# AInvestigation of Wikipedia Recent Changes Summary Field

By Thomas Martin

17th February 2019

## 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. 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`

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.

## Section title goes here

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. 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 |

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 compared to registered users are much more likely not to submit a summary along with their changes, 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.

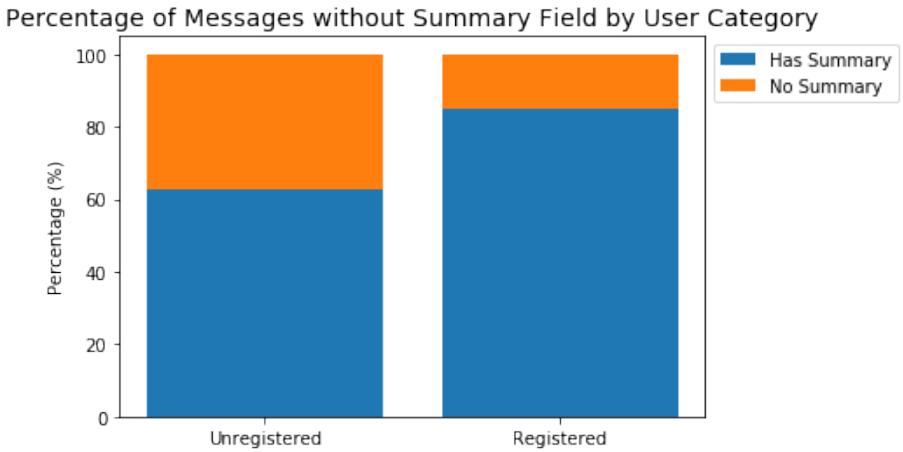


Figure 1: Percentage of Messages without Summary Field by User Category

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 summary field itself. 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.

To characterise the relative usage of key verbs between the two groups, I used inverse document frequency (IDF). This is to 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.

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 verbs to 94.

The following table gives the top 5 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 (4) |

**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.

Registered users are not necessarily more selective that unregistered users in their choice of words, but rather performing a different set of actions altogether.
...

| Top Verbs for Unregistered Users | Synonyms |
|---|---|
| add | N/A |
| fix | make, specify |
| remove | take, transfer |
| correct | adjust |
| change | transfer |

| Top Verbs for Registered Users | Synonyms |
|---|---|
| add | N/A |
| revert | return |
| use | N/A |
| fix | define, get, jam, make, repair, restore, specify |

| Top Verbs for Registered Users | Synonyms |
|---|---|
| remove | take |

Not repeated here, but the accompanying notebook

## Conclusion

## References

[1] Wikipedia:Welcome unregistered editing, https://en.wikipedia.org/wiki/Wikipedia:Welcome_unregistered_ed
last accessed 17th February 2019 [2] Vandalism on Wikipedia, https://en.wikipedia.org/wiki/Vandalism_on_Wiki
last accessed 17th February 2019 [3] Recent changes stream, https://meta.wikimedia.org/wiki/Research:Data#Re
last accessed 17th February 2019 [4] WikiMon, https://github.com/hatnote/wikimon,
last accessed 17th February 2019