# CE/CZ4073 : Data Science for Business

2017-2018, Semester 2 | Nanyang Technological University, Singapore

## Assignment 3

Posted on 25 March 2018 · Clarify by 3 April 2018 · Submit by 8 April 2018

**Submission :** Submit a single `R` file for the assignment – `assign3_FullName.R` – containing all the steps and comments in connection with the exploratory data analysis and model building for both the problems, where `FullName` is your full name, as per the official records at NTU.

Your submission should be emailed to `sg.sourav@ntu.edu.sg` by midnight of 8 April 2018 from your official NTU Email Address (do not use your personal email address). Late submissions will be penalized with lower grades, and submissions after 12 April 2018 will not be graded.

### Problem 1 [10 points]

**Background :** Consider the attached dataset `assign3_CuisineData.json`, which stores the recipes of multiple dishes in JSON format, where the first element of each recipe is a unique identifier (`"id"`) and the second element is the list of ingredients (`"ingredients"`) of the dish:

```
{  "id": 10259,
   "ingredients": [ "romaine lettuce", "black olives", "grape tomatoes", "garlic",
                    "pepper", "purple onion", "seasoning", "garbanzo beans" ]  }
```

The target is to find the *optimal* number of clusters (cuisines) that you can spot in the dataset.

**Task :** Import the dataset in JSON, convert it into a suitable Document-Term Matrix (DTM), and perform clustering with appropriate choice of algorithm and distance notion to identify the *optimal* number of clusters in the dataset `assign3_CuisineData.json`. Briefly comment (within the code) on the choices you make in the process of finding the *optimal* number of clusters.

### Problem 2 [10 points]

**Background:** Every individual has a unique preference for music, movies, hobbies, and interests, sometimes related to their health habits, phobias, personality, lifestyle, spendings, and even opinions. The Kaggle dataset `https://www.kaggle.com/miroslavsabo/young-people-survey` documents the responses from a survey, connecting individual preferences to the demography.

**Task :** You are required to find the *optimal* number of clusters that you can spot in the Kaggle dataset. Briefly comment (within the code) on the choices you make in the process of finding the *optimal* number of clusters. Based on the clusters you observe, find the *strongest* clustering parameters in this set of people. Do you think that `"gender"` plays a major role in clustering?

---

*This is an individual assignment. Properly acknowledge every source of information that you refer to, including discussions with your fellow students, if any. Verbatim copy from any source is strongly discouraged, and plagiarism will be heavily penalized. It is strongly recommended that you write the codes completely on your own. Feel free to write the codes in Python if you want.*