Why are you reading this article today?

Waarom lees je dit artikel vandaag?

為什麼你今天會讀這篇條目？

Warum lesen Sie diesen Artikel gerade?

למה אתה קורא את הערך הזה היום?

यह लेख आज आप क्यों पढ़ रहे हैं

Miért olvasod most ezt a szócikket?

De ce citiți acest articol anume astăzi?

あなたは今日何のためにこの項目を読んでいますか？

Почему вы читаете эту статью сегодня?

Чому Ви читаєте цю статтю сьогодні?

কেন এ নিবন্ধটি আজ পড়ছেন?

ماذا تقرأ هذه المقالة اليوم؟

Por qué estas leyendo este artículo hoy?

Why the World Reads Wikipedia

Florian Lemmerich, Diego Saez-Trumper, Robert West, Leila Zia

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

RWTH AACHEN UNIVERSITY

WIKIMEDIA FOUNDATION

2018-12-12

Who are they?

What are they trying to achieve?

6,000 pageviews per second

What languages they read in?

How do they learn?

Content

Content representation

Tool and feature development

Policy

Knowledge equity

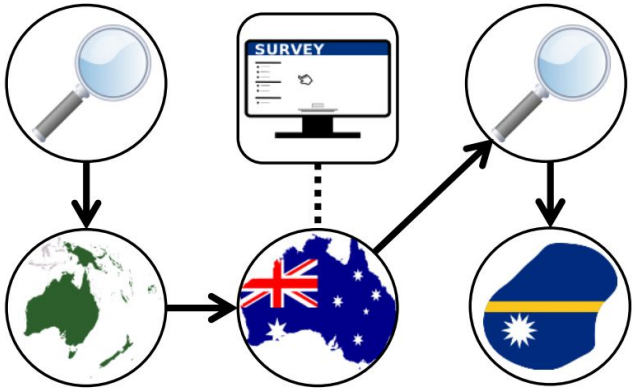
Wikimedia Movement Strategic Direction:
https://meta.wikimedia.org/wiki/Strategy/Wikimedia_movement/2017/Direction#Our_strategic_direction:_Service_and_Equity

By the end of this talk

- Are English Wikipedia reader behaviors and distribution of use cases representative of reader behavior in other Wikipedia languages?
- Are there commonalities between reader behaviors and distribution of use cases across Wikipedia languages?
- Do people read long articles or do they often come to Wikipedia to do quick look-ups?
- Are there readers that read Science, Education, Research, and Medicine articles more than others?
- ...

How do we do it?

Surveys



+

+

Webrequest logs

Country level statistics

Characterizing readers

While it is important to know what percentage of Spanish Wikipedia readers are students, we cannot stop once we learn the number. We need to understand what are the fundamental features that would allow us to characterize Wikipedia student readers across languages.

The Survey

Taxonomy of Wikipedia Readers

Singer, Philipp, et al. "Why we read wikipedia." *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2017. https://arxiv.org/abs/1702.05379

Why are you reading this article today?

Information need

I am reading this article to

- ◯ look up a specific fact or to get a quick answer.
- ◯ get an overview of the topic.
- ◯ get an in-depth understanding of this topic.

Prior knowledge

Prior to visiting this article

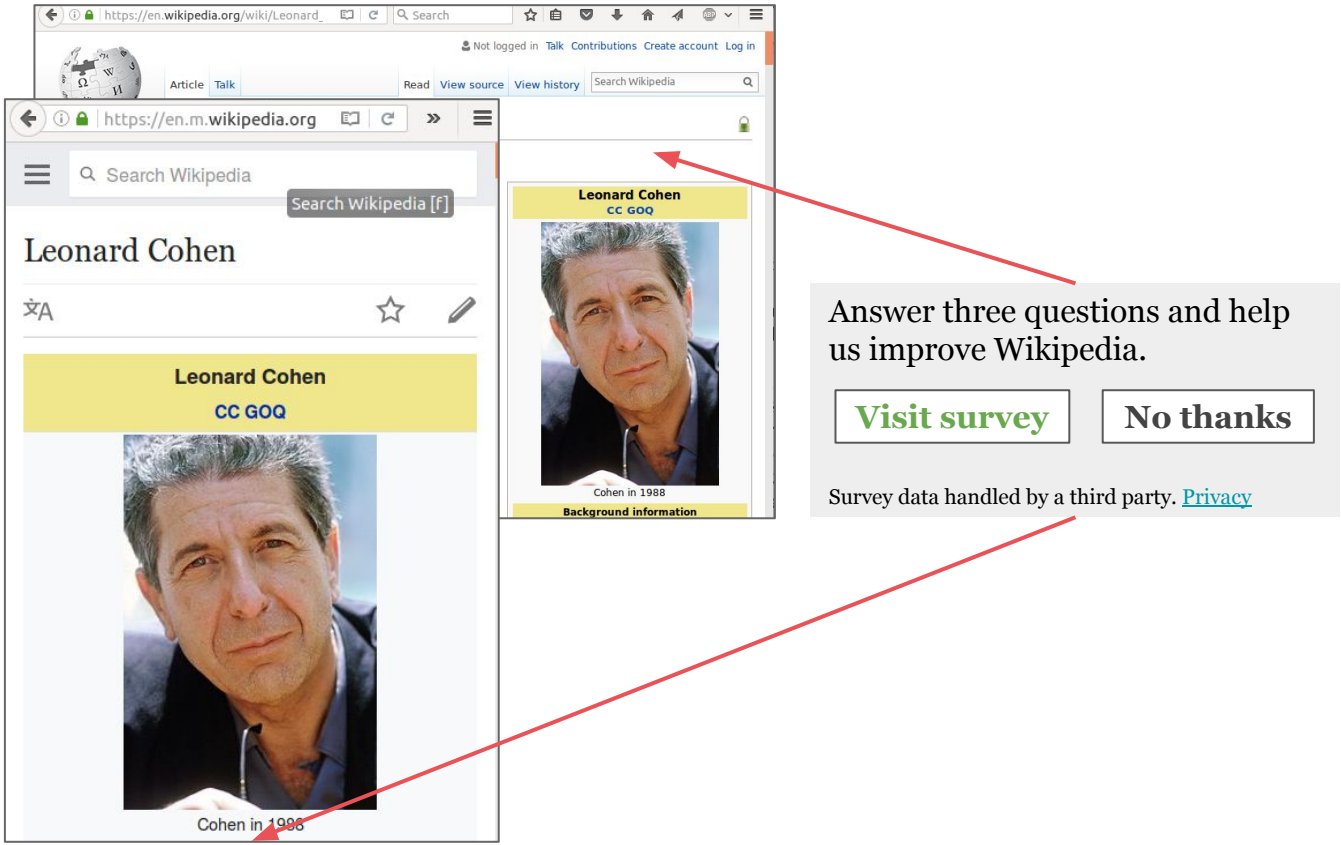
- ◯ I was not familiar with the topic and I am learning about it for the first time.
- ◯ I was already familiar with the topic.

motivation

I am reading this article because

Please select all answers that apply

- ☐ the topic was referenced in a piece of media (e.g., TV, radio, article, film, book).
- ☐ I need to make a personal decision based on this topic (e.g. to buy a book, choose a travel destination).
- ☐ I am bored and randomly exploring Wikipedia for fun.
- ☐ the topic came up in a conversation.
- ☐ I have a work or school-related assignment.
- ☐ I want to know more about a current event (e.g. a soccer game, a recent earthquake, somebody's death)
- ☐ this topic is important to me and I want to learn more about it. (e.g., to learn about a culture).
- ☐ Other ☐


https://en.wikipedia.org/wiki/Leonard_
Search
Not logged in Talk Contributions Create account Log in
Article Talk
Read View source View history
Search Wikipedia
https://en.m.wikipedia.org
Search Wikipedia
Search Wikipedia [f]
Leonard Cohen
Leonard Cohen
CC GOQ
Cohen in 1988
Background information
Leonard Cohen
CC GOQ
Cohen in 1988
Answer three questions and help us improve Wikipedia.
Visit survey
No thanks
Survey data handled by a third party. Privacy

The survey

- Duration: 1 week, June 22-29, 2017
- 14 languages:
  Arabic, Bengali, Chinese, Dutch, English, German, Hebrew, Hindi, Hungarian, Japanese, Romanian, Russian, Spanish, and Ukrainian
- Mobile and Desktop platforms
- Sampling rate: varied across languages (1:40 en, 1:1 bn)
- On article pages and to those with “Do not Track” off
- Responses: 215,000+

The data …

Survey
Motivation
Information need
Prior knowledge

Request
Country
Continent
Local time weekday
Local time hour
Host
Referer class

Article
In-degree
Out-degree
Pagerank
Text length
Pageviews
Topics
Topic entropy

Session/Activity
Session length
Session duration
Average dwell time
Average pagerank difference
Average topic distance
Referer class frequency
Session position
Number of sessions
Number of requests

Use of log data

- Bias correction
    - Compare response users with random sample
    - Correct for overrepresentation by weighting responses
    - Inverse propensity score weighting based on gradient boosting

- Find associations between behavior and use case

Survey results

Outline

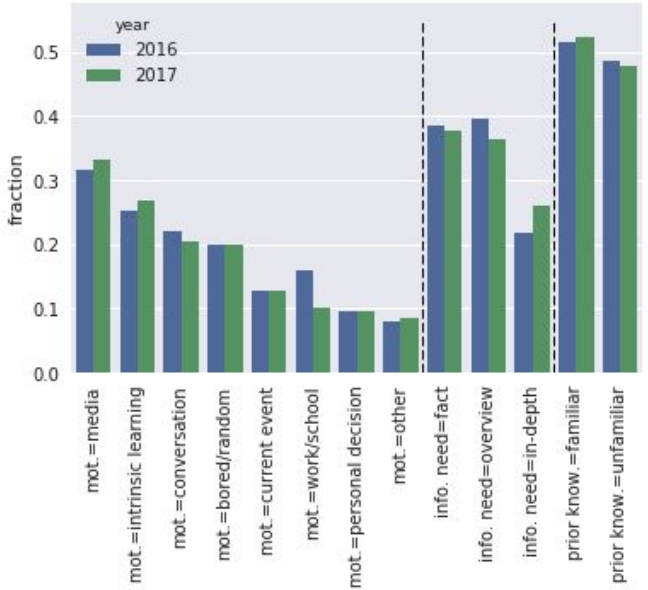
- Robustness
- Direct survey results
- Survey results and Webrequest Logs
- Survey results and country level data

Averaging over the 14 languages will hide important results!

Results I:

Robustness

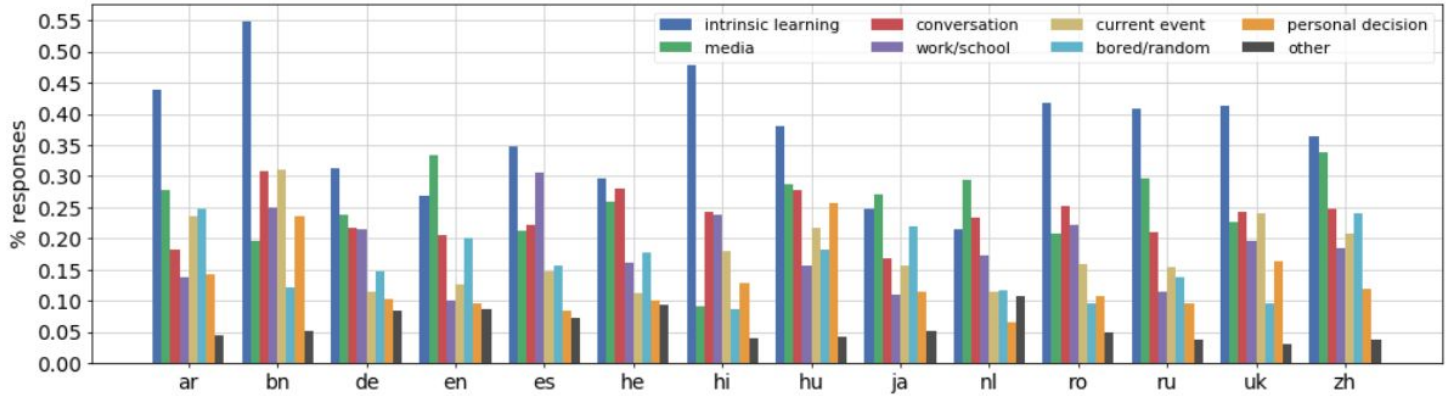
Robustness: “en” 2016 vs “en” 2017



- Overall: very similar results
- Difference for “motivation = work/school”

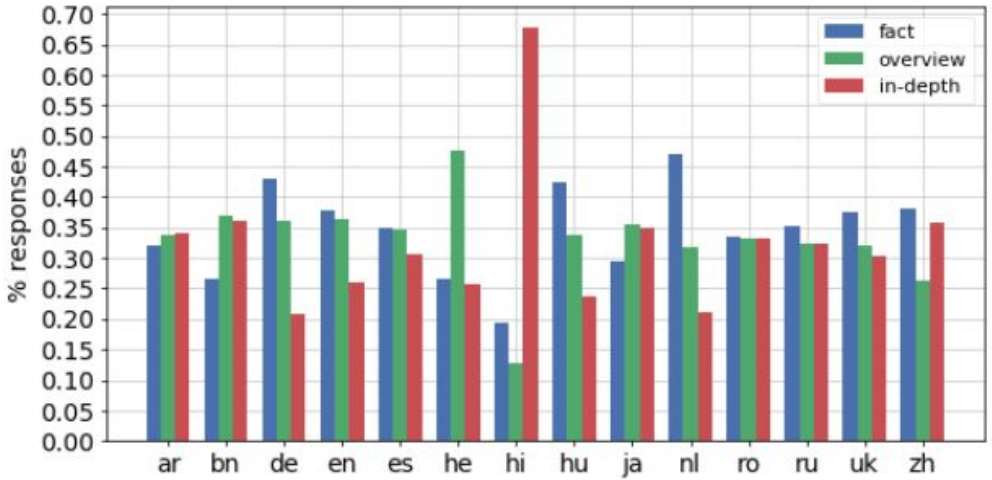
Results II:
Direct survey results

Results: Motivation



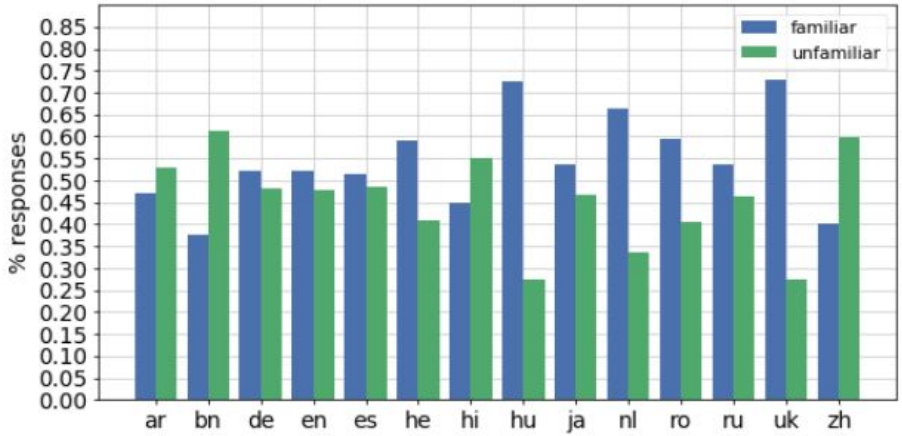
- Mixture of motivations in all languages
- Intrinsic learning:
  - Most prevalent in all but 3 languages
  - More common in Eastern and central asian languages
- Media is top motivation in the 3 languages (en, nl, ja)
- Partially strong differences:
  E.g.: work/school: en → ~10%, es → ~30%
  E.g.: bored/random: hi/ro → ~10%, en, ja → ~ 20%

Results: Information need



- Overall, the three information needs are relatively equally common
- In-depth less common in western/central European languages (de, en, es, hu, nl)
- More fact checking in these languages
- Hindi is a strong outlier

Results: Prior knowledge

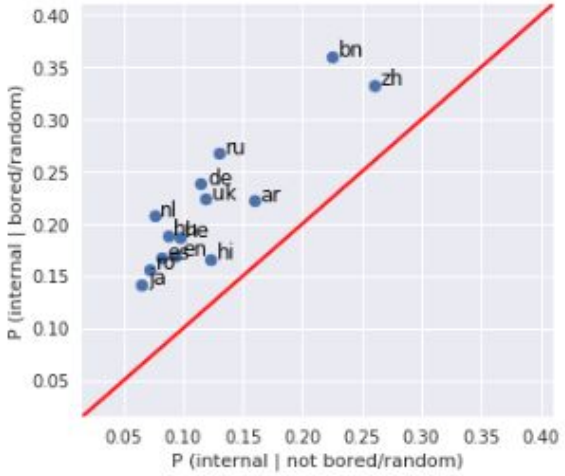


- Overall, roughly the same number of people feel familiar vs. unfamiliar
- Eastern European languages (hu, ro, ru, uk, also nl) feel more familiar
- Asian language (not ja) report to be more unfamiliar
  ➔ social desirability of humility?

Results III:

Survey Results and Webrequest Logs

Survey results and log patterns



- Take patterns of viewing behavior
  E.g., “at night”, “session_length > 3”, …
- Look at pairs of behavioral pattern and survey answers
- Plot for each language:
  likelihood of pattern in presence of survey answer
  vs. likelihood of pattern in absence of survey answer
- Point above red diagonal:
  association between pattern and answer
- Do this for all 247 pairs, sort by *effect*

Survey results and log patterns

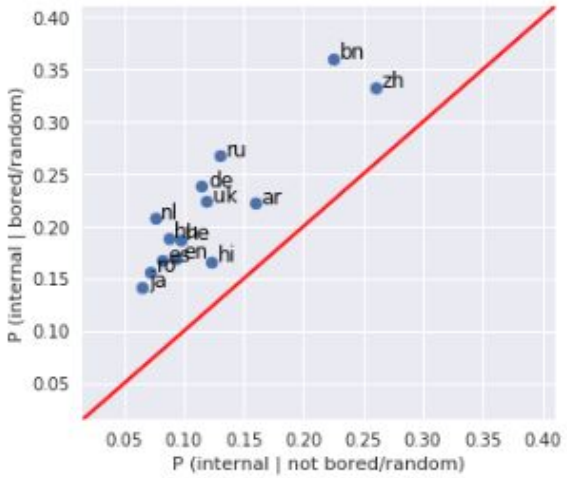
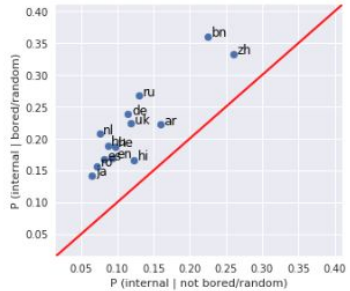


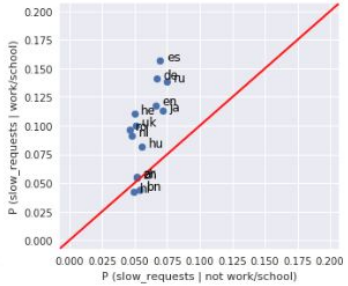
Table 2: Pairs of usage patterns and survey responses with the largest normalized mean effect $\bar{\mu}(E)$ across language editions. This table provides for each pair information on the mean share (likelihood of the pattern) $\mu(S)$, and the relative standard deviation of the share $rs(S)$ across language editions. Furthermore it displays the mean effect (increase of the pattern likelihood in presence of the response) $\mu(E)$, the normalized mean effect $\bar{\mu}(E)$, the standard deviation $\sigma(E)$ and the normalized standard deviation of the effect $\bar{\sigma}(E)$.

Pattern[4]	Response	$\mu(S)$	$rs(S)$	$\mu(E)$	$\bar{\mu}(E)$	$\sigma(E)$	$\bar{\sigma}(E)$
internal	mot.=bored/rand.	.136	.416	.095	.697	.028	.206
slow_requests	mot.=work/school	.065	.220	.038	.594	.030	.457
desktop	mot.=work/school	.342	.303	.187	.547	.122	.358
rapid_requests	mot.=bored/rand.	.102	393	.041	.405	.023	.229
long_sessions	mot.=bored/rand.	.252	.204	.097	.383	.047	.188
time:night	mot.=bored/rand.	.112	.541	.031	.281	.032	.289
long_article	prior knowl.=familiar	.143	.473	.036	.251	.032	.221
time:afternoon	mot.=work/school	.308	.116	.064	.207	.044	.142
time:night	mot.=media	.112	.541	.022	.197	.031	.281
internal	mot.=intrinsic learn.	.136	.416	.022	.163	.018	.131
long_sessions	info. need=in-depth	.252	.204	.040	.158	.019	.075
slow_requests	mot.=other	.065	.220	.009	.140	.021	.324
time:night	mot.=intrinsic learn.	.112	.541	.015	.131	.013	.114
weekday:Friday	mot.=bored/rand.	.113	.238	.015	.131	.018	.155
internal	info. need=in-depth	.136	.416	.017	.127	.015	.112
long_sessions	prior knowl.=familiar	.252	.204	.032	.126	.022	.088
desktop	mot.=other	.342	.303	.042	.124	.058	.169
long_sessions	mot.=intrinsic learn.	.252	.204	.030	.119	.024	.094
time:night	prior knowl.=familiar	.112	.541	.013	.118	.021	.192
long_article	mot.=current_event	.143	.473	.017	.117	.021	.144

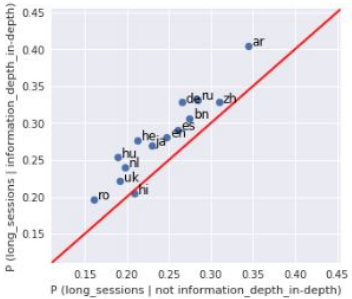
Survey results and log patterns



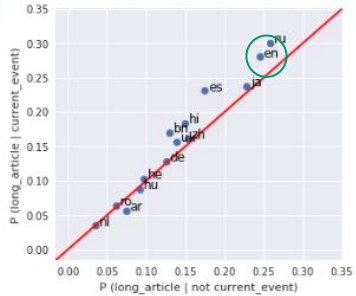
(a) Effect of motiv. = bored/random on internal referrer



(b) Effect of motiv. = work/school on slow requests



(c) Effect of info. need = in-depth on long sessions

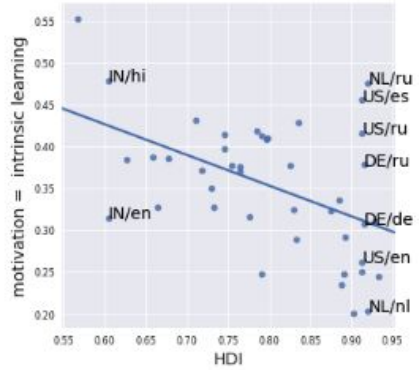


(d) Effect of motiv. = current event on long articles

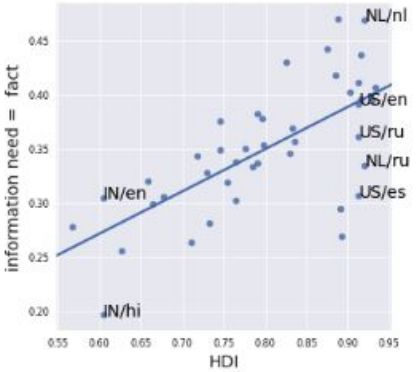
- Average effects across languages:
  - Bored/random associated with long sessions, internal navigation, night time requests, short time between requests
  - work/school associated with desktop usage, long intervals between requests, afternoon
  - conversation associated with less internal navigation, mobile platform, short dwelling times
- Mostly common effects across languages, exception work/school (b)
- Spread across languages stronger than effect of motivation (c+d)
- Not all patterns from en wiki hold across languages (d)

Results IV:
Survey Responses and Country Statistics

Country level statistics



(a) HDI *vs.* intrinsic learning



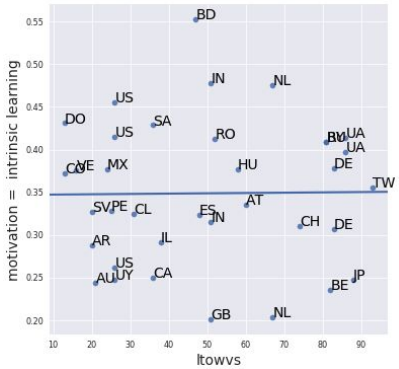
(b) HDI *vs.* fact checking

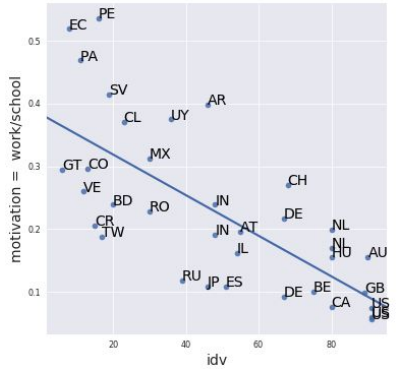
- Split data by language AND by country level
- 43 country/language pairs with at least 500 responses
- Get country level statistics, esp. Human Development Index
  HDI := geometric mean of life expectancy, education, income

- Compute correlations:

	Response	HDI	GDP p. cap.	Second. educ.
Motivation	media	0.63***	0.58***	0.42
	work/school	-0.55**	-0.54**	-0.40
	current event	-0.45*	-0.48*	-0.20
	intrinsic learning	-0.40	-0.43	-0.20
	personal decision	-0.28	-0.32	-0.08
	other	0.26	0.35	-0.08
	bored/random	0.21	0.25	-0.02
	conversation	-0.07	-0.12	-0.02
info. need	fact	0.66***	0.62***	0.55**
	in-depth	-0.60***	-0.57*	-0.46*
	overview	0.25	0.27	0.11
prior knowl.	familiar	0.44*	0.39	0.47*
	unfamiliar	-0.44*	-0.39	-0.47*

Cultural dimensions

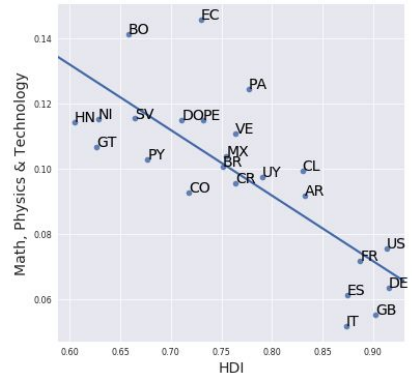


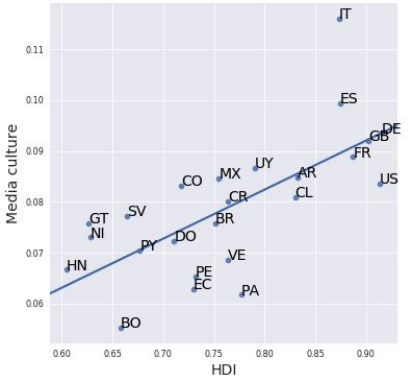


- Use country level data on Hofstede dimensions: Power distance, Individualism, Masculinity, Uncertainty Avoidance, Indulgence, Long term orientation
- Less clear correlations (exception: individualism)

	Response	LTO	IDV
Motivation	media	0.39	0.63***
	work/school	-0.37	-0.77***
	current event	0.13	-0.38
	intrinsic learning	0.00	-0.26
	personal decision	0.31	-0.14
	other	-0.37	0.04
	bored/random	-0.17	0.17
	conversation	0.22	0.13
info. need	fact	0.36	0.53*
	in-depth	-0.23	-0.43
	overview	-0.13	0.06
prior knowl.	familiar	0.27	0.42
	unfamiliar	-0.27	-0.42

Topics and HDI





- For Spanish Wikipedia specifically
- Look at topics (acquired by LDA) viewed from different countries (Can do this without survey data)

	feature	correlation_coeff	significance
17	Math, Physics & Technology	-0.753043	0.000434
15	Research & Education	-0.733913	0.000894
19	Medicine & Biology	-0.712174	0.001895
8	Media culture	0.712174	0.001895
7	Literature & Language	-0.646087	0.012967
6	Numbers, Lists & Sports	0.605217	0.034548
13	Sports & Teams	0.595652	0.042652

- “scientific/academic” topics more viewed in low HDI countries
- “leisure/entertainment” topics more viewed in high HDI countries
- For English: a less clear picture
- Future research warranted more fine grained topics/categories...

Summary

- Studied user motivation on Wikipedia with a large survey with more than 215,000 survey responses in 14 languages
- Combining survey data with webrequest logs allows for
  - Debiasing
  - Find associations between user behavior and use cases

- Found heterogeneous behavior across language editions
- Found globally valid links between use cases and log data
- Found correlations between socio-economic indicators and Wikipedia use cases

Discussion & future work

- Socio-demographics (age, gender, nationality, ...)
    - Does the characterizing of motivations change as a function of socio-demographics of readers?
    - Inequalities finer than country level?
    - ...
- Who is NOT reading Wikipedia?
- Language switching behavior
- A data challenge? Share your questions with us!

Thank you! :)

Paper link (arxiv):
https://arxiv.org/abs/1812.00474

Ongoing documentation
https://meta.wikimedia.org/wiki/Research:Characterizing_Wikipedia_Reader_Behaviour

TEC-9: Address Knowledge Gaps
https://www.mediawiki.org/wiki/Wikimedia_Technology/Annual_Plans/FY2019/TEC9:_Address_Knowledge_Gaps

Credits

- Page 1: Note that the logos used on the first slide belong to the corresponding institutions.

- Page 14: Leonard Cohen, 1988 01 by Gorupdebesanez, https://commons.wikimedia.org/wiki/File:Leonard_Cohen,_1988_01.jpg CC BY-SA 3.0 https://creativecommons.org/licenses/by-sa/3.0/deed.en