

The Department of Computer Science

**CIS4515**

**Practical Data Analysis**

Level 7

Portfolio Task 2

2023/2024

### Module Leader: Dr Muhammad Usman

🕿 01695 59 9718

**Email: usmanm@edgehill.ac.uk**

## *Administrators:*

### 🕿 01695 65 7603

|  |
| --- |
| Portfolio Task 2 |

Hand in: **Monday, 19/2/24, noon (Week 5), via a dropbox on BlackBoard**

Feedback: **Monday, 4/3/24**

Learning Outcomes Assessed:

1. Demonstrate proficiency in using collecting requirements, designing and modelling bespoke solutions for big data processing, analysis and presentation
2. Assess the usability, functionality and scalability of big data analysis solutions and select suitable solutions per case

You **must** complete each of the tasks listed in this document. You **must** answer each of the listed tasks in a separate file. Each file must be named “question\_X.EXT”, where X is the number of question and EXT the file extension. All files **must** be placed in a folder called “task 1”. Compress this folder using ZIP or RAR and then upload it to the dropbox on BlackBoard. The due date and time for this task can be found on the Blackboard dropbox and in the module handbook.

|  |
| --- |
| Background |

You are a member of Amazon's Data Analysis team and you have been assigned to a new project that aims to analyse the language used in positive customer reviews. More specifically, you are provided with a dataset of positive customer reviews relevant to different Video games which are available through Amazon's web store. You can download the dataset from BlackBoard: Module **Assessment -> Coursework Resources -> Portfolio Task 2 Dataset: Amazon Positive Reviews**. The dataset is a tab separated file where each line corresponds to a different customer review. Moreover, the dataset consists of three columns. The first column indicates the review score (which is always 3 because the dataset contains positive reviews only), the second column to the product id while the third column to the actual customer review.

In this task, you need to complete the following three questions:

|  |
| --- |
| Question 1 (question\_1.ipynb) |

Develop a python script in Jupyter that loads the third column of the Amazon dataset and stores the reviews into a single variable (e.g., document). Then use the NLTK library to complete the following tasks:

1. extract the words of the document (i.e., tokenisation);
2. identify the part-of-speech tags of the words;
3. extract the bi-grams of the document; and
4. identify the part-of-speech tags of the bigrams.

The example below shows the steps that you need to complete for this step:

PYTHON Example

document = load\_third\_column\_of\_amazon\_review

words = tokenise(document)

words\_with\_pos\_tags = pos\_tag(words)

bigrams\_with\_pos\_tags = bigrams(words\_with\_pos\_tags)

|  |
| --- |
| Question 2 (question\_2.ipynb) |

Develop a python script in Jupyter that extracts the 40 most important collocations from the Amazon dataset using the co-occurrence frequency algorithm (without filtering). Your script should compute the frequency of the bigrams (e.g., Question 1), sort the bigrams base on their frequency and display the top 40 items of the sorted list.

|  |
| --- |
| Question 3 (question\_3.ipynb) |

Develop a python script in Jupyter that improves upon the results of Question 2. Here, you are required to use the part-of-speech filtering to remove frequent bigrams that are not actually collocations. As in Question 2, you should display the 40 most important bigrams according to their frequency.

|  |
| --- |
| Question 4 (question\_4.ipynb) |

Develop a python script in Jupyter that extracts the 40 most important collocations from the Amazon dataset using the mutual information metric. Your script should compute the mutual information of each bigram, sort the bigrams according to the mutual information and display the top 40 items of the sorted list.

|  |
| --- |
| Question 5 |

Write a data analysis report (500-1,000 words) where you critically analyse the results that you obtained in Questions 2, 3 and 4. Your report should clearly explain what are the advantages and limitations of the different collocation extraction algorithms and you should also provide an appropriate justification of your analysis results (e.g., what did you observe when you applied part-of-speech filtering to the co-occurrence frequency metric? which algorithm is more efficient in extracting collocations from Amazon reviews?)