# Wikipedia Text Reuse: Within and Without

Milad Alshomary,[1] Michael Völske,[2] Tristan Licht,[2] Henning Wachsmuth,[1]
Benno Stein,[2] Matthias Hagen,[3] and Martin Potthast[4]

[1]Paderborn University
[2]Bauhaus-Universität Weimar
[3]Martin-Luther-Universität Halle-Wittenberg
[4]Leipzig University

**Abstract.** We study text reuse related to Wikipedia at scale by compiling the first corpus of text reuse cases within Wikipedia as well as without (i.e., reuse of Wikipedia text in a sample of the Common Crawl). To discover reuse beyond verbatim copy and paste, we employ state-of-the-art text reuse detection technology, scaling it for the first time to process the entire Wikipedia as part of a distributed retrieval pipeline. We further report on a pilot analysis of the 100 million reuse cases inside, and the 1.6 million reuse cases outside Wikipedia that we discovered. Text reuse inside Wikipedia gives rise to new tasks such as article template induction, fixing quality flaws, or complementing Wikipedia's ontology. Text reuse outside Wikipedia yields a tangible metric for the emerging field of quantifying Wikipedia's influence on the web. To foster future research into these tasks, and for reproducibility's sake, the Wikipedia text reuse corpus and the retrieval pipeline are made freely available.

## 1 Introduction

Text reuse is second nature to Wikipedia: *inside* Wikipedia, the articles grouped in a given category are often harmonized until informal templates emerge, which are then adopted for newly created articles in the same category. Moreover, passages may even be copied verbatim from one article to another when they form a hierarchical relationship. While the reuse of text inside Wikipedia has been a de facto policy for many years, neither the MediaWiki software nor tools developed by and for the Wikipedia community offer any reuse support. Unless a dedicated Wikipedia editor takes care of it, a copied passage will eventually diverge from its original, resulting in inconsistency. *Outside* Wikipedia, we distinguish reuse of Wikipedia's articles by third parties, and reuse of third-party content by Wikipedia. The former is widespread: passages of articles are manually reused in quotations and summaries, or automatically extracted to search result pages. Many sites mirror Wikipedia partially or in full; sometimes with proper attribution, other times violating Wikipedia's lenient copyrights.[1] The latter form of reuse is discouraged by Wikipedia's editing policies.[2]

[1] en.wikipedia.org/wiki/Wikipedia:Copyrights
[2] en.wikipedia.org/wiki/Wikipedia:Copyright_violations

With a few exceptions reviewed below, Wikipedia text reuse has not been analyzed at scale. This gap is due to the lack of open and scalable technologies capable of detecting text reuse, and the significant computational overhead required. Only recently, resulting from six consecutive shared tasks on plagiarism detection held at PAN to systematically evaluate reuse detection algorithms, new classes of algorithms emerged that specifically address the detection of various kinds of text reuse from large text corpora. To foster research into Wikipedia text reuse, we compiled the first Wikipedia text reuse corpus, obtained from comparing the entire Wikipedia to itself as well as to a 10%-sample of the Common Crawl. By scaling up the aforementioned detection algorithms, we render the computations feasible on a mid-sized cluster. A first exploratory analysis enables us to report insights on the nature of text reuse inside Wikipedia, and to quantify Wikipedia's influence on the web in terms of monetary exploitation of its content.

## 2 Related Work

Wikipedia's openness and success fuels tons of research about the encyclopedia[3] and how it can be exploited in different fields [8,12]. Wikipedia's influence on the web has recently become a focus of interest: for instance, posts on Stack Overflow and Reddit that link to Wikipedia have been found to outperform others in terms of interactions [20]. Other works have studied Wikipedia's role in driving research in the scientific community [19], and its importance to enrich search engines' result pages [11]. The ever increasing quality of Wikipedia drives the reuse of its content by third parties, but in a "paradox of reuse" reduces the need to visit Wikipedia itself [18], depriving the encyclopedia of potential new editors.

In general, text reuse detection is applied in many domains [2], such as the digital humanities [7], and in journalism and science (e.g., to study author perspectives [6] or to pursue copyright infringement and plagiarism [5]). Text reuse detection divides into the subtasks of *source retrieval* and *text alignment* [14,17], where the former retrieves a set of candidate reuse sources given a questioned document [9], and the latter aligns reused passages given a document pair. Approaches addressing each task have been systematically evaluated at PAN [14].

As for Wikipedia, text reuse detection has the potential to help improve the encyclopedia and to quantify its influence on the web. However, Wikipedia text reuse has only been targeted in two pioneering studies to date: Weissman et al. [21] use similarity hashing to identify redundant or contradictory near-duplicate sentences within Wikipedia that may harm article quality. Similarly, Ardi and Heidemann [1] employ hashing to detect near-duplicates of complete Wikipedia articles in the Common Crawl. Both studies neglect the text alignment step, restricting the ability to perform in-depth reuse analysis. Our text reuse detection pipeline incorporates similar hashing techniques for source retrieval but further filters and refines the results through text alignment to obtain the fine-grained actual reused text passages. In this respect, our corpus better captures the author's intent of reusing a given passage of text.

[3] en.wikipedia.org/wiki/Academic_studies_about_Wikipedia

## 3 Corpus Construction

Given two document collections $D_1$ and $D_2$, we aim to identify all cases of text reuse as pairs of sufficiently similar text spans. For within-Wikipedia detection, $D_1$ is the set of all English Wikipedia articles and $D_2 = D_1$, whereas otherwise $D_2$ is a 10%-sample of the Common Crawl (see Table 1 (left)). Our processing pipeline first carries out *source retrieval* to identify promising candidate document pairs, which are then compared in detail during *text alignment*.

### 3.1 Source Retrieval

In source retrieval, given a questioned document $d_1 \in D_1$, the task is to rank the documents in $D_2$ by decreasing likelihood of sharing reused text with $d_1$. An absolute cutoff rank $k$ and/or a relative score threshold $\tau$ may be used to decide how many of the top-ranked $D_2$-documents become subject to the more expensive task of text alignment with $d_1$. The parameters are typically determined in terms of the budget of computational capacity available as well as the desired recall level. An ideal ranking function would rank all documents in $D_2$ that reuse text from $d_1$ highest; however, the typical operationalization using text similarity measures does not reach this ideal. The higher the desired recall level, the lower the precision and the higher the computational overhead.

With a goal of maximizing recall, our budget was 2 months of processing on a 130 node Apache Spark cluster (12 CPUs and 196 GB RAM each). Since Wikipedia as a whole is questioned ($D_1$), we generalized source retrieval toward ranking all pairs $(d_1, d_2) \in D_1 \times D_2$ based on a pruned scoring function $\rho$:

$$\underbrace{\exists c_i \in d_1, c_j \in d_2 \colon \quad h(c_i) \cap h(c_j) \neq \emptyset}_{\text{Search pruning}} \quad \rightarrow \quad \rho(d_1, d_2) = \max_{\substack{c_i \in d_1 \\ c_j \in d_2}} (\varphi(c_i, c_j)),$$

where $c$ is a passage-length text chunk, $h$ is a locality-preserving hash function, and $\varphi$ is a text similarity measure. The idea is to view reuse as a passage-level phenomenon and to be lenient during pruning (a single hash collision suffices).

To select and fine-tune a suitable hash function $h$ and similarity measure $\varphi$, we compiled a ground truth training set, by sampling 1000 Wikipedia articles—each at least 2000 words long—and applying our text alignment approach described below to all their pairs with all other Wikipedia articles. The source retrieval "parameters" $h$ and $\varphi$ (and thus $\rho$) were optimized to maximize recall of these training set text alignment results in the source retrieval phase. We considered two hashing schemes for $h$: (1) random projections in the form of an instantiation of the data-independent locality-sensitive hashing (LSH) family [4], and (2) variational deep semantic hashing (VDSH), a data-dependent learning-to-hash technique [3]. We further considered four text similarity measures for $\varphi$: (a) cosine similarity on a $tf \cdot idf$-weighted word unigram representation, (b) Jaccard similarity on stop word $n$-grams [16], (c) cosine similarity on a simple additive paragraph vector model [13], and (d) a weighted average of (b) and (c).

**Table 1.** Overview of the input dataset characteristics (left), the source retrieval performance (middle), and the retrieved text reuse cases (right).

| Dataset | Count (million) |
|---|---|
| *Wikipedia* | |
| Articles | 4.2 |
| Paragraphs | 11.4 |
| *Common Crawl* | |
| Websites | 1.4 |
| Web pages | 591.0 |
| Paragraphs | 187.0 |

| Source Retrieval | Recall | Precision |
|---|---|---|
| *Search pruning* | | |
| (1) LSH | 0.32 | $9.8{\cdot}10^{-6}$ |
| **(2) VDSH** | **0.73** | $\mathbf{4.5{\cdot}10^{-4}}$ |
| *Ranking up to rank* $k = 1000$ | | |
| **(a)** $\boldsymbol{tf \cdot idf}$ | **0.87** | **0.007** |
| (b) Stop $n$-grams | 0.74 | 0.007 |
| (c) Par2vec | 0.67 | 0.008 |
| (d) Hybrid | 0.76 | 0.009 |
| **VDSH +** $\boldsymbol{tf \cdot idf}$ | **0.66** | **0.005** |

| Reuse | Within | Without |
|---|---|---|
| Cases | 110 million | 1.6 million |
| *Documents with Reuse Cases* | | |
| Articles | 360,000 | 1 million |
| Pages | – | 15,000 |
| *Words in Reuse Cases* | | |
| Min. | 17 | 23 |
| Avg. | 78 | 252 |
| Max. | 6200 | 1960 |

Table 1 (middle) shows our evaluation results for the two components of the source retrieval pipeline. In general, the low precision values are due to the high cut-off rank (k=1000) required to collect most of the few positive cases. For search pruning, we selected VDSH with a 16-bit hash, which reduces the number of required evaluations of the $\rho$ measure by three orders of magnitude compared to an exhaustive comparison, while retaining the majority of text reuse cases. To construct the ranking function $\rho$ itself, we settle on cosine similarity in the $tf \cdot idf$ space as the similarity measure $\varphi$ due to its superior recall compared to the other considered models.

### 3.2 Text Alignment

Given a candidate document pair, text alignment extracts spans of reused text—if any—through the steps of *seed generation* (identification of short exact matches), *seed extension* (clustering of short matches to form longer spans), and *post filtering*. The state of the art evaluated at PAN is determined on datasets orders of magnitude smaller than our setting, often using complex setups that turned out to be difficult to scale and to be reproduced (e.g., lacking open source implementations). We hence resorted to ideas from the literature that offer a reasonable trade-off between performance, robustness, and speed, and tuned their parameters[4] based on the standard PAN-13 training data. Our text alignment achieves a macro-averaged *plagdet* score of 0.64 (0.84 on just the unobfuscated subset) on the corresponding PAN-13 test data. In terms of raw detection performance, this is in the lower middle range of the PAN results [15].

In our pruned all-pairs search setting, an input to the text alignment step is formed by one document $d \in D_1$ and a list of all candidate documents from the other collection $D_2$ sorted by descending $\rho$-score. Text alignment is applied sequentially to this list until one of two stopping criteria is met: (1) the current candidate pair's $\rho$-score is below a threshold (0.025 in our implementation), or (2) the number of consecutive miss-cases (i.e., candidate pairs in which the text alignment finds no reuse) exceeds some other threshold (we use 250). Both thresholds can be configured based on the time available for text alignment; we experimentally extrapolated them from the aforementioned training set.

[4] We used word 3-gram seeds, extended via DBScan clustering ($\varepsilon = 150$, minPoints $= 5$), and filtered cases shorter than 200 words or with cosine similarity $< 0.5$.

## 4 Corpus Analysis

Table 1 (right) shows basic statistics of the reuse we uncovered. Most interestingly, we find nearly 70 times more reuse cases within the Wikipedia than in the 10%-sample of the Common Crawl, but involving only one third as many articles. Based on this insight, we identify two fundamentally different kinds of text reuse within Wikipedia—the first making up for the bulk of the discrepancy. When articles use the same structure but different facts (e.g., geographical locations described in terms of their surroundings), we refer to this as *structure reuse* (Table 2, top left) and consider such cases as non-problematic (perhaps unavoidable) redundancy. On the other hand, articles may contain factually nearly-identical passages, likely after copying from one to the other. We consider such *content reuse* likely to result in inconsistency and contradiction as the articles may diverge over time (Table 2, bottom left). Ideally, such redundant sections should be replaced with a single, authoritative source. In this sense, text reuse analysis can help the Wikipedia community locate and improve articles with undesirable redundancy.

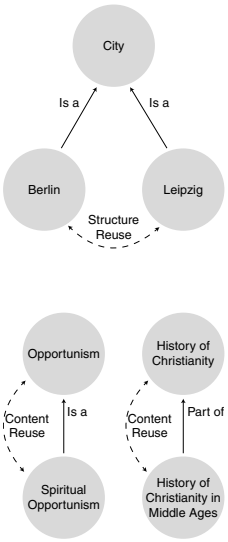
Further observations indicate that the ontological relationship between articles' topics correlates with the type of text reuse: Structure reuse occurs more frequently when articles represent concepts on the same level in the ontology tree (Table 2, top right), while two articles whose subjects are vertically aligned (e.g., "is a" or "part of" relationships) are more likely to exhibit content reuse (Table 2, bottom right). The latter association can also be envisioned as a solution to the sub-article matching task [10]: the occurrence of content reuse between articles can serve as an indicator of the ontological relationship between the concepts that they represent. However, automatically distinguishing content and structure reuse is not trivial. Our initial attempt at classifying reuse cases used a heuristic based on the ratio of reused to original text in the articles, as well as the Jaccard similarities between the sets of named entities and word 10-grams. Using two samples of 100 random structure reuse cases and 100 random content reuse cases, the heuristic achieved 100% precision for structure reuse, but only 57% for content reuse. While our heuristics identify 95.5 million (87%) of all within-Wikipedia reuse cases as structure reuse, the true number likely exceeds 100 million assuming our error estimates are accurate.

In the 10%-sample of the Common Crawl, 4,898 websites host at least one page that reuses text from a Wikipedia article for a total of 1.6 million cases.[5] We presume that Wikipedia's policy of avoiding reuse from third parties inside its articles is enforced by its editors, so that nearly all of the cases will be third parties reusing Wikipedia's articles instead. Most (94%) of the pages violate the terms of Wikipedia's license[6] by not referencing Wikipedia as a source (i.e., the term "Wikipedia" does not occur). With only a handful exceptions, such as un.org, all of the sites display advertisements, which extends to the pages containing the reuse. Furthermore, in nearly all of the cases, the reuse accounted for more than 90% of the main content, prompting usefulness questions.

[5] The top three being wikia.com (563), rediff.com (55), and un.org (28 reusing pages).
[6] en.wikipedia.org/wiki/Wikipedia:Reusing_Wikipedia_content

**Table 2.** Examples of the two types of text reuse within Wikipedia—structure reuse (top) and content reuse (bottom)—and corresponding ontological article relations (right).

| **Title:** Niedźwiedzie, Pisz County | **Title:** Zimna Woda, Zgierz County |
|---|---|
| Niedźwiedzie is a village in the administrative district of Gmina Pisz, within Pisz County, Warmian-Masurian Voivodeship, in northern Poland. It lies approximately south-east of Pisz and east of the regional capital Olsztyn. | Zimna Woda is a village in the administrative district of Gmina Zgierz, within Zgierz County, Łódź Voivodeship, in central Poland. It lies approximately north-west of Zgierz and north-west of the regional capital Łódź. |
| **Title:** Human tooth development | **Title:** Tooth eruption |
| Tooth eruption has three stages. The first, known as deciduous dentition stage, occurs when only primary teeth are visible. Once the first permanent tooth erupts into the mouth, the teeth are in the mixed (or transitional) dentition. [...] Primary dentition stage starts on the arrival of the mandibular central incisors, typically from around six months, and lasts until the first permanent molars appear [...] | The dentition goes through three stages. The first, known as primary dentition stage, occurs when only primary teeth are visible. Once the first permanent tooth erupts into the mouth, the teeth that are visible are in the mixed (or transitional) dentition stage. [...] Primary dentition starts on the arrival of the madibular central incisors, usually at eight months, and lasts until the first permanent molars appear [...] |




We conservatively estimate the potential advertisement revenue generated by the reused Wikipedia content. For simplicity, we assume that all reusing websites host only one ad per page and that advertisements are billed according to cost per mille (CPM), achieving a revenue per mille (RPM) of about half (1.4 USD) the average estimated CPM on the web in 2018 (2.8 USD).[7] Accounting for the fact that reusing pages are generally ranked lower than Wikipedia in search results, we use 10% of the monthly page view counts of reused articles (as per Wikipedia's API) as estimates for the page views of reusing pages. With these approximations, we arrive at an estimate of 45,000 USD monthly ad revenue generated by the detected 4,898 reusing sites. Extrapolated to the entire web (say, 600,000 reusing sites out of 180 million active sites as per netcraft.com), we arrive at 5.5 million USD estimated monthly ad revenue; which adds up to about 72.5% of Wikipedia's worldwide fundraising returns in the fiscal year 2016–2017.[8]

## 5 Conclusion

In an effort to bring text reuse analysis to very large corpora, we propose a scalable pipeline comprising the source retrieval and text alignment subtasks. We address challenges of scale primarily in the former via candidate filtering, and evaluate a set of hashing and text similarity techniques for this purpose. Our framework and the two compiled text reuse datasets—within Wikipedia and in a 10%-sample of the Common Crawl—are publicly available.[9] This way, we hope to stimulate future research targeting Wikipedia quality improvement (e.g., by template induction or automatic detection of reuse inconsistencies) and understanding Wikipedia's influence on the web at large.

[7] monetizepros.com/cpm-rate-guide/display/
[8] foundation.wikimedia.org/wiki/2016-2017_Fundraising_Report
[9] github.com/webis-de/ECIR-19, webis.de/data/webis-wikipedia-text-reuse-18.html

## References


1. Ardi, C., Heidemann, J.: Web-scale content reuse detection (extended). USC/Information Sciences Institute, Tech. Rep. ISI-TR-692 (2014)
2. Bendersky, M., Croft, W.: Finding text reuse on the web. In: Proceedings of WSDM 2009, pages 262–271.
3. Chaidaroon, S., Fang, Y.: Variational deep semantic hashing for text documents. arXiv preprint:1708.03436 (2017)
4. Charikar, M.S.: Similarity estimation techniques from rounding algorithms. In: Proceedings of STOC 2002, pages 380–388.
5. Citron, D.T., Ginsparg, P.: Patterns of text reuse in a scientific corpus. PNAS **112**(1), 25–30 (2015)
6. Clough, P.D., Wilks, Y.: Measuring text reuse in a journalistic domain. In: Proceedings of the CLUK Colloquium 2001.
7. Coffee, N., Koenig, J.P., Poornima, S., Forstall, C.W., Ossewaarde, R., Jacobson, S.L.: The Tesserae project: Intertextual analysis of Latin poetry. Literary and Linguistic Computing **28**(2), 221–228 (2012)
8. Generous, N., Fairchild, G., Deshpande, A., Del Valle, S., Priedhorsky, R.: Global disease monitoring and forecasting with Wikipedia. PLoS Computational Biology **10**(11), e1003892 (2014)
9. Hagen, M., Potthast, M., Adineh, P., Fatehifar, E., Stein, B.: Source retrieval for web-scale text reuse detection. In: Proceedings of CIKM 2017, pages 2091–2094
10. Lin, Y., Yu, B., Hall, A., Hecht, B.: Problematizing and addressing the article-as-concept assumption in Wikipedia. In: Proceedings of CSCW 2017, pages 2052–2067.
11. McMahon, C., Johnson, I.L., Hecht, B.J.: The substantial interdependence of Wikipedia and Google: A case study on the relationship between peer production communities and information technologies. In: Proceedings of ICWSM 2017, pages 142–151.
12. Mestyán, M., Yasseri, T., Kertész, J.: Early prediction of movie box office success based on Wikipedia activity big data. PLoS One **8**(8), e71226 (2013)
13. Mitchell, J., Lapata, M.: Vector-based models of semantic composition. In: Proceedings of ACL 2008, pages 236–244.
14. Potthast, M., Gollub, T., Hagen, M., Tippmann, M., Kiesel, J., Rosso, P., Stamatatos, E., Stein, B.: Overview of the 5th international competition on plagiarism detection. In: Working Notes Papers of the CLEF 2013 Evaluation Labs.
15. Potthast, M., Hagen, M., Beyer, A., Busse, M., Tippmann, M., Rosso, P., Stein, B.: Overview of the 6th international competition on plagiarism detection. In: Working Notes Papers of the CLEF 2014 Evaluation Labs.
16. Stamatatos, E.: Plagiarism detection using stopword $n$-grams. JASIST **62**(12), 2512–2527 (2011).
17. Stein, B., Meyer zu Eißen, S., Potthast, M.: Strategies for retrieving plagiarized documents. In: Proceedings of SIGIR 2007, pages 825–826.
18. Taraborelli, D.: The sum of all human knowledge in the age of machines: A new research agenda for Wikimedia. In: Proceedings of the ICWSM 2015 Workshop Wikipedia, a Social Pedia: Research Challenges and Opportunities.
19. Thompson, N., Hanley, D.: Science is shaped by Wikipedia: Evidence from a randomized control trial. MIT Sloan Research Paper No. 5238-17. (2018)
20. Vincent, N., Johnson, I., Hecht, B.: Examining Wikipedia with a broader lens: Quantifying the value of Wikipedia's relationships with other large-scale online communities. In: Proceedings of CHI 2018, pages 566:1–566:13.

21. Weissman, S., Ayhan, S., Bradley, J., Lin, J.: Identifying duplicate and contradictory information in Wikipedia. In: Proceedings of JCDL 2015, pages 57–60.