Lab exercise for the web mining lab on 19th September 2018

Part A:

You are given 9 one-line documents here. Consider the following keywords to represent the documents in the vector space model:

[1] Automotive [2] Car [3] motorcycles [4] self-drive [5] IoT [6] hire   [7] Dhoni
Represent the documents in vector space Model using these keywords and use it as input to cluster the documents using Manhattan distance as parameter. Ignore case differences.

You also need to do K- means clustering with K=4.

Documents for use in question

Doc1

Electric automotive maker Tesla Inc. is likely to introduce its products in India sometime in the summer of 2017.

Doc 2

Automotive major Mahindra likely to introduce driverless cars

Doc 3

BMW plans to introduce its own motorcycles in india

Doc 4

Just drive, a self-drive car rental firm uses smart vehicle technology based on IoT

Doc 5

Automotive industry going to hire thousands in 2018

Doc 6

Famous cricket player Dhoni brought his priced car Hummer which is an SUV

Doc 7

 Dhoni led india to its second world cup victory

Doc 8

IoT in cars will lead to more safety and make driverless vehicle revolution possible

Doc 9

Sachin recommended Dhoni for the indian skipper post

Part B

Use the same program which you have developed for part A to do " K-means clustering" of the following web documents means with K=4. Use the keywords.

[1] Tesla [2] Electric [3] Car/Vehicle/Automobile [4] pollution [5] de-monetisation [6] GST [7] black money

Download the webpage into a .txt file [ignore images,tables and limit the size of the document to 500 words Max] and build your vector space model using Term frequency.

Ignore case differences. Treat singular and plural of nouns as same. Treat Car/vehicle/automobile as one word [synonyms]. Treat "black money" as a single word.

List of webpages:

[1] https://www.zigwheels.com/newcars/Tesla

[2] https://www.financialexpress.com/auto/car-news/mahindra-to-launch-indias-first-electric-suv-in-2019-all-new-e-verito-sedan-on-cards/1266853/

[3] https://en.wikipedia.org/wiki/Toyota_Prius

[4] https://economictimes.indiatimes.com/industry/auto/auto-news/government-plans-new-policy-to-promote-electric-vehicles/articleshow/65237123.cms

[5] https://indianexpress.com/article/india/india-news-india/demonetisation-hits-electric-vehicles-industry-society-of-manufacturers-of-electric-vehicles-4395104/

[6] https://www.livemint.com/Politics/ySbMKTIC4MINsz1btccBJO/How-demonetisation-affected-the-Indian-economy-in-10-charts.html

[7] https://www.hrblock.in/blog/impact-gst-automobile-industry-2/

[8] https://inc42.com/buzz/electric-vehicles-this-week-centre-reduces-gst-on-lithium-ion-batteries-hyundai-to-launch-electric-suv-in-india-and-more/

[9] https://www.youthkiawaaz.com/2017/12/impact-of-demonetisation-on-the-indian-economy/

[10] https://indianexpress.com/article/india/demonetisation-effects-cash-crisis-mobile-wallets-internet-banking-4406005/

[11] https://www.news18.com/news/business/how-gst-will-curb-tax-evasion-1446035.html

[12] https://economictimes.indiatimes.com/small-biz/policy-trends/is-gst-helping-the-indian-economy-for-the-better/articleshow/65319874.cms