Predicting Edits to Wikipedia Articles

1. Who are your group members?

- Natalia Timakova, Nathaniel Weinman, Ugur Yildirim

2. What are the primary questions you want to answer with your project?

- How likely is a Wikipedia article to be edited in the next seven days?
- What is the predicted number of edits to a Wikipedia article in the next 24 hours?

The first question would be helpful for Wikipedia to surface to readers that the article is likely to change soon, while the second could be a useful feature to predict trends in current events.

3. What is the dataset you will use, and how will you process it?

According to our analysis, the normal distribution of articles edited to not edited in a given day is roughly 1 to 1000. Running SQL queries on WikiMedia's Quarry, we will collect data of Wikipedia articles in a 30-day period. We will use the MediaWiki database to access metadata about Wikipedia articles (e.g. namespace) from the first quarter of 2016. We'll specifically be collecting data from the "talk" and "main" namespace for pages without edit restrictions. We'll collect the number of edits in a range of periods (e.g. 1, 2, 5, 10, 30, 90 days). Additionally, we'll use Page View Statistics for Wikimedia Projects to access the view counts for these articles. This database also gives us the ability to compute the minimum, maximum, average, and changes in length over each time period. For a complete list of features, please, refer to 6.

As this data is challenging to obtain at scale, as a proof of concept we built a preliminary data set based on one day - connecting 4/2/2016 edits to 4/1/2016 features. In the process, we've also built generalizable mechanisms to make it easy to get this data over a larger timeframe. This proof of concept is enough to see if there is a correlation between edits and article views and size. For the results, see 7.

4. What analysis has been done on this dataset by others?

There are a few lines of research that use data on Wikipedia editors' activity and article popularity.

The first one focuses on editors' behavior and role identification. An example of such a research would be Diyi Yang et. al. [1], who proved that articles in different stages of quality require different types of editors, and there are eight types of such editors. What is especially relevant to our project, is the taxonomy of edit categories and a set of features Yang et. al. developed to classify an edit into one or more of the edit categories. Because "just as documents are mixtures of topics, editors are mixtures of roles," the authors use graphic model underlying the Latent Dirichlet Allocation (LDA) - method widely used for topic modeling. In the second section of the paper, the authors study how different editors' contributions influence article quality. They built 24 regression models to predict edit counts in each individual edit category from editors' role distribution.

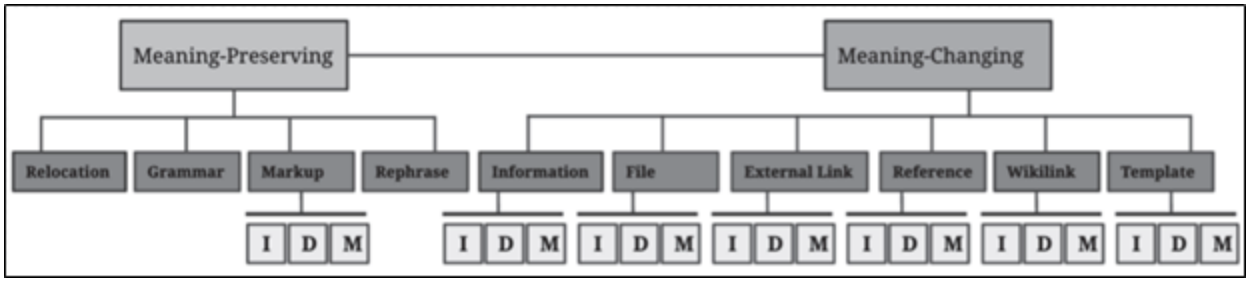


Figure 1: The Taxonomy of Edit Categories. Note: Insertion is abbreviated as I, Deletion as D and Modification as M. Diyi Yang et. al.[1]

The second group of research is focusing on predicting certain events, like disease outbreaks [2] or movie box offices [3]. Twitter data is usually used along with Wikipedia in this line of research. It is extremely relevant for our project to account that real life events significantly impact editor's activity on Wikipedia, and even a completed article which haven't seen revisions for months can become trending among editors if a certain event triggers its popularity.

The third, and quite massive frontier of research is Wikipedia vandalism detection. It's relevant to our project because vandalism is a type of edit, and since certain articles are especially prone to attacks and attract a lot of damaging edits, we might want to take them into account in our analysis. For example, Kiesel et. al. [6] use edit timestamps to understand spatio-temporal characteristics of the revisions made by vandals. In particular, the authors discovered that vandalism is more likely on weekdays than on weekends, and is equally of characteristic for all seasons except summer. Their findings are based on visual inspection, backed by analysis of the variances of the average vandalism ratios (with Cohen's *d*) and the significance test (Welch Two Sample t-test).

Finally, the fourth group of whitepapers is focusing on article quality. For example, Blumenstock [7] used word count to successfully predict whether an article is among the featured ones (one of the signs of article quality). A simple binary classification technique - count every article over 2,000-words long as a featured one - helped reach 96.31% accuracy. A multi-layer perceptron yielded 97.15%, k-nearest neighbor - 96.94%, logit model - 96.74%, and a random-forest classifier - 95.80% accuracy.

5. Sketch out the methods you will use to analyze your dataset.

Our first outcome variable is whether an article will be edited or not within 7 days. Since this is a binary variable (yes/no), we can use logistic regression with regularization using gradient descent to predict this. In addition to logistic regression, we will also use decision trees to predict whether an article is likely to be edited or not. Decision trees allow us to see the most important features of our model. We will also compare feature importances with the coefficients from logistic regression.

Our second outcome variable is the number of edits an article will have the following day. In this case, the outcome variable is count, and we can use either Poisson regression or linear regression with regularization using gradient descent. In addition to regression, we will also use k-nearest neighbors (using a cross-validated k). The forward selection algorithm will be especially useful in selecting which features to include here.

With all of the techniques that we have, we will use cross-validation to avoid the problem of overfitting and optimize our hyperparameters. Regularization with regression and finding the optimal k with knn are also attempts on our part to avoid overfitting our data.

6. Provide a table of summary statistics that indicates how many observations you have, what features you have for each observation, and basic information (min/mean/median/max/SD) for each of those features.

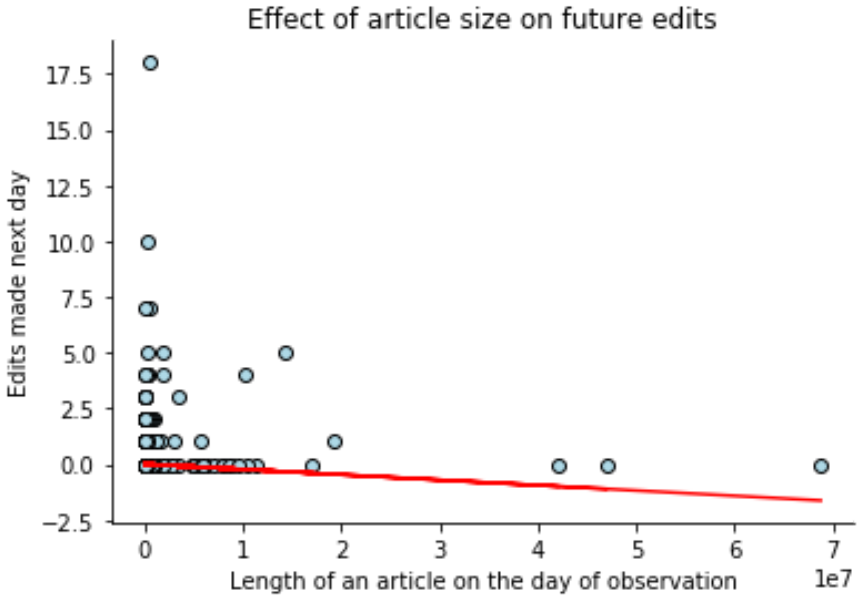
Our sample dataset has 30,697 observations, but the final dataset that we will train our models on will be much larger.

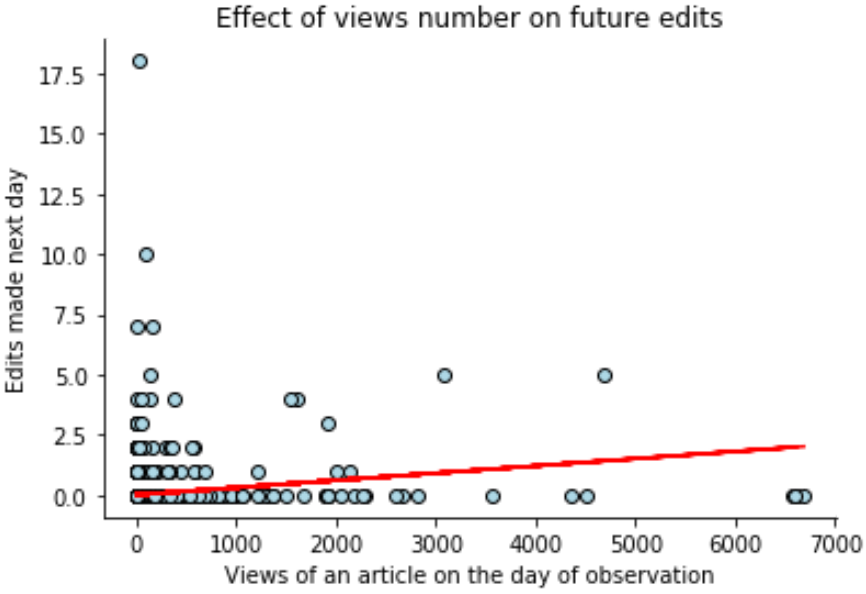
We plan to include the following predictors in our models:

Predictor	**Min**	**Max**	**Mean**	**Median**	**SD**
Number of views for the article	1	6688	14.27	2	116.57
Number of views for the talk page	1	110	1.33	1	2.28
Article size by the end of the period	872	68779248	45822.88	15902.5	604649.25
Max article size during the period	872	228428216	113413.75	20418	2153994.05
Min article size during the period	0	63595712	25770.71	11395	431823.74
Talk page size by the end of the period	3716	506984	15973.03	9854	21044.51
Max talk page size during the period	4953	526240	17842.44	9864.5	28001.75
Min talk page size during the period	0	506984	14507.68	9776.5	18123
Number of past edits of the article	0	27	0.01	0	0.24
Number of past edits for the talk page	1	2	1.05	1	0.21
Number of past minor edits for the article	0	18	0	0	0.12
Number of past minor edits for the talk page	0	1	0.14	0	0.35

We will use these features to generate various relevant aggregate features, such as number of talk page views in the last 90 days. Depending on future data set exploration, we may also be able to add more features.

7. Construct at least one figure to illustrate some interesting pattern in your data that relates to the main question you want to answer.





References

1. Diyi Yang, Aaron Halfaker, Robert Kraut, Eduard Hovy. Who Did What: Editor Role Identification in Wikipedia. Proceedings of the Tenth International AAAI Conference on Web and Social Media (ICWSM 2016)

2. Reid Priedhorsky, Dave Osthus, Ashlynn R. Daughton, Kelly R. Moran, Nicholas Generous, Geoffrey Fairchild, Alina Deshpande, Sara Y. Del Valle. Measuring Global Disease with Wikipedia: Success, Failure, and a Research Agenda. CSCW '17 Proceedings of the 2017 ACM Conference on Computer

Supported Cooperative Work and Social Computing, Portland, Oregon, USA — February 25 - March 01, 2017
3. Márton Mestyán, Taha Yasseri, János Kertész. Early Prediction of Movie Box Office Success Based on Wikipedia Activity Big Data. PLOS ONE, Volume 8, Issue 8, August 2013
4. Mirko Kämpf, Eric Tessenow, Dror Y. Kenett, Jan W. Kantelhardt. The Detection of Emerging Trends Using Wikipedia Traffic Data and Context Networks. PLOS ONE, December 31, 2015
5. Stewart Whiting, Joemon M. Jose, Omar Alonso. Wikipedia as a Time Machine. WWW '14 Companion Proceedings of the 23rd International Conference on World Wide Web, Pages 857-862, Seoul, Korea — April 07 - 11, 2014
6. Johannes Kiesel, Martin Potthast, Matthias Hagen and Benno Stein. Spatio-temporal Analysis of Reverted Wikipedia Edits. Association for the Advancement of Artificial Intelligence, 2017
7. Joshua E. Blumenstock. Size Matters: Word Count as a Measure of Quality on Wikipedia. WWW 2008 / Poster Paper April 21-25, 2008 · Beijing, China