Matching Ukrainian Red Links with English Wikipedia's Articles

https://github.com/Katerali/redlinks_linking

https://doi.org/10.6084/m9.figshare.11550774

Kateryna Liubonko, UCU (aloshkina@ucu.edu.ua)
Diego Sáez-Trumper, Wikimedia Foundation (diego@wikimedia.org)

Problem & Approach

817 892 articles*

ВікіпедіЯ
Вільна енциклопедія

5 719 743 articles*
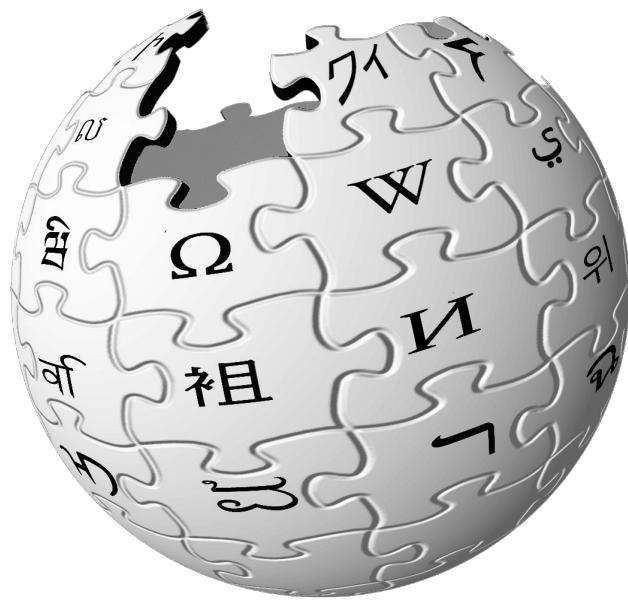
WIKIPEDIA
The Free Encyclopedia

* Wikipedia dumps from the 20th of September, 2018

Results & Contribution

Similarity metrics	fp	tn	tp	fn	P	R	F$_1$
BabelNet	4	65	108	456	0.964	0.191	0.32
Incoming links (top1, th=0.26)	385	149	64	37	0.143	0.634	0.233
Concurrent links (top1, th=0.1)	263	360	2	10	0.008	0.167	0.014
Levenshtein (top1, th=0.39)	295	24	41	275	0.122	0.13	0.126
Multi-factored model (top1)	31	1445	339	90	0.92	0.79	**0.85**

Table 2: Evaluation results on similarity metrics as independent models and on a multi-factored model based on the results of the independent ones.
The best F$_1$ score is in bold

Absolute value of logreg coefficients
(scaled by maximum coefficient)
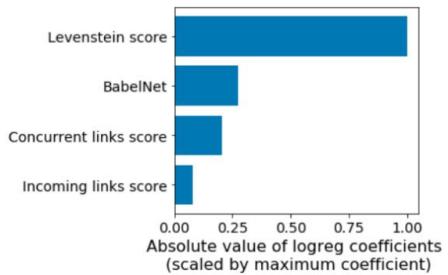
Dataset of 2 957 927 pairs of red links and the candidate articles in English Wikipedia for the most frequent 3 171 red links from Ukrainian Wikipedia.
Publicly released at
https://doi.org/10.6084/m9.figshare.11550774

Code publicly released at
https://github.com/Katerali/Red_links_Project_for_Wiki.