|  |  |
| --- | --- |
| Business School Port ID.JPG | Summer ResearchScholarship 2016/2017 **Otago Business School** |

**SUPERVISOR’S FINAL REPORT**

**(Note: This form activates the final scholarship payment to your student)**

1. **Name of student: William SANSON**
2. **Student ID number:** **4213239**
3. **Programme: BA Year of study: 6**
4. **Name of supervisor(s): Nigel Stanger, Brendon Woodford**
5. **Department: Information Science**
6. **Title of project or brief description (no more than 30 words)**:

Automatic Classification of Old Digital Document Formats

1. **The student has participated in a 10 week research project. Yes ☑ No ☐**
2. **Report from supervisor:**

The main goal of the project was to use machine learning techniques to more accurately identify the specific version of Microsoft Word that created a document (in the old .doc format). William carried out excellent work in four main areas:

1. Writing Visual Basic programs to automatically generate Word documents.
2. Assembling a well-curated data set of 1,360 test documents, each of which was created by a known version of Word.
3. Extracting useful identifying information from the highly complex Word document format.
4. Evaluating changes in extracted information across approximately 20 different major and minor revisions of Word, from Word 95 through to Word 2010.

Unfortunately, the complexities of the Word document format, which we had expected to be considerable, proved even more formidable than our initial expectations. Microsoft Word encodes its files in a way that totally obscures precisely the kinds of features we sought to use to classify documents, making it almost impossible to extract useful feature information. After a couple of weeks’ work on this issue it became clear that we would need more time than there was remaining to resolve it. We therefore directed William to thoroughly document and package up all the work that he had completed to that point.

Consequently, we were unable to proceed with the second phase of the research, i.e., extracting learning features from Word documents and constructing an automatic classifier. William was, however, able to extract and document some important structural elements of Word’s document format that enable us to make at least a coarse identification of the version of Word that created a document, and has tested the behaviour of this information across different Word versions, as noted above.

The results from the are not probably sufficiently novel to be publishable at present, but can form a basis for further work in future. We initially plan to write up the current results as a departmental discussion paper that will be published in OUR Archive.

Brendon Woodford and I would both like to thank William for his excellent work over the summer.



Nigel Stanger

23/3/2017

**SUPERVISOR SIGNATURE NAME** (Printed) **DATE**

**Return completed application form to Theresa Forbes, Information and Special Projects Officer, Dean’s Office, Otago Business School.**