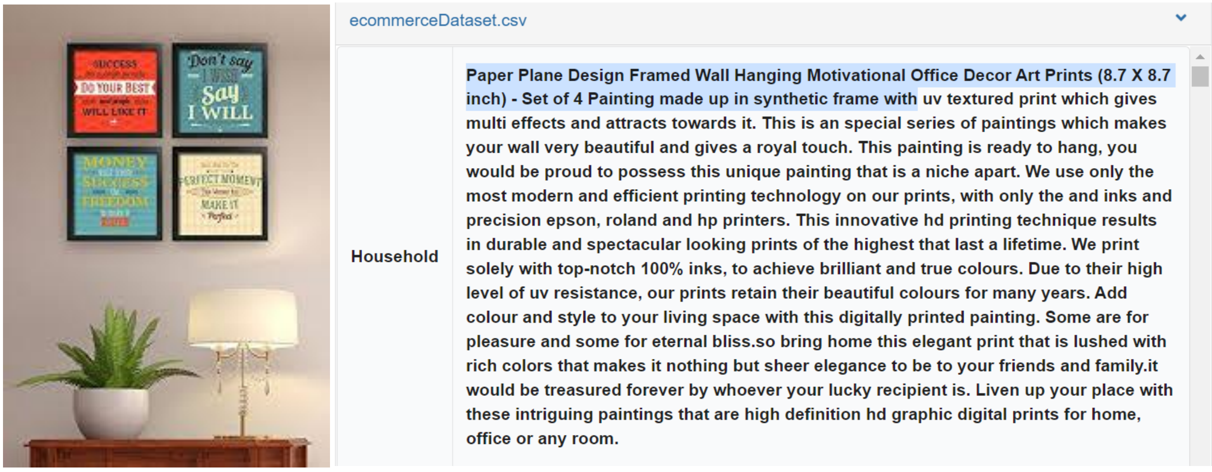
# COS30081: Fundamentals of Natural Language Processing [Credit Task]

| | |
|---|---|
| **Structure** | Script submission (.ipynb) |
| **Group or Individual** | **Individual** |
| **Learning Outcomes Assessed** | This assessment task is designed to test your achievement of learning outcomes 1,2,3 and 4 |
| **Task** | Credit Task |
| **Due Date** | **Sunday, 11:59PM (**Week 9) |

**Assessment Overview:**

The digital economy is growing exponentially even here in Sarawak. Dr Joel wants to create an online shopping platform specifically for Sarawak. He was disappointed by the heavy shipping charges for the current shopping platforms (Shipee and Leezada). He wants to create a platform called Sarapee Shopping. To get started on his idea he wants to do some competition analysis to view how many different types of products his future competitors are selling. He has tasked you to build an NLP classifier to classify product categories based on product descriptions.



ecommerceDataset.csv

Household

Paper Plane Design Framed Wall Hanging Motivational Office Decor Art Prints (8.7 X 8.7 inch) - Set of 4 Painting made up in synthetic frame with uv textured print which gives multi effects and attracts towards it. This is an special series of paintings which makes your wall very beautiful and gives a royal touch. This painting is ready to hang, you would be proud to possess this unique painting that is a niche apart. We use only the most modern and efficient printing technology on our prints, with only the and inks and precision epson, roland and hp printers. This innovative hd printing technique results in durable and spectacular looking prints of the highest that last a lifetime. We print solely with top-notch 100% inks, to achieve brilliant and true colours. Due to their high level of uv resistance, our prints retain their beautiful colours for many years. Add colour and style to your living space with this digitally printed painting. Some are for pleasure and some for eternal bliss.so bring home this elegant print that is lushed with rich colors that makes it nothing but sheer elegance to be to your friends and family.it would be treasured forever by whoever your lucky recipient is. Liven up your place with these intriguing paintings that are high definition hd graphic digital prints for home, office or any room.

**Figure 1: Example of Product & Description**

**Dataset:**

The dataset is in ".csv" format with two columns - the first column is the class name and the second one is the datapoint of that class. The data point is the product and description from the e-commerce website. The dataset has the following features :

Data Set Characteristics: Multivariate
Number of Instances: 50425
Number of classes: 4

You can find the dataset online:

- Gautam. (2019). E commerce text dataset (version - 2) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.3355823

**Specific Tasks**

- Exploratory Data Analysis
    - Charts/tables to describe the dataset
- Preprocessing
    - Removal of stopwords/punctuations
    - Tokenization
    - Vectorization/Word Embeddings
- Classification
    - Training/test data split
    - Building & compiling a CNN model
    - Training a CNN model
    - Showing performance of testing
        - Classification report or other measures
        - Confusion matrix
    - Test on real-world product descriptions (English) – minimum 3 from 1 class each
    - Saving model

**Credit Criteria**

A credit grade is achieved if the required specific tasks are achieved and all pass tasks are marked as complete.

**Other Issues**

Submission Requirements

- The report must be submitted via the Canvas, as a ipynb, using the Credit Task link
- Do include comments in your script to show your comprehension of the functions and libraries used.
- Do not email the assessment to either the convener or tutor.
- Submitted file should be named with your student id as following: "CreditTask_100XXXXX.ipynb"
- Keep a backup of your submission. If your assessment goes missing, a copy will be requested

Extensions and Late Submission

Please reread the section on Extensions and Late Submission that can be found in the Unit Outline. Extension requests must be directed to the unit convenor, using the Application for an Extension for a piece of Assessment form. Late submission of an assessment will result in a late penalty being appliedas required by Swinburne University assessment guidelines.

Plagiarism

Please reread the section on plagiarism that can be found in the Unit Outline. Any evidence of plagiarism will result in a Fail. Collaborative discussion with other participants in the unit aroundconcepts and additional examples is highly recommended, but don't copy.

Assessment Help

If you have any queries or concerns you may discuss it with the convenor and/or tutor in the Canvas discussion board in the appropriate discussion forum or by email.

**Marking Rubric:**

To obtain the credit grade, all items below must be marked complete:

| No. | Item | Criteria for Completion | Complete/ Incomplete |
|---|---|---|---|
| 1 | Exploratory Data Analysis | 1 Chart or 1 Table to describe the distribution of data and labels | |
| 2 | Preprocessing | • Tokenization<br>• Removal of stopwords + punctuations<br>• Vectorization/Word Embeddings | |
| 3 | Training/test data split | Splitting the data (training set must be larger than test set) | |
| 4 | Building & compiling a CNN model | • Appropriate hyper parameters (batch size, kernel size) & layers are added. (sequential, conv, activation, dense)<br>• Usage of model.compile()<br>• Showing summary of model | |
| 5 | Training a **CNN** model | • Usage of correct function model.fit() With a minimum of 5 epochs. | |
| 6 | Showing performance of testing | ▪ Classification report or other measures<br>▪ Confusion matrix | |
| 7 | Testing on real-world tweets | ▪ Test on **at least 3** real world product description of each class – link to each product must be given<br>▪ Display of result of inference vs actual label | |
| 8 | Saving the trained model | ▪ Saving the model architecture and weights | |