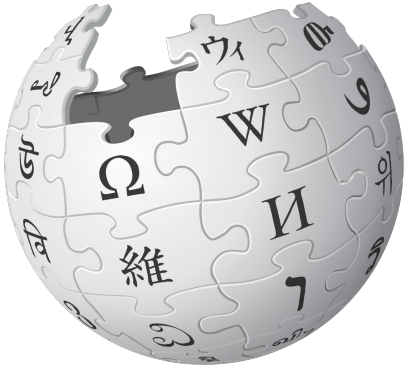
Growing Wikipedia Across Languages via Recommendations

Leila Zia

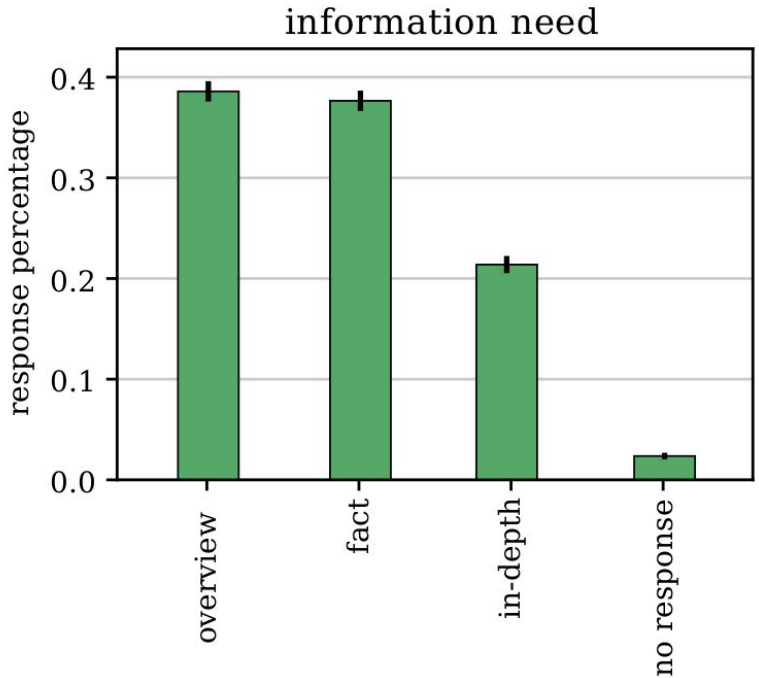
WIKIMEDIA
FOUNDATION

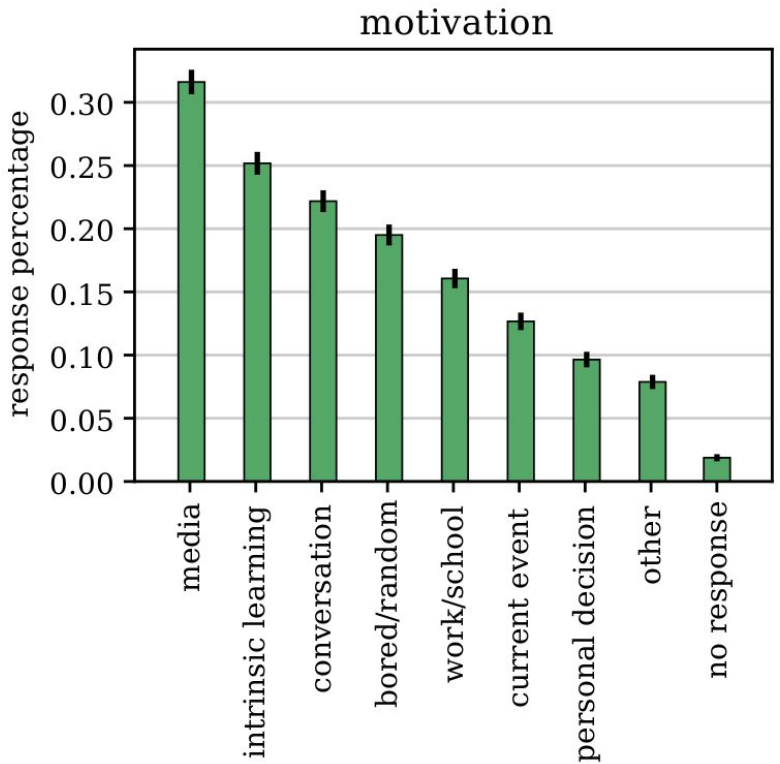
Wikipedia



- is the biggest encyclopedia in human history
- contains more than 43M articles in ~160 actively edited languages
- has around 60K active editors per month
- is viewed by humans 6000 times per second
- ...

Wikipedia

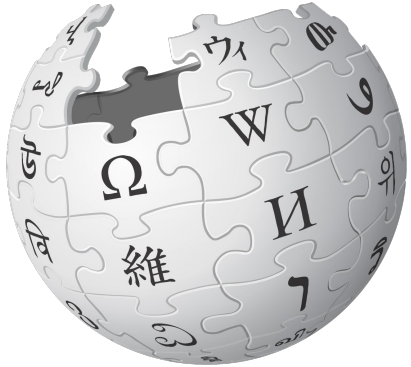




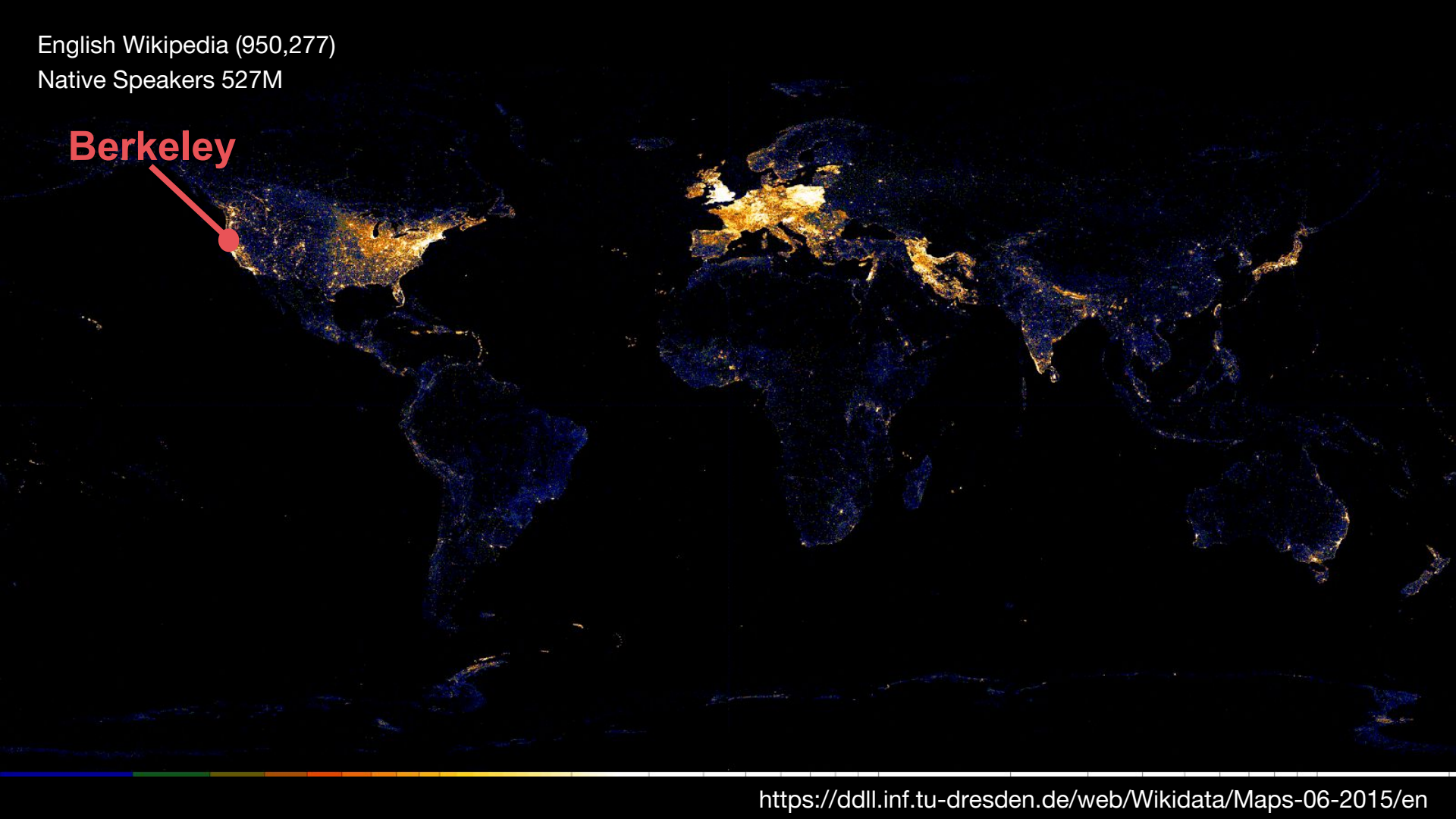
"Why We Read Wikipedia", Philipp Singer, Florian Lemmerich, Robert West, Leila Zia, Ellery Wulczyn, Markus Strohmaier, Jure Leskovec, to appear in International World Wide Web (WWW) Conference, Perth, Australia., 2017

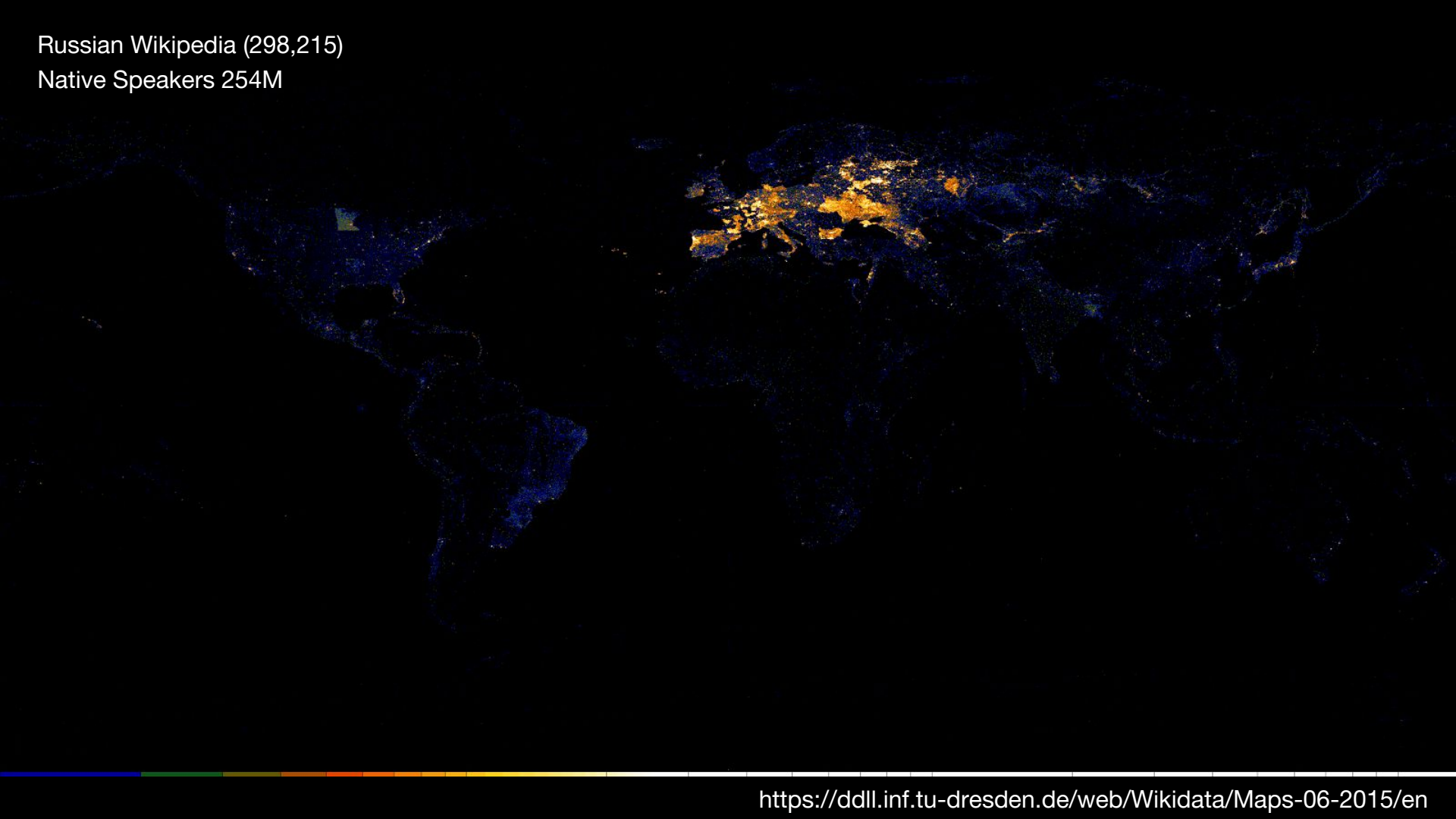
Wikipedia

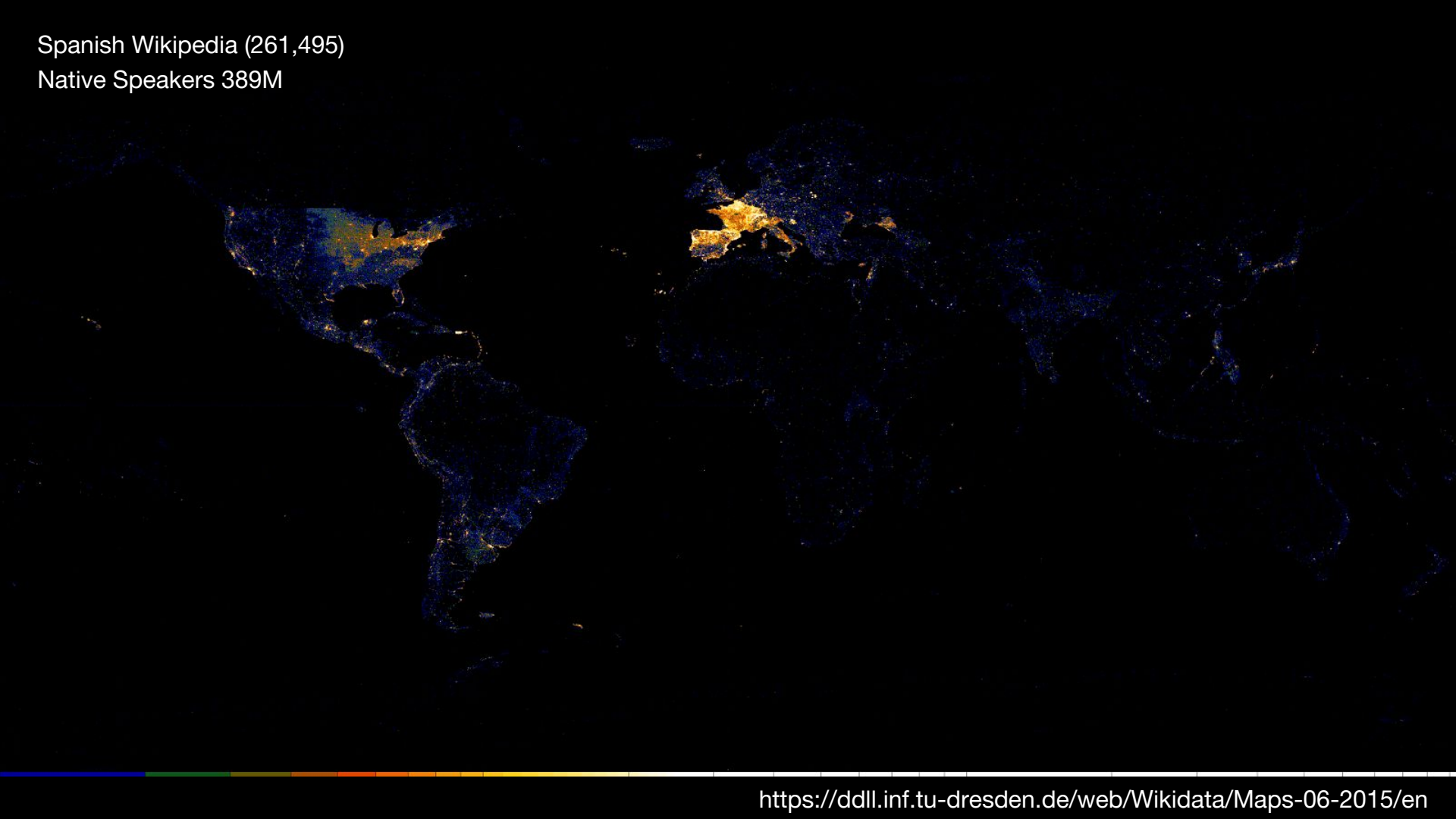
Wikipedia is our destination for accessing neutral point of view encyclopedic content in a variety of topics and across many languages.

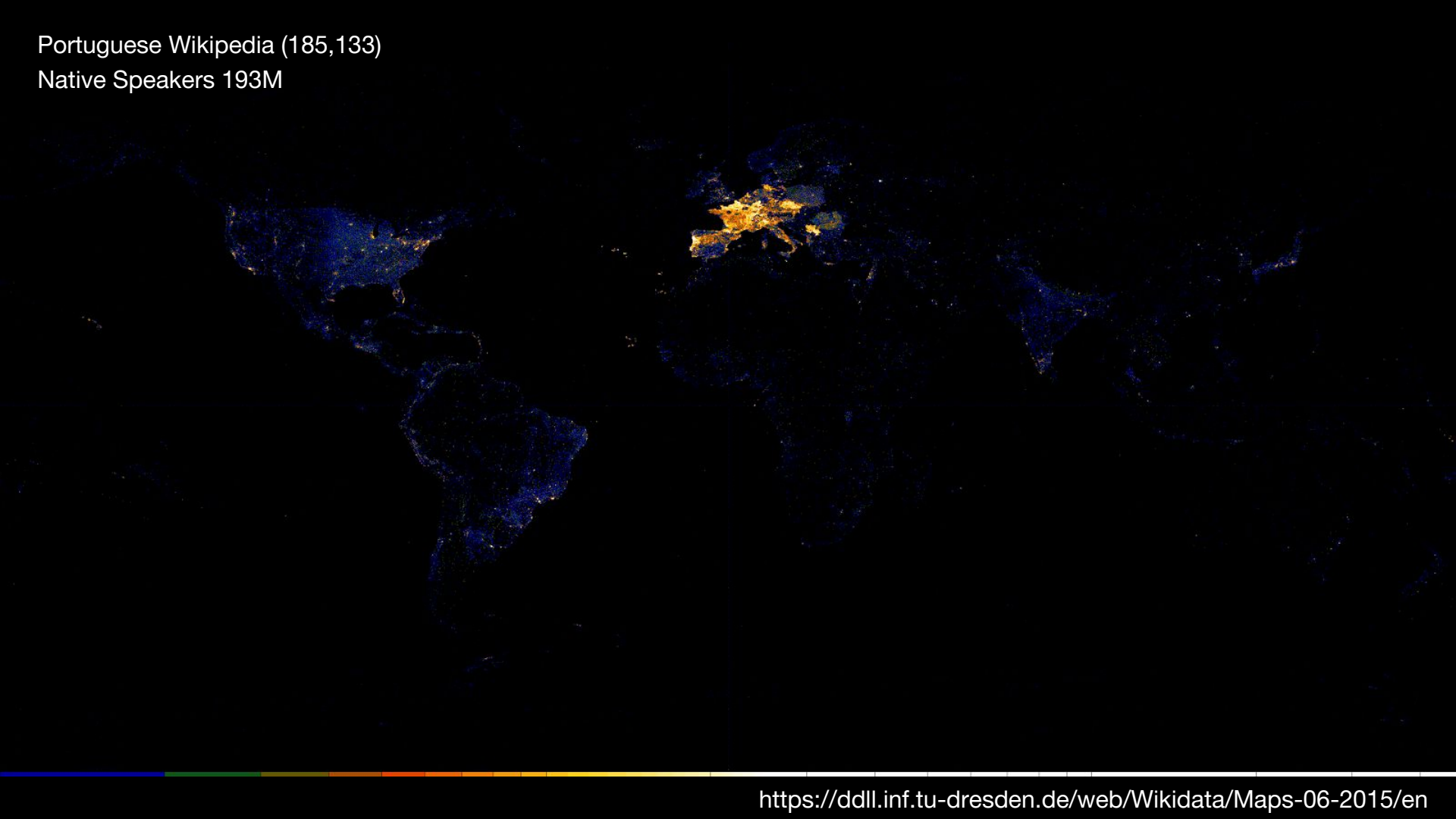


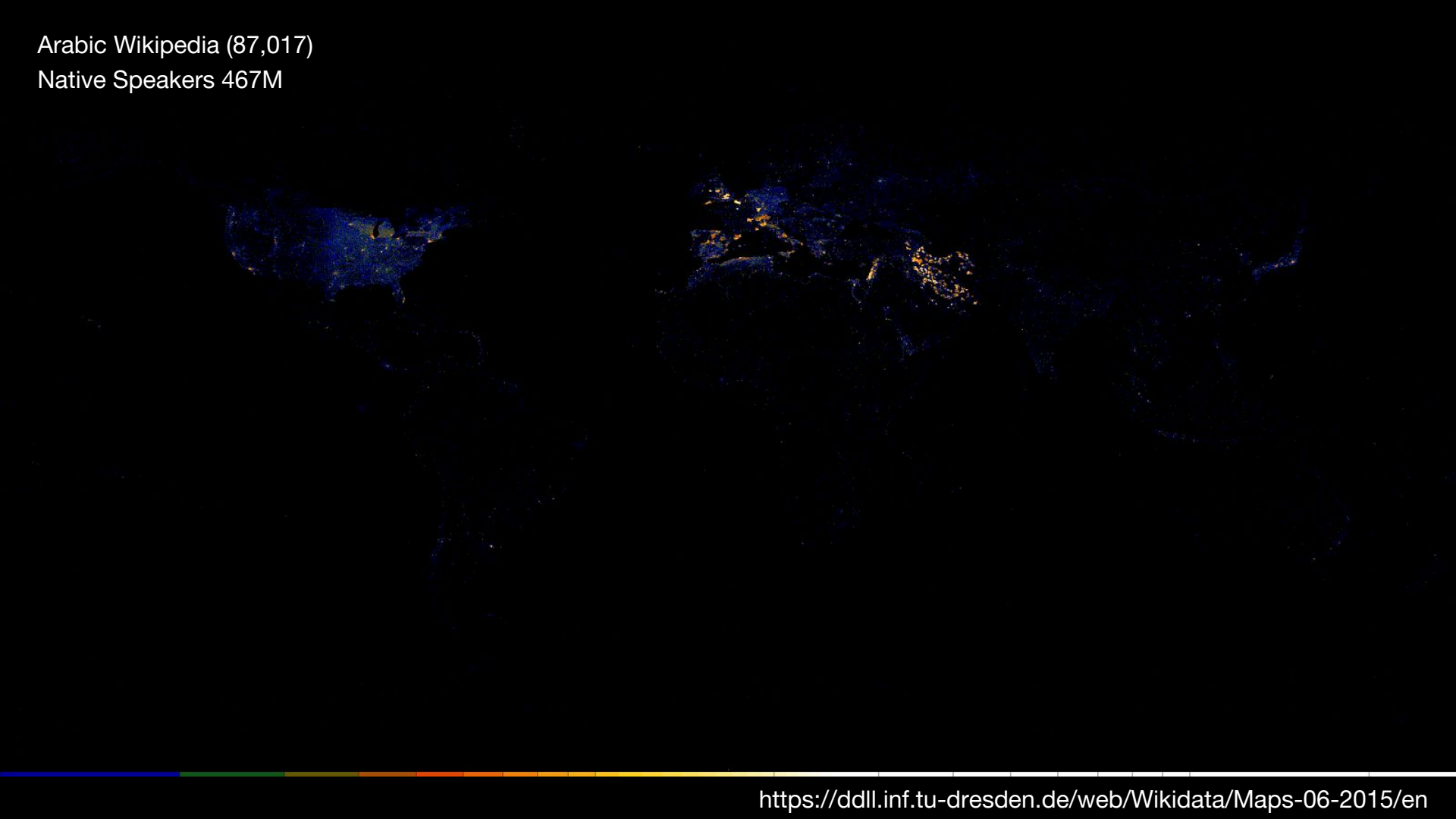
And yet, Wikipedia is incomplete!


English Wikipedia (950,277)
Native Speakers 527M
Berkeley
https://ddll.inf.tu-dresden.de/web/Wikidata/Maps-06-2015/en


Russian Wikipedia (298,215)
Native Speakers 254M
https://ddll.inf.tu-dresden.de/web/Wikidata/Maps-06-2015/en


Spanish Wikipedia (261,495)
Native Speakers 389M
https://ddll.inf.tu-dresden.de/web/Wikidata/Maps-06-2015/en


Portuguese Wikipedia (185,133)
Native Speakers 193M
https://ddll.inf.tu-dresden.de/web/Wikidata/Maps-06-2015/en


Arabic Wikipedia (87,017)
Native Speakers 467M
https://ddll.inf.tu-dresden.de/web/Wikidata/Maps-06-2015/en

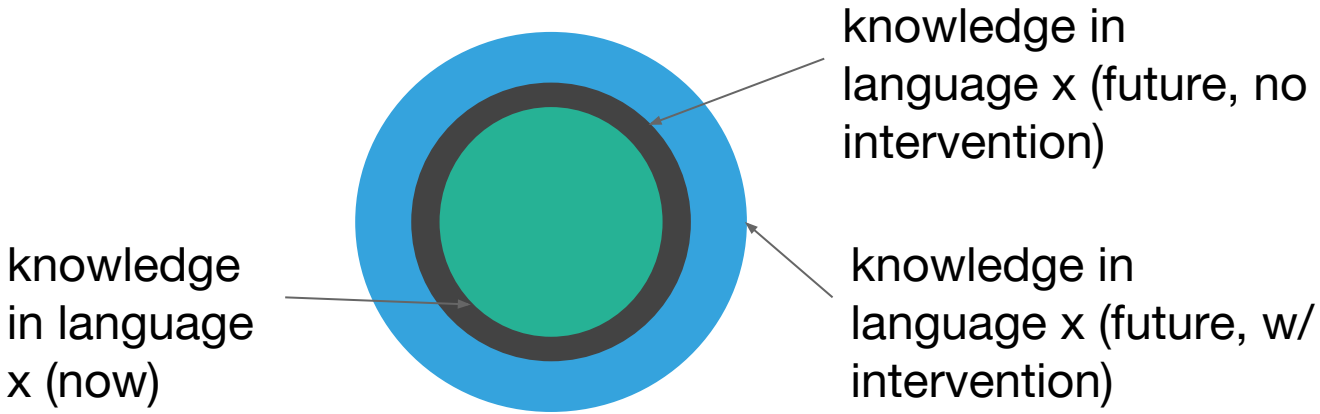
Let’s step back

- Demand
  - 2471 languages
  - More than 50% of the world’s population is monolingual
  - The next billion users are coming online in the coming 5 years
- Supply
  - Articles are created at a rate of 6500 per day
  - 60K active editors contribute to WP every month and this number has not changed significantly
  - 14K new accounts are created every month

How do we make access to all knowledge universal?

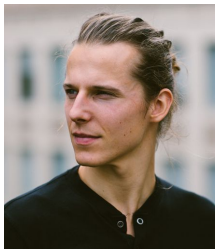
Goal

Increase article coverage in terms of the number of articles in different languages and the contents of the articles within a language by identifying and prioritizing missing content and routing attention where it's needed.



Growing Wikipedia Across Languages by Article Creation Recommendations

Ellery Wulczyn
Wikimedia Foundation



Bob West
Stanford

Jure Leskovec
Stanford

Leila Zia
Wikimedia Foundation

Growing Wikipedia Across Languages via Recommendation, Ellery Wulczyn, Robert West, Leila Zia, and Jure Leskovec. International World Wide Web (WWW) Conference, Montréal, Qué., 2016.

Article Creation Recommendations

Goal:

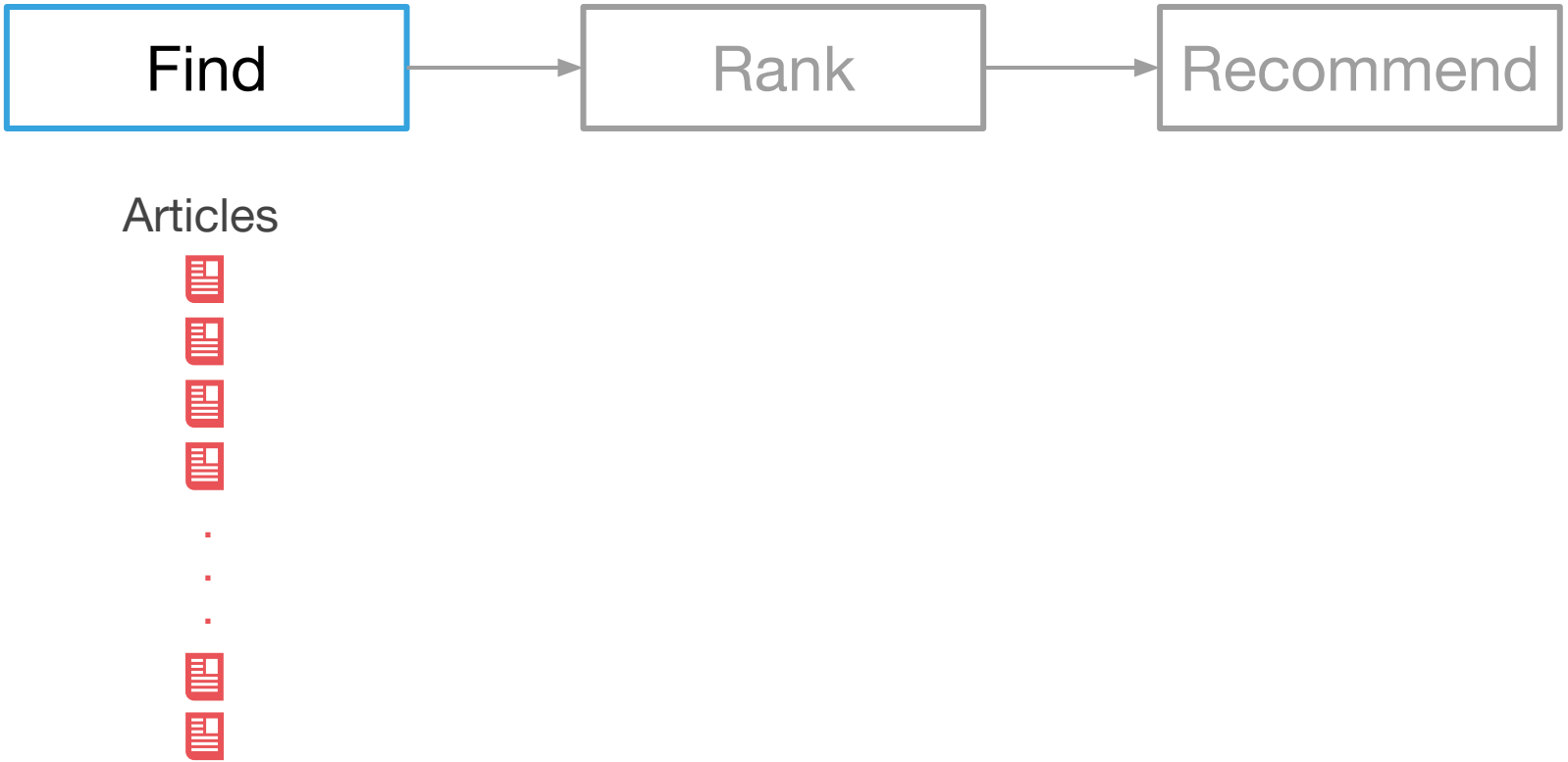
Recommend important missing articles to human editors

System overview:

Bottomline:

3.2x growth rate with no loss in article quality

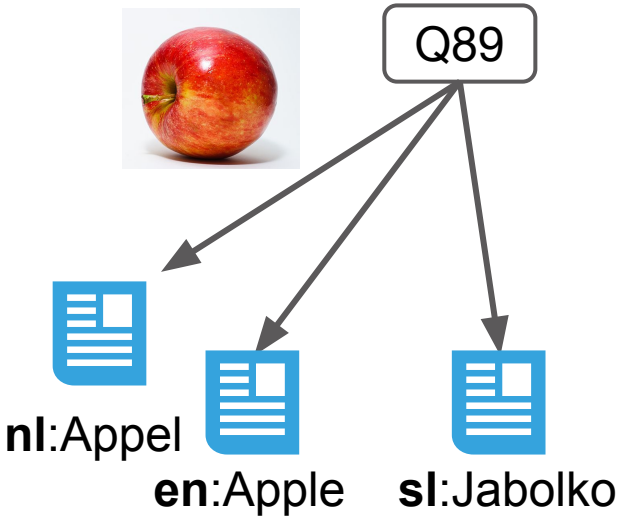
Find missing articles in a given language



Find missing articles in a given language

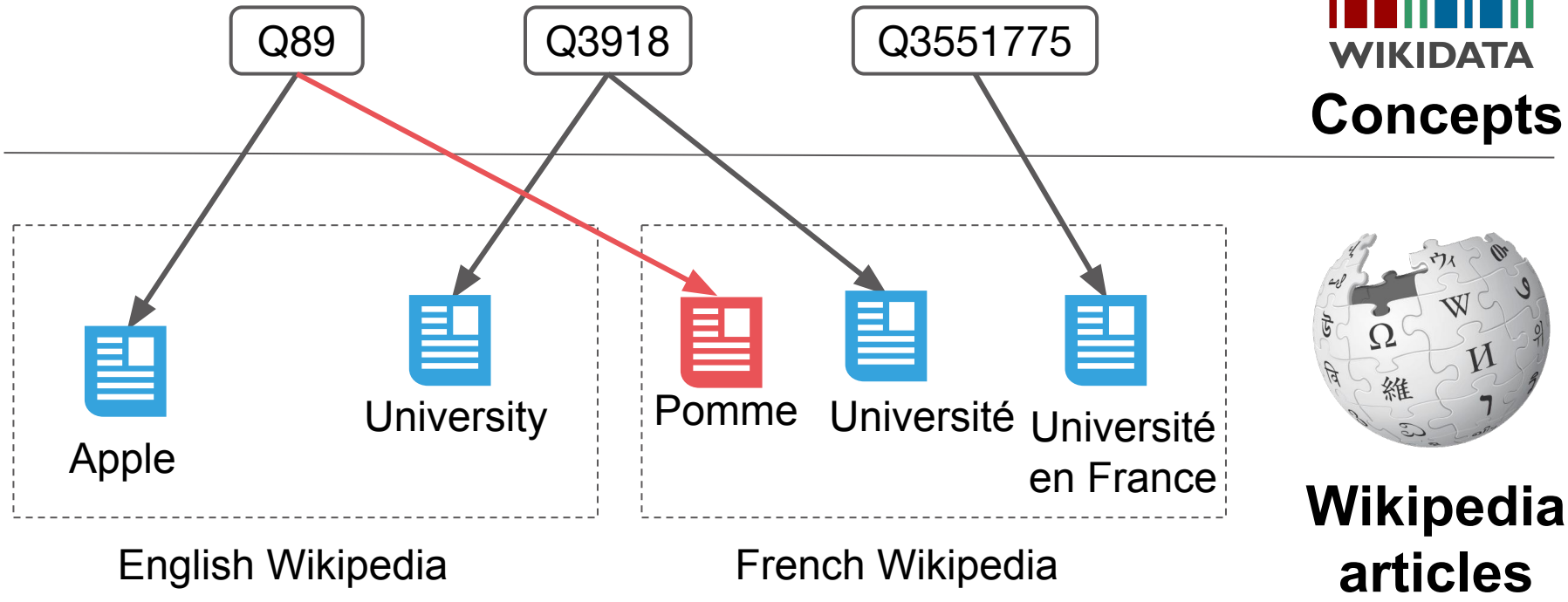
To find missing articles you can compare Wikipedia in language x:

- to external repositories, or
- with other Wikipedia languages
  - Wikidata: Knowledge base for Wikipedia
  - Every article in every language is mapped to a node in a concept graph



WIKIDATA

Find missing articles in a given language



Rank missing articles

Rank missing articles

Goal:

Given a missing article, predict popularity rank in the target language (once it's created)

Challenge:

The article doesn't exist, yet (so we don't have its features for prediction)

Rank missing articles

Approach:

Predict popularity of existing French Wikipedia articles using features extracted from the corresponding non-French articles

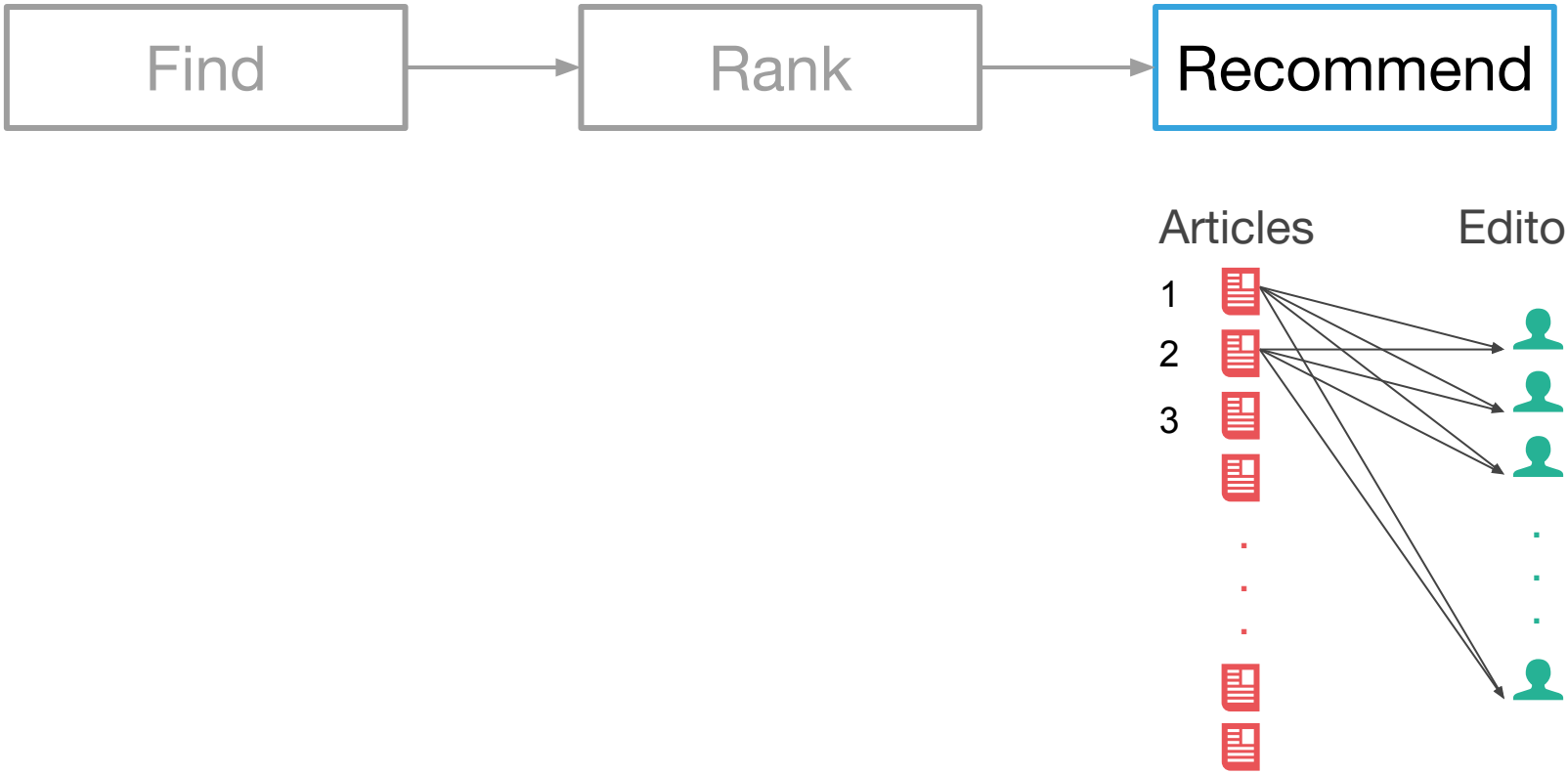
Features:

(Geo) Pageviews, Quality classes, topics, links, edit activity, number of Wikipedia languages with the article

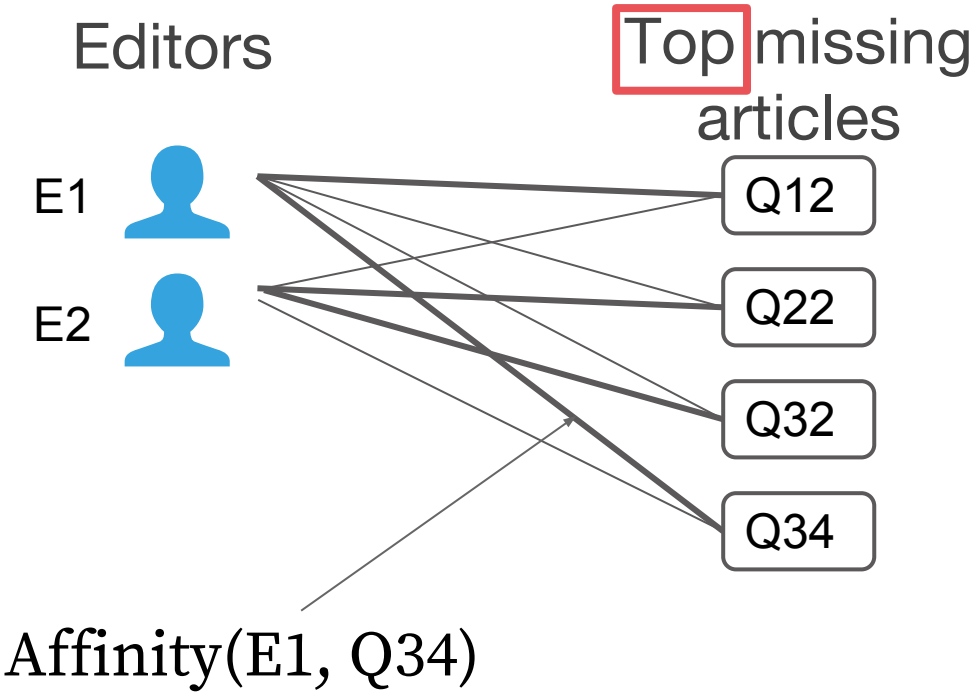
Classification algorithm:

Random Forest

Recommend missing articles



Recommend missing articles



Bipartite matching between editors and missing items

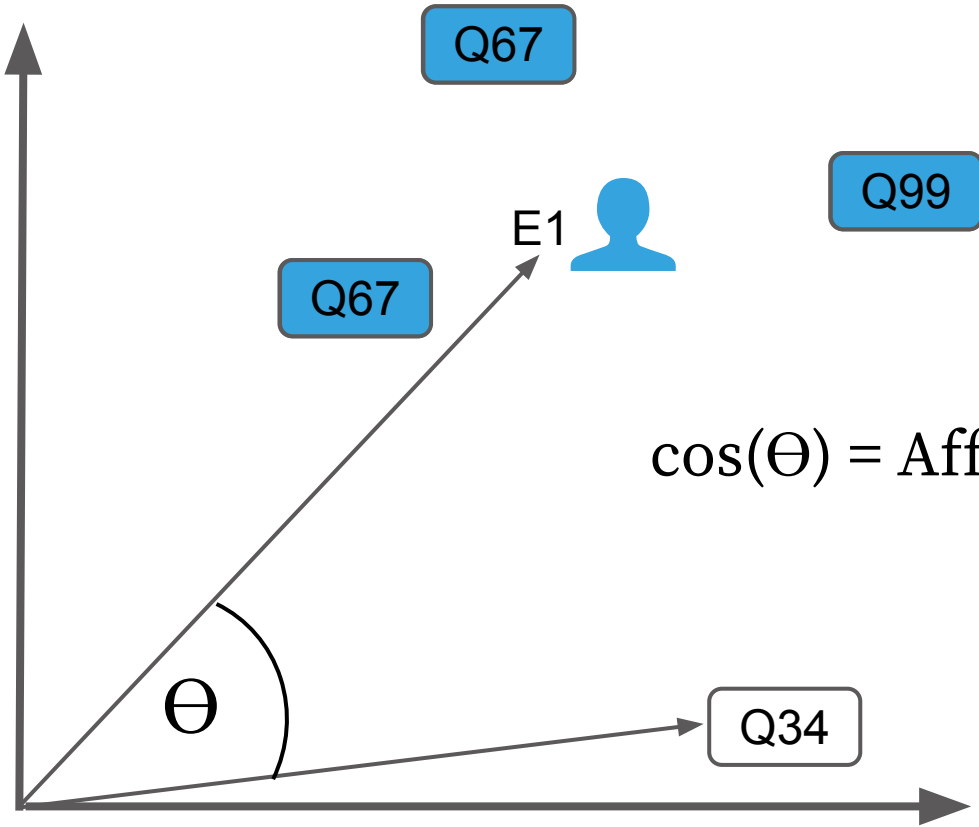
$$\max c \cdot x$$

$$\text{subject to} \quad \sum_i x_{i,j} \leq 1$$

$$\sum_j x_{i,j} = k$$

$$0 \leq x_{i,j} \leq 1$$

Computing affinities



$\cos(\Theta) = \text{Affinity}(E1, Q34)$

Methodology summary

Identify missing concepts using Wikidata knowledge base

Predict language-specific popularity

Match missing articles to editors considering their interest

(English, French) Wikipedia Experiment

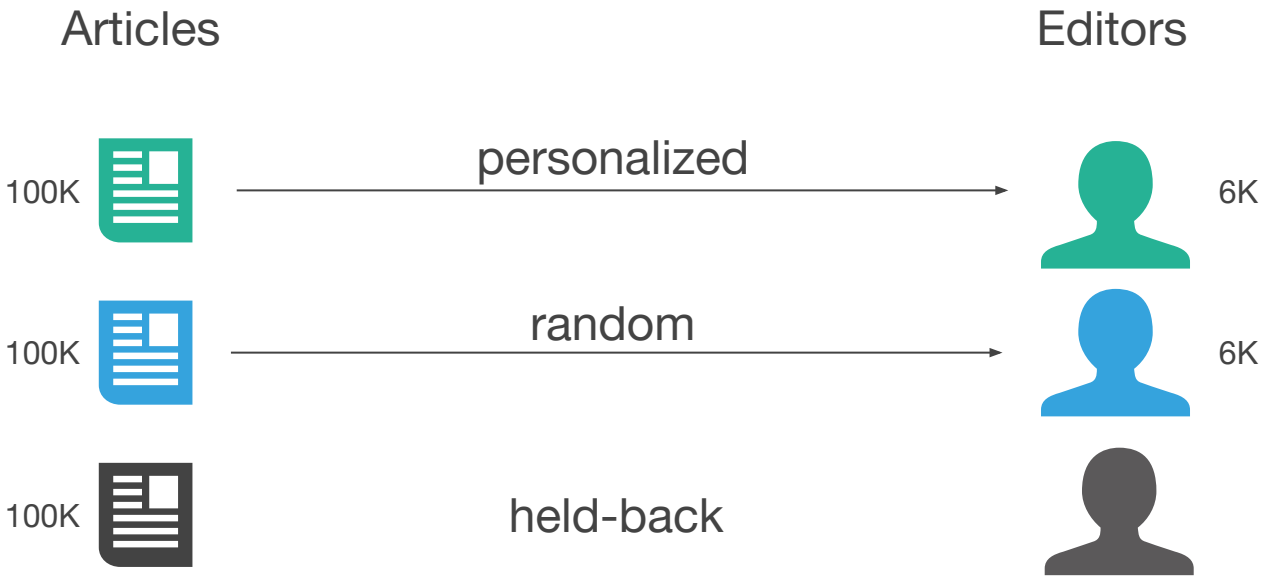
General setup:

- 12K emails were sent to French Wikipedia editors
- They each received five unique articles to create
- Translation was used as a way to create the article

Research questions:

- RQ1: Do recommendations boost article creation rate?
- RQ2: How much does personalization help?
- RQ3: How do recommendations affect article quality?

Design of the experiment



We contacted people

*“What are you doing, trying to scare contributors away from Wikipedia ?
Congratulations, you are doing a fine job -- great way to waste donor's money !”*

“Test language pairs in both directions. Only You can Stop Language Imperialism! If you're testing EN --> FR, also test FR --> EN (and either make it easy for anyone to try the experiment in their own language-pair, or try at least two lang-pairs yourself)”

“Merci pour me contacter. C'est une honneur d'avoir reçu votre courriel.”

“i know how hard it is to work with 70,000 back sit drivers, so i think you deserve some nice words”

“Oh, I forgot to tell you that I find the idea nice, perhaps even brilliant (let’s not be only negative).”

RQ1: boosting article creation rate?

	Personalized	**held-back**
Creation rate	1.05%	0.32%

*** $p < 0.00001$

Yes, personalized recommendations increase creation rate by a factor of 3.2.

RQ2: Does personalization help?

personalized

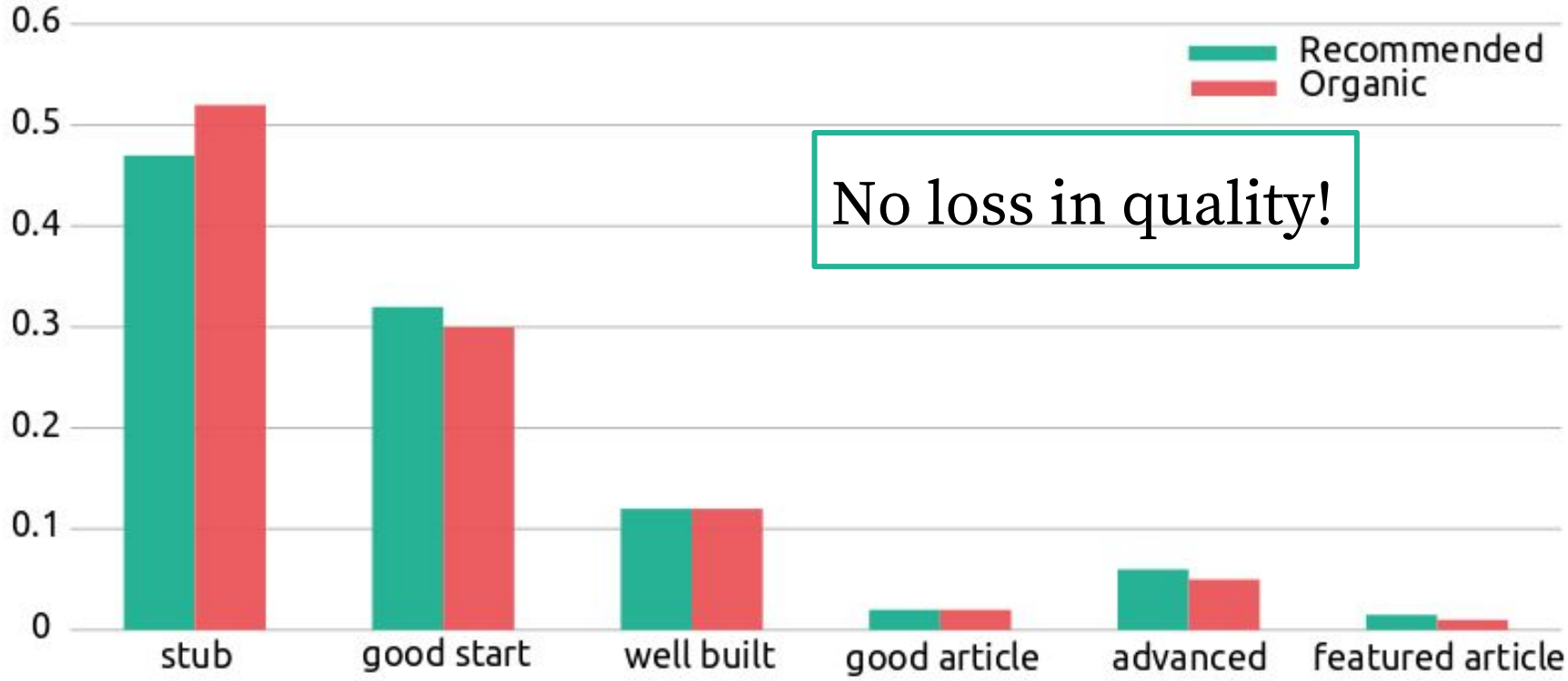
random

	Personalized	**random**
# Editors	6K	6K
Activation rate	2.3%	1.1%

*** p < 0.00001

Yes, personalized recommendations boost activation rate by a factor of 2.

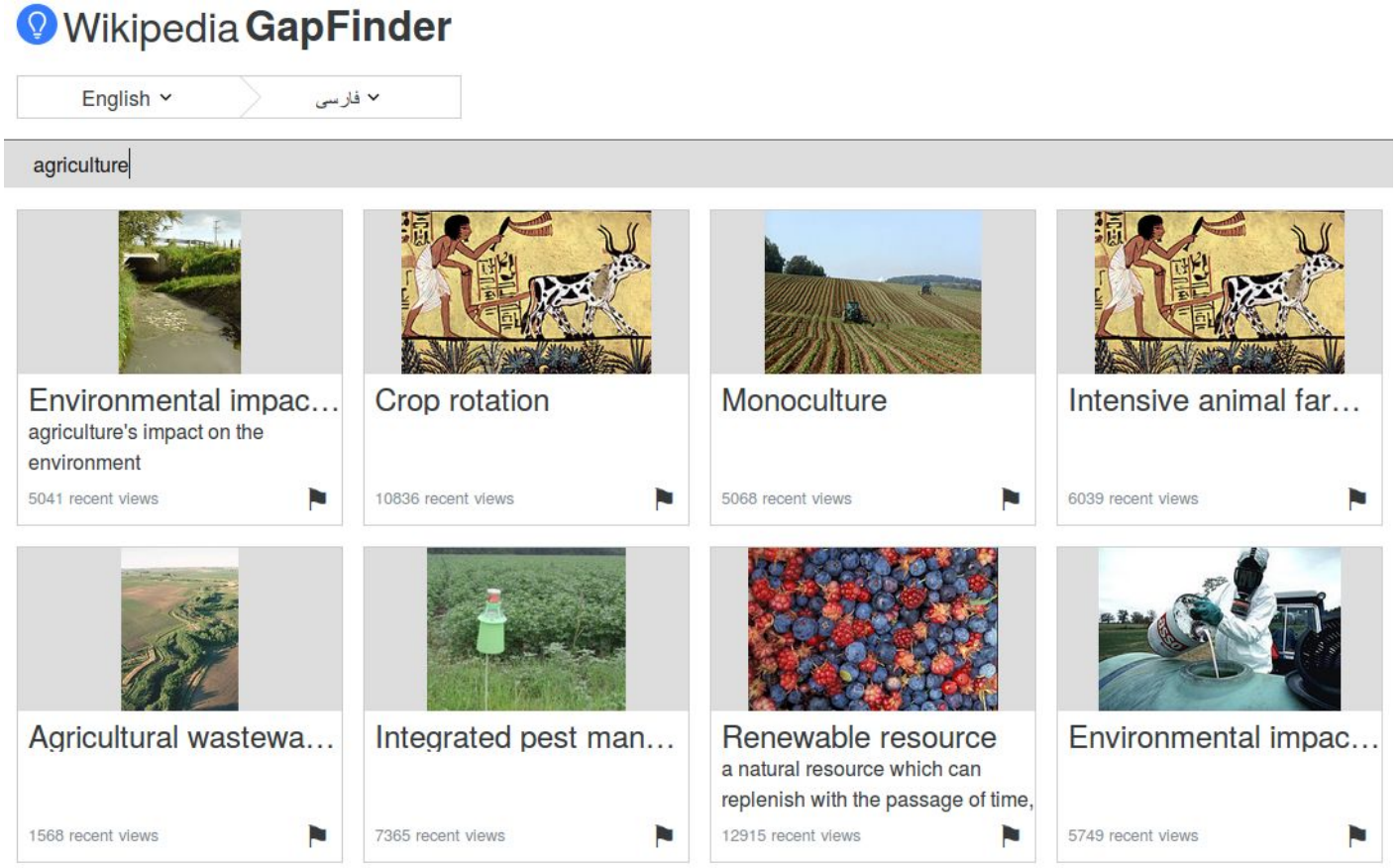
RQ3: Impact on article quality?



Deployment

- Public API for translation recommendations
- Serving Suggestions feature in Content Translation tool.
- GapFinder

Agriculture in Farsi Wikipedia



Climate change in Malagasy

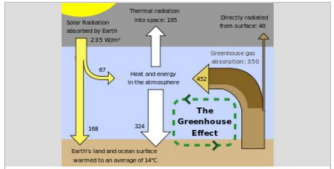
Wikipedia **GapFinder**

English → Malagasy

climate change

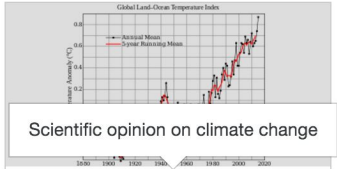
Intergovernmental P…
scientific intergovernmental body

10546 recent views



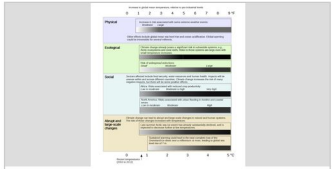
Greenhouse gas
gas in an atmosphere that absorbs and emits radiation within

51057 recent views



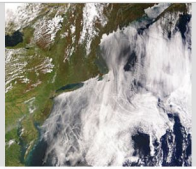
Scientific opinion on …

8047 recent views



Effects of global war…

22475 recent views

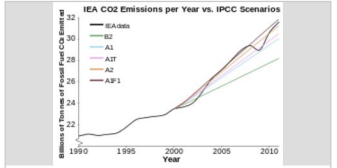


Climate change
significant time variation in long-term weather patterns

0 recent views

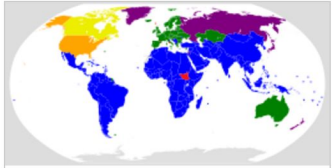
Tropical cyclone
storm system

27758 recent views



Climate change mitig…
actions to limit climate change in order to reduce the risks of global

4595 recent views

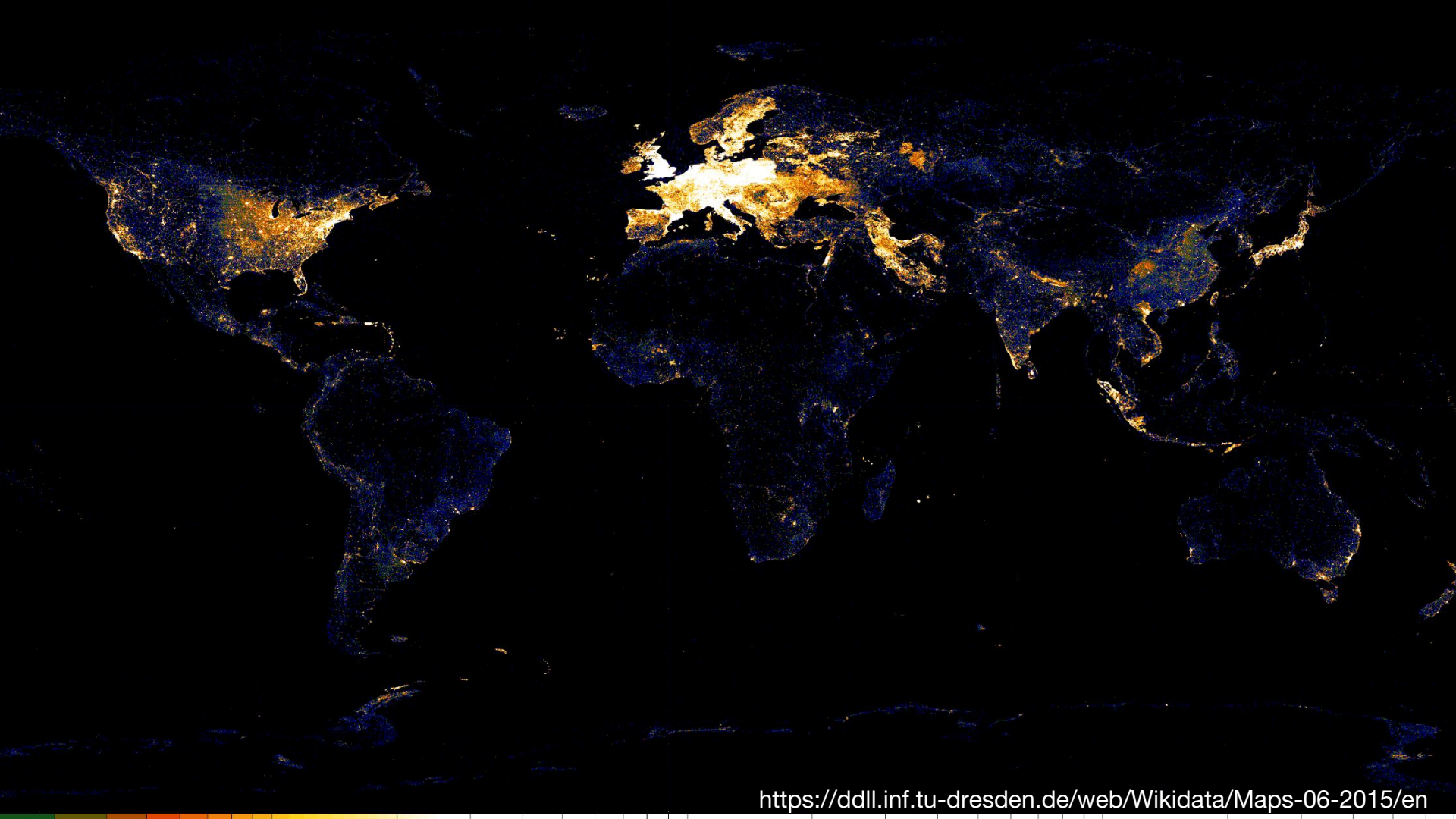


Kyoto Protocol
International Treaty to reduce greenhouse gas emissions

37088 recent views

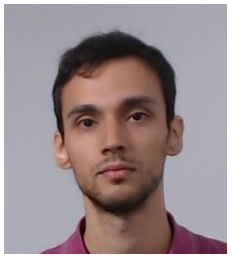
Conclusions

- There are massive knowledge gaps in Wikipedia when it comes to the existence of articles across languages
- We built an end-to-end system for recommending important missing articles
- The system can boost article creation rate by a factor of 3.2 with no loss in article quality
- We built an API and a tool. The system is now being used in Wikipedia.


https://ddll.inf.tu-dresden.de/web/Wikidata/Maps-06-2015/en

Growing Wikipedia Across Languages by Article Expansion Recommendations

Tiziano Piccardi
EPFL



Bob West
EPFL

Michele Catasta
EPFL

Leila Zia
Wikimedia Foundation

https://meta.wikimedia.org/wiki/Research:Expanding_Wikipedia_stubs_across_languages

Overview

- More than 2M articles (37%) in English Wikipedia are tagged as stubs.
- Many of the articles that are not stubs still miss content.

- The community of Wikipedia editors organizes hundreds of editathons across the year to:
  - expand articles' contents across language editions
  - onboard newcomers to Wikipedia

Challenges

Participants

- Newcomers and less experienced users need guidance on how to choose and expand articles.

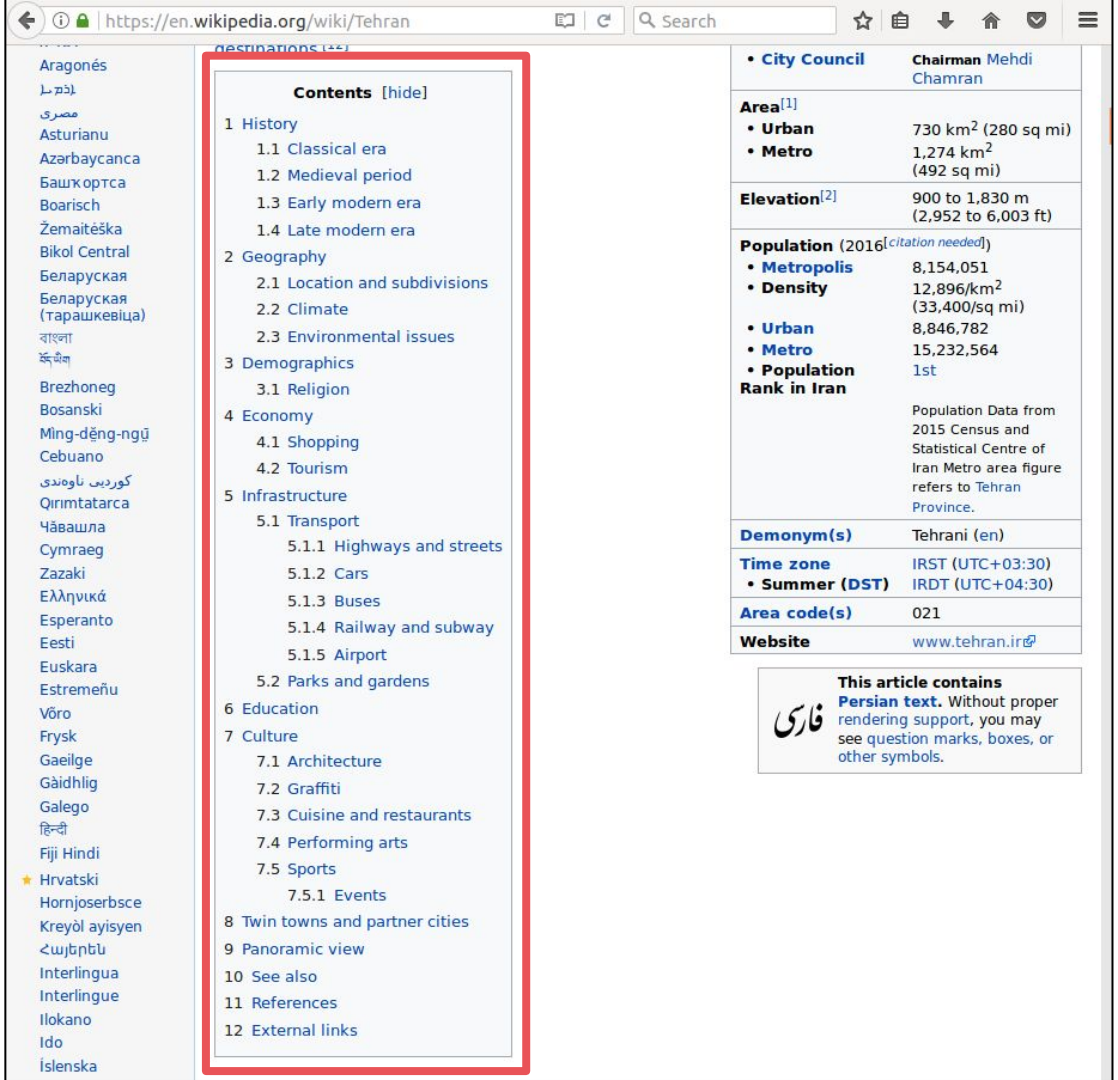
Organizers

- Topics or articles to be expanded need to be manually identified.
- Templates and guidelines for how to expand articles will need to be manually generated by organizers.
- There are far too many categories of articles to search through manually.

Problem statement

Lack of scalable and language independent methodologies for generating guidelines for contributors (especially newcomers) to help them expand articles is a big problem for editathon organizers.


https://en.wikipedia.org/wiki/Sanandaj
WIKIPEDIA
The Free Encyclopedia
Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store
Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page
Tools
What links here
Related changes
Article
Talk
Sanandaj
From Wikipedia, the free encyclopedia
For the administrative subdivision, see Sa
Sanandaj pronunciation (help·info) (Persian
2016 census, its population was 373,987 [1] s
capital of Kordestan province at Iran. Sana
twenty-third largest city in Iran. Sanandaj is
founded about 200 years ago, yet under its s
it has grown to become a center of Kurdish c
Contents [hide]
1 Society
2 Famous people connected to Sanandaj
3 References
4 External links


https://en.wikipedia.org/wiki/Tehran
Search
Aragonés
Asturianu
Azərbaycanca
Башҡортса
Boarisch
Žemaitėška
Bikol Central
Беларуская
Беларуская (тарашкевіца)
বাংলা
Brezhoneg
Bosanski
Mìng-dĕ̤ng-ngṳ̄
Cebuano
Qırımtatarca
Чăвашла
Cymraeg
Zazaki
Ελληνικά
Esperanto
Eesti
Euskara
Estremeñu
Võro
Frysk
Gaeilge
Gàidhlig
Galego
हिन्दी
Fiji Hindi
Hrvatski
Hornjoserbsce
Kreyòl ayisyen
Հայերեն
Interlingua
Interlingue
Ilokano
Ido
Íslenska
Contents [hide]
1 History
1.1 Classical era
1.2 Medieval period
1.3 Early modern era
1.4 Late modern era
2 Geography
2.1 Location and subdivisions
2.2 Climate
2.3 Environmental issues
3 Demographics
3.1 Religion
4 Economy
4.1 Shopping
4.2 Tourism
5 Infrastructure
5.1 Transport
5.1.1 Highways and streets
5.1.2 Cars
5.1.3 Buses
5.1.4 Railway and subway
5.1.5 Airport
5.2 Parks and gardens
6 Education
7 Culture
7.1 Architecture
7.2 Graffiti
7.3 Cuisine and restaurants
7.4 Performing arts
7.5 Sports
7.5.1 Events
8 Twin towns and partner cities
9 Panoramic view
10 See also
11 References
12 External links
City Council
Chairman Mehdi Chamran
Area[1]
Urban 730 km² (280 sq mi)
Metro 1,274 km² (492 sq mi)
Elevation[2] 900 to 1,830 m (2,952 to 6,003 ft)
Population (2016[citation needed])
Metropolis 8,154,051
Density 12,896/km² (33,400/sq mi)
Urban 8,846,782
Metro 15,232,564
Population Rank in Iran 1st
Population Data from 2015 Census and Statistical Centre of Iran Metro area figure refers to Tehran Province.
Demonym(s) Tehrani (en)
Time zone IRST (UTC+03:30)
Summer (DST) IRDT (UTC+04:30)
Area code(s) 021
Website www.tehran.ir
This article contains Persian text. Without proper rendering support, you may see question marks, boxes, or other symbols.

https://en.wikipedia.org/wiki/Sanandaj

WIKIPEDIA
The Free Encyclopedia

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia
Wikipedia store

Interaction
Help
About Wikipedia
Community portal
Recent changes
Contact page

Tools
What links here
Related changes

Article Talk

Sanandaj

From Wikipedia, the free encyclopedia

For the administrative subdivision, see Sa

Sanandaj pronunciation (help·info) (Persian
2016 census, its population was 373,987 [1]
capital of ***Kordestan*** province at Iran. Sana
twenty-third largest city in Iran. Sanandaj is
founded about 200 years ago, yet under its
it has grown to become a center of Kurdish c

Contents [hide]

1 Society
2 Famous people connected to Sanandaj
3 References
4 External links

https://fa.wikipedia.org/wiki/سنندج

Search

سال شهرشدن	دوره صفوی[۲] (۱۳۰۷ ح)[۳]
مردم	
جمعیت	۳۷۳٬۹۸۷ نفر
تراکم جمعیت	تراکم خالص شهر سنندج در حال حاضر[چه زمانی؟] ۳۲۱//۸۷ نفر در هکتار می‌باشد.[۴] نفر بر کیلومتر مربع
جغرافیای طبیعی	
مساحت	۳٬۰۳۳ [۷]
ارتفاع از سطح دریا	۱٬۴۵۰ تا ۱٬۵۳۸ متر[۸]
آب‌وهوا	
میانگین دمای سالانه	۱۲ درجهٔ سانتی‌گراد
میانگین بارش سالانه	۵۰۰ میلی‌متر
اطلاعات شهری	
شهردار	منوچهر فخری
ره‌آورد	قالی، گلیم، کلاش، منبت کاری، نازک‌کاری چوب،[۱۰]
پیش‌شماره تلفنی	۰۸۷۳۳ [۹]۰۸۷
وبگاه	شهرداری سنندج
شناسهٔ ملی خودرو	ایران ۵۱
تابلوی خوش‌آمد به شهر	

روی نقشه ایران
۳۵.۳۱۷۲° شمالی ۴۶.۹۹۸۹° شرقی

محتویات [نهفتن]

۱ نام
۲ پیشینه تاریخی
۲.۱ تاریخچه‌ای از منطقه سنندج
۲.۲ سنه‌دژ
۳ وضعیت طبیعی
۳.۱ جغرافیا
۳.۲ موقعیت جغرافیایی
۳.۳ آب و هوا
۴ مردم
۴.۱ زبان
۴.۲ جمعیت
۴.۳ مذهب
۴.۴ لباس
۵ فرهنگ و هنر
۵.۱ مشاهیر
۵.۲ جشن‌ها
۵.۳ سینماها
۵.۴ سینماهای فعال
۶ کتابخانه‌های عمومی
۶.۱ شهید برهان عالی
۶.۲ فردوسی
۶.۳ هه ژار
۶.۴ امام رضا
۶.۵ خاتم الانبیاء
۶.۶ مستوره اردلان
۷ جاذبه‌های تاریخی و مذهبی
۷.۱ عمارت و خانه‌های تاریخی
۷.۱.۱ بازارها
۷.۱.۲ حمام‌ها
۷.۱.۳ دیگر آثار تاریخی
۷.۲ جاذبه‌های مذهبی
۷.۲.۱ مساجد
۷.۲.۲ امامزاده‌ها
۷.۲.۳ مکان‌های مذهبی دیگر ادیان
۷.۳ جشن‌های مذهبی
۸ جاذبه‌های طبیعی
۸.۱ قلل مرتفع
۸.۲ مجموعه پارک تفریحی آبیدر
۸.۳ بزرگترین سینمای روباز جهان
۹ فهرست هتل‌های سنندج

تغییرات مرتبط
بارگذاری پرونده
صفحه‌های ویژه
پیوند پایدار
اطلاعات صفحه
آیتم ویکی‌داده
بادکرد پیوند این مقاله

زبان‌های دیگر
Deutsch
English
Español
Français
हिन्दी
Italiano
한국어
Русский
中文
۳۵ مورد دیگر
ویرایش پیوندها

Expand articles via recommendations

Goal:

Recommend to editors how to expand articles across languages

Unit of recommendation:

Article sections, content in info-boxes, media files, references, etc.

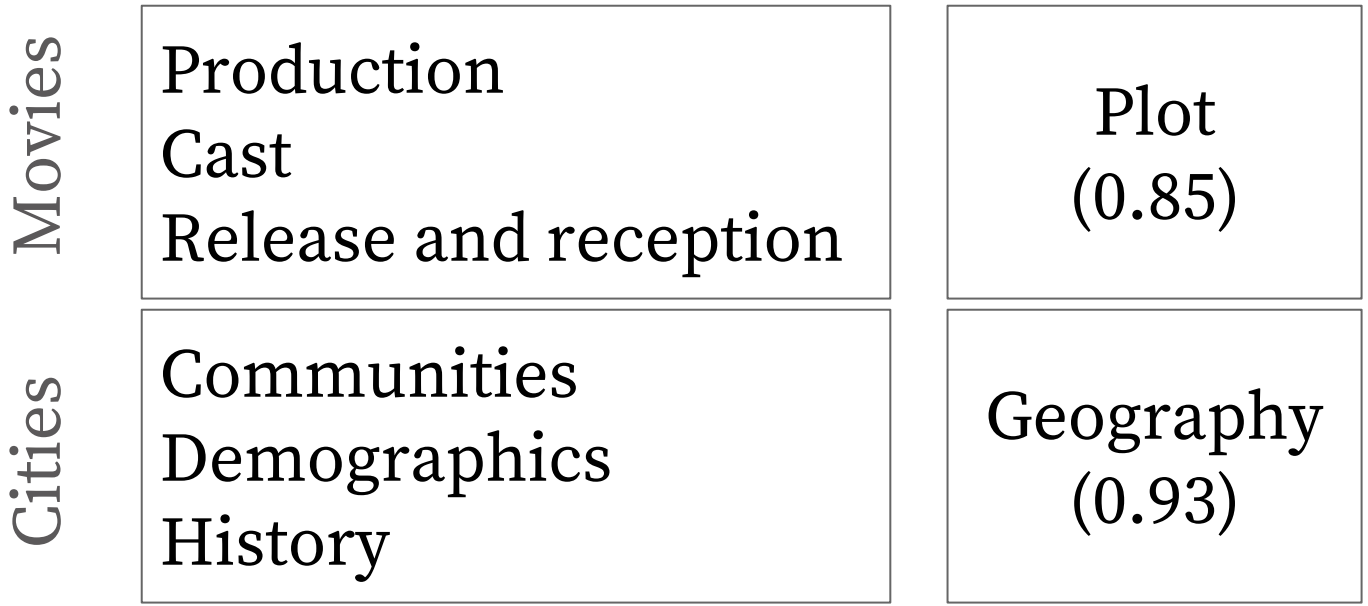
Intuition and current methodology

Intuition: Use the existing content on Wikipedia to infer recommendations for article expansions:

- Cooccurrences of items (e.g., sections) in a given topic within a language
- Using multilingual signals to infer missing sections by comparing section occurrences in fixed topics across languages

Methodology: Market basket analysis

Examples


Movies
Cities
Production
Cast
Release and reception
Plot
(0.85)
Communities
Demographics
History
Geography
(0.93)

Next steps and open questions

- Improving the algorithms by building a user feedback loop
- Article importance
- User interests beyond topical interests
- Expanding how we find missing articles and contents

Thank you!