



APPLIED  
SCIENCES  
FACULTY



# Matching Ukrainian Red Links with English Wikipedia's Articles



[https://github.com/Katerali/redlinks\\_linking](https://github.com/Katerali/redlinks_linking)



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

Kateryna Liubonko, UCU ([aloshkina@ucu.edu.ua](mailto:aloshkina@ucu.edu.ua))

Diego Sáez-Trumper, Wikimedia Foundation ([diego@wikimedia.org](mailto:diego@wikimedia.org))

# Problem & Approach

2002—2003	Ландскрун (Финляндия)	11 (3)
2003—2004	Анкона	13 (4)
2004—2007	Стандард	Анкона (футбольный клуб) (еще не написана)
2008—2009	Трогир	7 (1)

Named Entity Linking  
on word and graph features

## U.S. Ancona 1905

From Wikipedia, the free encyclopedia

Not to be confused with A.C. Ancona.

Unione Sportiva Ancona 1905 [2][3] commonly referred to as Ancona, was an Italian football club based in Ancona, Marche.

The club changed its name to U.S. Ancona 1905 from S.S. Piano San Lazzaro in 2010, due to the bankruptcy of the main football team of the city, A.C. Ancona. However, after a relegation in 2017, U.S. Ancona 1905 withdrew from the football pyramid.

Contents [hide]
1 History
1.1 2010–11 season
1.2 2011–12 season
1.3 2012–13 season
1.4 Promoted to professionalism and disestablishment

\* Wikipedia dumps from the 20th of September, 2018



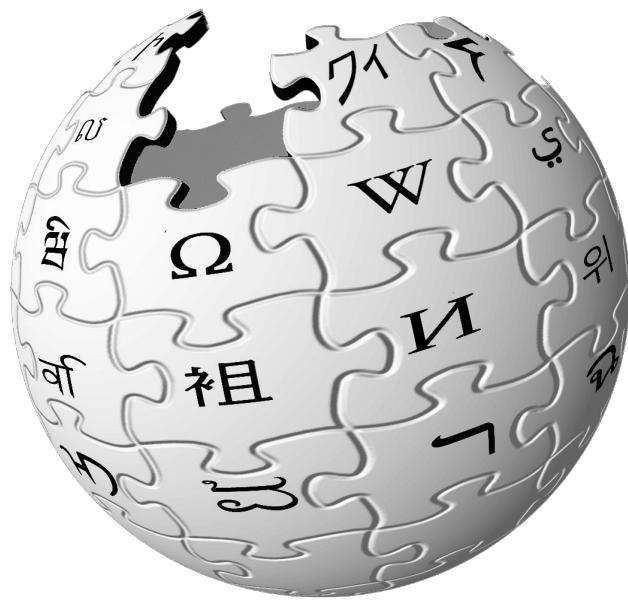
Source:  
<https://uk.wikipedia.org/wiki/Файл:Wikipedia-logo-v2-uk.svg>

817 892 articles\*



ВІКІПЕДІЯ  
Вільна енциклопедія

5 719 743 articles\*



WIKIPEDIA  
*The Free Encyclopedia*

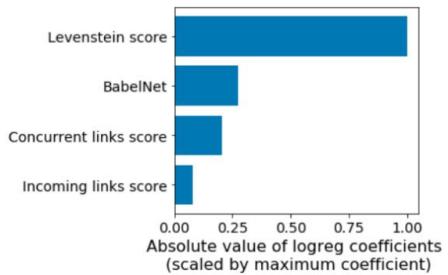
Source:  
<https://uk.wikipedia.org/wiki/Файл:Wikipedia-logo-v2-en.svg>

# Results & Contribution

Similarity metrics	fp	tn	tp	fn	P	R	F <sub>1</sub>
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	<b>0.85</b>

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



**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](https://github.com/Katerali/Red_links_Project_for_Wiki).