# Investigation of Wikipedia Recent Changes Summary Field

By Thomas Martin

17th February 2019

Repo: https://github.com/tpgmartin/wiki-stream

## Background & Motivation

In this report, I wanted to investigate how metadata related to edits on main article pages on Wikipedia vary between unregistered users (users who are not logged in) and registered users. In particular, I wanted to see if there was a difference between how these user groups used the "summary" field to detail their changes, and whether this is indicative of different behaviours between these two groups. Beyond the scope of this report, I believe such an approach could form the basis of moderating changes made by users: For example flagging changes that use particular keywords to admins by using a naive Bayes classifier. This has some immediate utility for Wikipedia, as while the organisation is very open to unregistered editing [1] this has led to problems such as vandalism [2].

The data collected for this report comes from the Wikimedia recent changes stream [3] accessed via the WikiMon WebSocket [4]. This was collected over the course of a single continuous, arbitrarily chosen, hour long period between 22:43 to 23:43, 14th February 2019. These messages relate to changes to the English language version of the website. In addition to the default fields contained in WebSocket messages, I added a timestamp to each message on its receipt. To view the code for this investigation, please visit the accompanying GitHub repo [5]. The script written to collect this data dump is located in `./scripts/get_data.py`, the data dump itself is in `./data/dump.json`. The remaining code used to compile this report can be found in either `./notebooks/Investigation of Wikipedia Recent Changes Summary Field.ipynb` or `./notebooks/Investigation of Wikipedia Recent Changes Summary Field.html`

## Edits by User Category

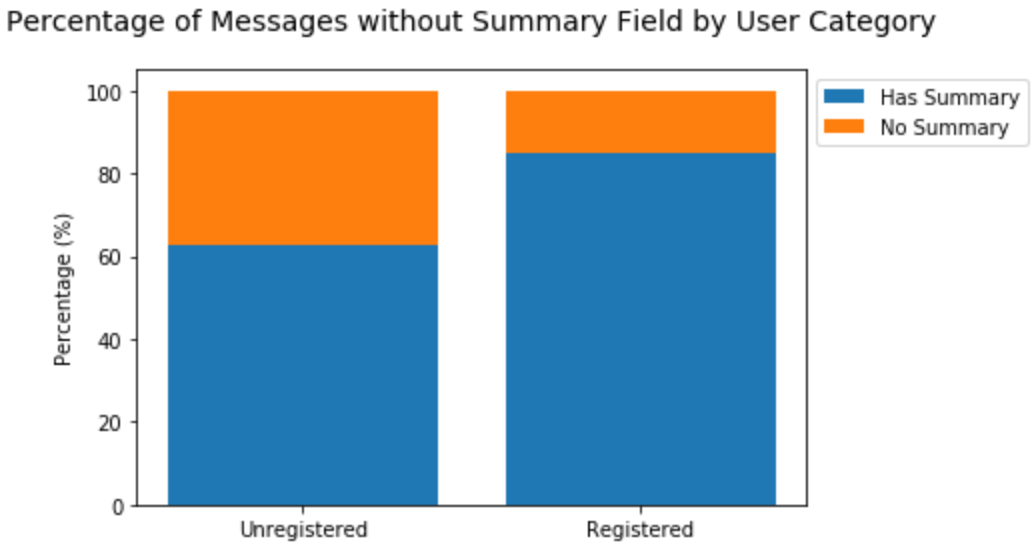
The changes reflected in the stream cover a range of actions and areas of the website, as indicated respectively by the "action", and "ns" flag of the broadcasted message. I chose to focus on edits to main article pages as these corresponded to the vast majority of messages received and demonstrate a richer variety of behaviour.

Of the 7,949 messages collected, 6,937 were produced by human authors as indicated by the `is_bot` flag. I chose to ignore messages related to bots, as I assumed the messages created would be mostly standardised and out of the scope of this investigation. Of the human authored messaged, 4,469 related to main article pages, with a split between registered and unregistered as follows,

| **User Category** | **Frequency** |
|---|---|
| Registered | 3,577 |
| Unregistered | 1,007 |
| Total | 4,469 |

# Messages without Summary Field by User Category

For changes made to articles, users are advised to provide a written summary to help explain the reasons for their change for the benefit of other users. The following figure shows the percentage of messages due to unregistered and registered users edits that contain no summary field. This shows that unregistered users  are much more likely not to submit a summary along with their changes compared to registered users, 37% vs 15% respectively. This could be due to the fact that unregistered users may not be as familiar with the editing process as registered users.

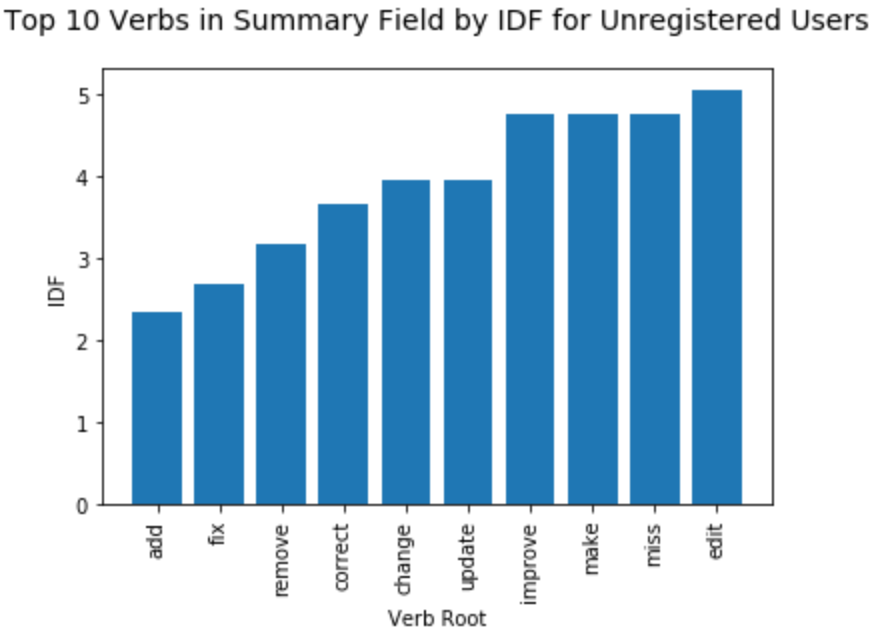


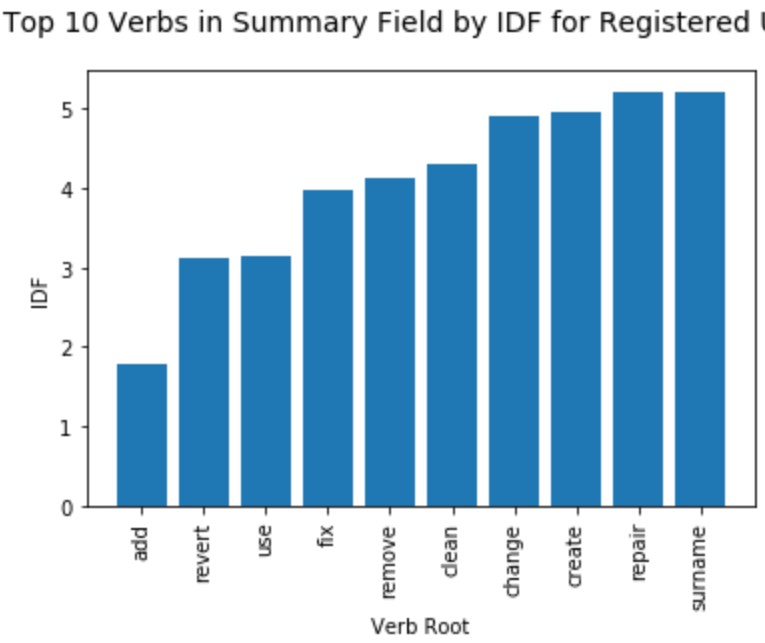
# Top Verbs used in Summary Field by User Category

The "edit" actions that appear in the recent changes stream correspond to rich variety of activities, which can only really be appreciated by investigating the text of the summary field directly. In order to to this, I wanted to find the verb roots used in the summary field, by both registered and unregistered users. Verb root or base form, usually corresponds to the present tense of a verb e.g. "did" and "does" have the base form "do". By focussing on verbs, we can understand the essence of the edit summary, largely separate of its original context. This approach does assume that users submit written summaries consistent with their actions i.e. "add" means adding material to an article. Please refer to the code segments 13 - 17 in the notebook for the specific steps in parsing, tokenising, and lemmatising the summary text fields.

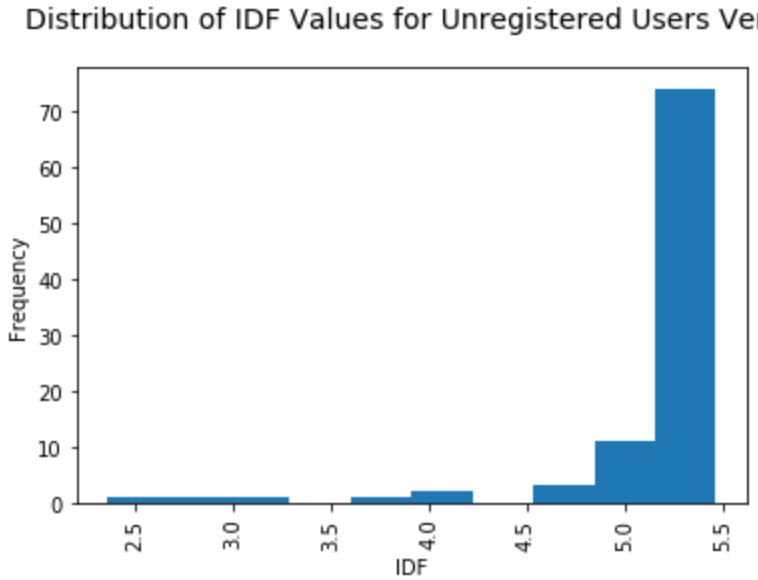
To characterise the relative usage of key verbs between the two groups, I used inverse document frequency (IDF). This is to be able to fairly compare between two groups of a different number of total messages. The lower the IDF value, the more frequent a given word occurs in a set of messages.

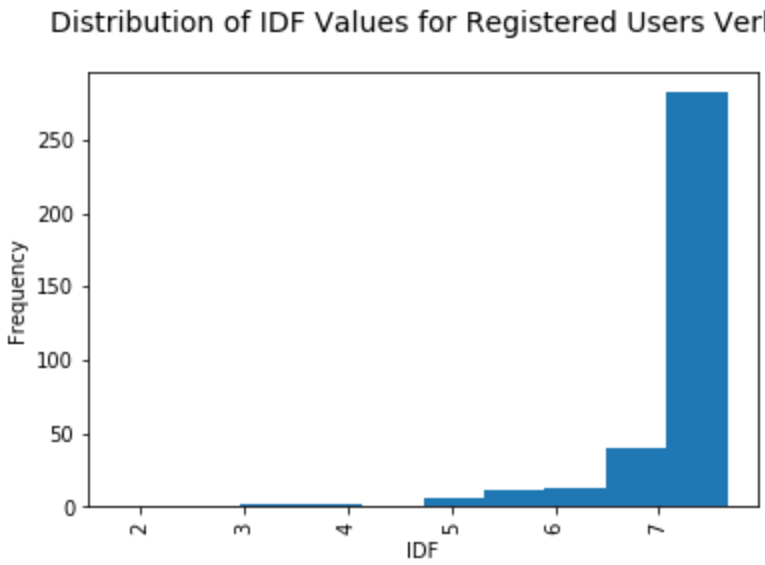
As shown by the figures below, for both registered and unregistered the most commonly used verb was "add", although this was much more widely used by registered users. This most likely indicates the greater incidence of such tasks rather than a "standardisation" of the vocabulary used amongst registered users as the total vocabulary used by registered users was much larger than that used by unregistered users: a total of 359 unique verbs compared to 94.






The following histograms plots, show the distribution of IDF values for both user categories. In both cases, the distributions are very heavily negatively skewed, with the unregistered and registered distributions having a skewness of -3.53 and -3.28 respectively. This pattern is quite common in NLP, where a small subset of words are used much more frequently than the remainder. These figures again demonstrate the wider variety of verbs used by registered users overall: the median IDF value for registered users is 7.7, compared to 5.5 for unregistered users.






The following table gives the top 10 verbs by IDF rank for both sets of users. The number in brackets give the rank of the verb in the other group e.g. "fix", which is ranked 2nd in the unregistered users' verbs is ranked 4th amongst the registered users' verbs.

| **Rank by IDF** | **Unregistered Users** | **Registered Users** |
|---|---|---|
| 1 | add (1) | add (1) |
| 2 | fix (4) | revert (N/A) |
| 3 | remove (5) | use (89) |
| 4 | correct (14) | fix (2) |
| 5 | change (7) | remove (3) |
| 6 | update (12) | clean (N/A) |
| 7 | improve (46) | change (5) |
| 8 | make (16) | create (N/A) |
| 9 | miss (17) | repair (N/A) |
| 10 | edit (22) | surname (N/A) |

It is interesting to see how few of the top ranked verbs for registered users are used by unregistered users. As noted above, this is in part due to the larger number of verbs used by registered users overall. However, given how high they rank in the list of registered users' verbs probably again are some evidence in the difference of actions taken by either group.

# Synonyms for Top Verbs by User Category

The following tables show detected synonyms found in the total set of extracted root verbs for each user category. Please refer to the notebook code cells 33 - 35 for the code samples that produced these matches. Apart from what has already be mentioned above about the use of words by either group, it is also suggestive that broader categories of edits exist due to the number of synonyms for certain words.

| **Top Verbs for Unregistered Users** | **Synonyms** |
|---|---|
| add | N/A |
| fix | make, specify |
| remove | take, transfer |
| correct | adjust |
| change | transfer |
| update | N/A |
| improve | amend |
| make | build, fix, form, take, work |
| miss | N/A |
| edit | N/A |

| **Top Verbs for Registered Users** | **Synonyms** |
|---|---|
| add | N/A |
| revert | return |
| use | N/A |
| fix | define, get, jam, make, repair, restore, specify |
| remove | take |
| clean | blank |
| change | alter, convert, modify |

| | |
|---|---|
| create | make, produce |
| repair | fix, restore |
| surname | N/A |

On a final note, the notebook also contains a set of verbs that a unique to either user category from cells 36 onwards.

# Conclusion

In this report, I showed there are notable differences in the use of edit summary field between registered and unregistered users. In particular, unregistered users are more than twice as likely to submit an edit without a summary. This may simply be due to a lack of familiarity with the submission guidelines. Typical verbs used by either group vary greatly, although this may simply be a consequence of the small number of messages collected overall. Registered users are much more likely to use the root verb "add", which given the larger number of verbs used overall, is perhaps indicative of a greater number of those tasks performed than unregistered users. A similar argument can be made for the verb "revert", which does not appear in an unregistered user's summary field.

The precise reasons for these differences are not clear from this investigation, however, is does suggest that there the different categories of users are performing broadly different actions with their edits.

A possible extension of this investigation could be to use topic modeling to identify related edits by the summary field verbs. This approach could identify broader subcategories of edits, as suggested by the number of synonyms found for the top ranked verbs. It is also suggestive of an approach to flag edits by their keywords.

# References

[1] Wikipedia:Welcome unregistered editing, https://en.wikipedia.org/wiki/Wikipedia:Welcome_unregistered_editing, last accessed 17th February 2019

[2] Vandalism on Wikipedia, https://en.wikipedia.org/wiki/Vandalism_on_Wikipedia, last accessed 17th February 2019

[3] Recent changes stream, https://meta.wikimedia.org/wiki/Research:Data#Recent_changes_stream, last accessed 17th February 2019

[4] WikiMon, https://github.com/hatnote/wikimon, last accessed 17th February 2019

[5] Wiki-stream, https://github.com/tpgmartin/wiki-stream, last accessed 17th February 2019