Your Task

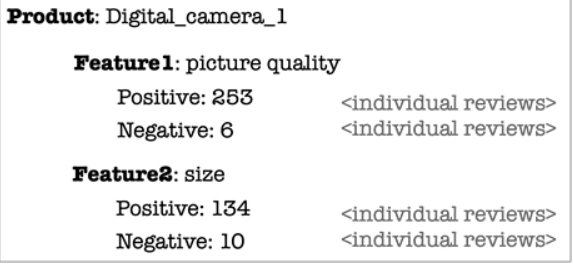
This final practical assignment is based on the material covered in lectures and previous practicals. The task that you will be addressing is Opinion Mining; it is closely related to sentiment analysis and involves using techniques related to information extraction. This is a real-world task that is popular in academic contexts, where opinion mining is an active area of research, and industrial contexts, where companies benefit from in-depth analysis of their customers’ reviews.

As the number of user reviews available online is constantly growing, automated extraction and summarisation of the useful information is of critical importance for both businesses and customers. For example, suppose you have hundreds of reviews on a specific model of a digital camera. Typically, such reviews would contain both positive and negative feedback on various features of the camera: for instance, the majority of the users might agree that picture quality is good, but battery life is poor. On the one hand, if you were a data analyst working for the camera producer, you would conclude that it is improvements in the battery life that the company should focus on primarily; on the other hand, if you were yourself a customer thinking of buying a new camera, the information about different positive and negative aspects of the product might influence your decision.

In this assignment, you will be working with widely used datasets for mining and summarising customer reviews. In a nutshell, the task can be broken down into three subtasks:

1. Identification of keyphrases that describe product features (e.g., picture quality and battery life for digital cameras).
2. Detection of polarity in the sentences that discuss each of the product features.
3. Summarisation of the features and sentiments associated with them.

For example, here is how an output for a set of reviews on digital cameras might look like:

  
**Figure 1:** Summary of the product features and sentiments

**Datasets**

You will be using three product reviews datasets from <https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html>. The reviews were extracted from Amazon and manually labelled with product features and sentiment polarity as well as sentiment strength. For more information on the notation and labels, check the README files provided with the datasets. As noted in these files, sentiment strength is quite subjective, so feel free to use only the binary sentiment annotation and ignore the numerical scores.

For background information on the task and the datasets, take a look at the following accompanying papers:

* Mining Opinion Features in Customer Reviews (<https://www.cs.uic.edu/~liub/publications/aaai04-featureExtract.pdf>)
* Mining and Summarizing Customer Reviews (<https://www.cs.uic.edu/~liub/publications/kdd04-revSummary.pdf>)

You can also check other publications on <https://www.cs.uic.edu/~liub/FBS/sentiment-analysis.html>. Feel free to replicate the approaches from the papers or apply techniques you have learned about in the course.

**Your Task**

Your task is to build a product review opinion miner using the data provided. Specifically, you will need to apply the steps from the pipeline discussed in the course:

* *Analyse the data and the task:*Familiarise yourself with the data and the task. Summarise your ideas in the Jupyter notebook.
* *Apply relevant data pre-processing steps:* Apply appropriate pre-processing steps that you have learned about in the course.
* *Extract relevant information:* Identify keyphrases that describe product features. Relevant techniques from the course that can be applied at this step include PoS tagging, chunking or parsing (e.g., for the identification of noun phrases). You can refer to the two papers above for some inspiration (note: the papers mention the use of association rules to identify meaningful phrases; instead, you could use pointwise mutual information, which is discussed in Week 7 of the course, as the word association measure). Identify sentences that contain opinions (i.e., are sentiment-bearing) about the product features.
* *Apply a relevant algorithm:* Using the identified product features and the extracted sentiment-bearing sentences, apply a sentiment analysis algorithm using the knowledge and skills acquired during the course. Note that if you experiment with different features, settings or algorithms in a machine learning framework, you need to split the dataset into training and test sets (or apply cross-validation) and run the experiments on the same splits to make different runs of your algorithm comparable.
* *Report evaluation results:* Apply relevant evaluation metrics and report the results at different steps of your implementation. For example, you are encouraged to report precision and recall of your keyphrase extraction step, and the metrics relevant at the sentiment analysis step, using the annotation in the provided datasets as the ground truth (or “gold standard”). Note that you will not be assessed on the basis of your evaluation results: i.e., if you implement a reasonable algorithm that attempts to solve the task but achieves low evaluation scores, you will not be penalised for that.

Finally, to present the output of your opinion miner, print out “summaries” presenting the features of the product and the break-down of the positive and negative opinion sentences related to each product feature (see Figure 1).

To get full marks on this project, you will need to perform a comparative experiment: i.e., implement a baseline model (e.g., the simplest or most straightforward algorithm) and experiment with at least one extension, comparing the results of your extended models to the baseline model using the same data splits. Pass mark on this project will be awarded if you write a good report on a baseline model.

1<https://en.wikipedia.org/wiki/Pointwise_mutual_information>



Guidelines

* This is a graded assignment.
* Submit your assignment by **Monday before 12:00 midday (UK time)**. Remember: it may take several minutes to upload your file so leave yourself plenty of time. Late assignments will be penalised.
* You should submit a Jupyter notebook providing a solution to the task together with the accompanying description of the steps you applied. The description of your work in the notebook should be within 2,500 words excluding tables, graphs and images.
* The assessment will be based on the clarity of the description and motivation of the work done, steps implemented and evaluated, demonstration of the skills and knowledge acquired during the course, and insights gained. Assessors may run your code, but you will not be assessed on the quality of your code writing, nor will you be assessed on the basis of where your system’s results rank amongst the results achieved by the systems submitted by other course participants or results reported in published papers.
* You should use the **Bath Harvard Referencing System**. Review the library’s website for more information on correct referencing.
* Save your assignment using the following **naming convention**: UnitTitle\_AssignmentTitle
* **Do not include your name** anywhere on your assignment.
* This assignment uses **plagiarism detection software**. Remember to reference your work.
* Your feedback and grade will be made available in the Grades page 15 working days after the assignment due date. Your assignment grade is provisional, subject to approval by the Faculty Board of Studies following the Board of Examiners meeting, where your overall unit result will be confirmed.