Experiment 8

**Hierarchial Clustering**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Name:** | **:** | **Abhishek N N** | **Register No** | **:** | **20BCE1025** |
| Faculty | : | Dr. Alok Chauhan | Slot | : | L51-L52 AB1-605B |
| Course | : | Web Mining Lab | Code | : | CSE3024 |
| Programme | : | B.Tech CSE Core | Semester | : | Win – 22 - 23 |



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 Euclidean distance as parameter. Ignore case differences.

You need to do hierarchical clustering with single-link, complete-link, average-link agglomerative clustering and show corresponding dendrograms.

**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**

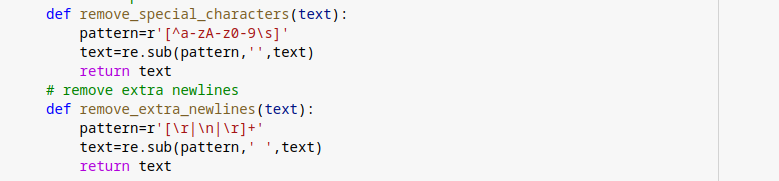
**Doc 8**

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

**Doc 9**

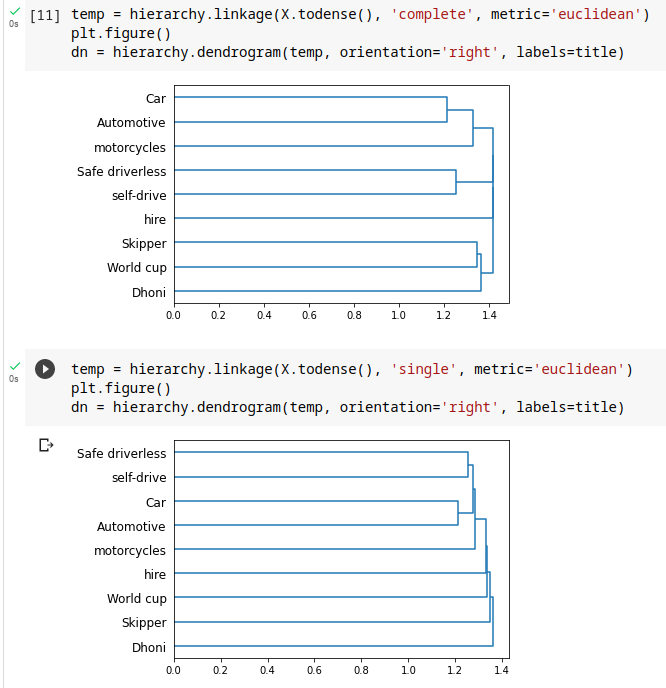
**Sachin recommended Dhoni for the indian skipper post**

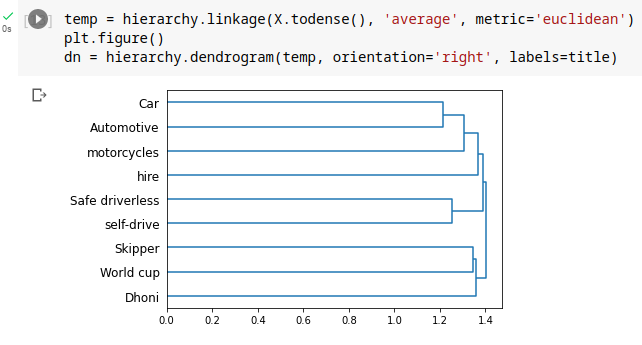












Part B

Use the same program which you have developed for part A to do “hierarchical clustering” of the following web documents. Use the keywords.

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

Download the webpage into a .txt file [ignore images, tables and limit the size of the document to 250 words Max] and build your vector space model using Tf-idf.

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.

