**Pearson correlation**

The Pearson correlation coefficient measures the linear relationship between two datasets. This varies between -1 and +1 with 0 implying no correlation. The p-value roughly indicates the probability of an uncorrelated system producing datasets that have a Pearson correlation at least as extreme as the one computed from these datasets. The p-values are not entirely reliable but are probably reasonable for datasets larger than 500 or so.

**Methodology**

**Additional information**

Additional information constitutes part of the REF output submission where a paragraph of text is given for each submission to describe how significant the output is, how earlier work (before 2008) has been reviewed to incorporate new findings (if any), and how much authors other than the lead author have been contributing to the work (REF, 2012). There are different requirements for additional information across various main panels and even various UOAs. Nevertheless, additional information provides textual descriptions that most closely reflect on the actual content of the submissions. It is therefore hypothesized that the more well-written additional information is, the higher the score would be. A clear definition for ‘well-written’ additional information is therefore required.

**Levenshtein distance vs TFIDF values**

Levenshtein distance measures how similar two strings are in a character-by-character manner. The distance is equavalent to the number of edits (e.g. deletion, insertion or substitution) required to transform one string to another (Gilleland, 2006). Given that Levenshtein distance is one of the easiest-to-compute string similarity measurements, it was used in an attempt to measure the distance from the additional information strings in the top-ranking institution to the strings in every other institutions, by treating the top institution as ‘gold standard’. The institutions were ranked based using all 4\*, 3\*, 2\*, 1\* and UC scores – they were firstly sorted by 4\* scores, if two institutions had the same 4\* score, they were further sorted by 3\* scores and so on - ‘\*’ scores were sorted in descending order, while UC scores were sorted in ascending order. It was noted that, for UOA 12, the top university ranked based on both overall and output scores was the same. When computing the Levenshtein distance between two institutions, all the strings (additional information paragraphs) in both institutions were compared against each other before an average was calculated as the final distance. As seen in Figure [], the averaged Levenshtein distance from the top institution to every other institution was consistent regardless of the scores. It was therefore questionable whether the additional information written by the top university could be served as gold standard; and even if it could, whether the edit distance between two strings of roughly 100-word long would be a suitable measure – for example, since the characters in strings were compared in order, reversing the order of two words would give rise to a large Levenshtein difference. Althernative approaches needed to be exploered.

[Figure of Levenshtein distance vs overall & output 4\* scores – UOA12]

[Figure of unusual word count vs overall & output 4\* scores – UOA12]

TFIDF, short for term frequency-inverse document frequency, provides a numerical measure to how important a word is in a collection of documents. The idea is to find a word that appears commonly in a particular document (high TF) but not often mentioned in other documents of the collection (high IDF), by assigning a weighting factor to each word as a product of TF and IDF (Rajaraman & Ullman, 2011). ‘Important’ words found with ‘high’ TFIDF values are considered to characterize the topic of the document to which they belong, it is therefore hypothesized that the more ‘important’ words a document (additional information paragraph) has, the more ‘well-written’ and informative the document is. Given that the documents in this case are paragraphs of 100-words, to cover the broad range of topics in which a research work is involved, the number of occurances of ‘important’ words in a document is expected to be small. The ‘important’ words are thus referred to ‘unusual’ words. The number of ‘unusual’ words in a document is counted as the number of words with TFIDF values above a threshold. Detailed explanation in threshold determination would be included later on in this report. As compared to the Levenshtein approach, the TFIDF method allowed a more postive correlation to be drawn (Figure []) thus was chosen to carry out further analysis.

**Computing TFIDF**

One typical question for TFIDF is how to calculate TF and IDF. It is easy to find standard formulae for such calculations in text books, however, adjustments are necessary for this specific application.

According to Rajaraman & Ullman (2011), TF and IDF are calculated as follows:

|  |  |  |
| --- | --- | --- |
|  |  | (1) |
|  |  | (2) |

where the term frequency of a word in document () is the number of occurrences of a word in document () normalised by the maximum number of occurrences of any word in the document (), and gives a measure for a word appearing in documents out of total number of documents . is therefore simply the product of and .

The first adjustment was associated with the computation of . It was noticed that not every submission was provided with a solid paragraph of description as its additional information, some only had one generic sentence which was not considered very descriptive. In such cases, the value of would be relatively small, giving rise to a relatively higher values comparing to more descriptive paragraphs which simply contained more words. These falsely large values failed to reflect on the importance of words, thus the normalisation factor was removed from the calculation leaving only .

For the computation of , one had to raise the question whether the collection of documents was defined per institution ( represents the number of submissions per institution) or as an overall pool of documents ( represents the total number of submissions in UOA). To answer this question, both normalisation scales were tested for UOA11, with results presented in Figures [][]. It can be seen that normalising by institution gives a much better correlation with overall 4\* scores than that by normalising across all submissions. This could be explained as each institution has its writing style or a templete provided for additional information, a more informative document should stand out from its peers - i.e containing more ‘unusual’ words comparing to documents of the same style (in the same institution). Thus, when normalised by institution, a better correlation is shown as higher scores are awarded to the institution with distinctively written additional information for each of its submissions. The weak correlation shown for normalisation across all submissions probably infers that even within a particular UOA, there seems not to be any absolutely powerful word that guarantees high scores when mentioned in additional information. Consequently, was computed by normalising documents per institution.

[Figure unusual word count vs overall 4\* - normalised by institution]

[Figure unusual word count vs overall 4\* - normalised across all submissions – 2%]

**Determining threshold**

A threshold defined such that words with a TFIDF value above the threshold are considered ‘unusual’. Based on the fact that the requirements for additional information differ between main panels and even between UOAs, it is difficult to set a uniform threshold across all sample UOAs. An althernative approach was to use the threshold that gave the best correlation between unusual word count and 4\* scores inside each UOA, and then to compare the best correlations across different UOAs or panels.

A document threshold was given as a certain percentile of the TFIDF values in a document, and the final threshold would be an average of document thresholds across all submissions in a UOA regardless of institutions. The final threshold was applied to count the number of unusual words per document, and the counted numbers were then averaged across each institution to give a unusual word count for every institution.

To obtain the threshold that gave the best correlation, a range of document thresholds were initially set from 90 to 99.5 percentile at a 0.5 percentile increment. Pearson correlations were computed between unusual word count and overall/output 4\* scores at each threshold level, and the best correlation was chosen to represent, when the optimal threshold was chosen, how well the unusual word count could give indication to the overall/output 4\* scores awarded to each insitution.

**Results**

[Unusual word count vs overall/output 4\* scores plots for every sample UOA]

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Main panel** | **UOA** | **% submissions with additional information** | **Overall 4\*** | | **Output 4\*** | |
| **Threshold (percentile)** | **Best correlation** | **Threshold (percentile)** | **Best correlation** |
| A | 2 | 50.75% | 98 | **-0.54** | 99.5 | -0.39 |
| 5 | 25.35% | 99 | 0.22 | 99 | 0.21 |
| B | 9 | 32.57% | 91 | **0.53** | 95 | **0.51** |
| 11 | 93.74% | 98 | 0.46 | 97 | 0.41 |
| 12 | 95.19% | 97.5 | **0.62** | 95.5 | **0.55** |
| C | 18 | 6.66% | 90.5 | 0.08 | 94 | 0.06 |
| 24 | 7.97% | 91 | 0.40 | 99 | 0.34 |
| 25 | 8.49% | 96 | 0.19 | 91.5 | 0.15 |
| D | 32 | 5.79% | 99 | 0.06 | 99 | 0.06 |
| 36 | 15.60% | 99.5 | -0.15 | 99.5 | -0.25 |

Table: Summary of best correlations between unusual word count and overall/output 4\* score in sample UOAs

As seen in Table[], correlations of moderate level and above are shaded while strong correlations are bold as well as shaded.

**Discussion**

Based on the results, different levels of correlations are obtained for various UOAs. This section attempts to discuss all the possible reasons behind such difference thus to conclude the level of influence additional information has on the scores awarded.

**Percentage of submissions with additional information**

In the first instance, one would suspect the difference in correlations is largely associated with the amount of data available – the more data there is, the more likely a correlation is going to be found. This could be illustrated in Figure[], which seems to show a trend that Pearson correlation coefficient becomes more positive as more data is given. However, there are two distictive cases where negative and even strongly negative correlations are seen. Since 2 out of 10 chosen UOAs are showing negative correlations, it is therefore hard to conclude that unusual word count correlates positively with 4\* scores. Further clarification should be made by considering more UOAs. However, given the chosen sample UOAs for this project, it is suggested to look at the main panels in which a UOA belongs on efforts to identify any pattern that gives rise to different directions of correlations.

**Main panels**

Table[] shows that the highest positive correlations are found in main panel B, moderate positive correlations are occasionally found in main panel C, while strong and week negative correlations are found in panels A and D respectively. This reflects on the fact that the UOAs in panel B generally provide the largest amount data in the form of additional information, while panel A being the second largest – strong correlations (both positive and negative) are found in these panels. The difference in the amout of information available could be explained by REF submission requirements – for additional information, panel A emphasizes the contributions of co-authors and co-researchers, panel B focuses on abstract-like description for the actual work, while panels C and D do not have any strict emphasis. This casts a doubt on the previous hypothesis whether a more “well-written” and informative additional information woud give rise to a better score in all main panels, given that the additional information in some panels does not summarise the submissions. Thus, this hypothesis could be modified such that more ‘well-written’ additional information only contributes positively to 4\* scores for UOAs in panel B. Since mainly co-author/co-researcher information is involved in additional information in panel A, it is reasonable to question to what extent this information contributes to the scores. Therefore, more investigations in panel A are required to clarify whether or not the negative correlation is a coincidence.

# References

Gilleland, M., 2006. *Levenshtein Distance, in Three Flavors.* [Online]   
Available at: http://people.cs.pitt.edu/~kirk/cs1501/Pruhs/Spring2006/assignments/editdistance/Levenshtein%20Distance.htm  
[Accessed 07 12 2017].

Rajaraman, A. & Ullman, J., 2011. Data Mining. In: *Minding of Massive Datasets.* s.l.:s.n.

REF, 2012. *Annex A Summary of additional information about outputs,* s.l.: s.n.