**DAB402: Assessment 3**

**Title**

**RECOMMENDATION OF WEBSITES BASED ON REVIEWS**

**Description:**

The afore mentioned project regarding recommending the websites based on reviews needs a huge amount of dataset of people reviews regarding items purchased. This cannot be specific to single category. So, we have done a research on dataset and requested it from an authorized person [julian.mcauley@gmail.com](mailto:julian.mcauley@gmail.com).

Secondly, the dataset is being generated by crawling through the websites by scrapy and parse hub to get the required dataset.

* Quality of the data is crucial as having all the columns in a dataset to complete the need of creating a model to recommend website based on reviews.
* Fitness, this is an important aspect as currently we are having more than 28000 rows and 18 columns in our dataset by merging dataset of three different websites. This will allow our product to have complexity in generating accuracy.
* Usability, proficiency of our merged dataset is near about 75% that we can use maximum of its columns in our project.

We have collected the data by following the ethics as to get the authorization and permission to use data in our project.

**Outcomes Assessed:**

Prepare an ethical data collection and management plan that addresses all the data needs for the proposed product/solution:

* Collection of the data set is itself a challenge for us as using parse hub, this is a long process to get the html data in csv format.

**5C’s** in ethical use of data are Consent, Clarity, Consequence, Control, Consistency

**Consent:**

This project, Recommendation of Websites based on review is chosen to give the genuine information to the customers. To implement this project, data is collected from e-commerce websites were the user’s opinions are given for the products/services with consent. Here we are using the data which is openly available for the better product and website recommendation.

**Clarity:**

This dataset provides information regarding product details, reviews, rating, purchase date, product URL. Review URL etc. with which we can analyze the data and build appropriate models for efficient recommendation engine.

**Consequences:**

In our dataset we are not retrieving any personal information of customers so there are no consequences to be deal with.

**Control and Transparency:**

We have merged the data collected from various ecommerce site and formed a dataset. Here the information is analyzed using sentimental analysis to filter out the classification and perform further process we assure that the data will not be manipulated or mis-used.

**Consistency:**

Our data is consistent as we are interpreting the intentions of customers through the reviews and analyzing their views on the product to build better recommendation model for search engine. As our dataset does not contain any personal information about the users so, we are consistent with the information taken from online sites to work on our project.

## **References:**

DataCamp Community. (2020). *Making Web Crawlers Using Scrapy for Python*. [online] Available at: https://www.datacamp.com/community/tutorials/making-web-crawlers-scrapy-python#overview [Accessed 15 Feb. 2020].