**Database Management & Database Design**

**INFO6210**

**Assignment 01**



**Submitted To:**

**Prof Nick B. Brown**

**Submitted By:**

|  |  |
| --- | --- |
| **Name** | **NU ID** |
| Mansi Nagraj | 001427101 |
| Dhawal Priyadarshi | 001401831 |

**Abstract:**

The assignment aims to gather the data from three different sources including the website, Twitter (or Facebook or Instagram) and raw data from the csv files, auditing, transforming data and developing a conceptual model based on the relationship amongst the various attributes and entities involved. Finally, the related data is collected in a csv file.

**Dataset**:

* Real world data related to the various food groups particularly chocolates is gathered from the United States Dept. of Agriculture Research Service. The data scraped from this site includes the below mentioned data:  
  **1**. ***DB***: It denotes the two different databases: Branded Food Products abbreviated as BF and Standard Reference abbreviated as SR. This assignment is mainly centered towards the chocolates from different brands which are under the “BF” category

**2**. ***Ndb-Id***: It is a unique identification no. given to each group of food-items  
**3.** ***Food Description***: It includes the different combinations of food-items that have been grouped together for study by the Agricultural research services  
**4.** ***Food Group/Manufacturer:*** It lists the name of different manufacturers for the chocolates mentioned in the Food-description column

* Raw data files include the csv’s that contain the specifications for the various nutrients such as water, Vitamins, Carbohydrates etc. involved in the food-group
* The data from twitter demonstrates the created date and the popular tweets for particular chocolate brands

**Conceptual Data Model:**

The data has been categorized into three different entities with some attributes in each.  
Below listed are the different entities and attributes present in the chocolates database:

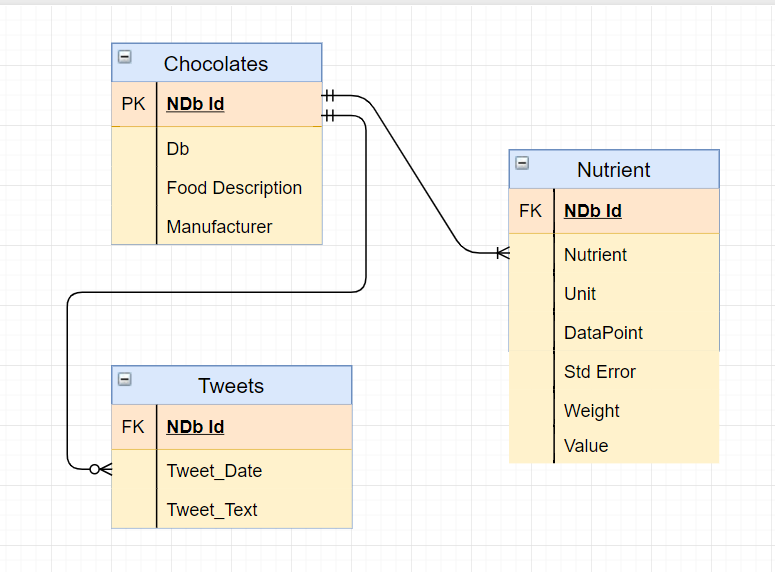
***Entity name:*** Chocolates  
***Attributes involved:*** Db, NDb Id, Food Description, Manufacturer

***Entity name:*** Nutrient  
***Attributes involved:*** NDb Id, Nutrient, Unit, DataPoint, StdError, Weight, Value

***Entity name:*** Tweets  
***Attributes involved:*** Tweet\_Date, Tweet\_text, NDb Id

***Relationship between different entities:***

Each entry in the chocolates entity will have a unique NDb-Id.   
For each Food-group(description), corresponding to a unique NDb-Id, there will be multiple nutrients. Therefore, there exists “*one mandate to many mandate*” relationship between these two entities.  
  
Similarly, for each entry in the chocolates Db there may/may not be a possible Tweet\_Text and Tweet\_Date in the Tweets table which implies “*one mandate to many optional”* relationship between these two entities.



**Conceptual Data Model for the Chocolates Database**

**Assumptions:**

1. Each entry in the “Chocolates” entity will have a unique NDb-Id.
2. The data from Twitter is collected for few NDb-Id’s due to the timeout feature. However, the code logic can be extended for additional entries.

**Data Quality/Audit results:**

**Audit** **validity/ accuracy:** **1.** The data collected from the csv file contains the nutrients pertaining to each food-group. An additional column of “NDb-Id” was added to this to be able to relate it with the data scraped from the website.  
 **2.** There are certain no. of empty values in the nutrients data (Details have been demonstrated in the AuditData.ipynb file)  
 **3.** There are no duplicate records present in any of the entities.  
**Audit completeness:**   
The dataset shows the different nutrient content and their amounts present in various brands of chocolates and how people around the world are discussing about the popular chocolate brands.

**Files involved:**

Following are the deliverables of the assignment:

1. Jupyter Notebook file: “DataMungingChocolates.ipynb”
2. Raw-data: CSV Files used: “CSV.zip”
3. Report file

***Folder-Structure:*** The parent folder named “Chocolates” contain the notebook file, AuditData file and the “data” folder holds the Input CSV-files and the output files generated by the code.

**Conclusion:**

The code successfully scrapes the data from three different sources and the conceptual model explains the relationship between the data gathered.

**Future Scope:**

A real-world database can be created by importing the final csv generated into any database to fit in a physical model.

**Citation and reference:**

* **Main website for Data-collection:** United States Department of Agriculture Research Service: USDA Food Composition Databases.  
  <https://ndb.nal.usda.gov/ndb/search/list>
* **Twitter related references:**
* Twitter standard search APIs and usage:  
  <https://developer.twitter.com/en/docs/api-reference-index>
* Tweepy Reference Index:  
  <https://tweepy.readthedocs.io/en/3.7.0/api.html>
* Using Tweepy based Authentication and Python Object Oriented structure:  
  <https://github.com/vprusso/youtube_tutorials/tree/master/twitter_python>
* Tweepy Cursor Usage Gist:  
  <https://gist.github.com/vickyqian/f70e9ab3910c7c290d9d715491cde44c>
* Capturing Rate Limit Error and Rate Limit Exception Handling:  
  <https://stackoverflow.com/questions/21308762/avoid-twitter-api-limitation-with-tweepy>
* Style guide for Python:

<http://google.github.io/styleguide/pyguide.html>

* Git-Ignore file for Python:

<https://github.com/github/gitignore/blob/master/Global/JetBrains.gitignore>

**Licensing:**

The text in the document by Nagraj, Mansi & Priyadarshi, Dhawal is licensed under CC BY 3.0 https://creativecommons.org/licenses/by/3.0/us/