

**CISC7201 – INTRODUCTION TO DATA SCIENCE PROGRAMMING**

**1st Semester, 2019/2020**

Report of

**Recommendation System of amazon product data**

***Submitted by***:

Group: K

Wong Kit Kan - mb95513

Chio Chon Hou - mb95540

**INTRODUCTION**

In this project, we are going to build a recommendation system based on the products and rating detail of amazon China. In the recommendation systems, there are three common algorithms that is content-based filtering, neighborhood methods and latent factor models. And the latter two are collectively referred to as collaborative filtering. In latent factor models we are already used in our assignment to train the neural network, so we decide using the transitional TF-IDF model as content-based filtering to rank the page (product name) and neighborhood methods to calculate the item-based and user-based ranking. An implementation of simple recommendation system and search engine.

**DATA SOURCE**

* From: Amazon China’s product data
* Data overview: (846 MB)
  + 20,000 products
  + 1,100 categories
  + 1.42 million users
  + 7.2 million reviews / rating data
* From: <https://github.com/SophonPlus/ChineseNlpCorpus/blob/master/datasets/yf_amazon/intro.ipynb>

**DATA CLEANING**

* Drop useless column
* Remove NaN product name/user id/rating data
* Remove punctuation(%$?) in product name
* Remove Extreme data (rating)
* Drop duplicate rating records

**DATA PROCESSING**

**DATA ANALYZE**

**DATA RECOMMENDATION**

**DATA VISUALIZSION**