The University Of Sheffield.

MSc in Data Analytics Project Themes 2018/19

MEDICINES DISCOVERY CATAPULT

Supervisor/Industry Advisor	**Medicines Discovery Catapult** **Matt Hodgkiss/Adam Poulston** **Supervised by:**
Project Title	**Classifying Figures from Scientific Publications**
Description	**The project is based on classifiction of figures from scientific literature** **The project will involve classifying figures extracted from journals in PubMed Central.** **A dataset containing both images and image labels will be supplied. Each image will have two classes. For example, a figure might be classified as a *graph* and as a *scatter plot.* Some figures also have multiple panels containing sub-figures and are referred to as compound-figures.** **Task 1- Build and evaluate a model to determine whether a figure is a compound figure or not using supplied images and labelled data.** **Task 2 - Build and evaluate a model to classify figures using supplied images and labelled data. Primary classifications are Plot / Image / Table. Secondary classifications are additional labels which classify the figures with more granularity.**

	Stretch goal **Identify locations of sub-figures for compound-figures which have been identified in Task 1. Apply model built in Task 2 to extracted sub-figures.**
Technical Requirements/Data Provision	**Data will be provided in the form of an archive file to download from MDC. This will contain the image files of figures and JSON files containing labelled data.**
Additional information or requirements	**A Non-Disclosure Agreement is required as MDC has commercial interests in this area.**

MEDICINES DISCOVERY CATAPULT

Supervisor/Industry Advisor	**Medicines Discovery Catapult** **Andrew Pannifer** **Supervised by:**
Project Title	**Chemistry**
Description	**Automated extraction of molecular structures from text documents using OCR, or from images of chemical structures, is prone to error. Automatic approaches to correcting these errors will be useful in reducing the number of errors. Two approaches will be used for correction; (1) identifying frequent substitutions/deletions/insertions directly from processed strings derived from the OCR/image processing and (2) the conditional probabilities of each character in the processed string using a recurrent neural network trained on a large number of drug-like molecules.**

Technical Requirements/Data Provision	**Data will be provided. Tensorflow, preferably implemented on GPUs, open source chemistry toolkits.**
Additional information or requirements	

THOMSON REUTERS

Supervisor/Industry Advisor	**Thomson Reuters** **Jochen Leidner** **Supervised by: Mark Stevenson**
Project Title	**A "Call for Papers" Information Extraction System**
Description	**One of the key applications of natural language processing is the automation of quasi-mechanical tasks in scenarios where the input is textual, but the input documents vary slightly with regards to how information with the same kind of meaning is linguistically expressed, and in the way documents are formatted.** **"Calls for papers" are email messages that encourage the reader to submit a scholarly paper for consideration to the programme committee of an upcoming conference, workshop or symposium. The purpose is to inform researchers there is an opportunity to disseminate their work and get it published. From the organizer's perspective, they need to put a compelling conference event together, so they need enough high-quality submissions; from the receiving researcher's perspective, they need to check the email's topic against their own research**

	area, and they need to find out whether they can complete any paper they may be working on by the time of the submission deadline (besides other things like required travel budget to attend the event, time availability to attend the conference, and prestige of the conference). **In this project, your task is to develop a system that performs information extraction from emails to populate a database with information about upcoming conferences (date of the event, title of the conference, location of the event, paper submission deadline, topics of the event etc). These are obtained from CfP emails, generally quasi-formulaic messages shared across multiple mailing lists and they typically get repeated ('first call for paper', '2nd CfP'). The project requires a data-driven approach, so one of the key tasks is the creation of an annotated collection of e-mail messages comprising so-called *Calls for Papers* (relevant documents) and Calls for Participation and other messages (non-relevant documents) and the hand-marking of relevant pieces of information in them. On the Internet, various Web sites offer archives of mailing lists, from which collections of Computer Science CfPs can be harvested.** **The project comprises the following tasks:** **Studying the relevant literature to learn about previous related work** **Construct a CfP Corpus (CFPC), a collection of N=500 emails (comprising 250 CfP messages and 250 no-CfPs) in the English language and on any topic within computer science from a diverse set of email sources (mailing lists).** **Create a relational database schema for all important pieces of information that one may want to extract.** **Mark up elements from the data model in a random sample of your CfPs, using an annotation tool like Brat or GATE.** **Develop a system for extracting the elements of a CfP automatically; either use JAPE rules (if using**

	GATE) or using any machine learning tool that offers support for Hidden Markov Models or Conditional Random Fields or LSTM neural networks). **Write a report that contains a qualitative and qualitative error analysis.** **Create a Web demo that permits the visitor to a Web page the pasting of an email into a CGI form that gets submitted to your system so that the extracted elements are shown using colour coding. Any extracted data is stored in a relational database so it can be searched using SQL queries.**
Technical Requirements/Data Provision	**Laptops to conduct data annotation, machine learning experimentation and programming - as in a typical industry R&D lab focusing on Natural Language Processing. The students will collect a couple of calls for conferences (from public mailing lists for example)**
Additional information or requirements	**Programming skills in either Python or Java Text processing (e.g. Linux commandline tools and/or regular expressions)** **Basic skills in databases like SQLite, MariaDB or PostgreSQL**

ADV

Supervisor/Industry Advisor	**Accelerated Digital Ventures** **Timo York** **Supervised by:**
Project Title	**Pitch Deck Data Extraction**
Description	**ADV is a patient Venture investor.** **As part of our investment process, ADV receives applications for funding and these usually include a 'pitch deck' that highlights the key attributes of the company (e.g. founders, problems space, solution,**

	financial forecasts, funding requirements etc). **All this information locked up in these decks currently needs to be reviewed by humans as part of the decision making process.** **ADV believes that using modern technologies and data science approaches, these decks can be reviewed by 'machine' and the relevant data extracted to create a structured data set pertaining to the company, their product and the funding they are applying for. Such a data set could then be used for business intelligence (e.g. growing ADV's BI for a given market) and for supporting the investment decision making process programmatically (not part of this project).** **The objective of this project is to prove/disprove the above hypothesis.** **Key requirements - scalable extraction of data from pdf 'pitch decks'.** **As a minimum: founder team** **Problem domain** **Opportunity size** **Competitor information** **Progress to date** **Creation of a structured model to represent a typical 'pitch deck' & infomation included.** **Population of model from submitted pitch decks**
Technical Requirements/Data Provision	**ADV will provide an overview of the ADV investment process (for context)** **ADV will provide sample pitch decks**
Additional information or requirements	**ADV will try to limit the need for an NDA based on the info/pitch decks shared for this project. However, an NDA may be required.**

Supervisor/Industry Advisor	**Accelerated Digital Ventures** **Timo York** **Supervised by:**
Project Title	**Review Outcome Prediction**
Description	**ADV is a patient Venture investor.** **As part of our investment process, members of the ADV team review and score funding applications to determine whether the application is taken through to the next stage. During the review, the team members are looking at attributes such as the size of the opportunity, the make-up of the founding team, timing of the opportunity with respect to the market/concept and alignment with the ADV investment thesis.** **ADV believes that machine learning (or other suitable technologies and approaches) can be deployed effectively such that machines can review applications and predict a score, ultimately allowing our review model to scale (through support for or replacement of human activity).** **The objective of this project is to prove/disprove this hypothesis.**
Technical Requirements/Data Provision	**ADV will provide an overview of investment process and in particular the current review mechanism (for context)** **ADV will provide access to model and training date with respect application and reviews.**
Additional information or requirements	**ADV will try to limit the need for an NDA based on the info/pitch decks shared for this project. However, an NDA may be required.**

SIEMENS

Supervisor/Industry Advisor	**Siemens** **Tony Latimer/Jason McGinty** **Supervised by:**
Project Title	**Health Index for gas turbine and major subsystems**
Description	**Gas Turbines are high value industrial equipment, prized for their high power output and availability. Whilst relatively simple in basic principle, they are made up of numerous subsystems, all of which are critical to the overall operation of the turbine.** **The purpose of the health index is to give an overall indication of the condition of the unit to allow the customer to engage with customer support prior to an event actually occurring in an unplanned fashion and causing considerable downtime.** **The index is not concerned with any individual event that occurs within its domain, but rather the aggregate effect of such events on the operation of the turbine as a package.** **For example: many events will be identified within the engine as occurring that will result in a period of downtime, the customer is concerned with the likelihood that a downtime incident will occur not the specific remedy which will be handled by the manufacturer.** **This problem represents a realistic unbalanced data set in that the number of events will be much smaller than the typical operating data, and will allow the student to explore strategies in order to produce models that operate on such data.** **If feasible, it would be useful if the different subsystems could also be scored as part of the index.** **Ideally any such technique could be transposed and tuned to different engines of the same type with**

	minimal new data. It would also be a benefit if the approach could be easily applicable to other similar product types.
Technical Requirements/Data Provision	**Data will be supplied in .csv format. Consists of the following elements:** **1: Config notes for units** **2: Multi year sensor data (typically 1 minute resolution, across 80+ sensors)** **3: Message logs for engines** **4: Event lists (service work, notified breakdowns etc)** **5: Classification of event types and sensors to subsystems** **Starting point will be approximately 30 units. More could be made available later if necessary.**
Additional information or requirements	**System should be able to operate on individual snapshots of data, ideally less than a day if possible. This reflects the fact that data can be returned intermittently and many sites are not permanently connected to data gathering equipment for either operational or security reasons.**

PTC

Supervisor/Industry Advisor	**PTC** **Tanveer Saifee** **Supervised by:**
Project Title	**Prediction of failure scenarios on industrial** **equipment**
Description	**In the Industrial Internet of Things sector we are seeing considerable interest from customers in the field of predictive maintenance. Due to affordability of** **sensors and widespread use of equipment monitoring solutions, people are now turning their attention to more sophisticated uses of data from connected assets, including prediction of failure.**

	In this project we offer a choice of 2 different datasets for students to select from: 1. Use sensor vibration data to predict low grease in a motor 2. Predict when a service call will be required for a cutter on a shipbuilding robot in a Mars Colony Students are invited to tackle these datasets using their machine learning skills, and then compare with results from an automated machine learning platform.
Technical Requirements/Data Provision	**Students will need access to their own machine learning toolkits (e.g. R, Python). For the automated machine learning part they will need to install Thingworx Foundation and Analytics, trial edition available here**
Additional information or requirements	

WANDISCO

Supervisor/Industry Advisor	**WANdisco International Ltd. Alex Shutt/Joe Dreimann/Drew McLaughlin Supervised by:**
Project Title	**Analysis of log file data produced by a distributed system**
Description	**Three possible activities were identified as possible master projects, those were: Trace Modelling Prediction of Throughput Prediction of Failure The overall basis of these projects comes from the distributed architecture of WANdisco Fusion and it's use for replicating "big data" datasets. This results in**

	a very high rate of events being logged, with significant repetition and correlation between the logs at each node. **The initial stages of each project are broadly similar, with different objectives for the data analytics in the later stages.** **For each project, the scope for identifying new insight is considerable, as there are multiple dimensions to the flow of events through Fusion logs that could be chosen for analysis:** **Different filesystem operations generate a different number of connected events, and there are many different operations that are received.** **Each event itself consists of multiple steps as it's processed from beginning to end. Each event will passthrough the same set of steps. Some of these steps will be logged at every node, but some will only be logged at a subset of nodes.** **There can be any number of zones (locations) in a system.** **Each zone can contain any number of nodes.** **The time taken for any step at any node can vary depending on external factors (networking etc) or internal efficiencies or limits.**
Technical Requirements/Data Provision	**WANdisco will provide log data sets**
Additional information or requirements	**An NDA is required to be signed.**

GWEEK

Supervisor/Industry Advisor	**Gweek** **Dr Saeid Mokaram / James Bryce** **Supervised by:**
Project Title	**Validation of scoring criteria and scoring thresholds within gweek's Speech Intelligence Analytics® levels**

Description	**A project to understand whether the score boundaries employed by gweek within its communication skills learning levels (SIA®) are optimal. And to explore the validation of proposed additional features.** **Specific behavioural features in spoken communication, such as filled pauses (umm...ah) and selected discourse markers (you know...like) are understood to be tolerated by listeners, yet only up to a point. Similarly too much or too little eye contact from speaker to listener will determine the quality of relationship between the two.** **Gweek's Speech Intelligence learning levels are arranged in a way to analyse and teach communication skills one step at a time. For instance, Level 1 (Si1, audio analysis) assists with root fundamentals, such as speech pacing (i.e. controlled use of empty pauses as opposed to excessive use of filled pauses). Level 2 (Si2, video analysis) then progresses the learner in terms of the relationship between natural eye movement and empty pauses. Each level is out of 100, and the success threshold is set at 95, as a result of trial and error with user bases and academic research. We ask that a learner achieves a score of 95+ three times in a row to progress to the next level.** **Our questions/tasks in hand are:** **1. For our existing Si1 and Si2 levels have we got the thresholds right? Is the bar too high, or too low? Have we weighted the behavioural features within the scoring algorithm appropriately and defensibly?** **2. And for new features we'd like to add:** **a. Speech rate - what is the sweet spot for listeners and how can we integrate a scoring component to reflect achievement of this sweet spot (too fast, or sweet spot or too slow)?** **b. Lexical variety - identifying what is deemed sufficient variety for listeners to remain focused and engaged and**

	whether a threshold based score against such variety can be justified. **c. Information density - identifying what is deemed reasonable density for listeners to stay focused and engaged and whether a threshold based score against density can be justified.**
Technical Requirements/Data Provision	**Good understanding (under NDA) of the Si levels and their constituent parts; provision of raw data from gweek in order to achieve the tasks**
Additional information or requirements	**This data science enquiry may benefit sutdents with interests in social science, perceptual studies, psychology**

Supervisor/Industry Advisor	Haiping Lu
Project Theme	Kaggle Data Analytics - Amazon Fine Food Reviews
Description	**Analysing 500,000 food reviews from Amazon** **https://www.kaggle.com/snap/amazon-fine-food-reviews/home** **Context** This dataset consists of reviews of fine foods from amazon. The data span a period of more than 10 years, including all ~500,000 reviews up to October 2012. Reviews include product and user information, ratings, and a plain text review. It also includes reviews from all other Amazon categories. **Contents**

- Reviews.csv: Pulled from the corresponding SQLite table named Reviews in database.sqlite
- database.sqlite: Contains the table 'Reviews'

Analytic tasks

- Clustering for community detection, market segmentation, and review helpfulness prediction
- Link prediction
- Recommendation system
- Node classification (e.g., sentiment analysis)
- Visualisation
- System integration

Analytic methods

- Network/graph embedding, i.e., representation learning
- Clustering
- Graph convolutional network
- Convolutional neural network
- Matrix/tensor factorisation

Technical Requirements/Data Provision	● You can use all resources that you can find and build on top of them, e.g., **Kernels** and **Discussion** available at the Kaggle page, GitHub open source software, and research papers. An example is https://www.kaggle.com/gpayen/building-a-prediction-model. ● You will work on individual tasks and then integrate them into a product analytic system suitable for Amazon. **Data includes:** - Reviews from Oct 1999 - Oct 2012 - 568,454 reviews - 256,059 users - 74,258 products - 260 users with > 50 reviews
Additional information or requirements	Nil.