Peer-reviewed version available at *Computers* **2019**, *8*, 60; doi:10.3390/computers8030060

Article

Multilingual Ranking of Wikipedia Articles with Quality and Popularity Assessment in Different Topics

Włodzimierz Lewoniewski[1,†,‡]*, Krzysztof Węcel[1,‡] and Witold Abramowicz[1,‡]

[1] Poznań University of Economics and Business;
{wlodzimierz.lewoniewski,krzysztof.wecel,witold.abramowicz}@ue.poznan.pl
* Correspondence: wlodzimierz.lewoniewski@ue.poznan.pl; Tel.: +48 (61) 639-27-93
† Current address: al. Niepodległości 10, 61-875 Poznań, Poland
‡ These authors contributed equally to this work.

Abstract: In Wikipedia, articles about various topics can be created and edited independently in each language version. Therefore, quality of information about the same topic depends on language. Any interested user can improve an article and that improvement may depend on popularity of the article. The goal of this study is to show what topics are best represented in different language versions of Wikipedia using results of quality assessment for over 39 million articles in 55 languages. In this paper, we also analyze how popular are selected topics among readers and authors in various languages. We used two approaches to assign articles to various topics. First, we selected 27 main multilingual categories and analyzed all their connections with sub-categories based on information extracted from over 10 million categories in 55 language versions. To classify the articles to one of the 27 main categories we took into account over 400 million links from articles to over 10 million categories and over 26 million links between categories. In the second approach we used data from DBpedia and Wikidata. We also showed how the results of the study can be used to build local and global rankings of the Wikipedia content.

Keywords: Wikipedia; Information quality; Popularity; Topics identification; Wikidata; DBpedia; WikiRank

1. Introduction

Nowadays, in order to make the right economic decisions, one needs to analyze and interpret vast amount of information. The quantity and quality of information to a large extent determine the quality of decisions in various branches of the economy. On the one hand, one must take care of access to proper sources of information. On the other hand, the quality of information determined by various characteristics is also important. High-quality information is essential for effective operation and decision-making in organizations [1]. Inaccurate and incomplete information may have a negative impact on a company's competitive edge [2].

The Internet enables cooperation and exchange of information on a global scale. Useful information can be found both in specialized sources as well as in general online resources. Nowadays, everyone can also contribute to the development of common human knowledge on the Internet. One of the best examples of such online repositories is Wikipedia, in which content can be created from the level of a web browser. This online encyclopedia has been available for approximately 20 years as a freely available resource, and anyone willing can co-create content. Wikipedia relatively quickly became an important source of information around the world. It contains over 50 million articles in over 300 different languages [3]. The English language version is the largest and contains over 5.8

million articles. Currently, Wikipedia is placed on the fifth place in the ranking of the most visited websites on the Internet [4], giving way only to Google, YouTube, Facebook, and Baidu.

The popularity of Wikipedia is even reflected language that scientists use in their works [5]. Despite its popularity, Wikipedia is often criticized for the low quality of content [6]. Articles on a specific subject (a thing, a human, an event etc.) can be created and edited independently in each language version. Therefore, quality of information about the same subject often varies depending on the language [7–10]. It should also be noted that the topic described in one language version can be translated into other languages. However, a relatively small number of users with knowledge of two or more languages take up such an initiative by transferring content between different language versions [11].

Even the largest English Wikipedia does not contain information about all subjects. As we can see in Figure 1, there are over 15 million unique subjects described in at least one of 55 considered language versions. This can be explained by the fact that some issues may be more common in smaller geographical areas, hence the probability of finding more information on a given topic in the relevant language versions (other than English). Overall, we can find almost 10 million subjects that are not covered in English and appear in less-developed versions of Wikipedia [7,12].

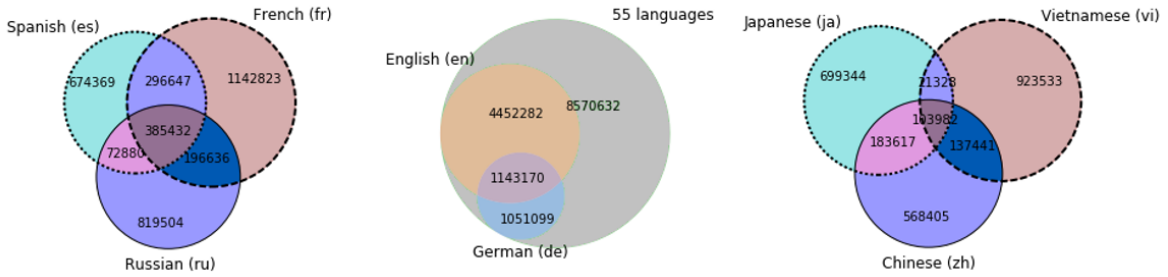

Figure 1. Subjects overlaps of articles in various language versions of Wikipedia. Source: own calculation based on Wikipedia dumps in April, 2019. Over 175 thousand of interactive combinations of these Venn diagrams can be found on the Web page: http://data.lewoniewski.info/computers/vn1/

When a subject is not described in the analyzed language version or information about the subject is of low quality, we can try to find information about it in other Wikipedia languages. However, identifying a language version best describing the subject may require significant effort from user – popular subjects are available in several dozen language versions.

Automatic quality assessment of Wikipedia articles is a known challenge in the scientific community. Existing works have some limitations, e.g. they focus mostly on the biggest edition (English) or other popular language versions of Wikipedia. Usually the measurement of quality is reduced to analysis of volume of content – number of important elements that the article must contain (such as references, images, sections). However, for quality assessment content must be checked by other users in terms of the neutral point of view, timeliness, quality of sources and other important elements that can be challenging even with current approaches. Therefore, the popularity of the article may be another factor to be considered for quality assessment – the more users read the content, the greater probability of introducing amendments to the article, especially when incorrect or outdated information is detected.

In this paper, we present the assessment of quality and popularity of Wikipedia articles in different languages related to selected topics. This assessment was performed for articles on two levels: within each considered language version (local) and for all languages combined (global).

For the purpose of this study we selected 55 language versions of Wikipedia that in 2018 and 2019 had at least 100 thousand articles and the depth indicator was at least 5. The depth (or editing depth) shows how frequently articles are updated in a specific language version [13]. Table 1 presents basic statistics about 55 language versions of Wikipedia that were considered in the study.

Peer-reviewed version available at Computers 2019, 8, 60; doi:10.3390/computers8030060

Table 1. 55 language versions of Wikipedia with articles count, views from unique devices and total page views (based on dump April 2019)

No.	Language version	Abbr.	Articles	Authors	Total page views	Unique devices
1	English	en	5 835 946	36 031 942	7 846 676 922	866 456 515
2	Swedish	sv	3 748 546	664 601	102 423 252	12 597 043
3	German	de	2 288 148	3 158 210	975 590 897	114 380 633
4	French	fr	2 094 723	3 405 365	742 709 055	96 553 550
5	Dutch	nl	1 962 531	986 565	155 136 113	23 873 475
6	Russian	ru	1 539 411	2 500 221	896 358 323	96 537 026
7	Italian	it	1 518 702	1 803 513	544 481 445	53 459 817
8	Spanish	es	1 514 431	5 375 409	1 090 438 930	180 071 200
9	Polish	pl	1 329 622	949 766	278 226 329	29 262 659
10	Vietnamese	vi	1 205 176	660 020	68 454 735	16 396 173
11	Japanese	ja	1 145 838	1 462 052	1 043 323 322	98 636 732
12	Chinese	zh	1 051 874	2 709 195	412 676 457	52 328 429
13	Portuguese	pt	1 007 942	2 230 598	352 570 671	69 605 320
14	Ukrainian	uk	896 476	448 345	62 906 361	10 849 975
15	Arabic	ar	715 850	1 643 146	188 230 435	39 994 487
16	Persian	fa	671 576	812 855	142 075 761	21 993 488
17	Serbian	sr	618 230	240 802	27 054 615	4 776 849
18	Catalan	ca	610 217	319 681	21 121 481	3 439 969
19	Norwegian (Bokmål)	no	506 510	457 767	36 974 998	6 017 919
20	Indonesian	id	458 034	1 047 391	146 481 271	33 774 831
21	Finnish	fi	454 859	413 533	65 437 832	7 372 105
22	Korean	ko	450 896	559 608	83 623 819	19 933 158
23	Hungarian	hu	448 744	133 232	54 741 921	8 298 454
24	Serbo-Croatian	sh	447 790	409 910	5 900 087	2 372 396
25	Czech	cs	425 852	448 816	73 574 810	9 338 114
26	Romanian	ro	393 439	470 902	39 466 674	7 711 157
27	Basque	eu	332 997	98 920	9 067 706	446 209
28	Turkish	tr	325 627	233 118	25 389 323	3 076 606
29	Malay	ms	325 592	1 028 128	12 291 727	3 960 414
30	Esperanto	eo	256 487	156 711	1 981 767	263 084
31	Bulgarian	bg	254 272	84 451	27 272 998	4 093 761
32	Danish	da	250 890	249 638	30 667 722	5 190 512
33	Armenian	hy	248 278	349 917	6 013 622	918 474
34	Hebrew	he	240 943	507 618	58 213 949	6 344 428
35	Slovak	sk	229 146	171 238	16 854 614	3 117 661
36	Min Nan	zh-min-nan	228 102	37 919	572 773	84 788
37	Kazakh	kk	223 881	85 934	11 562 925	2 142 268
38	Croatian	hr	204 240	216 016	21 779 929	4 497 371
39	Lithuanian	lt	194 537	131 095	12 276 882	1 984 922
40	Estonian	et	189 742	125 754	11 502 319	1 187 671
41	Belarusian	be	166 775	84 971	1 711 658	253 243
42	Slovenian	sl	164 036	178 042	8 497 867	1 491 437
43	Greek	el	160 482	271 125	34 866 919	6 330 938
44	Galician	gl	155 573	96 617	2 533 863	512 368
45	Azerbaijani	az	145 060	172 093	12 826 807	1 748 834
46	Urdu	ur	144 942	93 377	2 916 140	506 414
47	Simple English	simple	144 053	823 355	19 179 047	9 071 802
48	Norwegian (Nynorsk)	nn	142 635	95 945	1 733 721	563 079
49	Uzbek	uz	130 990	44 264	3 256 673	569 355
50	Thai	th	130 723	349 695	63 983 646	14 758 190
51	Hindi	hi	130 443	444 004	56 017 398	17 087 729
52	Latin	la	130 327	117 110	1 086 052	173 591
53	Georgian	ka	127 899	109 531	8 642 199	1 147 871
54	Volapük	vo	122 757	26 048	266 020	38 888
55	Tamil	ta	121 501	152 024	8 357 708	2 295 703

2. Topic Classifications of Wikipedia Articles

2.1. Category Classification

Wikipedia has extensive category network and each article can be annotated with multiple categories, organized into an "ontology of topics" [14]. Each language version can define own structure and hierarchy of categories. Moreover, in some language versions that structure is often too fine-grained to be directly analyzed [15]. All this may make it difficult to determine the number of possible topics to deal with.

Category structure and alignment of articles to each category can be analyzed based on files from Wikipedia dumps. There are three files that has to be used (example for English Wikipedia):

- **enwiki-latest-category.sql.gz** – category information; here we use category identifiers and their names;
- **en-latest-categorylinks.sql.gz** – wiki category membership link records; here we use information about source page ID and destination category name;
- **en-latest-page.sql.gz** – base per-page data; here we use pages ID, title and information about namespaces to identify articles (ns 0) and category (ns 14) pages.

For further research we extracted information about over 10 million articles in 55 language versions and analyzed about 400 million links from articles to categories and over 26 million links between categories. General statistics about categories are presented in Table 2. Category ratio shows the number of unique categories per number of articles in a particular language version. The highest value of this indicator has Urdu Wikipedia - 1.23. The largest English Wikipedia is in the middle in the ranking regarding the value of this indicator.

Table 2. Number of categories, number of links from articles to categories and between categories in 55 lanquage versions of Wikipedia (sorted by category density). Source: own calculations in April, 2019

Wikipedia language	Number of categories		Category ratio	Number of links		Average number of categories per article
	all	without page		from articles to categories	between categories	
Urdu (ur)	178271	8836	1.230	1048967	775590	7.237
Arabic (ar)	576872	6368	0.806	21548319	1982157	30.102
Persian (fa)	499231	37	0.743	9748824	1568018	14.516
Turkish (tr)	226145	10383	0.694	2322792	542366	7.133
Belarusian (be)	115205	33807	0.691	1182398	193168	7.090
Norwegian (Nynorsk) (nn)	88804	18156	0.623	789450	158280	5.535
Korean (ko)	268761	20773	0.596	4462341	652764	9.897
Thai (th)	73106	25130	0.559	922356	118369	7.056
Georgian (ka)	65047	15317	0.509	435646	103973	3.406
Slovenian (sl)	77146	21649	0.470	1078180	119567	6.573
Azerbaijani (az)	65627	2104	0.452	906108	127144	6.246
Hindi (hi)	54785	30507	0.420	593496	50673	4.550
Indonesian (id)	186977	102406	0.408	5279994	185266	11.528
Galician (gl)	62109	577	0.399	689762	120190	4.434
Chinese (zh)	395448	101111	0.376	12793208	716798	12.162
Greek (el)	60056	3826	0.374	1218241	156199	7.591
Armenian (hy)	87522	25729	0.353	1601227	136013	6.449
Czech (cs)	140757	665	0.331	2730698	333870	6.412
Esperanto (eo)	83331	15727	0.325	1136030	184428	4.429
Portuguese (pt)	316318	11293	0.314	9346482	751718	9.273
Slovak (sk)	70586	76	0.308	919689	199717	4.014
Russian (ru)	469180	53068	0.305	17351449	929165	11.271
Hebrew (he)	71150	25	0.295	2310076	170736	9.588
Norwegian (Bokmål) (no)	148816	6509	0.294	4182237	340251	8.257
English (en)	1711545	97	0.293	127118195	5545938	21.782
Latin (la)	38187	89	0.293	628280	76726	4.821
Romanian (ro)	115325	26231	0.293	3398779	274858	8.639
Malay (ms)	91578	62870	0.281	1393588	59264	4.280
Simple English (simple)	40052	477	0.278	778386	101112	5.403
Ukrainian (uk)	248614	46181	0.277	7008669	538437	7.818
Bulgarian (bg)	68898	2624	0.271	1291378	150452	5.079
Spanish (es)	398828	23074	0.263	9103226	903999	6.011
Tamil (ta)	30477	7661	0.251	483546	41080	3.980
Danish (da)	62490	5005	0.249	1861533	156608	7.420
Vietnamese (vi)	276936	101173	0.230	7745566	476364	6.427
Italian (it)	348216	32	0.229	14715516	847583	9.690
Basque (eu)	73827	19206	0.222	1497904	170504	4.498
French (fr)	425707	76	0.203	38654880	2583394	18.453
Japanese (ja)	232881	20231	0.203	8060212	551980	7.034
Kazakh (kk)	45512	23083	0.203	1660294	41958	7.416
Estonian (et)	29889	441	0.158	553027	53933	2.915
Finnish (fi)	72006	280	0.158	2707673	157913	5.953
German (de)	354701	29	0.155	12255563	886269	5.356
Polish (pl)	205391	206	0.154	5310093	399299	3.994
Min Nan (zh-min-nan)	32592	14516	0.143	608969	46280	2.670
Hungarian (hu)	60203	30	0.134	2895750	111067	6.453
Lithuanian (lt)	24721	316	0.127	541911	45874	2.786
Catalan (ca)	75951	168	0.124	2672097	179483	4.379
Serbo-Croatian (sh)	45527	374	0.102	1520947	101515	3.397
Serbian (sr)	59254	10899	0.096	4355457	106286	7.045
Swedish (sv)	354075	16	0.094	20002023	639059	5.336
Croatian (hr)	19065	53	0.093	503920	32903	2.467
Uzbek (uz)	12026	4001	0.092	832321	12758	6.354
Dutch (nl)	114899	10	0.059	10060345	320354	5.126
Volapük (vo)	2440	269	0.020	353343	2878	2.878

Another measure that can be useful to analyze how often Wikipedia users assign different categories to describe each article is the average number of categories per article. Based on data from Table 2 we can define top three leaders: Arabic with 30, English with 21, and French with 18 categories per article.

We can also notice that in some language versions of Wikipedia there is a large number of categories that do not have own page that describes these categories and point to the parent category. The highest values has Vietnamese, Chinese and Indonesian Wikipedia - about 100 thousand categories without pages. For first two languages with about 1 million articles this is one fourth and one third of all categories respectively. In Indonesian with about 460 thousand articles it is about half of all categories. For comparison, the largest English version with over 5 million articles has only 97 categories without a page.

The so called main categories are present in majority of considered languages. This applies mainly to those categories that are at highest levels in the polyhierarchy. One of the main categories are presented at special page "Category:Main topic classifications" [16]. Based on this page, we can identify 38 categories on specific topics in the English Wikipedia. Table 3 shows names of these categories with number of the considered language versions. As we can see, some topics may be not available in all languages.

Table 3. List of the categories in "Category:Main topic classifications" in English Wikipedia with number of the considered language versions (April 2019)

No.	Category name	Number of considered language versions
1	Education	55
2	Geography	55
3	History	55
4	Mathematics	55
5	Music	55
6	Philosophy	55
7	Religion	55
8	Science	55
9	Society	55
10	Sports	55
11	Arts	54
12	Organizations	54
13	People	54
14	Politics	54
15	Culture	53
16	Law	53
17	Technology	53
18	Health	52
19	Military	52
20	Entertainment	51
21	Events	51
22	Food and drink	51
23	Government	49
24	Nature	49
25	Crime	48
26	Business	47
27	Life	47
28	Academic disciplines	45
29	Human behavior	44
30	Knowledge	44
31	Concepts	43
32	Language	39
33	Objects	37
34	Mind	28
35	Humanities	27
36	World	27
37	Economy	17
38	Universe	5

As mentioned before, the category structure is a complex and ever-changing, as it can be edited by any person – users can add or change a category assignment to other category. The resulting category structure is noisy [14], sparse and it contains duplications and oversights [15]. So, we can also face the situation that categories are repeated at different levels of the tree, in which the root can be another main category (one of the 27 considered). In order to avoid such situations, we cut off those branches that were found at higher levels. Figure 2 shows an example of such procedure, when subcategory "Food and Drink" is found at different levels of the tree and only one remains, which is at the highest level.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2019 doi:10.20944/preprints201905.0144.v2

Peer-reviewed version available at Computers 2019, 8, 60; doi:10.3390/computers8030060

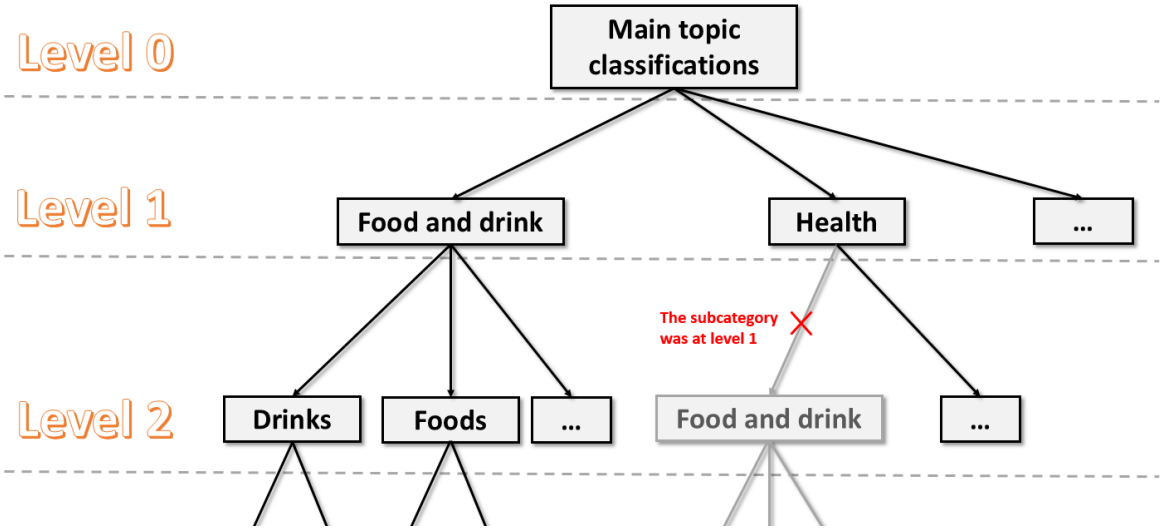

Figure 2. Occurrence of similar sub-categories in the English Wikipedia category polyhierarchy. Source: own work based on Wikipedia dumps from April 2019.

By counting articles in English Wikipedia in each of considered main categories we discovered that almost 15% of them are about people. Pie chart in Figure 3 shows shares of articles in English Wikipedia in 27 considered categories.

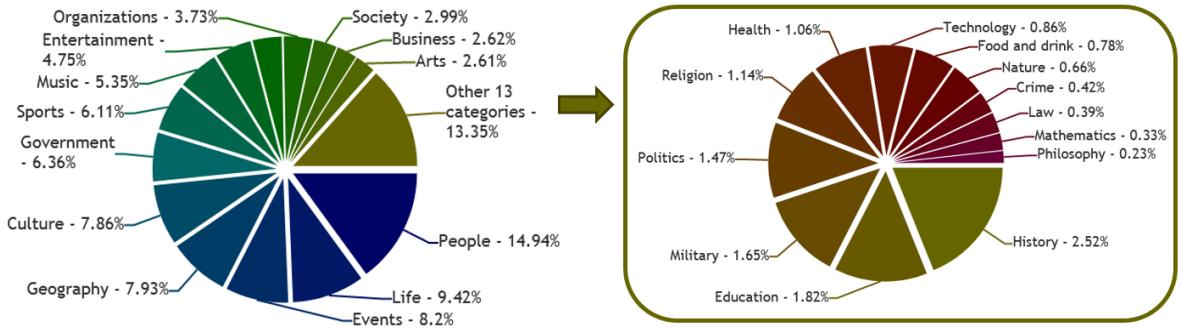

Figure 3. Shares of articles in each category in English Wikipedia. Source: own calculation based on Wikipedia dumps in April, 2019.

Figure 4 shows distribution of articles by category within each considered language version of Wikipedia. Darker colors in the heatmap represent higher share of articles in particular main category within the selected Wikipedia languages.

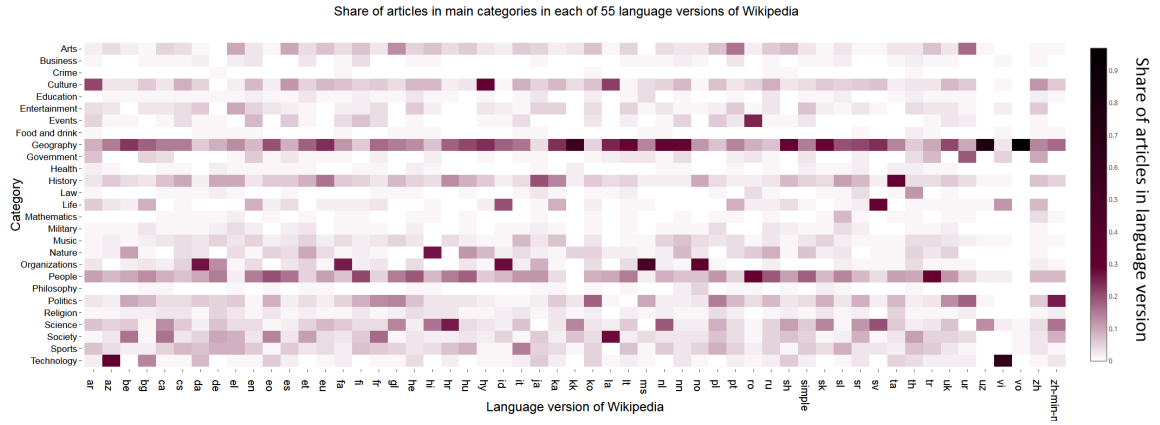

Figure 4. Share of articles in main categories within each of 55 language versions of Wikipedia. Source: own calculation based on Wikipedia dumps in April, 2019. More detailed and interactive chart can be found on the Web page: http://data.lewoniewski.info/computers/heatmap-cat-art

After combining articles from all considered language versions to particular category we concluded that the largest number of articles are in one of two categories: Geography (12.68%) and People (11.48%). Pie chart in Figure 5 presents how articles in all considered Wikipedia languages are distributed among 27 main categories.

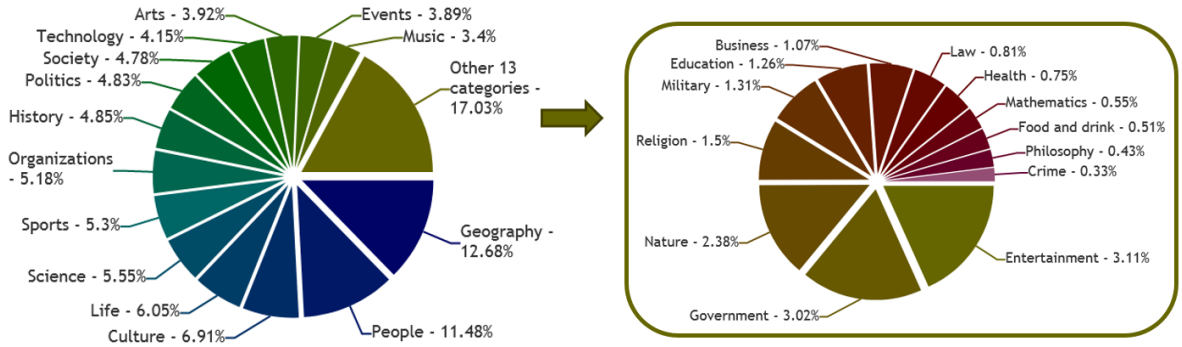

Figure 5. Shares of articles in each category in 55 language versions of Wikipedia. Source: own calculation based on Wikipedia dumps in April, 2019.

As we mentioned before, in some language versions there is a relatively high average number of categories assigned to each article. This may increase the possibility of an article falling into more than one main category. We studied this issue for the leading language versions (Arabic, English, French) with regard to the number of categories per article. Results are presented in Figure 6.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2019 doi:10.20944/preprints201905.0144.v2

Peer-reviewed version available at Computers 2019, 8, 60; doi:10.3390/computers8030060

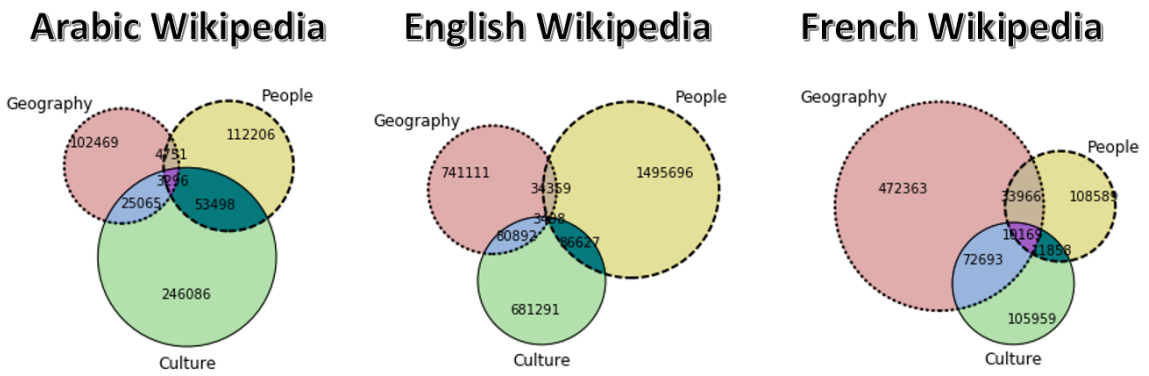

Figure 6. Overlap of articles between selected main categories in Arabic, English and French Wikipedia. Source: own calculation based on Wikipedia dumps in April, 2019. Over million of interactive combinations of these Venn diagrams (each main categories and language versions) can be found on the Web page: http://data.lewoniewski.info/computers/vn2/.

2.2. Semantic Classification

The second approach to category assignment to Wikipedia articles is based on Wikidata and DBpedia. Wikidata is a collaboratively edited knowledge base [17]. DBpedia is the semantic database resulting from extraction of structured, multilingual knowledge from Wikipedia [18,19]. The data from this open databases are widely used in a number of domains: web search, life sciences, maritime domain, art market, digital libraries, business networks and others [20–23].

DBpedia uses its own ontology with defined properties and classes organized into a hierarchy. DBpedia provides English names to each class, such as “Place”, “Species”, “Person” etc. Wikidata gives unique identifier to each class, for example class “city” is marked as Q515, “human” as Q5, “Organization” as Q43229. Another difference between these databases lies in the number of classes and placing these classes in an ontology. Wikidata has over 300 thousand classes [24], while DBpedia ontology consist of about 800 classes [25].

A significantly larger number of classes in Wikidata can lead to difficulties in finding a list of objects on a particular topic. For example, if we want to find all cities, it is not enough to take into account only one class Q515 (city), because city can also be described by Q1637706 (city with millions of inhabitants), Q5119 (capital), Q2264924 (port city), Q58339717 (city of India), Q174844 (megacity) and other identifiers. This variety of classes leads to significantly fewer instances in each class in Wikidata than in DBpedia [24].

We should consider also way of assigning a class to objects in these semantic databases. DBpedia extracts information from Wikipedia infoboxes and identifies classes based on name of the infobox and values of some special parameters. Thus, articles with the same infobox name often go to the same class. In Wikidata, items can be edited by everyone, therefore different classes can be assigned to similar objects.

There are some papers that study differences between DBpedia and Wikidata [24,26,27]. Each has own advantages, so we decided to use combined data to divide articles into separate classes: actor, automobile, business, city, film, football player, human, programming, university, videogame, and website. One of the advantages of such a classification approach by topic is that we are dealing here with more explicit assignment of articles to specific classes and each language version has at least several representatives of each class.

3. Quality Measures

In order to discern the quality of content, the Wikipedia community created a grading system for articles. However, each language version can use its own standards and grading scale [28,29]. For example, in English Wikipedia, articles can get one of 7 grades (from highest to lowest): Featured

Articles (FA), Good Article (GA), A-class, B-class, C-class, Start, Stub. Russian Wikipedia has also 7 quality grades but with other names and criteria: Izbrannaja Stat'ja (similar to FA), Horoshaja Stat'ja (similar to GA), Dobrotnaja Stat'ja, I, II, III, IV (similar to Stub). German Wikipedia uses only two quality grades (Exzellente Artikel and Lesenswerte Artikel) which has similar criteria to FA and GA grades respectively. Polish Wikipedia defined 5 quality grades: Artykuł na Medal (similar to FA), Dobry Artykuł (similar to GA), Czwórka (A-Class), Start, and Zalążek (similar to Stub).

Even though the grading system is available, still the big challenge is a large number of unassessed articles. For example, German and Polish Wikipedia has less than 1% of articles with quality grades. Moreover, articles about the same topic in different languages can also be graded using different criteria. The above facts not only pose problems for comparing the quality of articles in the same language but also for evaluating and comparing different language versions of articles on the same topic.

Using machine learning techniques it is possible to solve the problem of quality assessment of Wikipedia articles as a classification task. In order to build such models, various features can be taken into the account, for example length of an article, number of references, number of images or sections [30–35].

One of the universal approaches for quality assessment of multilingual articles is Objective Revision Evaluation Service (ORES) [36]. This service automates tasks like detection of vandalism and removal of edits made in bad faith [37]. Additionally the service can evaluate articles on a scale between 0 and 1 in some language versions. However, automatic quality assessment of an article by the ORES is currently limited to nine language version of the Wikipedia and it does not include such developed language chapters as German, Spanish, Italian, Polish, Japanese, or Chinese.

In our previous studies [28,38] we defined the synthetic measure to combine several features of articles to allow ranking of Wikipedia articles on a scale between 0 and 100. It is based on the most universal features inferred from machine learning models built for several languages. In the paper we present conclusions from an assessment of over 39 million articles. Additional focus of this work is analysis of demand for information about various topics in different languages from the point of view of readers, as well as from the authors of Wikipedia content. The intersection of those two dimensions is also considered.

Our previous study [39] showed that popularity of the Wikipedia articles can be measured by different SEO metrics from other websites. Such indicators as social signals from Facebook, Twitter, Pinterest, Youtube and others can help to determine also the quality the content in multilingual encyclopedia from the external sources. In this work we decided to use internal popularity measures from the point of view of readers and writers of the Wikipedia articles. Additionally we decided to provide cumulative (global) values of these measures over the language versions about various subjects.

Diverse approaches to defining information by researchers lead also to inconsistencies in defining the notion of its quality. According to the most popular definition, quality of information can be defined as fitness for use [40,41].

In order to define the quality dimensions in Wikipedia, one should take into account the similarity of this website with traditional encyclopedias and Web 2.0 services. On the one hand, content in Wikipedia is created to be a reference point, in an encyclopedic style. According to various studies it has comparable accuracy to other traditional encyclopedias [42,43]. The quality of an article in a traditional encyclopedia can be defined by 7 dimensions: authority, completeness, format, objectivity, style, timeliness, uniqueness [44,45]. On the other hand, Wikipedia is built in a way to allow collaboration between users. It it therefore based on Web 2.0 technologies, which have the following quality dimensions: accessibility, completeness, credibility, involvement, objectivity, readability, relevance, reputation, style, timeliness, uniqueness, usefulness [45,46].

Considering the quality criteria adopted by the Wikipedia community and previously described characteristics of traditional encyclopedia and Web 2.0 documents, we can choose the following quality

Peer-reviewed version available at Computers 2019, 8, 60; doi:10.3390/computers8030060

dimensions for the Wikipedia articles: completeness, credibility, objectivity, readability, relevance, style, timeliness. Figure 7 shows coverage between quality dimensions of the Web 2.0, traditional encyclopedia and Wikipedia.

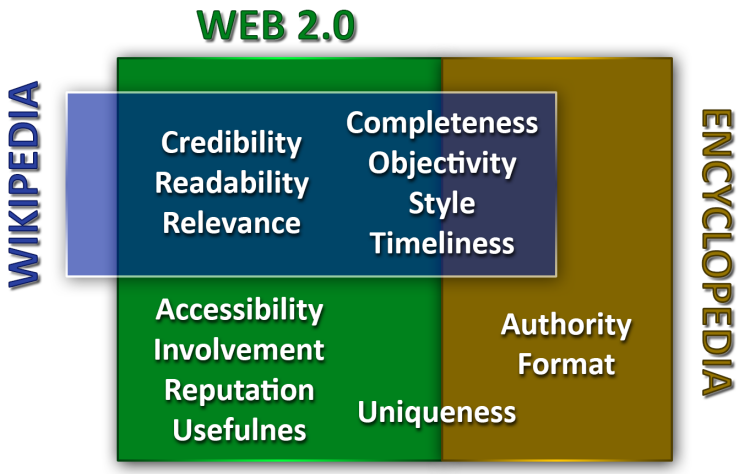

Figure 7. Quality dimensions of Web 2.0 portals, encyclopedias and Wikipedia. Source: own work based on [45]

Each quality dimension contains a specific set of features (measures). Some features can be related to multiple quality dimensions. There are different ways to define and extract features of the Wikipedia articles. Based on the literature and own experiments, we focused on one of the important features, which can show quality of Wikipedia article from different dimensions.

Length of text can be measured in various ways – most often it is represented by the length in bytes, the number of letters or words [28,38,47–58]. Length of an article is related to completeness and may indicate the presence of relevant facts and details in its articles.

High-quality articles are expected to use reliable sources [59]. Readers of encyclopedias must be able to check where the information comes from [60]. Therefore, one of the most commonly used reliability measures is the number of references in a Wikipedia article [28,34,38,48–50,56,58,61–64]. References are related to credibility of the article. Our previous research has shown that it is advantageous to analyze not only the quantity but also the quality of the references [39].

Length of text can be positively correlated with the number of references but it is important that all relevant facts in Wikipedia should be supported by reliable sources. For this purpose, the reference density can be calculated as the number of references divided by the length of text.

Wikipedia articles must provide information in a fair and impartial manner. In this case, we can take into account information presented graphically – images [28,34,38,47,50,55–57,61,62,65,66]. On the one hand, pictures can help to assess the objectivity of the presented material. On the other hand we can also measure completeness (because articles on a specific topic should contain images) and style (because the authors decided to add more photos instead of writing long text).

High-quality content must be prepared in accordance with the guidelines of Wikipedia regarding the style that applies to, among others, organization and structure of the article. Therefore, one of the simplest and most popular measures of this dimension is the number of sections in the article [28,32,34,50,52,56,58,61–63].

Quality measures mentioned before can be combined to build a synthetic measure for evaluation of Wikipedia articles. Unlike most methods in this domain, the synthetic measure can assess the quality of Wikipedia articles on a scale from 0 to 100 [38]. Thus, we can compare quality of articles between different language versions, which can have own quality grading scheme.

Synthetic measure encompasses normalized values of the following five features: length, number of references, reference density, number of images, and number of sections. Every considered language

Peer-reviewed version available at Computers 2019, 8, 60; doi:10.3390/computers8030060

of Wikipedia has a special distinction for articles of the highest quality – equivalents to FA and GA grades in English version. Normalization of the 5 selected features depends on language chapter of Wikipedia, since it uses thresholds, which depend on the best articles in the considered language version [38].

Normalization of each feature was conducted according to the following rule: if value of a given feature in a given language exceeded the threshold of median value of the best articles in the same language version, it was set to 100 points; otherwise its value was linearly scaled to reflect the relation of the value to the median value. For example, if the median for the number of references in Polish Wikipedia was 97, any article with a larger number of references would score 100 for this feature; an article with 59 references would score proportionally 60.82 (59/97) points after normalizing. Changing the value of any metric in a particular Wikipedia language version would have a different effect on the normalized value.

For each language version of Wikipedia, each feature could play an important role in assessing the quality; therefore we first counted the normalized metrics average (NMA) by the following formula:

$$NMA = \frac{1}{c}\sum_{i=1}^{c}\hat{m}_i, \tag{1}$$

where $\hat{m}_i$ is a normalized measure m_i and c is the number of measures.

Next we took into account the number of quality flaw templates (QFT) in the considered article (if they existed) and our final formula for the quality measure reads as follows:

$$QualityScore = NMA \cdot (1 - 0.05 \cdot QFT) \tag{2}$$

Previous research [29] revealed that the synthetic measure was one of the most significant among 100 variables used in quality model of Wikipedia.

4. Popularity Measures

Popularity of an article can be determined with measures reflecting the demand for information contained in it by the readers and Wikipedia authors. Popularity can play an important role in quality estimation in specific language versions of Wikipedia [29,34]. Larger number of users reading an article can contribute to faster identification and correction of errors, therefore amendments can be made more often (including update of the information).

Popularity of an article can be measured based on the number of visits [34,38]. For example, one of the studies compared reptiles species' page view numbers across languages and in their spatial distribution along with various biological attributes [67].

For assessment of popularity we decided to use features available in Wikipedia database – page views and number of unique authors of an article. We also provided local and global measurements characterizing articles, which took into account semantic links between language versions.

For each page of Wikipedia, daily page views statistics are available in a dedicated online service [68] and Wikimedia dumps [69]. We used dumps to analyze popularity of over 39 million articles in considered language versions of Wikipedia.

Popularity measure in this study were calculated as a median of number of page visits per day, as it was proposed in the previous study [38]. If the measurement concerns only selected language version, then we call it **local** popularity. We can also calculate the **global** popularity, which takes into account popularity of articles about the same topic in different languages (the so called *interwiki* links are considered). The global popularity of an article is calculated according to the following formula:

$$PopGlobal(article) = \sum_{lang=1}^{n} PopLocal_{lang}(article), \tag{3}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2019 doi:10.20944/preprints201905.0144.v2

Peer-reviewed version available at Computers 2019, 8, 60; doi:10.3390/computers8030060

where *PopLocal* means local popularity of the article, *lang* is the index of specific language version and *n* is number of the language versions of the selected *article*.

For quality improvement even more important than the number of page views is the number of real edits. Authors' interest (AI) can be measured as the number of unique authors of the Wikipedia articles. Each user editing articles on Wikipedia has own experience, level of knowledge and can adhere to a certain world view. In this regard, it can be assumed that larger number of authors can positively influence the objectivity of the article, since it may contain different points of view on a particular question. At the same time, the number of authors of an article can also indicate the relevance of the article to the Wikipedia community. To sum up, articles created by a larger number of people may be more objective, hence one of the measures leveraged in our research is the number of unique authors [28,34,47,55–58,63–65,70–75].

The number of authors can be extracted from article history. Figure 8 shows part of the article history about Game of Thrones (season 8) in English and German Wikipedia with highlighted authors.

Figure 8. Part of the article history about Game of Thrones (season 8) in English (en) and German (de) Wikipedia with highlighted authors. Source: [76,77]

Similarly to measuring popularity, AI can also be calculated for a specific language version (local AI) and as a cumulative value for all languages (global AI). Authors are identified by names or IP addresses. So, if the same user edited the article in different language versions, in the global AI it will be counted as one author. Calculation of this measure can be carried out using the flowing formula:

$$GlobalAI(article) = \left| \bigcup_{lang=1}^{n} Authors_{lang}(article) \right|, \quad (4)$$

where *Authors* means a set of authors' names, *lang* is the index of specific language version and *n* is the number of language versions of the *article*.

5. Quality and Popularity Assessment

Following the procedures described in previous sections, we extracted over 100 million values of features characterizing articles in all analyzed languages. These values were then used to calculate the synthetic measure that assesses quality of the content. We next grouped articles by 27 main categories and 55 languages. Within each of obtained groups (almost 1500) we calculated sum of all synthetic measure values and divided it by the number of articles. The resulting average quality of articles is presented in Figure 9. Darker colors in the heatmap represent higher values of average quality of articles in specific category and language version.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2019 doi:10.20944/preprints201905.0144.v2

Peer-reviewed version available at Computers 2019, 8, 60; doi:10.3390/computers8030060

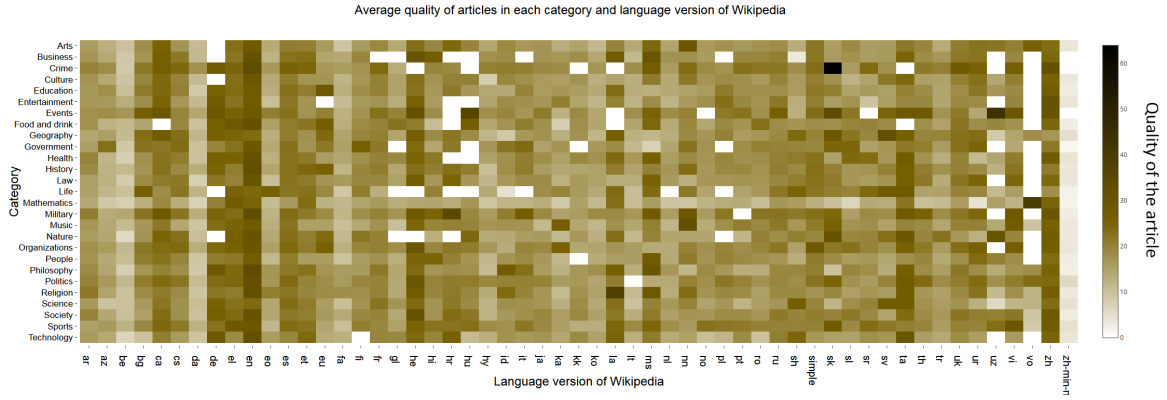

Figure 9. Average quality of articles in each category and language version of Wikipedia. Source: own calculation based on Wikipedia dumps in April, 2019. More detailed and interactive chart can be found on the Web page: http://data.lewoniewski.info/computers/heatmap-cat-quality.

The highest average quality have articles in category Crime in Slovak Wikipedia (sk) - 63.92 points. This is due to the fact that in this language version only a few articles fall into this category and they are generally well written according to studied features. Articles about crime also have relatively higher quality scores in English (en) and Chinese (zh) Wikipedia.

Second place in the ranking are taken by articles about events in Uzbek Wikipedia (uz) - 43.96 points. Again, this main category does not contain much content – there are only 31 articles. If we take into account the development of the Uzbek Wikipedia (about 130 thousand articles), we can conclude that this category is rather important for local community of editors. Articles about events also have relatively higher quality scores in Hungarian (hu), Slovak (sk), Hebrew (he), and Chinese (zh) Wikipedia.

Third place regarding the quality is taken by articles about mathematics in Volapük Wikipedia - 39.63 points. However, in this language chapter the category contains only 2 articles. Latin Wikipedia (la) has the fourth place with average quality of articles about religion - 37.77.

If we take into account the most developed English Wikipedia, the highest average quality of articles can be found in categories: Philosophy, Crime, Military, and History. Generally, we can conclude that English Wikipedia articles usually have high value of average quality measure in different topics.

Figure 10 shows average number of page views per article in year 2018 for each category and language version of Wikipedia. Darker colors in the heatmap represent higher average number of page views of articles in specific category and language version.

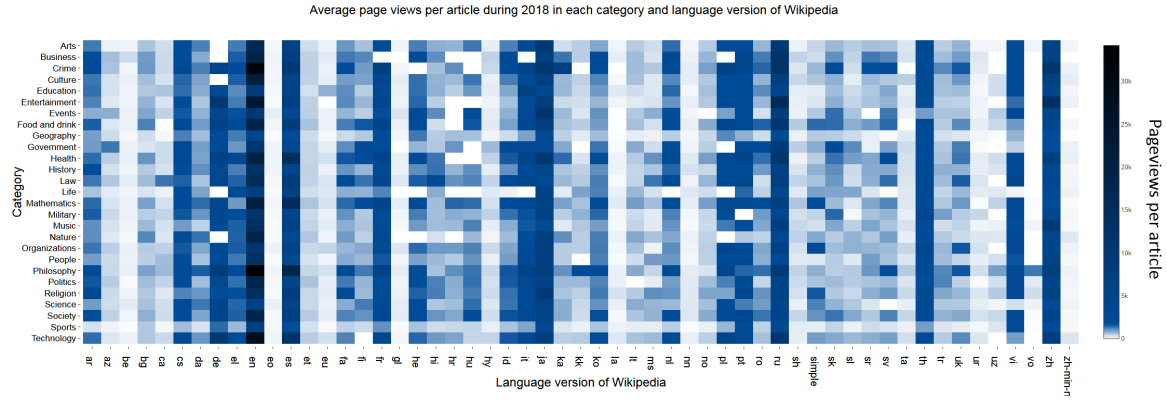

Figure 10. Average page views per article in year 2018 for each main category and language version of Wikipedia. Source: own calculation based on Wikipedia dumps. More detailed and interactive chart can be found on the Web page: http://data.lewoniewski.info/computers/heatmap-cat-views

Generally, page views values are higher for the most popular languages. This led to the fact that the first 11 positions in the rank are occupied by English (en) Wikipedia. The most popular topic in this language is Philosophy. One of the highest average popularity in this language characterizes also articles about crime, technology, entertainment, mathematics, culture, and health. All these categories had at least 20 thousand page views in year 2018.

Second most popular language version is Spanish (es). Similarly to English, the most visited category is Philosophy. It is also worth to mention two other popular categories in this language: Mathematics and Health. Articles in three mentioned main categories of Spanish Wikipedia have at least 14 thousand page views per year.

Third place is taken by Russian (ru) Wikipedia and category Entertainment, with about 16 thousand page views per year. Entertainment is also the most popular topic in Chinese (zh) Wikipedia.

Finally, Figure 11 shows average number of authors (authors' interest) per article in 2018 in each category and language version of Wikipedia. Darker colors in the heatmap represent higher values of average number of authors of articles in specific category and language version.

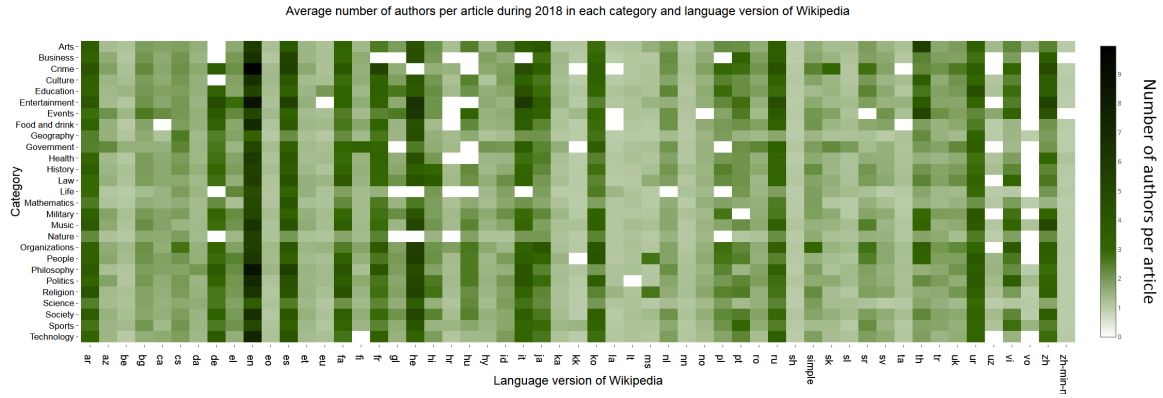

Figure 11. Average number of authors per article during 2018 in each main category and language version of Wikipedia. Source: own calculation based on Wikipedia dumps. More detailed and interactive chart can be found on the Web page: http://data.lewoniewski.info/computers/heatmap-cat-authors

As in the case of the popularity of page views, in the ranking of authors' interests categories in English Wikipedia topped the ranking. Here we have such popular categories as Crime, Philosophy, Entertainment. Articles about topics were edited at least by 8 authors during the 2018 year.

Second language version that has most active authors is Hebrew (he) Wikipedia with articles about entertainment. During a year at least 6 authors have edited each article in this topic. Entertainment is also popular among authors in Italian (it), Spanish (es) and Chinese (zh) Wikipedia. At the same time Italian Wikipedia we can met as the third language in the authors' interest ranking.

Table 4 presents main categories that have the highest value of average quality, average popularity and authors' interest in each language version of Wikipedia.

Peer-reviewed version available at Computers 2019, 8, 60; doi:10.3390/computers8030060

Table 4. Main category of articles with the highest value of average quality, average popularity and authors' interest in each language version of Wikipedia. Source: own calculations.

Language version	Quality	Popularity	Authors' interest
Arabic (ar)	Religion	Religion	Religion
Azerbaijani (az)	Government	Government	Government
Belarusian (be)	Government	Business	Events
Bulgarian (bg)	Events	Food and drink	Life
Catalan (ca)	Events	Law	Events
Czech (cs)	Organizations	Health	Crime
Danish (da)	Philosophy	Philosophy	Crime
German (de)	Entertainment	Entertainment	Events
Greek (el)	Entertainment	Health	Food and drink
English (en)	Crime	Philosophy	Philosophy
Esperanto (eo)	Philosophy	Events	Life
Spanish (es)	Philosophy	Philosophy	Crime
Estonian (et)	Crime	Food and drink	Crime
Basque (eu)	Education	Education	Education
Persian (fa)	Religion	Philosophy	Religion
Finnish (fi)	Government	Government	Government
French (fr)	Crime	Crime	Crime
Galician (gl)	Education	Events	Food and drink
Hebrew (he)	Entertainment	Events	Events
Hindi (hi)	Law	Law	Business
Croatian (hr)	Organizations	Mathematics	Military
Hungarian (hu)	Events	Events	Events
Armenian (hy)	Government	Government	Crime
Indonesian (id)	Arts	Business	Philosophy
Italian (it)	Entertainment	Education	Military
Japanese (ja)	Organizations	Events	Events
Georgian (ka)	Government	Crime	Music
Kazakh (kk)	Sports	Philosophy	Health
Korean (ko)	People	Business	Military
Latin (la)	Religion	Religion	Religion
Lithuanian (lt)	Education	Mathematics	Sports
Malay (ms)	People	Law	Business
Dutch (nl)	Education	Philosophy	Events
Norwegian (Nynorsk) (nn)	History	History	Music
Norwegian (Bokmål) (no)	Crime	Mathematics	Sports
Polish (pl)	Crime	Crime	Entertainment
Portuguese (pt)	Business	Health	Crime
Romanian (ro)	Government	Government	Food and drink
Russian (ru)	Entertainment	Entertainment	Events
Serbo-Croatian (sh)	Music	Mathematics	Science
Simple English (simple)	Organizations	Organizations	Organizations
Slovak (sk)	Crime	Crime	Crime
Slovenian (sl)	Government	Government	Government
Serbian (sr)	Crime	Crime	Life
Swedish (sv)	Events	Health	Geography
Tamil (ta)	Entertainment	Philosophy	Technology
Thai (th)	Arts	Military	Events
Turkish (tr)	Events	Politics	Nature
Ukrainian (uk)	Crime	Philosophy	Crime
Urdu (ur)	Education	Military	Organizations
Uzbek (uz)	Events	Philosophy	Events
Vietnamese (vi)	Organizations	Law	Sports
Volapük (vo)	Sports	Philosophy	Mathematics
Chinese (zh)	Entertainment	Entertainment	Crime
Min Nan (zh-min-nan)	Health	Technology	Politics

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2019 doi:10.20944/preprints201905.0144.v2

Peer-reviewed version available at Computers 2019, 8, 60; doi:10.3390/computers8030060

Depending on Wikipedia language version, we observed different categories with the highest average quality, popularity and AI. For example in English Wikipedia articles in category "Crime" have the highest average quality, but articles from category "Philosophy" has the highest average popularity and AI. Another example: Arabic Wikipedia has the articles from Religion category as the best for these three measures. Similar applies to Latin Wikipedia. In Persian Wikipedia there is also a similar situation, with exception to popularity – here category "Philosophy" has the highest values. Articles in Russian Wikipedia from category "Entertainment" are the most popular and has the highest average quality, at the same time from authors point of view is most popular "Events" category. Similar applies to German Wikipedia. Category "Government" Azerbaijan, Finnish, Slovenian Wikipedia occupies a leading position.

Finally, we do the similar calculations for articles in semantic classes: actor, automobile, business, city, film, football player, human, programming, university, videogame, website. Figure 12 shows average quality, authors interest and page views in 2018 per article in each semantic class and language version of Wikipedia. Darker colors in heatmaps represent higher values of the selected measures.

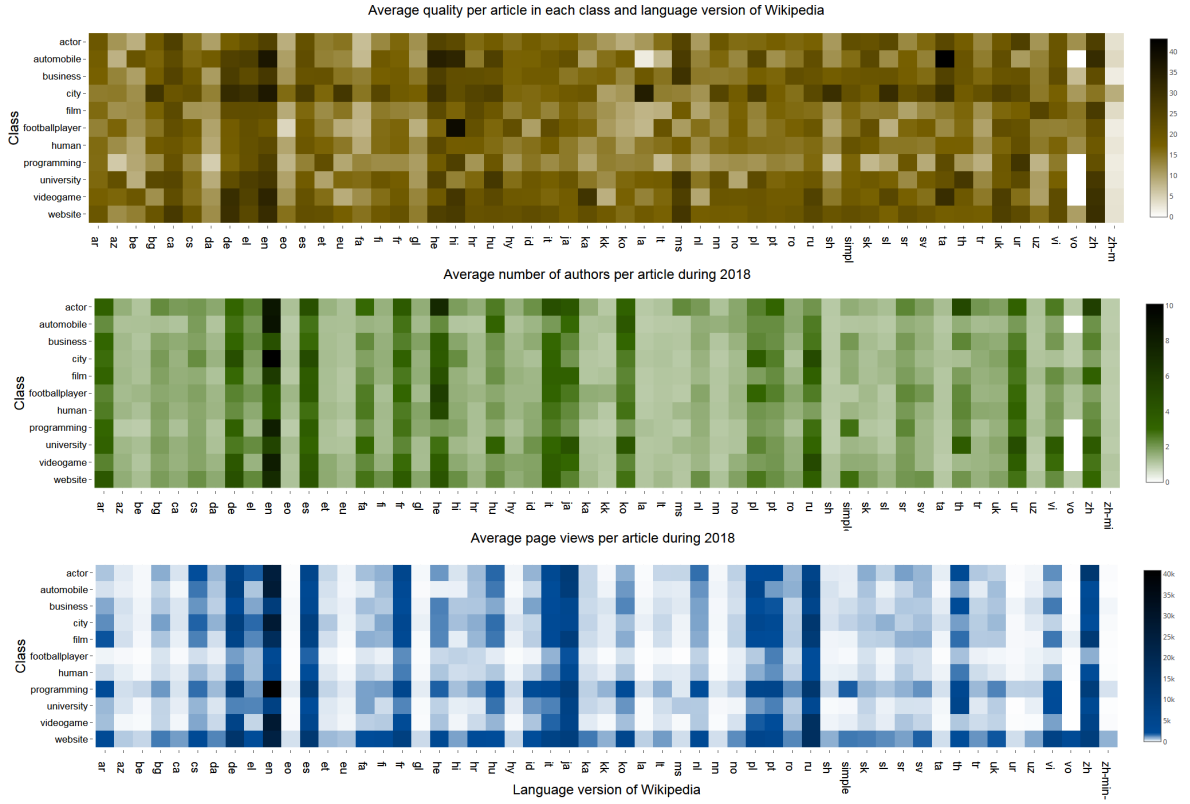

Figure 12. Average quality, authors interest and page views during 2018 per article in each class and language version of Wikipedia. More detailed and interactive chart can be found on the Web page: http://data.lewoniewski.info/computers/heatmap-classes

The leader in terms of the value of average quality is Tamil (ta) Wikipedia with articles that describe cars (automobiles) - 43.22 points. The second place in this ranking occupy articles about football players in Hindi (hi) Wikipedia - 40.35 points for quality per article. The third place in quality took English (en) Wikipedia with articles about cars - 37.39 points. Articles about cars have also relative high quality un Hebrew (he), Hindi (hi) and Chinese (zh) Wikipedia - over 31 points. In this quality ranking most often we can met articles about cities in English (en), Latin (la), German (de), Slovenian (sl), Serbo-Croatian (sh), Greek (el) Wikipedia - over 30 points per article.

As for page views, we have similar situation as it was in the case of main category classifications - English Wikipedia has here the highest values. The most popular class in this language versions is

programming, which has over 40 thousand page visits per article during 2018. Next the most popular classes with over 23 thousand visits per articles during a year are related to video games, cities, cars, actors, and web sites. Second language version that we can met in the top of the popularity ranking - Russian (ru) Wikipedia with articles about web sites and video games. Next is German (de) version with articles about web sites.

Authors' interest ranking of the classes also shows a leading position of English (en) Wikipedia. Here the highest number of authors per article in 2018 have articles about cities - over 10 authors edited each article during a year. Popular among authors are also articles about cars, actors, video games and programming languages - over 8 authors per article during a year. Following are articles from Hebrew (he) Wikipedia describing actors - over 7 authors per article during past year. Relatively high interest among authors we can observe also in Chinese (zh), Thai (th), Italian (it), Spanish and Japanese (ja) Wikipedia - over 4 authors per article about an actor during 2018. Articles about universities has similar values of average authors' interest in English (en), Urdu (ur), Japanese (ja) and Korean (ko) Wikipedia.

Table 4 presents classes that have the highest value of average quality, average popularity and authors' interest in each language version of Wikipedia.

Table 5. Classes of articles with the highest value of average quality, average popularity and authors' interest in each language version of Wikipedia. Source: own calculations.

Language version	Quality	Popularity	Authors' interest
Arabic (ar)	website	website	website
Azerbaijani (az)	website	website	university
Belarusian (be)	footballplayer	programming	automobile
Bulgarian (bg)	actor	website	city
Catalan (ca)	actor	website	website
Czech (cs)	city	website	city
Danish (da)	actor	website	automobile
German (de)	city	website	city
Greek (el)	actor	website	city
English (en)	city	programming	automobile
Esperanto (eo)	footballplayer	website	city
Spanish (es)	city	website	city
Estonian (et)	website	website	programming
Basque (eu)	website	website	city
Persian (fa)	university	website	university
Finnish (fi)	website	website	city
French (fr)	actor	website	website
Galician (gl)	business	website	city
Hebrew (he)	actor	website	automobile
Hindi (hi)	city	website	footballplayer
Croatian (hr)	actor	website	city
Hungarian (hu)	university	website	university
Armenian (hy)	videogame	website	footballplayer
Indonesian (id)	actor	programming	website
Italian (it)	actor	website	footballplayer
Japanese (ja)	university	actor	automobile
Georgian (ka)	footballplayer	website	videogame
Kazakh (kk)	footballplayer	website	website
Korean (ko)	university	website	automobile
Latin (la)	programming	website	city
Lithuanian (lt)	website	website	footballplayer
Malay (ms)	actor	university	business
Dutch (nl)	website	website	website
Norwegian (Nynorsk) (nn)	automobile	website	city
Norwegian (Bokmål) (no)	website	website	videogame
Polish (pl)	city	website	city
Portuguese (pt)	actor	programming	website
Romanian (ro)	website	website	business
Russian (ru)	videogame	website	videogame
Serbo-Croatian (sh)	website	website	city
Simple English (simple)	website	programming	actor
Slovak (sk)	website	website	automobile
Slovenian (sl)	website	website	city
Serbian (sr)	actor	actor	website
Swedish (sv)	website	website	city
Tamil (ta)	actor	website	automobile
Thai (th)	actor	university	university
Turkish (tr)	actor	website	city
Ukrainian (uk)	actor	website	videogame
Urdu (ur)	university	programming	programming
Uzbek (uz)	film	website	film
Vietnamese (vi)	university	website	videogame
Volapük (vo)	film	website	film
Chinese (zh)	actor	actor	automobile
Min Nan (zh-min-nan)	videogame	website	city

6. Local and Global Rankings of Wikipedia Articles

Based on assessment of over 39 million articles we built rankings of articles in each language version of Wikipedia separately and also leveraged knowledge about links between languages to build multilingual global rankings. Page views and authors' interest can change in time, therefore we also conducted calculations for individual months – from January 2018 till March 2019. This allows interesting analyses of changes of preferences of Wikipedia authors and readers.

Measurement of popularity can be carried out for specific language version of article. In this case results are used to create local ranking of the article in selected Wikipedia language, while combining popularity measurements from all the surveyed language versions of the same article was used to create a global ranking. As it was mentioned before, popularity was measured based on median value of the daily visits in selected month. For the purpose of ranking, if median is not sufficient to sort articles we use additional criterion – total number of visits in selected month is considered.

Another measure – authors' interest – is calculated as a number of unique authors who provided changes to an article during selected period (e.g. month). If the number of authors for selected articles is the same, we further sort based on total number of the page visits.

Popularity and AI measures can be used to build ranking on various topics and for a specific periods. Thus, we can examine which articles are popular from the point of view of their authors and readers in each selected month. Global measures can show these results, taking into account several different language versions.

Tables 6, 7 and 8 present top three articles about cars, films, and video games respectively with the highest values of page views and authors' interest in each period in all considered language versions.

Peer-reviewed version available at Computers 2019, 8, 60; doi:10.3390/computers8030060

Table 6. Top 3 articles about cars with highest number of page views and authors' interest in multilingual ranking, monthly. Source: own calculations.

Month	Page views	Authors' interest
January 2018	Volkswagen Golf BMW 3 Series Audi A4	Honda Accord Honda Ridgeline Toyota Avalon
February 2018	BMW 3 Series Volkswagen Golf Audi A4	Honda Civic Type R Tesla Model X Nissan GT-R
March 2018	BMW 3 Series Ford Mustang Volkswagen Golf	Honda Civic Type R Subaru Impreza Tesla Model X
April 2018	Ford Mustang BMW 3 Series Volkswagen Golf	Honda Civic Type R Subaru Impreza BMW M5
May 2018	Ford Mustang BMW 3 Series Volkswagen Golf	DMC DeLorean Subaru Impreza McLaren P1
June 2018	Ford Mustang BMW 3 Series Volkswagen Golf	Acura RDX LaFerrari Ford Model T
July 2018	BMW 3 Series Ford Mustang Volkswagen Golf	Honda Accord Volvo 850 Chevrolet Impala
August 2018	BMW 3 Series Ford Mustang Volkswagen Golf	Pontiac GTO Honda Accord BMW M3
September 2018	BMW 3 Series Ford Mustang Volkswagen Golf	Porsche 997 Opel Combo Ford Falcon (AU)
October 2018	BMW 3 Series BMW 3 Series (F30) Volkswagen Golf	Toyota Land Cruiser Lamborghini Aventador Lincoln Continental
November 2018	BMW 3 Series Tesla Model S Volkswagen Golf	Toyota Land Cruiser Honda Accord Mitsubishi Triton
December 2018	BMW 3 Series Volkswagen Golf Tesla Model S	Honda Civic Type R Toyota Land Cruiser Subaru Impreza
January 2019	BMW 3 Series Toyota Supra Volkswagen Golf	Toyota Prius Toyota Corolla Ford F-Series
February 2019	BMW 3 Series Volkswagen Golf Ford Mustang	BMW 3 Series (E36) Lincoln Continental Honda Accord
March 2019	BMW 3 Series Tesla Model S Ford Mustang	Toyota Prius Tesla Model X BMW 3 Series (E36)

Monthly multilingual ranking of Wikipedia articles about cars shows that depending on the period under consideration, various car models may be at the forefront. From readers' point of view, in the period of 2018-2019 the most interesting automobiles were: BMW 3 Series, Volkswagen Golf, Ford Mustang, Tesla Model S, Audi A4, BMW 3 Series (F30), and Toyota Supra. However, if we look from authors' point of view, there are other Wikipedia articles about cars in the lead: Honda Accord, Honda Civic Type R, Subaru Impreza, Toyota Land Cruiser, Tesla Model X, BMW 3 Series (E36), and Lincoln Continental.

Table 7. Top 3 articles about films with highest number of page views and authors' interest in multilingual ranking, monthly. Source: own calculations.

Month	Page views	Authors' interest
January 2018	Black Mirror The End of the F***ing World Star Wars: The Last Jedi	Pokkiri Dhoom 3 Street Lights
February 2018	Black Panther (film) Altered Carbon (TV series) Money Heist	The Ghost of Hui Family Children of Men Bairavaa
March 2018	Black Panther (film) The Shape of Water Avengers: Infinity War	Bairavaa A Night to Remember (1958 film) Acrimony (film)
April 2018	Avengers: Infinity War A Quiet Place (film) Money Heist	Jason X Traffik (2018 film) Crazy Rich Asians (film)
May 2018	Avengers: Infinity War Deadpool 2 Black Panther (film)	Bairavaa War for the Planet of the Apes Masterpiece (2017 film)
June 2018	Jurassic World: Fallen Kingdom Avengers: Infinity War Westworld (TV series)	Bairavaa Hello (2017 film) Crazy Rich Asians (film)
July 2018	Ant-Man and the Wasp Avengers: Infinity War The Handmaid's Tale (TV series)	Bairavaa Antenna (film) Bean (film)
August 2018	Story of Yanxi Palace Avengers: Infinity War Crazy Rich Asians (film)	Rangasthalam White Boy Rick Happy Death Day
September 2018	Story of Yanxi Palace The Nun (2018 film) The Matrix	Jaws 2 Bean (film) Instant Family
October 2018	Venom (2018 film) A Star Is Born (2018 film) The Haunting (TV series)	Doctor Sleep (2019 film) Escape Room (film) Jawani Phir Nahi Ani 2
November 2018	Bohemian Rhapsody (film) Fantastic Beasts: The Crimes of Grindelwald Fantastic Beasts and Where to Find Them (film)	Doctor Sleep (2019 film) Enai Noki Paayum Thota Scooby-Doo! and the Curse of the 13th Ghost
December 2018	Aquaman (film) Spider-Man: Into the Spider-Verse Bohemian Rhapsody (film)	Unda (film) Escape Room (film) Bairavaa
January 2019	Glass (2019 film) You (TV series) Aquaman (film)	Bairavaa Vaagai Sooda Vaa Bros: After the Screaming Stops
February 2019	Alita: Battle Angel The Umbrella Academy (TV series) Green Book (film)	Doctor Sleep (2019 film) Kanne Kalaimaane 8 Mile (film)
March 2019	Captain Marvel (film) Us (2019 film) Game of Thrones	Kanne Kalaimaane Son of Kashmir: Burhan 8 Mile (film)

In the multilingual ranking of Wikipedia articles related to films, we can also observe fluctuations among leaders in each considered month. Readers of this encyclopedia preferred such movies as Avengers: Infinity War, Black Panther, Bohemian Rhapsody, Story of Yanxi Palace, Money Heist, Aquaman, The Umbrella Academy, You, The Haunting, The Matrix, Venom, Game of Thrones, Green Book. It was not overlapping with authors' preferences who contributed mostly to films: Bairavaa, Doctor Sleep, Escape Room, Kanne Kalaimaane, 8 Mile, Bean, Crazy Rich Asians, Jaws 2, War for the Planet of the Apes, The Ghost of Hui Family, Traffik. Only one title appeared in both rankings - Crazy Rich Asians.

Table 8. Top 3 articles about video games with the highest number of page views and authors' interest in multilingual ranking, monthly. Source: own calculations.

Month	Page views	Authors' interest
January 2018	Assassin's Creed Devilman PlayerUnknown's Battlegrounds	Celeste (video game) Unreal Tournament Lego Marvel Super Heroes 2
February 2018	Assassin's Creed Kingdom Come: Deliverance Fortnite	Celeste (video game) Little Witch Academia: Chamber of Time Fire Emblem: The Binding Blade
March 2018	Fortnite Assassin's Creed Call of Duty	Ace Combat 7: Skies Unknown The Crew 2 Detective Pikachu
April 2018	God of War (2018 video game) Fortnite Far Cry 5	FortniteAce Combat 7: Skies Unknown H1Z1 Skynet (video game)
May 2018	Fortnite God of War (2018 video game) Assassin's Creed	Spider-Man 3 (video game) AirAttack Imperator: Rome
June 2018	Detroit: Become Human Fortnite Assassin's Creed	Ace Combat 7: Skies Rules of Survival Totally Accurate Battlegrounds
July 2018	Fortnite Detroit: Become Human Assassin's Creed	Ace Combat 7: Skies MicroVolts Aliens: Colonial Marines
August 2018	Fortnite Assassin's Creed World of Warcraft	Spider-Man 3 (video game) H1Z1 Shovel Knight
September 2018	Borderlands: The Pre-Sequel Spider-Man (2018 video game) Fortnite	Rules of Survival Nickelodeon Kart Racers H1Z1
October 2018	Borderlands: The Pre-Sequel Assassin's Creed Red Dead Redemption 2	RuneScape H1Z1 Starlink: Battle for Atlas
November 2018	Borderlands: The Pre-Sequel Red Dead Redemption 2 Fallout 76	Call of Duty: Black Ops III Spider-Man 3 (video game) Dragon Ball Xenoverse 2
December 2018	Borderlands: The Pre-Sequel Fortnite Red Dead Redemption 2	Marvel: Ultimate Alliance PewDiePie: Legend of the Brofist Yo-kai Watch
January 2019	Borderlands: The Pre-Sequel Fortnite Minecraft	Portal 2 Dick Vitale's "Awesome Baby" College Hoops Fire Emblem Warriors
February 2019	Borderlands: The Pre-Sequel Apex Legends Fortnite	Dick Vitale's "Awesome Baby" College Hoops Wargroove Fire Emblem Warriors
March 2019	Borderlands: The Pre-Sequel Fortnite Sekiro: Shadows Die Twice	Assassin's Creed II Dance Dance Revolution A20 Subnautica

Analysis of leading articles about video games in multilingual ranking shows similar tendencies. Readers preferred Wikipedia articles about such games as Fortnite, Assassin's Creed, Borderlands: The Pre-Sequel, Red Dead Redemption 2, Detroit: Become Human, God of War, Sekiro: Shadows Die Twice, Fallout 76, Spider-Man (2018 video game), Minecraft, PlayerUnknown's Battlegrounds, Devilman, Kingdom Come: Deliverance, Call of Duty, Far Cry 5, World of Warcraft, Apex Legends. Wikipedia authors have other priorities of games in the same period: H1Z1, Spider-Man 3 (video game), Ace Combat 7: Skies, Celeste (video game), Rules of Survival, Dick Vitale's "Awesome Baby" College Hoops, Fire Emblem Warriors, Ace Combat 7: Skies Unknown, MicroVolts, Call of Duty: Black Ops III,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2019 doi:10.20944/preprints201905.0144.v2

Peer-reviewed version available at Computers 2019, 8, 60; doi:10.3390/computers8030060

RuneScape, Aliens: Colonial Marines, Unreal Tournament, Portal 2. There is no overlap between top titles from readers' and authors' point of view.

These ranking show that the most popular articles from readers' point of view usually do not match with the priorities of the community of Wikipedia authors. This may be due to the fact that popular articles are sufficiently developed and do not require significant revisions. Nevertheless, we also found examples when popular articles are blocked for editing by anonymous users or users with low experience.

Such global quality rankings can show how specific product is popular worldwide. Tables 6, 7 and 8 show limited number of leading titles of the Wikipedia articles in some of the categories. Therefore, we implemented various multilingual rankings in WikiRank service [78], where it is possible to analyze how the position of a particular article has changed in rankings in comparison with the previous period, what is the most popular language version, what is the quality of the popular language version article etc. Figure 13 presents example of the ranking of the articles about films with different parameters.

Figure 13. List of the most popular articles about films in multilingual Wikipedia in WikiRank service. Source: [79]

Combination of measures from different languages makes it possible to create global rankings of all articles. Additionally, for each language version it is possible to generate local rankings – here measures from one language can be taken into account. Example of the local ranking with quality distribution of all articles in English Wikipedia is shown in Figure 14.

Calculated measures can be gathered to create individual profile for each article in each language version. For example, Figure 15 presents such a profile for article "Fortnite" in English Wikipedia on WikiRank with information about places in local and global rankings, quality and popularity scores, and also history of popularity rank.

Each Wikipedia article in WikiRank service can have information about local and global measurements of popularity, AI and their historical ranks for the last period (Figure 15 shows such data monthly from January 2018 to April 2019 on the right side).

If an article is written in more than one language, additional ranking of the most popular language versions as well as languages with the highest quality are displayed. Additionally, it is marked, which language versions were edited by the largest number of authors. Figure 16 shows an example of such ranking of the best language versions about Minecraft.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2019 doi:10.20944/preprints201905.0144.v2

Peer-reviewed version available at Computers 2019, 8, 60; doi:10.3390/computers8030060

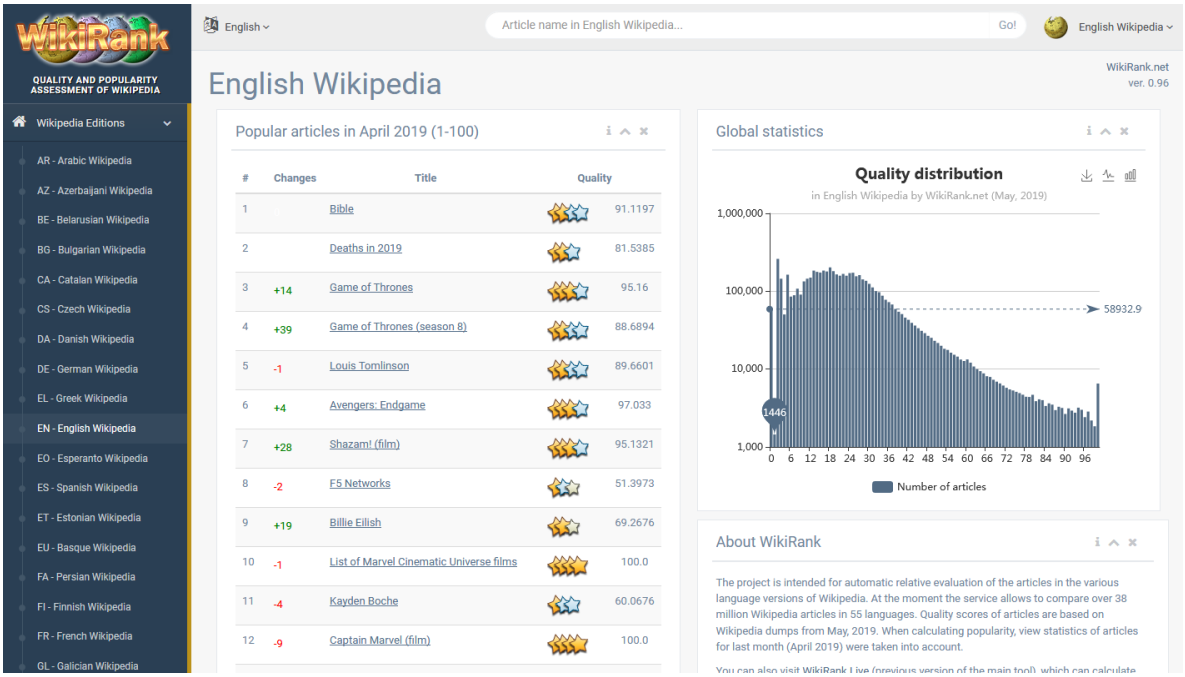

Figure 14. Local ranking with quality distribution of all articles in English Wikipedia in WikiRank service. Source: [80]

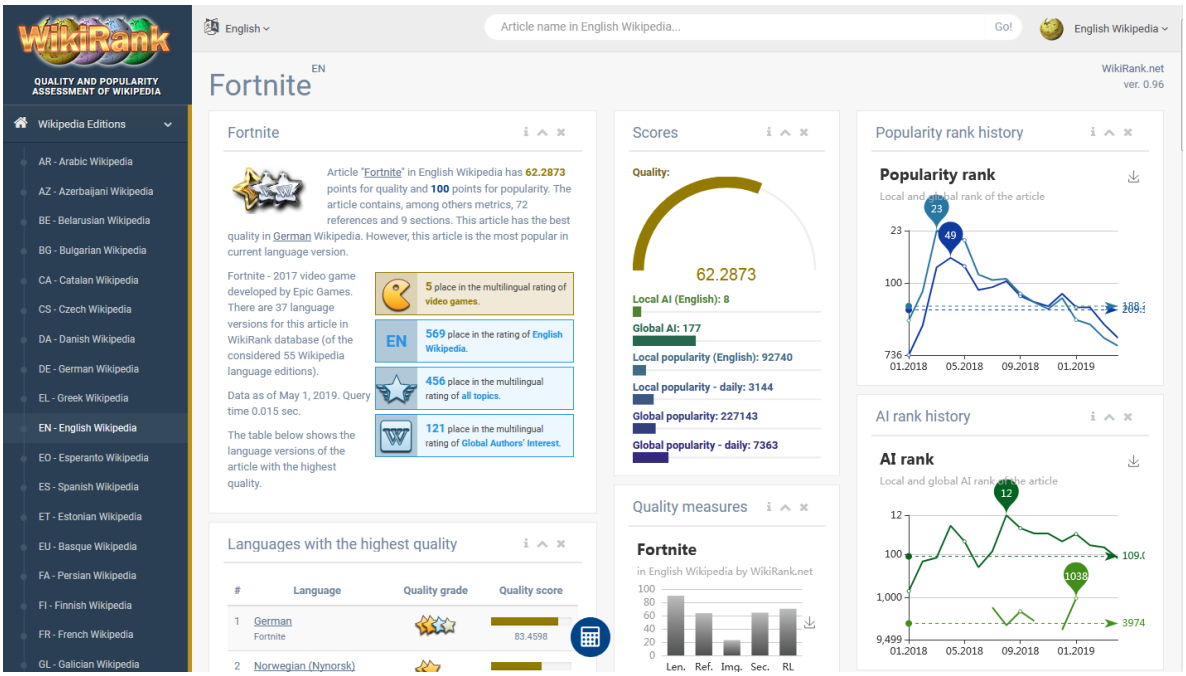

Figure 15. Profile on WikiRank of the article about Fortnite in English Wikipedia with information about places in local and global rankings, quality and popularity scores, history of popularity rank. Source: [81]

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2019 doi:10.20944/preprints201905.0144.v2

Peer-reviewed version available at Computers 2019, 8, 60; doi:10.3390/computers8030060

Figure 16. The most popular language versions, languages with the highest quality and language versions with the highest AI value for article about Minecraft on WikiRank. Source: [82]

Profiles of Wikipedia articles can also be used to compare the demand for a specific product between various language communities. For example video game Dota 2 is the most popular in English, Russian, Chinese, German, and Spanish [83]. Based on obtained measures for the action-adventure video game Grand Theft Auto V (GTA 5) we can see relatively large demand from English, Russian, Arabic, Spanish, and Chinese language community [84].

7. Results and Discussion

During the research we encountered several restrictions, mainly related to the differences between language versions of Wikipedia. For example, as we showed in Table 3, some main categories do not have links to all considered language versions. This is also true for developed languages. For example, category "Art" in English Wikipedia does not have direct equivalent in German Wikipedia, which uses category "Kunst und Kultur" [85] ("Arts and Culture") to describe part of this topic.

Regarding categories, our experiments showed that each language version has specific ratio between number of articles and number of categories. Additionally, some language versions can have a lot of undefined pages for the categories. There is also a difference in the number of categories that are assigned to each article. Some languages can use an average of 30 categories to describe one article, while the others are limited to 2-3 categories per article.

Depending on Wikipedia language version, we observed different categories with the highest average quality, popularity, and authors' interest. For example in English Wikipedia articles in category "Crime" have the highest average quality, but articles from category "Philosophy" have the highest average popularity and AI. Another example, Arabic Wikipedia has the articles from Religion category as the best for these three measures. Articles in Russian Wikipedia from category "Entertainment" are the most popular and have the highest average quality, while from authors point of view the most popular is "Events" category.

Results for authors popularity can be sometimes biased due to temporal or permanent restrictions. According to one of the main principles of Wikipedia anyone can edit content. However, in some particular situations this right can be revoked to protect content from unwanted changes (vandalism) [86]. Each language version can define own levels of page protection. For example, in English Wikipedia there is a full protection, where only administrators can edit an article, and semi-protection, which prevents editing by unregistered users or users that are not confirmed. Each article can be protected for a specified period. Figure 17 shows an example of the protected Wikipedia article about Bitcoin with a marked level of protection. As a result, some articles can have less authors' interest than it would in the situation without protection.

In our work, we provided classification of articles by main categories according to structure of categories in English Wikipedia. However, each language can have own definition of main categories. In future, we plan to develop more sophisticated methods to take into account refined category structures.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 5 August 2019 doi:10.20944/preprints201905.0144.v2

Peer-reviewed version available at Computers 2019, 8, 60; doi:10.3390/computers8030060

Figure 17. Wikipedia article about Bitcoin with a marked level of protection. Source: [87]

Supplementing research results are available online at WikiRank service [78]. In research we used some tools that are available on GitHub [88].

8. Conclusions and Future Work

In this paper we presented results of quality and popularity assessment of articles in multilingual Wikipedia. For this purpose we calculated over 200 million values characterizing quality and popularity of articles in 55 language versions of Wikipedia. Additionally, we analyzed over 10 million categories, over 26 million links between them, and about 400 million links from articles to categories in order to determine assignment of articles to one of the topics in main classification. In order to assign articles from different languages to various topics we also used semantic databases – Wikidata and DBpedia. We combined data from these sources to obtain more comprehensive classifications of articles.

Results of the research showed not only how quality and popularity differ for articles from various topics and languages but also how the same topic is developed in different languages of Wikipedia in terms of quality and popularity of content. We observed that articles from topics that are popular in a given language are characterized by a relatively higher quality. For instance articles related to main category 'Religion' have relatively higher quality and popularity in Arabic and Latin Wikipedia. Likewise, articles from main category 'Goverment' have relatively higher quality and popularity in Azerbaijani, Finnish, Armenian, Romanian, and Slovenian language version of Wikipedia. Articles related to main category 'Entertainment' are more popular in Chinese, Russian, German Wikipedia. At the same time, articles in those three language versions has relatively the highest quality compared to other main categories.

Additionally to categories, we also studied semantic classes as defined by DBpedia ontology and their relation to quality and popularity. The highest average number of page views among different classes in almost all considered language versions had articles that described websites, e.g. Facebook, YouTube, Google. However, popular articles from this class rarely were assessed as articles of high quality. Articles about cities were relatively better described in English, German, Czech, Hindi, Polish, and Spanish Wikipedia. Actors were described better than other classes in Bulgarian, Catalan, Danish, Greek, French, Hebrew, Croatian, Indonesian, Italian, Malay, Portuguese, Serbian, Tamil, Thai, Turkish, Ukrainian, and Chinese language versions.

With regard to popularity, we proposed to pay attention not only to how often users visits certain articles but also what is authors' interest in them. The authors' interest measure can be calculated for a language version or can be combined across studied languages. Sometimes both popularity measures show similar leader in main categories and semantic classes. For example, Slovenian Wikpedia has the

most popular articles related to main category 'Government', while for readers and authors of English Wikipedia articles have higher preference related to 'Philosophy'. If we consider semantic classes, we can conclude that among analyzed languages the most popular articles for Wikipedians are related to cities and automobiles. We also aggregated numbers for all considered languages so that global demand for specific products, such as films, video games, cars, can be studied.

Additional analyses of popularity measures allowed to find priorities and preferences of Wikipedians and readers in relation to temporal dimension. Often the most popular subjects of the readers differed from leading subjects from authors point of view in the same periods of time. This can be explained by the fact that popular articles are protected and cannot be edited by anonymous users. Additionally, some Wikipedia authors may choose articles based on various initiatives related to improvement of specific topics at certain period of time.

Presented results can be used to build more complex models for quality assessment of information in Wikipedia in different languages and topics. In the future, they can help not only to automatically enrich less-developed language versions of Wikipedia but also can be used to build massive semantic databases with powerful inference system, creating new knowledge for humanity in a relatively short time.

The work towards more precise assessment of Wikipedia quality will be continued, especially different measures and approaches for quality assessment in Wikipedia and other collaborative knowledge bases will be studied. As of April 2019, based on our calculations, there were over 70 thousand wiki services in the Internet, which potentially can be used to enrich various knowledge bases used in enterprises. Additionally, there are over 1300 linked databases [89] that use data from open sources. We can also take into account dedicated web portals that allow companies and individuals to share their databases for research, such as Kaggle [90]. Local and global AI measurements can be improved by including different additional features. For example, it is possible to divide all users into three categories: anonymous users, registered users, and bots. We can also take into account reputation and experience of each author of the article. For this purpose we can use information provided by services like GUC [91] or WikiTop [92].

Author Contributions: K.W. and W.L. conceived the research problem; W.L. conducted state of the art analysis; K.W. proposed research methodology and designed the experiments, starting from hypotheses to be verified statistically; W.L. collected data and performed the analysis; W.L. and K.W. interpreted the results; W.A. provided an overall guidance.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Price, R.; Shanks, G. A semiotic information quality framework: development and comparative analysis. In *Enacting Research Methods in Information Systems*; Springer, 2016; pp. 219–250.
2. Xu, H.; Koronios, A. Understanding information quality in e-business. *Journal of Computer Information Systems* **2005**, *45*, 73–82.
3. Wikipedia Meta-Wiki. List of Wikipedias. https://meta.wikimedia.org/wiki/List_of_Wikipedias, accessed on 2019-05-05.
4. Alexa. wikipedia.org Traffic Statistics. https://www.alexa.com/siteinfo/wikipedia.org, accessed on 2018-10-08.
5. Thompson, N.; Hanley, D. Science is shaped by wikipedia: Evidence from a randomized control trial. *Social Science Research Network* **2018**.
6. Osman, K. The role of conflict in determining consensus on quality in Wikipedia articles. Proceedings of the 9th International Symposium on Open Collaboration. ACM, 2013, p. 12.
7. Callahan, E.S.; Herring, S.C. Cultural bias in Wikipedia content on famous persons. *Journal of the American society for information science and technology* **2011**, *62*, 1899–1915.
8. Laufer, P.; Wagner, C.; Flöck, F.; Strohmaier, M. Mining cross-cultural relations from Wikipedia: a study of 31 European food cultures. Proceedings of the ACM Web Science Conference. ACM, 2015, p. 3.

9. Gieck, R.; Kinnunen, H.M.; Li, Y.; Moghaddam, M.; Pradel, F.; Gloor, P.A.; Paasivaara, M.; Zylka, M.P. Cultural differences in the understanding of history on Wikipedia. In *Designing Networks for Innovation and Improvisation*; Springer, 2016; pp. 3–12.
10. Samoilenko, A.; Karimi, F.; Edler, D.; Kunegis, J.; Strohmaier, M. Linguistic neighbourhoods: explaining cultural borders on Wikipedia through multilingual co-editing activity. *EPJ data science* **2016**, *5*, 9.
11. Kim, S.; Park, S.; Hale, S.A.; Kim, S.; Byun, J.; Oh, A.H. Understanding editing behaviors in multilingual Wikipedia. *PloS one* **2016**, *11*, e0155305.
12. Bao, P.; Hecht, B.; Carton, S.; Quaderi, M.; Horn, M.; Gergle, D. Omnipedia: bridging the wikipedia language gap. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. ACM, 2012, pp. 1075–1084.
13. Wikimedia Meta-Wiki. Wikipedia article depth. https://meta.wikimedia.org/wiki/Wikipedia_article_depth, accessed on 2019-04-26.
14. Kittur, A.; Chi, E.H.; Suh, B. What's in Wikipedia?: mapping topics and conflict using socially annotated category structure. Proceedings of the SIGCHI conference on human factors in computing systems. ACM, 2009, pp. 1509–1512.
15. Boldi, P.; Monti, C. Cleansing wikipedia categories using centrality. Proceedings of the 25th International Conference Companion on World Wide Web. International World Wide Web Conferences Steering Committee, 2016, pp. 969–974.
16. English Wikipedia. Category:Main topic classifications. https://en.wikipedia.org/wiki/Category:Main_topic_classifications, accessed on 2019-04-27.
17. Vrandečić, D. Wikidata: A new platform for collaborative data collection. Proceedings of the 21st international conference on world wide web. ACM, 2012, pp. 1063–1064.
18. Auer, S.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; Ives, Z. DBpedia: A nucleus for a web of open data. In *The semantic web*; Springer, 2007; pp. 722–735.
19. Lehmann, J.; Isele, R.; Jakob, M.; Jentzsch, A.; Kontokostas, D.; Mendes, P.N.; Hellmann, S.; Morsey, M.; Van Kleef, P.; Auer, S.; others. DBpedia–a large-scale, multilingual knowledge base extracted from Wikipedia. *Semantic Web* **2015**, *6*, 167–195.
20. Abramowicz, W.; Auer, S.; Heath, T. Linked Data in Business. *Business & Information Systems Engineering* **2016**, *58*, 323–326. doi:10.1007/s12599-016-0446-0.
21. Lewańska, E. Towards Automatic Business Networks Identification. International Conference on Business Information Systems. Springer, 2017, pp. 389–398. doi:doi.org/10.1007/978-3-319-52464-1_36.
22. Filipiak, D.; Filipowska, A. Improving the Quality of Art Market Data Using Linked Open Data and Machine Learning. Business Information Systems Workshops; Abramowicz, W.; Alt, R.; Franczyk, B., Eds.; Springer International Publishing: Cham, 2017; pp. 418–428.
23. Stróżyna, M.; Eiden, G.; Abramowicz, W.; Filipiak, D.; Małyszko, J.; Węcel, K. A framework for the quality-based selection and retrieval of open data - a use case from the maritime domain. *Electronic Markets* **2018**, *28*, 219–233. doi:10.1007/s12525-017-0277-y.
24. Färber, M.; Bartscherer, F.; Menne, C.; Rettinger, A. Linked data quality of dbpedia, freebase, opencyc, wikidata, and yago. *Semantic Web* **2018**, *9*, 77–129.
25. DBpedia. Ontology Classes. http://mappings.dbpedia.org/server/ontology/classes/, accessed on 2019-05-05.
26. Ringler, D.; Paulheim, H. One knowledge graph to rule them all? Analyzing the differences between DBpedia, YAGO, Wikidata & co. Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz). Springer, 2017, pp. 366–372.
27. Ismayilov, A.; Kontokostas, D.; Auer, S.; Lehmann, J.; Hellmann, S.; others. Wikidata through the Eyes of DBpedia. *Semantic Web* **2018**, *9*, 493–503.
28. Węcel, K.; Lewoniewski, W. Modelling the Quality of Attributes in Wikipedia Infoboxes. In *Business Information Systems Workshops*; Abramowicz, W., Ed.; Springer International Publishing, 2015; Vol. 228, *Lecture Notes in Business Information Processing*, pp. 308–320. doi:10.1007/978-3-319-26762-3_27.
29. Lewoniewski, W. The method of comparing and enriching information in multlingual wikis based on the analysis of their quality. Phd, Poznań University of Economics and Business, 2018.
30. Xu, Y.; Luo, T. Measuring article quality in Wikipedia: Lexical clue model. Web Society (SWS), 2011 3rd Symposium on. IEEE, 2011, pp. 141–146. doi:10.1109/SWS.2011.6101286.

31. Anderka, M.; Stein, B.; Lipka, N. Predicting quality flaws in user-generated content: the case of wikipedia. Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval. ACM, 2012, pp. 981–990.
32. Warncke-wang, M.; Cosley, D.; Riedl, J. Tell Me More : An Actionable Quality Model for Wikipedia. WikiSym 2013, 2013, pp. 1–10. doi:10.1145/2491055.2491063.
33. Su, Q.; Liu, P. A Psycho-Lexical Approach to the Assessment of Information Quality on Wikipedia. 2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT), 2015, Vol. 3, pp. 184–187. doi:10.1109/WI-IAT.2015.23.
34. Lewoniewski, W.; Węcel, K.; Abramowicz, W. Quality and Importance of Wikipedia Articles in Different Languages. In *Information and Software Technologies: 22nd International Conference, ICIST 2016, Druskininkai, Lithuania, October 13-15, 2016, Proceedings*; Springer International Publishing: Cham, 2016; pp. 613–624. doi:10.1007/978-3-319-46254-7_50.
35. Dang, Q.V.; Ignat, C.L. Quality assessment of Wikipedia articles without feature engineering. 2016 IEEE/ACM Joint Conference on Digital Libraries (JCDL), 2016, pp. 27–30.
36. Halfaker, A.; Taraborelli, D. Artificial intelligence service 'ORES' gives Wikipedians X-ray specs to see through bad edits. https://blog.wikimedia.org/2015/11/30/artificial-intelligence-x-ray-specs/, accessed on 2017-12-31.
37. Wikimedia Foundation. ORES. https://ores.wikimedia.org/, accessed on 2019-05-05.
38. Lewoniewski, W.; Węcel, K.; Abramowicz, W. Relative Quality and Popularity Evaluation of Multilingual Wikipedia Articles. *Informatics* **2017**, *4*. doi:10.3390/informatics4040043.
39. Lewoniewski, W.; Härting, R.C.; Węcel, K.; Reichstein, C.; Abramowicz, W. Application of SEO Metrics to Determine the Quality of Wikipedia Articles and Their Sources. Information and Software Technologies; Damaševičius, R.; Vasiljevienė, G., Eds.; Springer International Publishing: Cham, 2018; pp. 139–152. doi:10.1007/978-3-319-99972-2_11.
40. Kahn, B.K.; Strong, D.M.; Wang, R.Y. Information quality benchmarks: product and service performance. *Communications of the ACM* **2002**, *45*, 184–192.
41. Tayi, G.K.; Ballou, D.P. Examining data quality. *Communications of the ACM* **1998**, *41*, 54–57.
42. Giles, J. Internet encyclopaedias go head to head, 2005.
43. Holman Rector, L. Comparison of Wikipedia and other encyclopedias for accuracy, breadth, and depth in historical articles. *Reference services review* **2008**, *36*, 7–22.
44. Crawford, H. Encyclopedias. *Reference and information services: An introduction* **2001**, pp. 433–459.
45. Lewoniewski, W. Measures for Quality Assessment of Articles and Infoboxes in Multilingual Wikipedia. International Conference on Business Information Systems. Workshops; Abramowicz, W.; Paschke, A., Eds. Springer, Springer International Publishing, 2019, pp. 619–633. doi:doi.org/10.1007/978-3-030-04849-5_53.
46. Dalip, D.H.; Gonçalves, M.A.; Cristo, M.; Calado, P. A general multiview framework for assessing the quality of collaboratively created content on web 2.0. *Journal of the Association for Information Science and Technology* **2017**, *68*, 286–308.
47. Yaari, E.; Baruchson-Arbib, S.; Bar-Ilan, J. Information quality assessment of community generated content: A user study of Wikipedia. *Journal of Information Science* **2011**, *37*, 487–498.
48. Dang, Q.V.; Ignat, C.L. Measuring Quality of Collaboratively Edited Documents: The Case of Wikipedia. Collaboration and Internet Computing (CIC), 2016 IEEE 2nd International Conference on. IEEE, 2016, pp. 266–275.
49. Shen, A.; Qi, J.; Baldwin, T. A Hybrid Model for Quality Assessment of Wikipedia Articles. Proceedings of the Australasian Language Technology Association Workshop 2017, 2017, pp. 43–52.
50. Zhang, S.; Hu, Z.; Zhang, C.; Yu, K. History-Based Article Quality Assessment on Wikipedia. Big Data and Smart Computing (BigComp), 2018 IEEE International Conference on. IEEE, 2018, pp. 1–8.
51. Warncke-Wang, M.; Ranjan, V.; Terveen, L.G.; Hecht, B.J. Misalignment Between Supply and Demand of Quality Content in Peer Production Communities. ICWSM, 2015, pp. 493–502.
52. Lerner, J.; Lomi, A. Knowledge categorization affects popularity and quality of Wikipedia articles. *PloS one* **2018**, *13*, e0190674.
53. Blumenstock, J.E. Automatically Assessing the Quality of Wikipedia Articles. Technical report, UC Berkeley, 2008. doi:10.1080/17439880802324251.

687 54. Dalip, D.H.; Gonçalves, M.A.; Cristo, M.; Calado, P. Automatic Assessment of Document Quality
688 in Web Collaborative Digital Libraries. *Journal of Data and Information Quality* **2011**, *2*, 1–30.
689 doi:10.1145/2063504.2063507.
690 55. Stvilia, B.; Twidale, M.B.; Smith, L.C.; Gasser, L. Assessing information quality of a community-based
691 encyclopedia. *Proc. ICIQ* **2005**, pp. 442–454.
692 56. Wu, K.; Zhu, Q.; Zhao, Y.; Zheng, H. Mining the factors affecting the quality of Wikipedia articles.
693 Information Science and Management Engineering (ISME), 2010 International Conference of. IEEE, 2010,
694 Vol. 1, pp. 343–346.
695 57. Stvilia, B.; Twidale, M.B.; Smith, L.C.; Gasser, L. Assessing information quality of a community-based
696 encyclopedia. *Proc. ICIQ* **2005**, pp. 442–454.
697 58. Conti, R.; Marzini, E.; Spognardi, A.; Matteucci, I.; Mori, P.; Petrocchi, M. Maturity assessment of Wikipedia
698 medical articles. Computer-Based Medical Systems (CBMS), 2014 IEEE 27th International Symposium on.
699 IEEE, 2014, pp. 281–286.
700 59. Wikipedia. Featured article criteria. https://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria,
701 accessed on 2019-05-05.
702 60. Wikipedia. Verifiability. https://en.wikipedia.org/wiki/Wikipedia:Verifiability, accessed on 2019-05-05.
703 61. Blumenstock, J.E. Size matters: word count as a measure of quality on Wikipedia. WWW, 2008, pp.
704 1095–1096. doi:10.1145/1367497.1367673.
705 62. Dalip, D.H.; Gonçalves, M.A.; Cristo, M.; Calado, P. Automatic quality assessment of content created
706 collaboratively by web communities: a case study of wikipedia. Proceedings of the 9th ACM/IEEE-CS
707 Joint Conference on Digital Libraries, 2009, pp. 295–304. doi:10.1145/1555400.1555449.
708 63. Ferschke, O.; Gurevych, I.; Rittberger, M. FlawFinder: A Modular System for Predicting Quality Flaws in
709 Wikipedia. CLEF (Online Working Notes/Labs/Workshop), 2012, pp. 1–10.
710 64. di Sciascio, C.; Strohmaier, D.; Errecalde, M.; Veas, E. WikiLyzer: interactive information quality assessment
711 in Wikipedia. Proceedings of the 22nd International Conference on Intelligent User Interfaces. ACM, 2017,
712 pp. 377–388.
713 65. Liu, J.; Ram, S. Using big data and network analysis to understand Wikipedia article quality. *Data &*
714 *Knowledge Engineering* **2018**.
715 66. Shang, W. A Comparison of the Historical Entries in Wikipedia and Baidu Baike. International Conference
716 on Information. Springer, 2018, pp. 74–80.
717 67. Roll, U.; Mittermeier, J.C.; Diaz, G.I.; Novosolov, M.; Feldman, A.; Itescu, Y.; Meiri, S.; Grenyer, R. Using
718 Wikipedia page views to explore the cultural importance of global reptiles. *Biological conservation* **2016**,
719 *204*, 42–50.
720 68. Wikimedia Toolforge. Pageviews Analysis. https://tools.wmflabs.org/pageviews/, accessed on
721 2019-05-05.
722 69. WMF Analytics. Wikistats pageview files. https://dumps.wikimedia.org/other/pagecounts-ez/, accessed
723 on 2019-05-05.
724 70. Lih, A. Wikipedia as Participatory Journalism: Reliable Sources? Metrics for evaluating collaborative
725 media as a news resource. *5th International Symposium on Online Journalism* **2004**, p. 31.
726 71. Wilkinson, D.M.; Huberman, B.a. Cooperation and quality in wikipedia. *Proceedings of the 2007 international*
727 *symposium on Wikis WikiSym 07* **2007**, pp. 157–164. doi:10.1145/1296951.1296968.
728 72. Kittur, A.; Kraut, R.E. Harnessing the wisdom of crowds in wikipedia. *Proceedings of the ACM 2008*
729 *conference on Computer supported cooperative work - CSCW '08* **2008**, p. 37. doi:10.1145/1460563.1460572.
730 73. Wilkinson, D.M.; Huberman, B.a. Cooperation and quality in wikipedia. *Proceedings of the 2007 international*
731 *symposium on Wikis WikiSym 07* **2007**, pp. 157–164. doi:10.1145/1296951.1296968.
732 74. Kane, G.C. A multimethod study of information quality in wiki collaboration. *ACM Transactions on*
733 *Management Information Systems (TMIS)* **2011**, *2*, 4.
734 75. Flekova, L.; Ferschke, O.; Gurevych, I. What makes a good biography?: multidimensional quality analysis
735 based on wikipedia article feedback data. Proceedings of the 23rd international conference on World wide
736 web. ACM, 2014, pp. 855–866.
737 76. German Wikipedia. „Game of Thrones/Staffel 8" – Versionsgeschichte. https://de.wikipedia.org/w/
738 index.php?title=Game_of_Thrones/Staffel_8&action=history, accessed on 2019-06-01.

77. English Wikipedia. Game of Thrones (season 8): Revision history. https://en.wikipedia.org/w/index.php?title=Game_of_Thrones_(season_8)&action=history, accessed on 2019-06-01.
78. WikiRank. Quality and popularity assessment of Wikipedia. https://wikirank.net/, accessed on 2019-04-27.
79. WikiRank. Films multilingual ranking. https://wikirank.net/top/film, accessed on 2019-06-01.
80. WikiRank. English Wikipedia. https://wikirank.net/en/, accessed on 2019-06-01.
81. WikiRank. Fortnite. https://wikirank.net/en/Fortnite, accessed on 2019-06-01.
82. WikiRank. Minecraft. https://wikirank.net/en/Minecraft, accessed on 2019-06-01.
83. WikiRank. Dota 2. https://wikirank.net/en/Dota_2, accessed on 2019-05-05.
84. WikiRank. Grand Theft Auto V. https://wikirank.net/en/Grand_Theft_Auto_V, accessed on 2019-05-05.
85. Deutschsprachige Wikipedia. Kategorie: Kunst und Kultur. https://de.wikipedia.org/wiki/Kategorie:Kunst_und_Kultur, accessed on 2019-05-05.
86. English Wikipedia. Wikipedia:Protection policy. https://en.wikipedia.org/wiki/Wikipedia:Protection_policy, accessed on 2019-05-05.
87. English Wikipedia. Bitcoin. https://en.wikipedia.org/wiki/Bitcoin, accessed on 2019-06-01.
88. GitHub. Lewoniewski - user profile. https://github.com/lewoniewski, accessed on 2019-05-05.
89. The Linked Open Data Cloud. Datasets. https://lod-cloud.net/datasets, accessed on 2019-05-05.
90. Kaggle. Datasets. https://www.kaggle.com/datasets, accessed on 2019-05-05.
91. Wikimedia Toolforge. Global user contributions. https://tools.wmflabs.org/guc/, accessed on 2019-06-01.
92. WikiTop. Wikipedians Top. http://wikitop.org/, accessed on 2019-06-01.