21COP509: Natural Language Processing Coursework

Module leader: Dr Georgina Cosma, Department of Computer Science, Loughborough University
g.cosma@lboro.ac.uk

1 Dataset

Download the Jewellery datasets provided in the Coursework folder found on the module's Learn Page. Setup the following data-path in your Google Drive: "/content/drive/My Drive/Colab Notebooks/COP509cw/Datasets/".
Add the files in your Datasets folder. You can thereafter use the data as follows:
file1 = "/content/drive/My Drive/Colab Notebooks/COP509cw/Datasets/JewelleryReviewsLSA.csv"
file2 = "/content/drive/My Drive/Colab Notebooks/COP509cw/Datasets/JewelleryReviewsQueryRelevantID.csv"
file3 = "/content/drive/My Drive/Colab Notebooks/COP509cw/Datasets/JewelleryReviewsSummarisationTargets.csv"
file4 = "/content/drive/My Drive/Colab Notebooks/COP509cw/Datasets/JewelleryReviewsDeepLearning.csv"

2 Task Descriptions: Semantic Analysis, Sentiment Analysis and Summarisation

For Tasks 1 and 2 use the **JewelleryReviewsLSA.csv** dataset and the **JewelleryReviewsQueryRelevantID.csv** file that contains queries and their relevant IDs.

1. [10 marks] Pre-process the dataset. Comment the code to explain your pre-processing steps.
2. [20 marks] Use the pre-processed version of the dataset to perform the following tasks.
   (a) Develop a Latent Semantic Indexing (LSI) model. [5 marks]
   (b) Develop functionality such that for each query, the LSI model retrieves the top 10 most similar reviews from the dataset. The set of queries and relevant IDs have been provided in the JewelleryReviewsQueryRelevantID.csv file. [5 marks]
   (c) Empirically tune your LSI model (weighting scheme and SVD dimensions). Present the evaluation results of your best tuned LSI model. [10 marks]
3. [20 marks] For Task 3 use the **JewelleryReviewsLSA.csv** dataset, and the **JewelleryReviewsSummarisationTargets.csv** file that contains the targets for summarisation. Develop an approach (either neural or non-neural) to summarise the reviews of each class. Evaluate the performance of the model using suitable evaluation measures. Explain the findings (max 200 words). (Mark allocation: 10 marks for the model, 10 marks for evaluation and comparison).
4. [20 marks] For Task 4 use the file entitled **JewelleryReviewsDeepLearning.csv** that contains the reviews and ratings for each review. Compare a deep learning bag-of-words model to one that uses word embeddings for the task of sentiment analysis (review rating classification). If using BERT, you will need to implement a solution to process the large dataset of reviews. Use suitable evaluation measures to analyse and compare their performances. Explain the findings (max 200 words). (Mark allocation: 10 marks for the models, 5 for evaluation, and 5 for explanations).
5. [30 marks] For Task 5 use the **JewelleryReviewsLSA.csv** dataset, and the **JewelleryReviewsQueryRelevantID.csv** file that contains queries and their relevant IDs. Develop a model to perform neural information retrieval. Extra marks will be given for allowing the user to type in their own query. (Mark allocation: Model development (10 marks); evaluations and graphs using the queries (10 marks), additional functionality (10 marks)).

3 Submission information

- **Submission deadline:** Friday 18 March 2022.
- **What to submit?**
  1. A zip file containing Five (5) colab (.ipynb) files, one file for each task. Each file should contain the entire answer to the question and outputs. Add your student ID and Task number at the top of each file.
  2. Save each colab file as a PDF showing the code and the outputs. To do this: File –> Print and then save as PDF.
- **Submission:** Submit your coursework to Gradescope (instructions provided on the module's Learn page).
- **Source-code quality:** Source-code that does not execute will receive 0 marks. Source-code that executes but does not answer the question will receive 0 marks. Source-code that executes and partially answers the question will receive partial marks.
- **Source-code execution**: Ensure that the code runs without the marker/lecturer having to make any changes to the source-code. Code that does not run will receive 0 marks, therefore ensure that the paths pointing to the data are correctly set in your solutions. You have been warned and therefore excuses will not be accepted.
- **Plagiarism:** All submissions (report and source-code) will be checked for plagiarism. Please refer to this resource: https://www.lboro.ac.uk/students/exam-support/dont-risk-your-degree/
- **Use of Libraries**: You may use the material provided in the labs or your own choice of Python libraries when answering the questions. The choice is yours, but make sure to provide references where needed by adding a comment above the relevant code fragments.
  If you have copied and adapted the code add:
  [This code was adapted from:] (add the link)
  If you have copied and NOT adapted the code add:
  [This code was copied from:] (add the link)