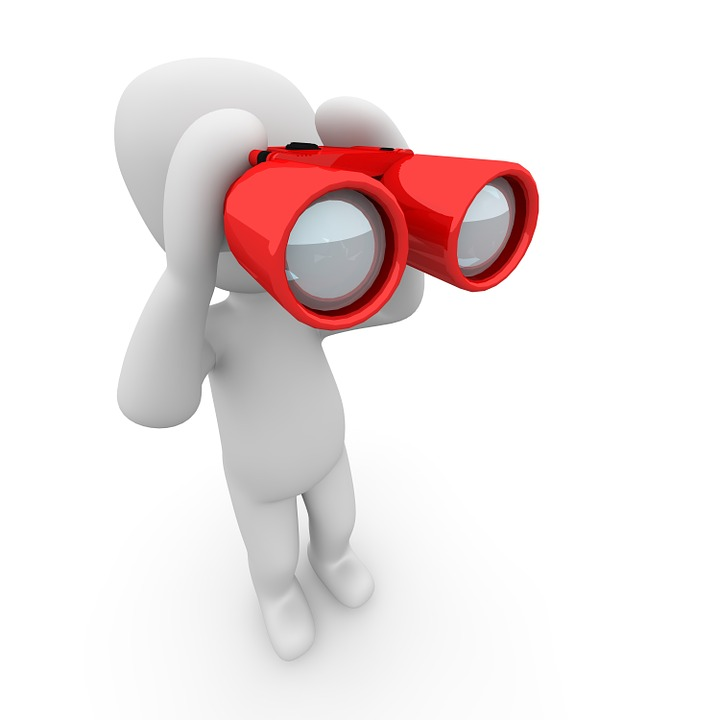
**Text-Mining Final Assessment**

This assessment is individual [No groups allowed]. This assignment was updated on 28 November 2022.

With this assessment, you will complete exercises for each of the main topics covered in the module. These can be used to showcase your work as a Data Scientist to future employers.

**The goal of the assignment:**

The important thing in this assignment is not so much if you use python or you use Orange. The main point is to find useful information that can help the company to run their business in a more efficient way. Imagine that you are a company in the pharmaceutical sector that provides all the drugs in the dataset and you are getting feedback from the clients on those drugs. To this end, finding insights that can help the company and convert them into useful strategies is what is more valuable. The main question you should be asking is what useful information in this dataset can you find for the company.

**Required Tasks:**

You are to complete a full example of applying multiple approaches to the following dataset, including details of analysis and reflections on findings.

There are multiple tasks you can apply to this dataset (classification, regression, clustering, feature selection, etc). You should write a document explaining all the steps you did in this assignment. To simplify classification it is ok if you divide the dataset into two groups (Positive 6-10) or Negative (1-5). You can also divide the dataset into three groups (Positive 8-10), Negative (1-3), and Neutral (4-7).

**Dataset**:

The dataset provides patient reviews on specific drugs along with related conditions: [UCI Machine Learning Repository: Drug Review Dataset (Druglib.com) Data Set](https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Druglib.com%29)

More information on the dataset is: <https://dl.acm.org/doi/10.1145/3194658.3194677>

The original file is in TSV format but you can easily transform it into CSV using the following script:

<https://www.geeksforgeeks.org/python-convert-tsv-to-csv-file/>

**Tasks to perform in the dataset:**

There are multiple tasks we have developed during the course. You can use python or orange to create your data mining analysis.

Here are some resources you can use for developing your project:

* [Introduction to Orange Datamining](https://www.youtube.com/watch?v=OY5KYOhnMls&list=PLYW0LRZ3ePo5jW_mhbvHV5ciNkhe61dTT) and the repository <https://github.com/HussamHourani/HussamHourani/tree/Orange-Datamining/English>
* [Getting Started with Orange 01: Welcome to Orange](https://www.youtube.com/watch?v=HXjnDIgGDuI&list=PLmNPvQr9Tf-ZSDLwOzxpvY-HrE0yv-8Fy)
* <https://www.youtube.com/@OrangeDataMining>
* <https://www.javatpoint.com/text-data-mining>
* [NLTK Sentiment Analysis Tutorial: Text Mining & Analysis in Python | DataCamp](https://www.datacamp.com/tutorial/text-analytics-beginners-nltk)

Some of the tasks like for example showing a dendrogram are not feasible due to the large volume of instances in the dataset. For this case, you can use a sample of the dataset.

**Software:**

Use Orange Data Mining or Python to complete all the Data Mining tasks. For Forecasting and Text Mining, you can use Python.

**Fonts, size, and separation:**

For doing the assignment the fonts should be 10.5, the font Calibri, and the separation between the lines (1.15). Leave one line between paragraphs and leave the margins as they are in the template.

**References & Citations**:

Include a References section for each Task. Each reference should be cited in your text. Use a referencing/citation method you are used to or select APA or Numeric Style. You can use an article in Google Scholar and then click on “Cite” and then include that on the assignment ([Adding Citations & References Using MS Word](https://www.youtube.com/watch?v=CnVq_BpwP2E)).

* https://tudublin.libguides.com/APA\_quick\_guide
* https://www.dit.ie/media/library/documents/Numeric.pdf

**Deliverables:**

You will be required to document your work. I've provided a **template document**, including a plagiarism cover page, for this assessment on Brightspace. Use this **template document**. You can alter/change the suggested format/headings.

Only include details and images which are important for your findings and narrative. Do not fill your report with reports/charts/etc which do not add to your discussion.

**Document / Report:**

See the sample document/report (on BrightSpace). This document contains generic sections for you to complete for each part of this assignment. These sections are just suggestions! You can modify or use your own section headings. Whatever makes sense to you for completing the assignment and telling your story.

**Submission Details:**

The assignment is due on 6th January at @23:00. You should create one document/report containing all the material for each part of the assignment. For presenting your assignment submit a word document.

Make sure you complete the Plagiarism Report on the first page.

You will need to submit your assignment on **BrightSpace VLE**. You cannot submit your assignment via email.

**Q&A and Support**

You should commence work on this portfolio immediately. This will allow you to ask questions and gain guidance on each task during the weekly lab sessions.

Week 13 will be used as a session to discuss any questions you may have.

I will be available until 22nd December to answer any questions via email. All questions asked and their answers will be posted on the FAQ for the module. Make sure to check the FAQ regularly.

**Post-Christmas & New Year Break:**

I'll be available on the 5th of January to answer any questions via email. Again, updates will be made to FAQ.

**Important:**

I will have no access to emails during the Christmas and New Year breaks. It is vital you commence your work early on this assessment and raise any questions as they arise, ideally before the Christmas break.

**Marking Scheme**

See Marking Rubric. The documentation for your assignment must contain the name, student number, class, course (**TU??**) and year information for each student in the group. **Failure to give this information will incur a 10% penalty.**

Each submission must be original work as plagiarism will result in a **zero** mark (0%). **There will be a 10% penalty deduction will be applied for each day the assignment is late.** There is no penalty for submitting early.

DIT Plagiarism Policy:

* https://tudublin.libguides.com/c.php?g=674049&p=4794713
* https://www.tudublinsu.ie/advice/exams/breachesofregulations/

**Assignment Feedback**

Feedback will be via Brightspace VLE. This will consist of a mark for your assignment and a short comment on the assignment.

**Marking Rubric**

| **Achievement** | Excellent | Satisfactory | Basic | Unsatisfactory |
| --- | --- | --- | --- | --- |
| **% of Marks Available** | >75% | 55-75% | 40-55% | <40% |

|  | Weighting |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Problem Definition and presentation | 10 | Well defined problem definition, justifications for selections. Excellent presentation and clarity. | Problem definition, medium to well defined. Clear and defined presentation. | Problem definition  missing details and  Impact. Presentation has a few mistakes. | Poorly defined  problem, and poor presentation. |
| Data Insights & Data Preparation | 20 | Good focused insights from dataset select, good explanation, no trivial data analysis, selection of  appropriate data  preparation,  explanations given, | Useful insights with  explanations, impact of these on problem  solution, selection of some appropriate data preparation,  explanations given | Some insights given, limited details, limited data preparations,  appropriateness of  data prep, minimum explanations given | No or poorly selected data insights, limited or no data  preparation |
| Application of  Algorithms | 15 | Suitable algorithms  selected, good details on these and why, good details on  experimentation,  insights from  experimentation,  reflections, and  discussion | Suitable algorithms  selected, some detail of selection and why,  some details of  algorithm  experimentation, some discussion of  experimentation | Suitable algorithms  selected, limited details of selections given,  limited details of  application of  algorithms given,  limited details of  algorithm settings and tuning | Limited or no details of selection and  application of  algorithms for data and problem. No  explanations |
| Analysis of  Results | 20 | Excellent detailed  analysis of results and excellent insights of these. Clearly  demonstrates impact and outcomes | Well detailed analysis of results, good level of insights on these, what then mean, their  impact and outcomes | Some discussion of  results, at a basic level with little insights | Little, no or very  limited analysis of  results and outcomes from tasks |
| Learning for  work | 15 | Excellent level of  insights, brings  together details  through work, 4-8  citations comparing related work in each task, clear  identification of  improvements,  reflection on learning outcomes from tasks | Good level of  discussion of results, identify some areas for improvement, 3-5  citations used to  compare results in each task, identification of some improvements with limited discussion | Some discussion and evaluation of work,  some comparison with related research,  a limited number of  citations used | Little or no discussion or work, no  comparison with  related research |

| Ethics & Legal | 20 | Excellent level of  discussion and insights for both case studies. Clear well defined  ethical and legal issues. Clear well defined  impact on role and  tasks | Good level of  discussion of ethical and legal issues for  both case studies.  Clearly reflection on these and the roles and tasks | Some discussion of  ethical and/or legal  aspects. Limited  discussion with little or no reflection, and  impact on roles and  tasks | Little or no discussion, simple overviews  given, Little or no  ethical and legal  aspects considered |
| --- | --- | --- | --- | --- | --- |