

School of Computing, Napier University
Assessment Brief

1. Module number	SET11521
2. Module title	Data Wrangling
3. Module leader	<i>Dimitra Gkatzia</i>
4. Tutor with responsibility for this Assessment	Dimitra Gkatzia (<i>D.Gkatzia@napier.ac.uk</i>)
5. Assessment	Coursework: Development of a data-driven recommendation system in Python.
6. Weighting	100% of module assessment
7. Size and/or time limits for assessment	1500 word individual report (500 for Part A, and 1000 for Part B), developed code for all questions and extra data used.
8. Deadline of submission Your attention is drawn to the penalties for late submission	Part A: 24/02/17 at 2300 UK time Part B: 21/04/17 at 2300 UK time
9. Arrangements for submission	Your Coursework must be submitted via the Moodle. Further submission instructions are included in the attached specification, and on the Moodle Webpages
10. Assessment Regulations	All assessments are subject to the University Regulations.
11. The requirements for the assessment	See Attached
12. Special instructions	See Attached
13. Return of work	Feedback and marks will be provided within three weeks of submission.
14. Assessment criteria	Your coursework will be marked using the marking sheet attached as Appendix A. This specifies the criteria that will be used to mark your work. Further discussion of criteria is also included in the coursework specification attached.

Assessment Brief

The assignment aims to cover the learning outcomes specified for the module:

- LO1: Critically evaluate the tools and techniques of the data storage, interfacing, aggregation and processing
- LO2: Select and apply a range of specialised data types, tools and techniques for data storage, interfacing, aggregation and processing
- LO3: Employ specialised techniques for dealing with large data sets
- LO4: Design, develop and critically evaluate data driven applications in Python

For this assignment you will require to use the following dataset: <http://grouplens.org/datasets/movielens/1m/> You can choose a different dataset if it contains details about ratings and users similar to this one. *You will need to check with the instructors before you do so.* If you have any questions, please email Dimitra Gkatzia at D.Gkatzia@napier.ac.uk.

Part A - 30%. Deadline: Friday 24 February at 11pm (UK time).

1. Describe the dataset (e.g. how many files, what type of data, what type is each feature etc.) - (5% of final mark)
2. Read files using one of the approaches you learnt. Which one did you choose and why (what are the benefits over the other approaches etc.)? - (5% of final mark).
3. Perform descriptive statistics analysis and answer the following questions (brief answers):
 - ❖ How many movies have an average rating over 4? (5% of final mark).
 - ❖ How many movies have been rated by men over 4 on average? (5% of final mark)
 - ❖ How many movies have been rated by women below 4 on average? (5% of final mark)
 - ❖ Which are the top-10 movies? You should also provide a definition of what makes a movie “top”. (5% for definition of “top” movies and for code)

You will submit:

1. The source code with your name and comments, so it is easy to identify how you calculated the above answers.
2. Answers to the above questions as a “.pdf” document, maximum 500 words. The document should include your name, matriculation number and contact details.

Part B - 70%. Deadline: Friday 21 April at 11pm (UK time).

For the second part of the assignment you will need **to develop a recommendation system**. You will use the data provided for Part A, but you can also identify and

incorporate data from other sources not used in the module, such as website reviews, ratings from [imdb.com](https://www.imdb.com) etc. You can choose one of the approaches you were taught in Unit 10 or identify one from the literature. The system should be able to recommend 5 movies for a given user. The system can take as **input** a user's info and / or ratings of movies and **output** a set of 5 recommended movies. You will also need to provide answers to the following questions.

1. Which recommendation approach did you choose and why? Please cite relevant literature (5% of final mark).
2. What knowledge does your recommender system need in order to function? Did you use any external data sources, e.g datasets etc.? (20% of final mark: 15% for extra sources and 5% for describing the knowledge needed).
3. Describe how your algorithm / code works (25% of final mark:: 5% for the description and 20% for the actual code).
4. Evaluate your algorithm, either offline or with a user study. Describe the evaluation setup and results (10% of final mark).
5. Reflect on your work: How does your algorithm perform? How can you improve on it in the future? (10% of final mark).

You will need to submit:

1. The source code of the recommendation system with your name and comments.
2. Answers to the above questions as a “.pdf” document - maximum 1000 words. The document should include your name, matriculation number and contact details.

***Tip:** If you refer to scientific articles and/or books, make sure you use an appropriate referencing style such as APA, Harvard style or other.*

Appendix A: Marking Scheme

	No Submission	Very poor	Inadequate	Adequate	Good	Very good	Excellent	Outstanding
A1	No work submitted	Dataset not described adequately, i.e. described only the topic of the data	Dataset not described adequately, leaving most features unexplained	Dataset described partially: half of its elements covered	Dataset described partially, but most attributes covered	Dataset described almost fully or word limit was not followed	Dataset fully described within word limit	Dataset fully described and further investigation was performed, e.g. other similar datasets
	0 points	1 point	1.5 points	2.5 points	3 points	3.5 points	4 points	5 points
A2	No work submitted	Code with many bugs but the approach not justified	Code with many bugs but the approach is justified well	Code with a few bugs but the approach is justified well	Code is correct but justification of approach not appropriate	Code is correct and justification covers a few points	Code is correct and justification covers most points	Code is correct and justification covers all relevant points
	0 points	1 point	1.5 points	2.5 points	3 points	3.5 points	4 points	5 points
A3	No work submitted	None of the questions answered correctly and methodology somewhat correct	None of the questions answered correctly but methods are mostly correct	None of the questions answered correctly but methodology correct for all of them	One question answered correctly and methodology correct for the rest	Two questions answered correctly and methodology correct for the other two	Three questions answered correctly but methodology correct for the rest	All questions answered correctly
	0 points	4 points	6 points	10 points	12 points	14 points	16 points	20 points
B1	No work submitted	Justification was not offered.	Poor justification, leaving most aspects unexplained	Justification only covered a few aspects	Justification was given partially, half of its aspects were discussed	Justification was given partially, but mostly covered all aspects	All relevant aspects were covered but word limit was not followed.	All relevant aspects covered within word limit
	0 points	1 point	1.5 points	2.5 points	3 points	3.5 points	4 points	5 points

	No Submission	Very poor	Inadequate	Adequate	Good	Very good	Excellent	Outstanding
B2	No work submitted	Extra sources not used and knowledge not described adequately	Extra sources not used and knowledge described partially	Extra sources not used and knowledge not described in detail	Extra sources not used but knowledge described in great detail	Extra sources used but was not described adequately	Extra sources used but input was not described in detail	Extra sources used and knowledge described in detail
	0 points	4 points	6 points	10 points	12 points	14 points	16 points	20 points
B3	No work submitted	Code with bugs and algorithm not well described	Code with bugs but algorithm well described	Code with a minor bug but algorithm not well described	Code with a minor bug but algorithm well described	Code without bugs but algorithm not described	Code without bugs but algorithm not described in great detail	Code without bugs and algorithm described in detail
	0 points	2.5 points	7.5 points	10 points	12.5 points	17.5 points	20 points	25 points
B4	No work submitted	Work out of topic	Neither the evaluation setup nor the results are described appropriately	Evaluation setup is not justified but almost correctly executed and results are mentioned	Evaluation setup is not justified but correctly executed and results are mentioned	Evaluation setup is somewhat justified and results are somewhat mentioned and discussed	Evaluation setup is somewhat justified and results fully described and discussed	Evaluation setup is justified and results fully described and discussed
	0 points	1 point	1.5 points	3 points	5 points	7 points	8.5 points	10 points
B5	No work submitted	Reflection and future work suggestions did not make sense	Not adequate reflection provided neither suggestions for future work	Either only reflection or suggestions for future work submitted	Average reflection and suggestions for future work	Good reflection and suggestions for future work	Very good reflection and suggestions for future work	Excellent reflection and suggestions for future work
	0 points	1 point	1.5 points	3 points	5 points	7 points	8.5 points	10 points

Late submission policy

Coursework submitted after the agreed deadline will be marked at a maximum of 40% (undergraduate) or P1 (postgraduate). Coursework submitted over five working days after the agreed deadline will be given 0% (although formative feedback will be offered where requested).

Extensions

If you require an extension, please contact the module leader **before** the deadline.