**2021 COMP5840M coursework specification draft**

**Lecture: Eric Atwell, School of Computing, University of Leeds**

**Eric personal introduction：**[**http://www.comp.leeds.ac.uk/eric**](http://www.comp.leeds.ac.uk/eric)

**Assessment: 10% presentation (due XX/XX/XX), 10% test (due XX/XX/XX), 20% report (due XX/XX/XX), 60% exam**.

You should work in a **TEAM** of up to **3** students: choose your partners, choose a team name, then write your **TEAM NAME** next to **YOUR NAME** on the Class List in the MSc Lab. If you don’t have partners, look at the Class List to find others without a team, and ask them to join you. You can also use the Differ chat forum to find partners and discuss the coursework. Each team must produce a joint team cw1 and cw3 submission; all partners get the same grade. Each student must also take an individual test. If you cannot work in a team and need to work alone, you can do so, but you must let me know:

[e.s.atwell@leeds.ac.uk](mailto:e.s.atwell@leeds.ac.uk)

**Learning objectives:** this exercise will enable you to

- Investigate theory, methods and terminology used in Data Mining and Text Analytics;

- Experience how to apply algorithms, resources and techniques for implementing and evaluating Data Mining and Text Analytics in a practical exercise set by the Kaggle research contest;

- Summarise and present your achievements to a peer audience.

**Coursework Background**

**Title:** Real or Not? NLP with Disaster Tweets

**Aim:** Predict which Tweets are about real disasters and which ones are not

**Description:** Twitter has become an important communication channel in times of emergency. The ubiquitousness of smartphones enables people to announce an emergency they’re observing in real-time. Because of this, more agencies are interested in programmatically monitoring Twitter (i.e. disaster relief organizations and news agencies).

**Coursework 1** **Presentation, deadline XX/XX/XX (summative, 10% of module grading)**

recording of 6-minute team presentation: a summary of your solution, covering CRISP-DM phases Business Understanding, Data Understanding, Data Preparation, Modelling and Evaluation for the Prediction Challenge. To record the presentation, make a PowerPoint file, add audio narration to each slide, Save As **ppt / pptx**, then Save As video format, upload to a website eg: YouTube or VideoLeeds. **ONE of your team must SUBMIT via Minerva:** the Powerpoint file with recorded audio narration, AND as a text comment, the TEAM NAME and names of team members, and the URL of online video recording e.g. on YouTube

(do NOT submit the video file, just the pptx and a URL where I can view it)

**Coursework 2 test XX/XX/XX (summative, 10% of module grading):** you MUST attend the timetabled lecture to take an individual in-class test.

**Coursework 3 report, deadline 7.5.20 (summative, 20% of module grading):** 6-8 page research workshop paper (ACL format PDF file) on your initial and revised solutions.

For coursework1, your team developed a solution to the Kaggle competition, in coursework3 your team aim is develop a better solution. After cw1 submission, you can view other solutions by other groups, and learn from them. You may also try other data-mining tools, e.g. Python tools such as Gensim. Your team must write and submit a research workshop paper comparing your initial and final solutions, covering CRISP-DM phases Business Understanding, Data Understanding, Data Preparation, Modelling and Evaluation for the kaggle task. Your cw3 paper must comply with ACL’2019 format, Microsoft Word or LaTeX

You MUST keep to limits: 6 minutes maximum per presentation video, 8 pages workshop paper.

I want to show some of the 5-minute team presentations to current and future Data Mining and Text Analytics students, and also to others, for example on Open Days and/or Applicant Days. If you do NOT want your presentation on public view, EMAIL ME:

[e.s.atwell@leeds.ac.uk](mailto:e.s.atwell@leeds.ac.uk)

**Marking scheme:**

In your cw1 presentation, I will assess:

Business Understanding: state objectives and requirements, and data mining problem definition (0-2)

Data Understanding: explain data format and content, note data quality issues and expected results (0-2)

Data Preparation: how the data format was converted for valid input to mining tool, with evidence (0-4)

Modelling: Classifiers, features and parameter settings investigated, with example outputs (0-6)

Evaluation: how to produce output for evaluation, tables of results, best classifier(s) (0-4)

In addition: Presentation quality: overall clarity and impressiveness of the presentation (0-2)

TOTAL: up to 20 marks

In your cw3 research workshop paper, comparing cw1 method with new method, I will assess:

1. Business Understanding: state objectives & requirements, and data mining problem definition (0-2)
2. Data Understanding: explain data format a content, note data quality issues and expected results (0-2)
3. Data Preparation: how the data format was converted for mining tools, with evidence (0-4)
4. Modelling: Classifiers, features and parameter settings investigated, with example outputs (0-6)
5. Evaluation: how to produce output for evaluation, tables of results, best classifier(s) (0-4)
6. In addition: Format: conforming to ACL’2019 format and organization (0-2)

TOTAL: up to 20 marks