

GikiCLEF topics and Wikipedia articles: Did it blend?



Nuno Cardoso
Faculty of Sciences, University of Lisbon, LaSIGE,
ncardoso@xldb.di.fc.ul.pt



Motivation

Using Wikipedia to answer GikiCLEF topics proved to be quite difficult. Where is the problem, in the task or in the systems? Did the systems mined Wikipedia conveniently, or they did not manage to scratch the surface? Are we aware of the difficulties in finding the right answers in Wikipedia? Let's have a look on the Wikipedia collection and in the GikiCLEF task.

GikiCLEF task

- Find answers and justifications in Wikipedia articles, for 50 geographically-challenging topics.
- 50 topics, 10 Wikipedia languages
- 8 participant systems, 20 runs submitted, totaling 21251 answers (18152 unique). Correct answers: 1008.

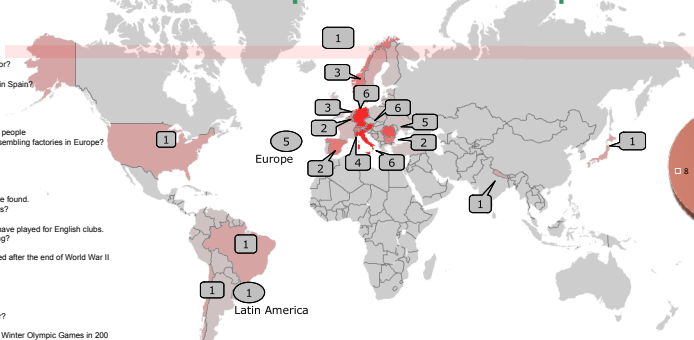
GikiCLEF topics

- [IT] 01. List the Italian places where Ernest Hemingway visited during his life.
- [] 02. Which countries have the white, green and red colors in their national flag?
- [BG] 03. In which countries outside Bulgaria are there published opinions on Petar Duno's ideas?
- [RO] 04. Name Romanian poets who published volumes with ballads until 1941.
- [] 05. Which written fictional works of non-Romanian authors have as subject the Carpathian mountains?
- [NL] 06. Which Dutch violinists held the post of concertmaster at the Royal Concertgebouw Orchestra in the twentieth century?
- [NL] 07. What capitals of Dutch provinces received their town privileges before the fourteenth century?
- [DE] 08. Which authors were born in and write about the Bohemian Forest?
- [DE] 09. Name places where Goethe fell in love.
- [NL] 10. Which Flemish towns hosted a restaurant with two or three Michelin stars in 2007?
- [NL] 11. What Belgians won the Ronde van Vlaanderen exactly twice?
- [] 12. Present monarchies in Europe headed by a woman.
- [] 13. Romantic and realist European novelists of the XXth century who died of tuberculosis.
- [] 14. Name rare diseases with dedicated research centers in Europe.
- [IT] 15. List the basic elements of the cassata.
- [] 16. In which European countries is the bialé commonly used?
- [IT] 17. List the 5 Italian regions with a special statute.
- [IT] 18. In which Tuscan provinces is Chianti produced?
- [ES] 19. Name mountains in Chile with permanent snow.
- [] 20. List the name of the sections of the North-Western Alps.
- [IT] 21. List the left side tributaries of the Po river.
- [ES,PT] 22. Which South American national football teams use the yellow color?
- [EN] 23. Name American museums which have any Picasso painting.
- [ES] 24. Which countries have won a total European championship played in Spain?
- [ES] 25. Name Spanish drivers who have driven in Mirami.
- [BG] 26. Which Bulgarian fighters were awarded the "Diamond belt"?
- [NL] 27. Which Dutch bands are named after a Bulgarian footballer?
- [PT] 28. Find coastal states with Petrobras refineries.
- [] 29. Places above the Arctic circle with a population larger than 100,000 people.
- [] 30. Which Japanese automakers companies have manufacturing or assembling factories in Europe?
- [IT] 31. Which countries have Italian as an official language?
- [RO] 32. Name Romanian writers who were living in USA in 2003.
- [] 33. What European Union countries have national parks in the Alps?
- [] 34. What eight-thousanders are at least partially in Nepal?
- [RO] 35. Which Romanian mountains are declared biosphere reserves?
- [RO] 36. Name Romanian caves where Paleolithic human fossil remains were found.
- [NO,NB] 37. Which Norwegian musicians were convicted for burning churches?
- [NO,NB] 38. Which Norwegian waterfalls are higher than 200m?
- [NO,NB] 39. National team football players from Scandinavia with sons who have played for English clubs.
- [DE] 40. Which rivers in North Rhine Westphalia are approximately 10km long?
- [DE] 41. Chefs born in Austria who have received a Michelin Star.
- [DE] 42. Political parties in the National Council of Austria which were founded after the end of World War II
- [DE] 43. Austrian ski resorts with a total ski trail length of at least 100 km
- [DE] 44. Find Austrian grape varieties with a vineyard area below 100 ha.
- [] 45. Find Swiss casting show winners.
- [DE] 46. German writers who are Honorary Citizens in Switzerland.
- [DE] 47. Which cities in Germany have more than one university?
- [DE] 48. Which German-speaking movies have been nominated for an Oscar?
- [] 49. Formula One drivers who moved to Switzerland?
- [] 50. Which Swiss people were Olympic medalists in snowboarding at the Winter Olympic Games in 200

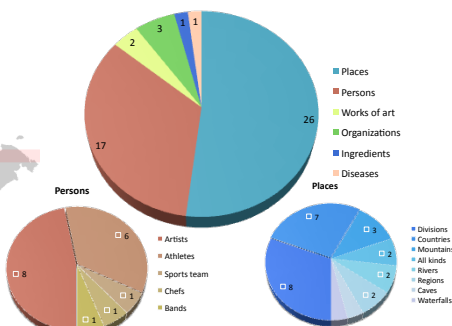
Language bias of topics

none	de	it	nl	ro	es	no	bg	pt	en
13	10	6	5	5	4	3	2	2	1

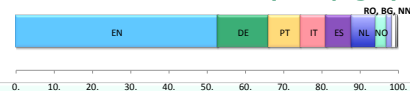
Geoscope distribution of topics



Expected Answer types



GikiCLEF collection sizes (nr of pages)

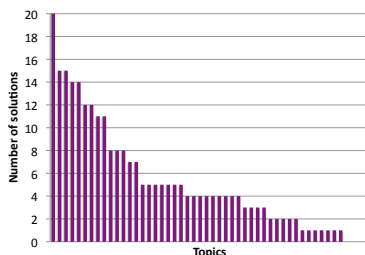


Wikipedia solutions

Solutions: entities found in the pool of all GikiCLEF participant's answers, merged to a single identifier
Example: Italy is one solution for a topic, regardless of the different answers given by Wikipedia articles from different languages.

Italy, Itàlia, Italia, Italien, Италия → **Italy** (country)

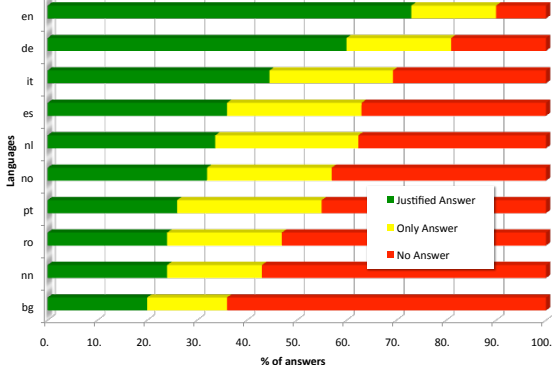
Number of solutions per topic



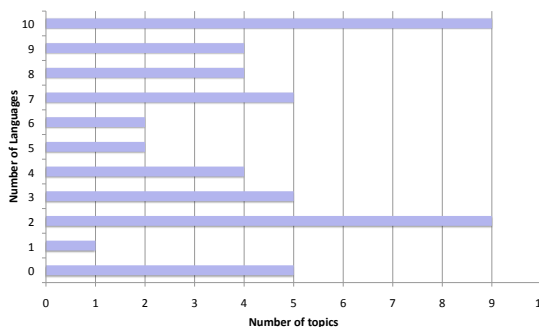
Best languages per topic bias

		bg	de	en	es	it	nl	nn	no	pt	ro
none	13	7	10	11	8	9	7	5	7	7	7
bg	2	2	0	2	0	0	0	0	0	0	0
de	10	3	9	9	4	5	5	3	4	4	2
en	1	0	1	1	1	1	1	0	1	1	0
es	4	1	3	3	2	1	1	0	1	2	1
it	6	3	6	6	5	6	4	3	5	5	3
nl	5	2	3	4	2	3	3	1	3	2	1
nn,no	3	0	2	2	1	1	2	2	2	1	1
pt	2	0	2	2	1	0	0	0	0	2	0
ro	5	1	3	5	3	3	2	1	2	2	4

Language coverage over the solutions



Language autonomy on topics



Solution location in Wikipedia

- Answer in the body text:
EN: 156 PT: 43

- Answer in the infobox:
EN: 30 PT: 14

In a non-answer page in the same language:
EN: 45 PT: 76

No answer:
EN: 26 PT: 117

Salient blending disturbers:

- Wikipedia is still too geographically on English language, which performed better on biased topics.
- The solutions of GikiCLEF 2009 topics were mostly found on the text body. Structured elements, like infoboxes, tables or ordered lists, are helpful but are rare and still machine-unfriendly.
- There was a significant amount of translated pages (EN / X or X / EN) as well as solutions that are normally on two languages (EN and X), thus no information gain of processing multiple languages.
- Wikipedia categories were difficult to work on – they differ on usage patterns per language, they are noisy and too structured (ex: Cities of Germany).