Big Data with ML systems HW1b

NetID: yl5090

**Question 3: Write-up**

1. The elbow plot :

**When length of signature = 5**



As seen from the plot, 10 seems to be a reasonable optimal number of clusters.

**When length of signature = 50**



Seems n = 18 might be a good choice as it is a local mimum.

**When length of signature = 100**



It seems choose 22 clusters, might be the optimal choice. Because it is a local minimum.

1. Flow of computation in my code and assumptions

There are three main steps: text processing (min-Hash), k-means, and intra-cluster variance computation.

Design of Min-hash:

* First convert text document to words and remove punctuation.
* Then, convert words to n-grams (shingles). Here I choose a trigram -- the idea comes from the MMDS book, while for convenience I did not select stop words.
* Then, use CRC32 hashing to hash the n-gram (shingles) to integers. The reason is, although we could use Python dictionaries to perform the shingle hashing, for large number of documents and shingles this is slow.
* Then the min-hash algorithm, which takes a 32-bit integer and hash it to a different integer with no collisions.
* The min-hash function is ../../../../random-hash-eq.png. a and b are random coefficients generated by Numpy, c is a prime that is larger than the largest shingle ID. The reason for c to be a prime is that in this case the random hash will not have collisions.
* Size of signature is 100, as recommended by the homework.

Implementation of Min-Hash:

* It is performed on Spark.
* The input to this phrase is a csv file. The output is an RDD object, with element being [Heading, (shingle\_tuple)]. I use a tuple here because later reduceByKey needs immutable object as key.
* This is created by reading the whole csv file into memory in Python, and then parallelize to Spark. The reason I’m not using the Spark csv reader, is the csv file has a field containing ‘/n’ which creates trouble in splitting. The python csv reader works.
* Each of the above step in the design of min-Hash is performed use a RDD.map()

K-means:

* Two main steps:
  + for input data, compute distances with all centroids and select the closest one
  + for centroids, compute intra-cluster mean and update as new centroids
* Selection of K: use a list of range (0, 20) and produce the elbow chart – total intra-cluster SSE vs selection of k. See from the graph above, after
* The final output is a labeled dataset represented as an RDD object. Each element being the format [(centroid), [heading, shingle\_set]]

Compute total intra-cluster SSE and plotting :

* For each point in RDD, use a map function to produce [1, error-to-centroid]. The error-to-centroid is the square of distance of this point to its clustered centroid.
* Then use a reduce function to sum over all errors and get total SSE.

1. Performance of my model

Signature:

1. signature = 50 produces SSE at 1E20, same scale as signature = 100. However signature = 100, when cluster = 22 the SSE is smaller than the SSE for all cluster in signature = 50. Thus, signature 100 is better than signature = 50
2. Compare signature = 5 and signature = 100. When signature = 100, the error is 10 times more than signature = 5. Considering the length of vector is also in the scale of ten, the performance is hard to compare numerically.
3. Now manually check examples for signature = 5 and signature = 100.

Bad examples

The clusters seem to perform quite good. The headings are sometimes incomplete.

This is for cluster 5, looks like about oil, price, international politics, and financial market. But with some bad examples, highlighted in yellow.

['oil prices fall in new yor', 'budget to be presented on june 5 finance minister', 'oil prices up in asian trad', 'despite govt announcement cng stations yet to reopen in punjab', 'seven power projects to reach financial close by year end minister', 'fbr runs rs63 bn short of tax collection targ', 'pakistan korea ink 500 million framework agr', 'Govt hikes prices of petrol di', 'Fresh milk prices increased by Rs5 per litr', 'Gold eyes worst week since Mar', 'Pakistan keen to expand trade and economic cooperation with Portugal NA Speaker', 'Pakistan sell 40 percent stock exchange end 2016early 2017', 'US oil prices end above 50', 'Pakistan drill 116 Oil Gas wells next fiscal year', 'POL prices to remain unchanged for July', 'Government decides against raising petroleum pri', 'PSO received Rs 293 bln payment against Rs 308 bln supplies16 ', 'PSO received Rs 293 bln payment against Rs 308 bln supplies16 ', 'CNG supply resumes Sindh three day long stri', 'PSO received Rs 293 bln payment against Rs 308 bln supplies16 ', 'PSO received Rs 293 bln payment against Rs 308 bln supplies16 ', 'PSO received Rs 293 bln payment against Rs 308 bln supplies16 ', 'CNG supply resumes Sindh three day long stri', 'Coming soon Pakistans biggest online shopping festiv', 'Pakistans biggest online shopping festival begins Wednesday', 'Few hours left for Pakistans biggest online shopping festival to begi', 'England turn to Croft for spin tips in South Afri', 'Judo team for South Asian Games announced', 'National Ranking Tennis from January 18', 'Williamson Guptill record stand helps NZ to level seri', 'Australia bans womens cricketer for betting on ', 'Venus vanquishes Doi to claim inaugural Taiwan ti', 'Lahore set PSL record target for Qu', 'PSL Sharjeels ton takes Islamabad to 176 3 score against Peshawar', 'Bangladesh put India into bat in Asia Cup opener', 'Islamabad United become first PSL champi', 'Former New Zealand captain Crowe dies aged 53', 'World T20 Pakistan security team to visit India on Monday', 'Afghanistan win toss bat against Scotland', 'Afghanistan amass 186 6 against Zimbab', 'Tamim ton helps Bangladesh set 181 for O', 'Sri Lankas Malinga ruled out of World T20', 'Kohli guides India to 6 wkt win against Paki', 'Windies record easy 7 wkt win over Sri L', 'Womens World T20 NZealands comfortable win over Aussi', 'Bangladesh urge ICC to lift Taskin b', 'Bangladesh restrict India to 146 7 in World T20', 'West Indies beat South Africa to reach semi', 'Sri Lanka win toss bowl against England', 'West Indies win toss bowl against Afghani', 'Pakistan beat Sri Lanka in 1st match of Blind cricket seri', 'Sarfraz Ahmed named as Pakistans T20 captai', 'Knee issue rules Lasith Malinga out of IPL', 'Herath retires from ODI T20 International cri', 'Determined Yuvraj sets 2019 World Cup g', 'Younis Khan decides to rejoin Pakistan Cu', 'Ahmed Shehzad Pakistan Cricket Younis K', '7 Eleven worker first woman climb Everest seven ti', 'Sri Lanka Kulasekara quits tests focus ODIs T20', 'Ronaldo leads Portugal squad for Euro 2016', 'Heather Knight leads England womens team challenge Paki', 'Pakistan 17 member squad England tour', 'Mohammad Amir receives UKs traveling document Pakistan Cri', 'Barcelona expect Neymar to stay at Nou C', 'Motorcycling Dovizioso tops practice times in A', 'Maradona urges Messi to stay with Argenti', 'Pak India cricket very important says Salahuddi', 'Pistorius defence team say will not appeal six year ', 'Czech veteran Cech announces international retir', 'Pakistan team black arm bands mourning Edhis d', 'Mourinho hopes United players will forge close bond in Chi', 'Yasir Shah six gives Pakistan 67 run lead England', 'Shaharyar Khan salutes restorative Pakistan triu', 'Barcelona pair Munir and Samper sign new d', 'India Ashwin regains number one spot test ranking', 'Taylor helps New Zealand build massive lead', 'IOC chief urges total review of world anti doping sy', 'Erdogan wants trade with Russia China in local currenci', 'China forex reserves fall by 320 bn in 2016', 'Worlds longest flight lands in New Zealand', 'New Development Bank plans joint investments in economies along Belt Road']

1. Reference :

Chris McCormick’s blog on MinHash:

http://mccormickml.com/2015/06/12/minhash-tutorial-with-python-code/