Data Mining: Reseach Analyser

Matthew Powell - Computing Science

This project looks at how machine learning and data mining techniques can be used to analyse and classify the meta-data of computer science research projects in order to determine their quality based on current measures and data available.

UK Research

The budget for research and innovation in the UK is over £6 billion. Universities and higher education institutions rely on this funding to carry out vital research in a great number of fields.

Competition for grants is fierce, with thousands of institutions conducting research, therefore a way of assessing the quality of research was devised to make sure that funding is going to the projects that are the highest standard and that have the biggest impact or effect on the world.

There are four funding councils in the UK:

- Research England
- The Scottish Funding Council (SFC)
- The Higher Education Funding Council for Wales (HEFCW)
- The Department for the Economy, Northern Ireland (DfE)

The four came together to create and carry out the **Research Excellence Framework**.

The Research Excellence Framework

REF2014
Research Excellence Framework

The REF replaced the previous Research Assessment Exercise in 2014 and its aim is to make sure UK research is of a world-class standard. The REF takes a three-point approach to this:

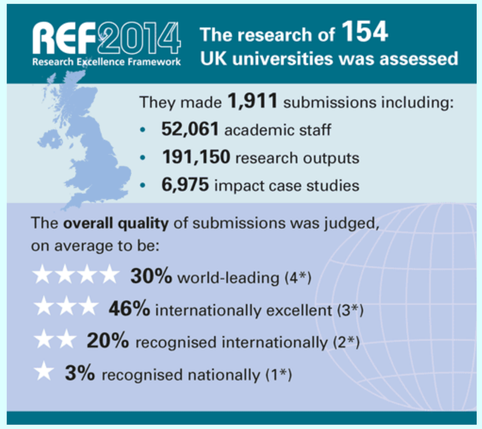
- To provide accountability for public investment in research and produce evidence of the benefits of this investment.
- To provide benchmarking information and establish reputational yard-sticks, for use within the HE sector and for public information.
- To inform the selective allocation of funding for research.

REF2021 Research Excellence Framework

With two out of three points based on money, it is easy to see why institutions take the REF so seriously with over £120m spent preparing for the 2014 REF.

Submissions are split into one of thirty-four units of assessment each with its own expert panel. For each submission, three distinct elements are assessed: the quality of outputs, their impact beyond academia, and the environment that supports research.

Focussing on the quality of the outputs, the submissions are then graded from 1 to 4 star. 4* being world-leading and 1* being only nationally recognised.



Analysing the Data

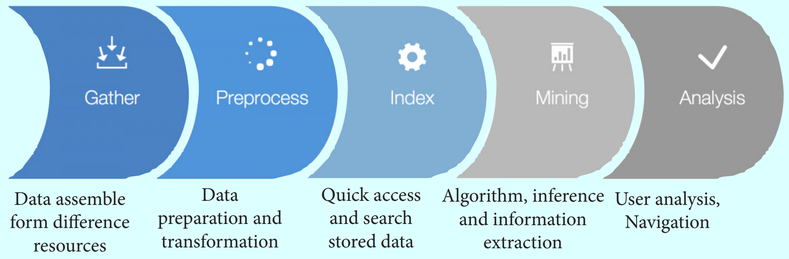
In the data from the 2014 REF, individual submissions are not given grades, it is the institution that is given the result of how many papers achieve each of the four grades.

In order to establish the grade of an individual submission, the institutions and the submissions must be analysed to determine what is contained in a 3* or 4*submission and what is contained in a 1* or 2* submission.

The most comprehensive data field from the 2014 REF in Computer Science is the additional information which contains 100-word summaries of the submissions. Data mining can be applied to these 100-word summaries to discover characteristics that may define high and low-quality submissions.

Data and Text Mining

As text is unstructured, amorphous, and difficult to deal with, the algorithms and techniques that can be used to interpret the data are more limited. Therefore, text has to be broken up using text mining so that it is in a format easy for a computer to understand.



The idea behind text mining is to analyse the information and discover patterns and trends so that different levels of explanation can be interpreted and defined.

Using the results of the mined text data, structured machine learning can then be applied to determine the grade of other submissions. There are tools that can do this processing and can be programmed to carry out this task.

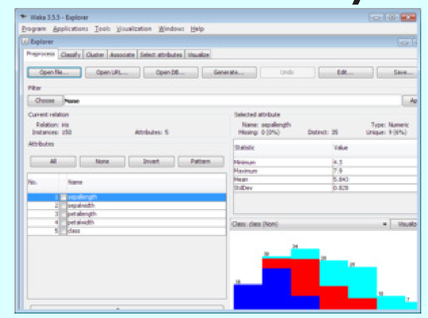
Sketch Engine

The Sketch engine tool is a lexicography tool that can be used to analyse large sets of text data. The text is made into a corpus and can be compared with other corpora to find patterns, differences and similarities. It can also display word frequencies. Multiword expressions (N-grams) and keywords.

SKETCH ENGINE

This information can then be used to train a model to recognise the features that the Sketch Engine has produced.

Waikato Environment for Knowledge Analysis (Weka)



This tool is a collection of machine learning algorithms developed at the University of Waikato. It is a Java-based tool that provides a GUI to handle data.

Weka provides a visual representation of data in a number of formats. It mainly uses arff files which represent datasets attributes, class and values. Data can then be classified with a large number of techniques including tree sorts, Bayes methods and many more. Text data like the outputs provided by the REF are not suitable to be analysed by Weka and will require pre-processing into an arff file format.

Outcomes

The final product of this project will be a tool that analyses research submissions for the REF 2021 and classifies them into a grade from 1* to 4*. This will provide money and time-saving opportunities to institutions as well as the chance to improve any lower graded submissions. It could also be used by experts to grade the papers at the click of a button.