gesis Leibniz Institute for the Social Sciences

Cultural Relation Mining on Wikipedia

According to Wikipedia…

Most similar to Ukrainian music?

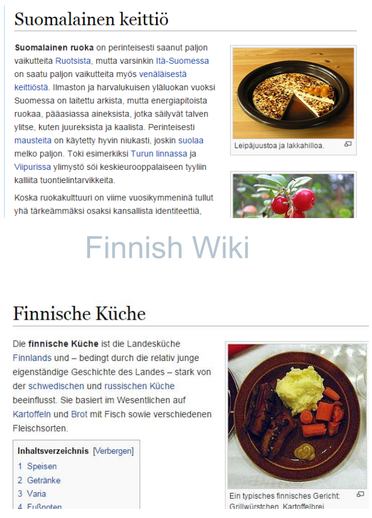
Best understanding of Spanish cuisine?

Most culturally interesting in Europe?

Most interesting cultural aspects?

Objective

Assess cultural relations between European countries → by analyzing different language editions of Wikipedia and their ***description of & interest in*** several cultural aspects of other countries.



Data

Wikipedia Articles Dataset: 9 different cultural aspects over 33 European countries in 27 different language editions

Outlinks: All linked WP concepts contained in the articles about a cultural aspect

Views Counts: Page views of the cultural aspect articles in each separate language edition

Method (metrics)

Cultural Similarity

$$sim_{tf-idf}(A,B) = \frac{\sum_{j=1}^{N} \overbrace{log\frac{N}{df(t_j)}}^{\text{inv. doc. freq.}} \times \overbrace{tf(t_j,A)}^{\text{j concept freq. in A}} \times \overbrace{tf(t_j,B)}^{\text{j concept freq. in B}}}{\| A \| \times \| B \|}$$

Example:
Similarity between German and Italian music
A – set of concepts from German music articles, aggregated over all languages
B – set of concepts from Italian music articles, aggregated over all languages

Cultural Interest (bias)

$$bias(l,o) = \overbrace{\frac{f(l,o)}{\sum_{\tilde{o}\in O} f(l,\tilde{o})}}^{\text{attention towards all resources}} - \frac{1}{|L|-1} \underbrace{\sum_{\bar{l}\in L/l} \frac{f(\bar{l},o)}{\sum_{\tilde{o}\in O} f(\bar{l},\tilde{o})}}_{\text{normalized relative attention of others}}$$

attention towards resource o

Example:
Interest of Italian language community in German music
l – Italian Wikipedia, $\bar{l}$ – other Wikipedia,
(l,o) – set of articles from “Musica tedesca" article in Italian Wikipedia,
O – set of articles about country specific music, L – set of Wikipedias

Cultural Understanding

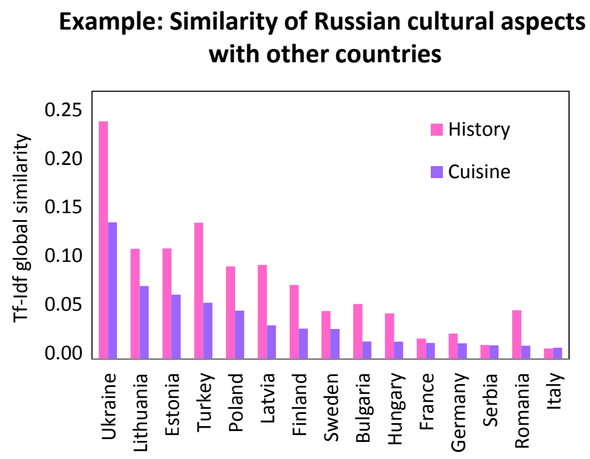
$$und_{tf-idf}(A',A'') = sim_{tf-idf}(A',A'')$$

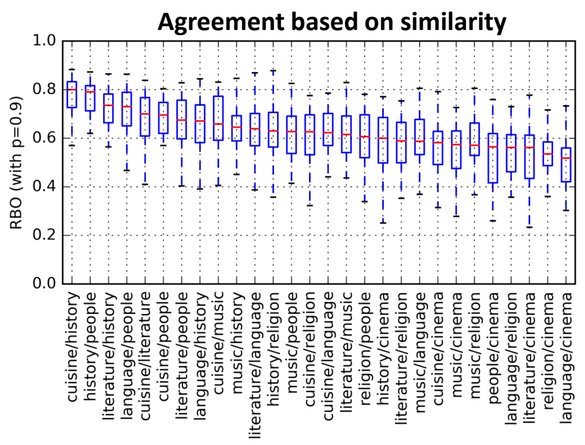
Example:
Understanding of German music by Italian language community
A' – set of concepts from German Wikipedia article "Musik in Deutschland" (native definition)
A'' – set of concepts from Italian Wikipedia article "Musica tedesca" (external definition)

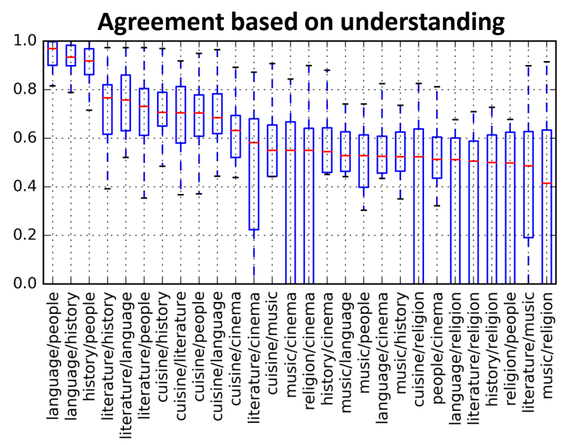
Results

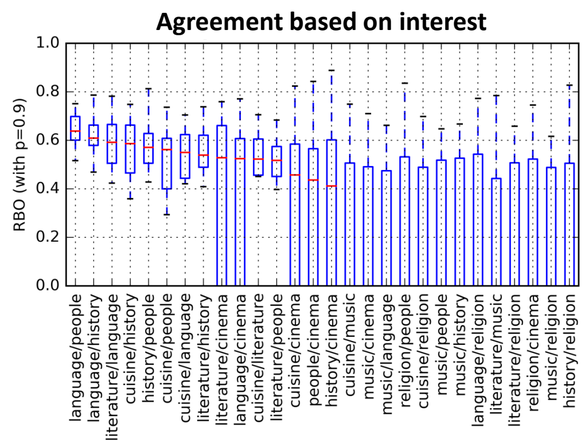
Agreement: How similar are our results for different cultural aspects?

For each language/country, we produce a ranked list of (i) most similar, (ii) best understood and (iii) most interesting countries based on each cultural aspect and based on each metric.









- In several cases, high levels of agreement can be observed between cultural aspects, pointing to a consistent relation being measured.
- While for understanding, the agreement tends to be medium/high, for similarity and interest it is notably lower for many aspect pairs.

Validity: comparing with further (external) data

The European Social Survey

Interest between countries ←→ Similarity score based on European Social Survey items on shared values. All aspects tested separately, linear regression.

	coef	Sig.
Cuisine	0.180	0.00
Literature	0.084	0.14
Music	0.007	0.9
Language	0.209	0.00
History	0.090	0.03
Religion	0.055	0.39
Cinema	0.131	0.02
World Heritage Sites	0.074	0.14
People	0.170	0.00
Economy	0.096	0.32
Politics	0.0605	0.47

R^2 of a complete regression model controlled for multicollinearity is 0.132

Migration Data

Pearson's correlation, for cuisine only:

	coef	Sig.
understanding – migration	0.36	0.00

Proximity and similarity (acc. to Tobler)

For all language editions: Difference of similarity scores with neighboring countries w.r.t. average similarity of non-neighbours (Example: Music)

Music	Sim	Music	Sim
Belgian	+7.49	Italian	+1.63
Bosnian	+10.82	Latvian	+4.88
British	+2.43	Lithuanian	+3.54
Bulgarian	+6.25	Norwegian	+3.99
Catalan		Polish	+1.79
Croatian	+5.06	Portuguese	+6.52
Czech	+4.27	Romanian	+3.77
Danish	+2.36	Russian	+2.82
Dutch	+5.49	Serbian	+6.78
English		Slovak	+4.87
Estonian	+2.48	Spanish	+2.46
Finnish	+3.30	Swedish	+7.09
French	+3.29	Turkish	+3.77
German	+2.55	Ukrainian	+3.90
Hungarian	+4.10		
Irish	+8.09		
Average			+4.50
Std.Dev			2.10

- Consistently higher similarity for neighbors, but not consistent regarding understanding and affinity
- Weak relation with ESS and migration data
- Turkers identified 99% of high/low similarity pairs correctly (cuisines)

P. Laufer et.al., ***Mining cross-cultural relations from Wikipedia - a study of 31 european food cultures***, *ACM Web Science 2015*

Olga Zagovora, GESIS & Univ. Koblenz
Tatiana Sennikova, GESIS & Univ. Koblenz
Claudia Wagner, GESIS & Univ. Koblenz
Fabian Flöck, GESIS (fabian.floeck@gesis.org)

Member of the Leibniz Association