 is about doctors give the medication.

4.6.2 **Birch**

|  |  |
| --- | --- |
| T1: Euclidean best | ['you', 'rt', 'ar', 'readi', 'need', 'man', 'workout', 'us', 'tip', 'todai', 'marathon', 'rule', 'amp', 'check', 'surviv', 'alarm', 'now', 'home', 'diseas', 'pound'] |
| T2:  Euclidean 2nd best | ['rt', 'fight', 'someon', 'worri', 'cancer', 'lung', 'tackl', 'bed', 'my', 'second', 'let', 'treatment', 'kid', 'world', 'grappl', 'diseas', 'approv', 'shouldnt', 'coach', 'real'] |
| T3:  Manhattan best | ['outbreak', 'us', 'brain', 'measl', 'cancer', 'grow', 'state', 'decai', 'pinpoint', 'ie', 'why', '12', 'warn', 'rt', 'men', 'help', 'nose', 'med', 'safe', 'kind'] |
| T4:  Manhattan 2nd best | ['rt', 'you', 'eat', 'cancer', 'healthi', 'todai', 'we', 'fit', 'research', 'ar', '10', 'dai', 'make', 'tip', 'weight', 'know', 'sai', 'thing', 'want', 'breast'] |

T1 is about the marathon and disease. T2 is about cancer and disease. T3 is about the safe nose. T4 is about eat, health and cancer

4.6.3 **Hierarchical Clustering**

|  |  |
| --- | --- |
| T1: Euclidean best | ['you', 'eat', 'follow', 'less', 'tip', '11', 'loss', 'mai', 'realli', 'rt', 'bad', 'ar', 'todai', 'weight', 'drop', 'watch', 'us', 'dont', 'drink', 'healthi'] |
| T2: Euclidean 2nd best | ['superfood', 'you', 'princeton', 'mening', 'rt', 'ar', 'flu', 'vaccin', 'want', 'closer', 'job', 'toddler', 'world', 'dai', 'eat', 'cold', 'thing', 'calori', 'student', 'amaz'] |
| T3:Manhattan best | ['you', 'hi', 'fat', 'correspond', '230', 'kid', 'us', 'ar', 'obes', 'amp', 'don', 'hand', 'law', 'festiv', 'et', 'light', 'chat', 'coke', 'medic', 'favorit'] |
| T4: Manhattan 2nd best | ['you', 'chat', 'love', 'sai', 'us', 'mom', 'mental', 'join', 'new', 'report', 'woman', 'pressur', 'rt', 'send', 'fat', 'ill', 'allergi', 'wed', 'blood', 'idea'] |

T1 is about the relationship between diet and health. T2 is about the flu vaccine. T3 is about obesity. T4 is the relationship between pressure and healthy situations (ill, fat, mental, etc.)

All team members agree with the same effort