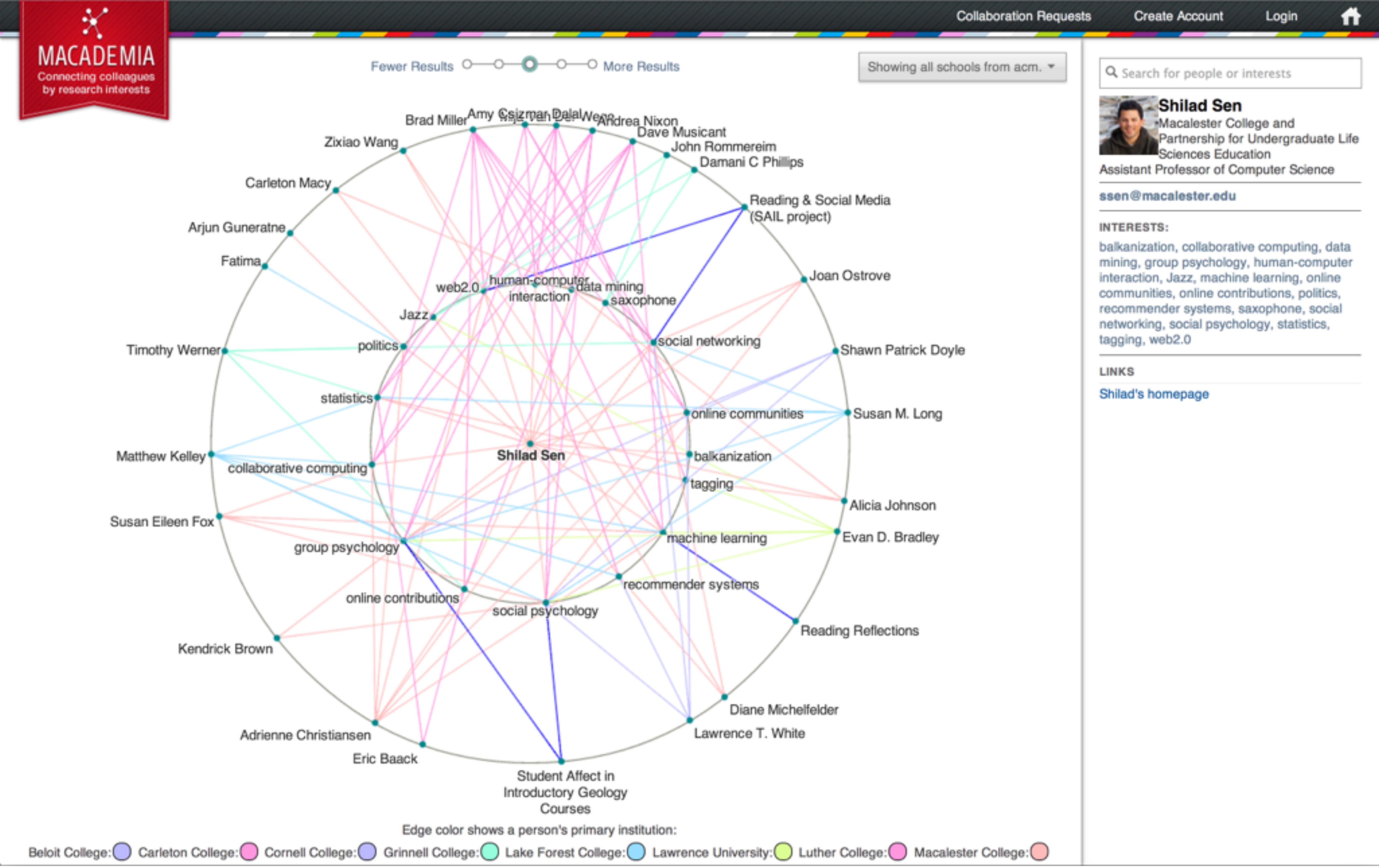
Computationally Harnessing Wikipedia's Knowledge

Shilad Sen
http://www.shilad.com
ssen@macalester.edu

MACALESTER COLLEGE

1. WikiBrain
2. WikiBrain-driven projects



Edge color shows a person's primary institution:

Beloit College: ● Carleton College: ● Cornell College: ● Grinnell College: ● Lake Forest College: ● Lawrence University: ● Luther College: ● Macalester College: ●

Search for people or interests

Shilad Sen
Macalester College and Partnership for Undergraduate Life Sciences Education
Assistant Professor of Computer Science

ssen@macalester.edu

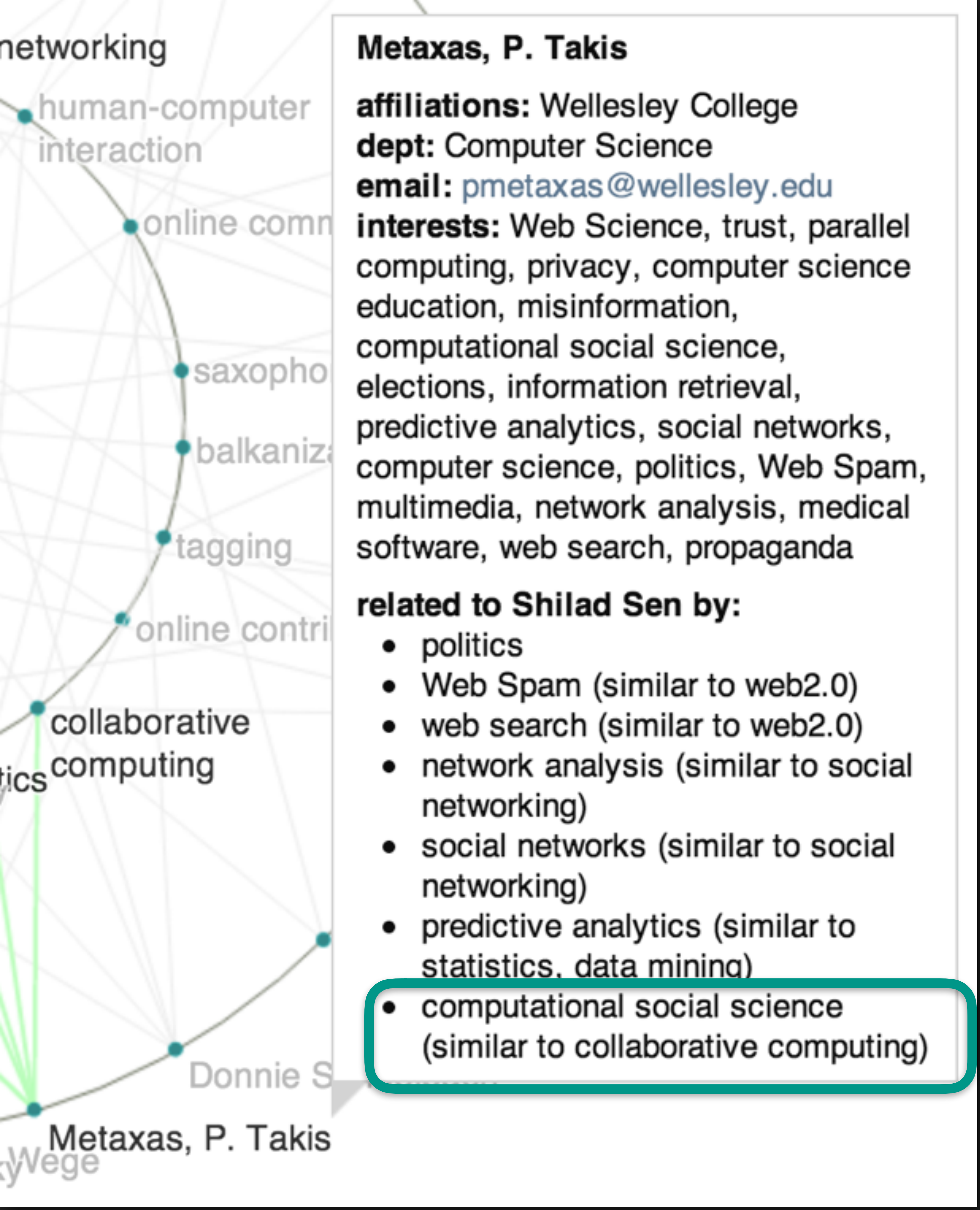
INTERESTS:

balkanization, collaborative computing, data mining, group psychology, human-computer interaction, Jazz, machine learning, online communities, online contributions, politics, recommender systems, saxophone, social networking, social psychology, statistics, tagging, web2.0

LINKS

Shilad's homepage

Shilad's Macademia profile - http://macademia.macalester.edu

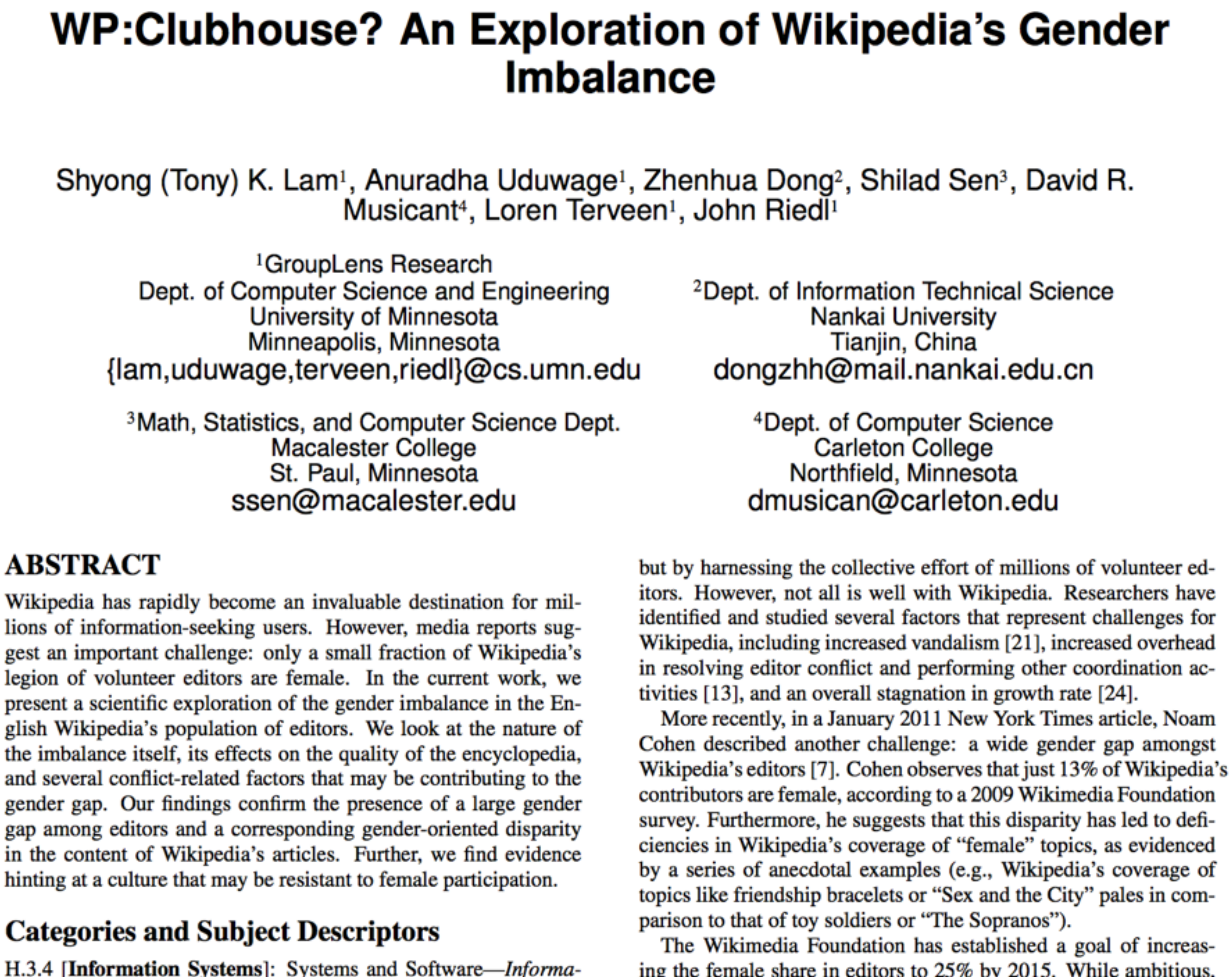


Metaxas, P. Takis

affiliations: Wellesley College
dept: Computer Science
email: pmetaxas@wellesley.edu
interests: Web Science, trust, parallel computing, privacy, computer science education, misinformation, computational social science, elections, information retrieval, predictive analytics, social networks, computer science, politics, Web Spam, multimedia, network analysis, medical software, web search, propaganda

related to Shilad Sen by:

- politics
- Web Spam (similar to web2.0)
- web search (similar to web2.0)
- network analysis (similar to social networking)
- social networks (similar to social networking)
- predictive analytics (similar to statistics, data mining)
- computational social science (similar to collaborative computing)



WP:Clubhouse? An Exploration of Wikipedia's Gender Imbalance

Shyong (Tony) K. Lam[1], Anuradha Uduwage[1], Zhenhua Dong[2], Shilad Sen[3], David R. Musicant[4], Loren Terveen[1], John Riedl[1]

[1]GroupLens Research
Dept. of Computer Science and Engineering
University of Minnesota
Minneapolis, Minnesota
{lam,uduwage,terveen,riedl}@cs.umn.edu

[2]Dept. of Information Technical Science
Nankai University
Tianjin, China
dongzhh@mail.nankai.edu.cn

[3]Math, Statistics, and Computer Science Dept.
Macalester College
St. Paul, Minnesota
ssen@macalester.edu

[4]Dept. of Computer Science
Carleton College
Northfield, Minnesota
dmusican@carleton.edu

ABSTRACT

Wikipedia has rapidly become an invaluable destination for millions of information-seeking users. However, media reports suggest an important challenge: only a small fraction of Wikipedia's legion of volunteer editors are female. In the current work, we present a scientific exploration of the gender imbalance in the English Wikipedia's population of editors. We look at the nature of the imbalance itself, its effects on the quality of the encyclopedia, and several conflict-related factors that may be contributing to the gender gap. Our findings confirm the presence of a large gender gap among editors and a corresponding gender-oriented disparity in the content of Wikipedia's articles. Further, we find evidence hinting at a culture that may be resistant to female participation.

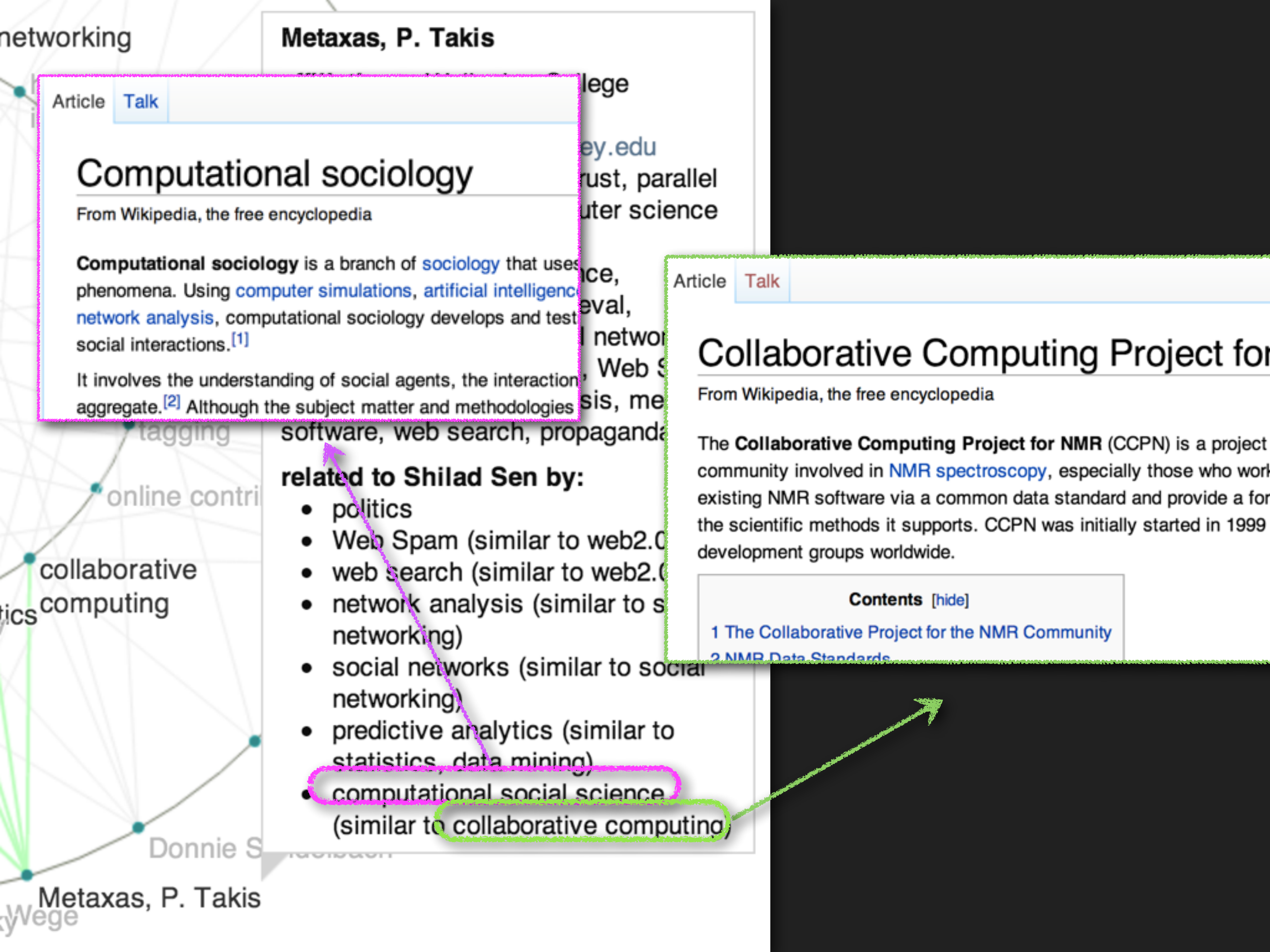
Categories and Subject Descriptors

H.3.4 [**Information Systems**]: Systems and Software—*Informa-*

but by harnessing the collective effort of millions of volunteer editors. However, not all is well with Wikipedia. Researchers have identified and studied several factors that represent challenges for Wikipedia, including increased vandalism [21], increased overhead in resolving editor conflict and performing other coordination activities [13], and an overall stagnation in growth rate [24].

More recently, in a January 2011 New York Times article, Noam Cohen described another challenge: a wide gender gap amongst Wikipedia's editors [7]. Cohen observes that just 13% of Wikipedia's contributors are female, according to a 2009 Wikimedia Foundation survey. Furthermore, he suggests that this disparity has led to deficiencies in Wikipedia's coverage of "female" topics, as evidenced by a series of anecdotal examples (e.g., Wikipedia's coverage of topics like friendship bracelets or "Sex and the City" pales in comparison to that of toy soldiers or "The Sopranos").

The Wikimedia Foundation has established a goal of increasing the female share in editors to 25% by 2015. While ambitious,



Metaxas, P. Takis

lege

ey.edu

rust, parallel

uter science

ce,

eval,

l networ

Web S

sis, me

software, web search, propaganda

related to Shilad Sen by:

- politics
- Web Spam (similar to web2.0
- web search (similar to web2.(
- network analysis (similar to s networking)
- social networks (similar to social networking)
- predictive analytics (similar to statistics, data mining)
- computational social science (similar to collaborative computing)

Article Talk

Computational sociology

From Wikipedia, the free encyclopedia

Computational sociology is a branch of sociology that uses phenomena. Using computer simulations, artificial intelligenc network analysis, computational sociology develops and test social interactions.[1]

It involves the understanding of social agents, the interaction aggregate.[2] Although the subject matter and methodologies

Article Talk

Collaborative Computing Project for

From Wikipedia, the free encyclopedia

The **Collaborative Computing Project for NMR** (CCPN) is a project community involved in NMR spectroscopy, especially those who work existing NMR software via a common data standard and provide a for the scientific methods it supports. CCPN was initially started in 1999 development groups worldwide.

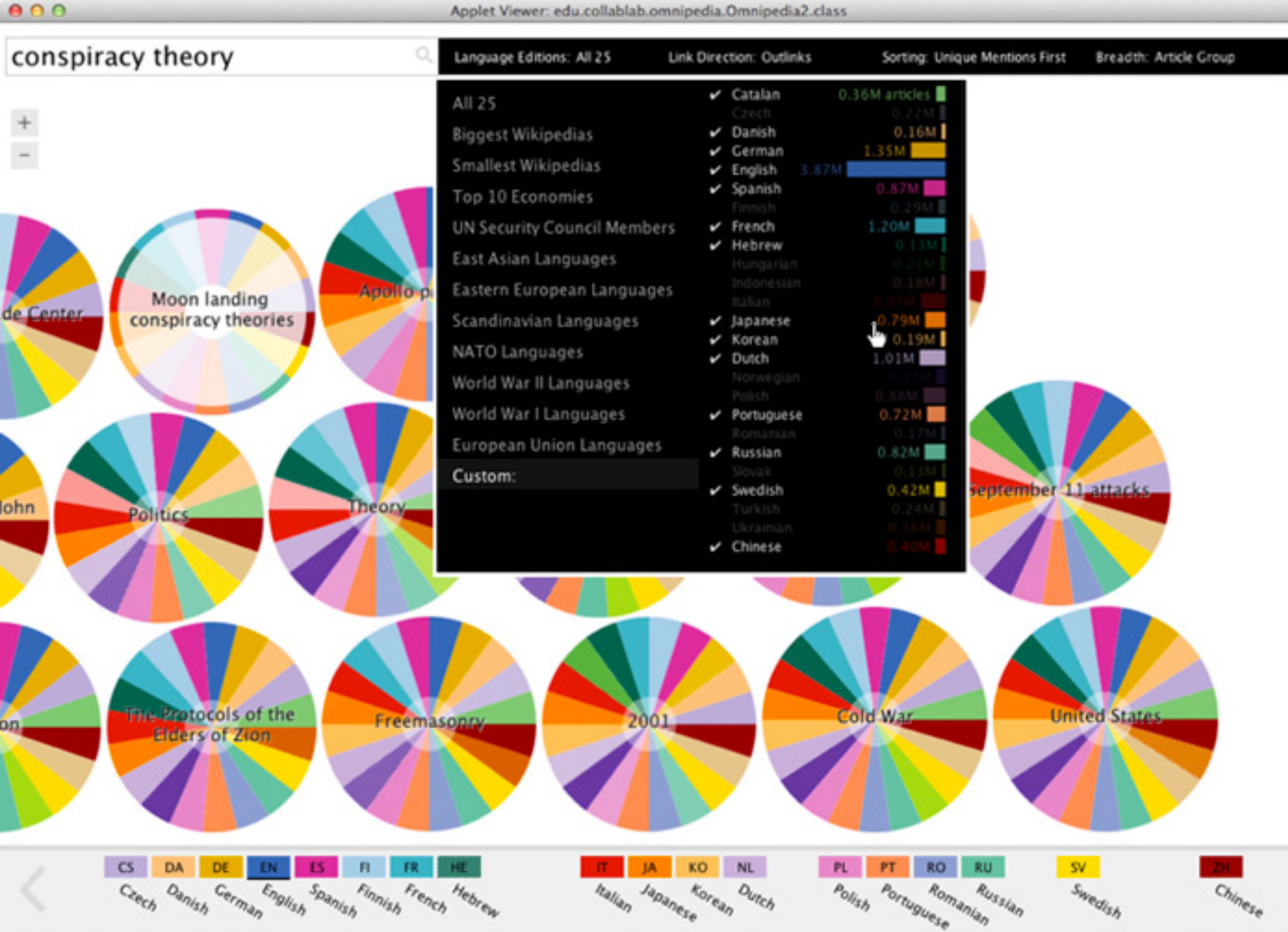
Contents [hide]

1 The Collaborative Project for the NMR Community

2 NMR Data Standards

Brent Hecht

Photo by Brent Hecht, CC-BY-SA 3.0

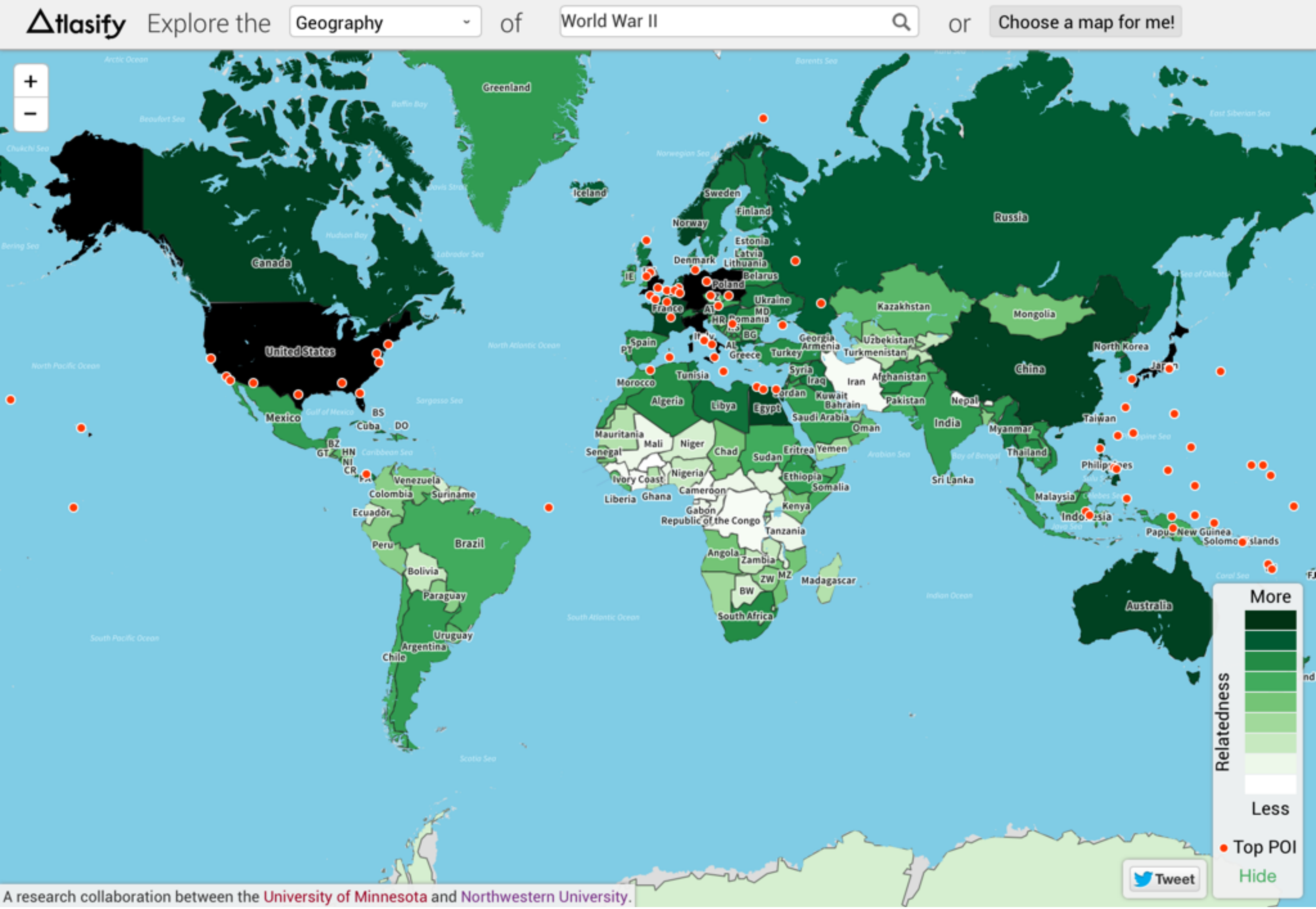


conspiracy theory

Language Editions: All 25 | Link Direction: Outlinks | Sorting: Unique Mentions First | Breadth: Article Group

✓	Catalan	0.36M articles
	Czech	0.22M
✓	Danish	0.16M
✓	German	1.35M
✓	English	3.87M
✓	Spanish	0.87M
	Finnish	0.29M
✓	French	1.20M
✓	Hebrew	0.13M
	Hungarian	0.21M
	Indonesian	0.18M
	Italian	0.97M
✓	Japanese	0.79M
✓	Korean	0.19M
✓	Dutch	1.01M
	Norwegian	0.37M
	Polish	0.88M
✓	Portuguese	0.72M
	Romanian	0.17M
✓	Russian	0.82M
	Slovak	0.13M
✓	Swedish	0.42M
	Turkish	0.24M
	Ukrainian	0.36M
✓	Chinese	0.40M

Omnipedia: Bridging the Wikipedia Language Gap. Bao, Hecht, Carton, Quaderi, Horn, and Gergle. CHI ’12



A research collaboration between the University of Minnesota and Northwestern University.

Hecht, B., Carton, S., Quaderi, M., Schöning, J., Raubal, M., Gergle, D., Downey, D. 2012. "Explanatory Semantic Relatedness and Explicit Spatialization for Exploratory Search". *SIGIR 2012*.

The challenge of Wikipedia algs

Wikipedia is crucial to NLP, AI, and geospatial algs.

but…

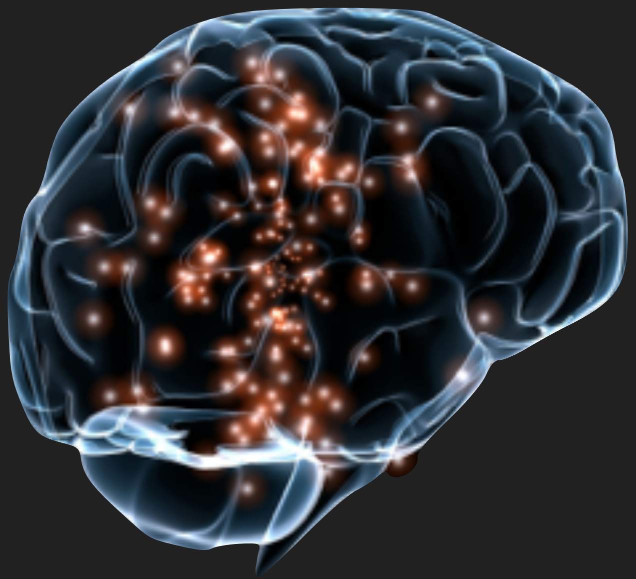
Wikipedia is big.

Wikipedia is messy.

Robust implementations of algorithms are rare.

Research is difficult to reproduce.

Enter WikiBrain



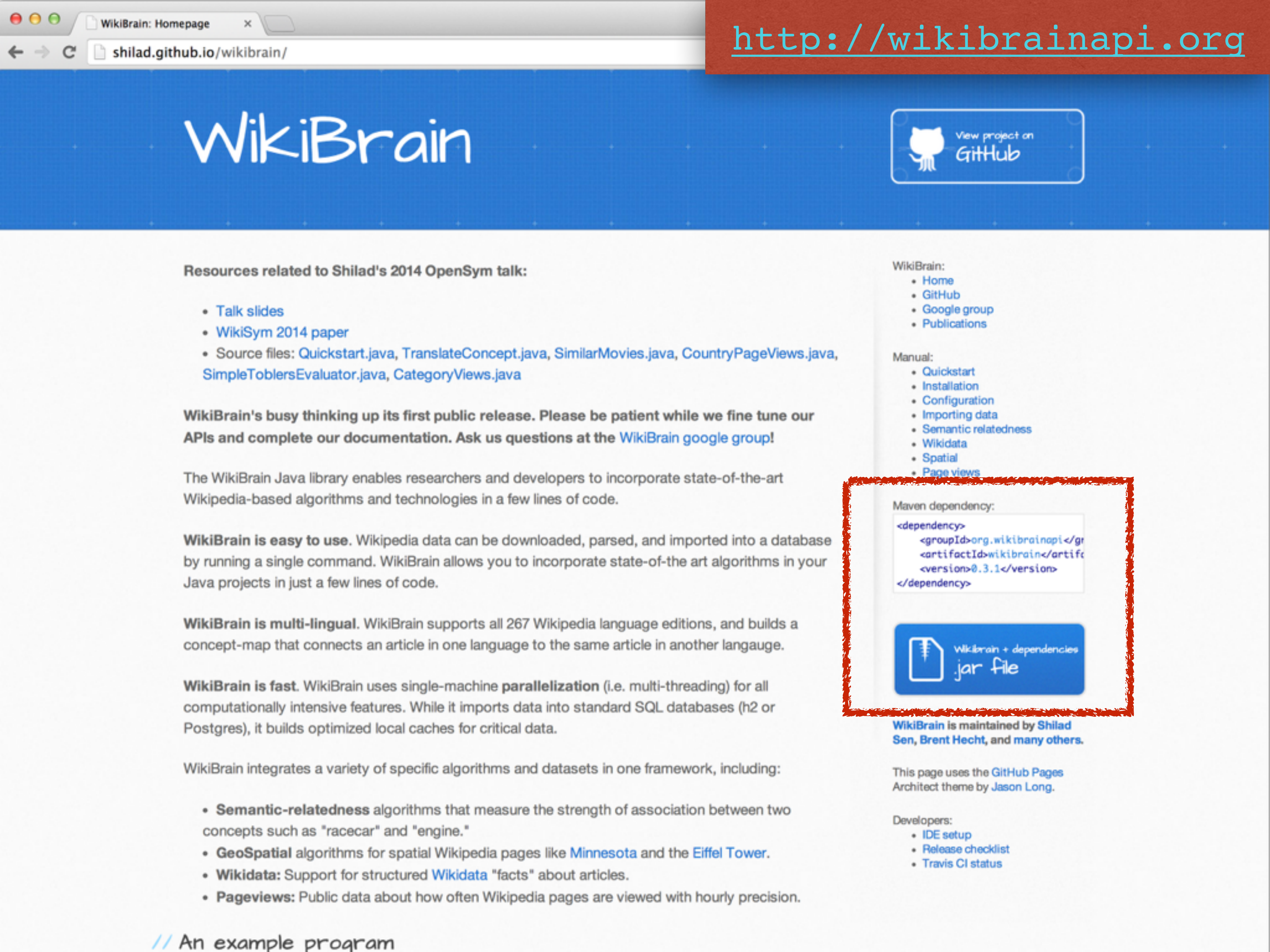
Mission: Democratize access to state-of-the-art Wikipedia algorithms and technologies.

Audience: Programmers with basic Java (for now).

Focus: Core data structures, AI, NLP, Geospatial.

Design goals: Fast, flexible, easy to use (3rd gen).

brain image in public domain - https://commons.wikimedia.org/wiki/File:Neuronal_activity_DARPA.jpg http://wikibrainapi.org


WikiBrain: Homepage
shilad.github.io/wikibrain/
View project on GitHub
Wikibrain + dependencies jar file

WikiBrain

Resources related to Shilad's 2014 OpenSym talk:

- Talk slides
- WikiSym 2014 paper
- Source files: Quickstart.java, TranslateConcept.java, SimilarMovies.java, CountryPageViews.java, SimpleToblersEvaluator.java, CategoryViews.java

WikiBrain's busy thinking up its first public release. Please be patient while we fine tune our APIs and complete our documentation. Ask us questions at the WikiBrain google group!

The WikiBrain Java library enables researchers and developers to incorporate state-of-the-art Wikipedia-based algorithms and technologies in a few lines of code.

WikiBrain is easy to use. Wikipedia data can be downloaded, parsed, and imported into a database by running a single command. WikiBrain allows you to incorporate state-of-the art algorithms in your Java projects in just a few lines of code.

WikiBrain is multi-lingual. WikiBrain supports all 267 Wikipedia language editions, and builds a concept-map that connects an article in one language to the same article in another langauge.

WikiBrain is fast. WikiBrain uses single-machine **parallelization** (i.e. multi-threading) for all computationally intensive features. While it imports data into standard SQL databases (h2 or Postgres), it builds optimized local caches for critical data.

WikiBrain integrates a variety of specific algorithms and datasets in one framework, including:

- **Semantic-relatedness** algorithms that measure the strength of association between two concepts such as "racecar" and "engine."
- **GeoSpatial** algorithms for spatial Wikipedia pages like Minnesota and the Eiffel Tower.
- **Wikidata:** Support for structured Wikidata "facts" about articles.
- **Pageviews:** Public data about how often Wikipedia pages are viewed with hourly precision.

// An example program

WikiBrain:
- Home
- GitHub
- Google group
- Publications

Manual:
- Quickstart
- Installation
- Configuration
- Importing data
- Semantic relatedness
- Wikidata
- Spatial
- Page views

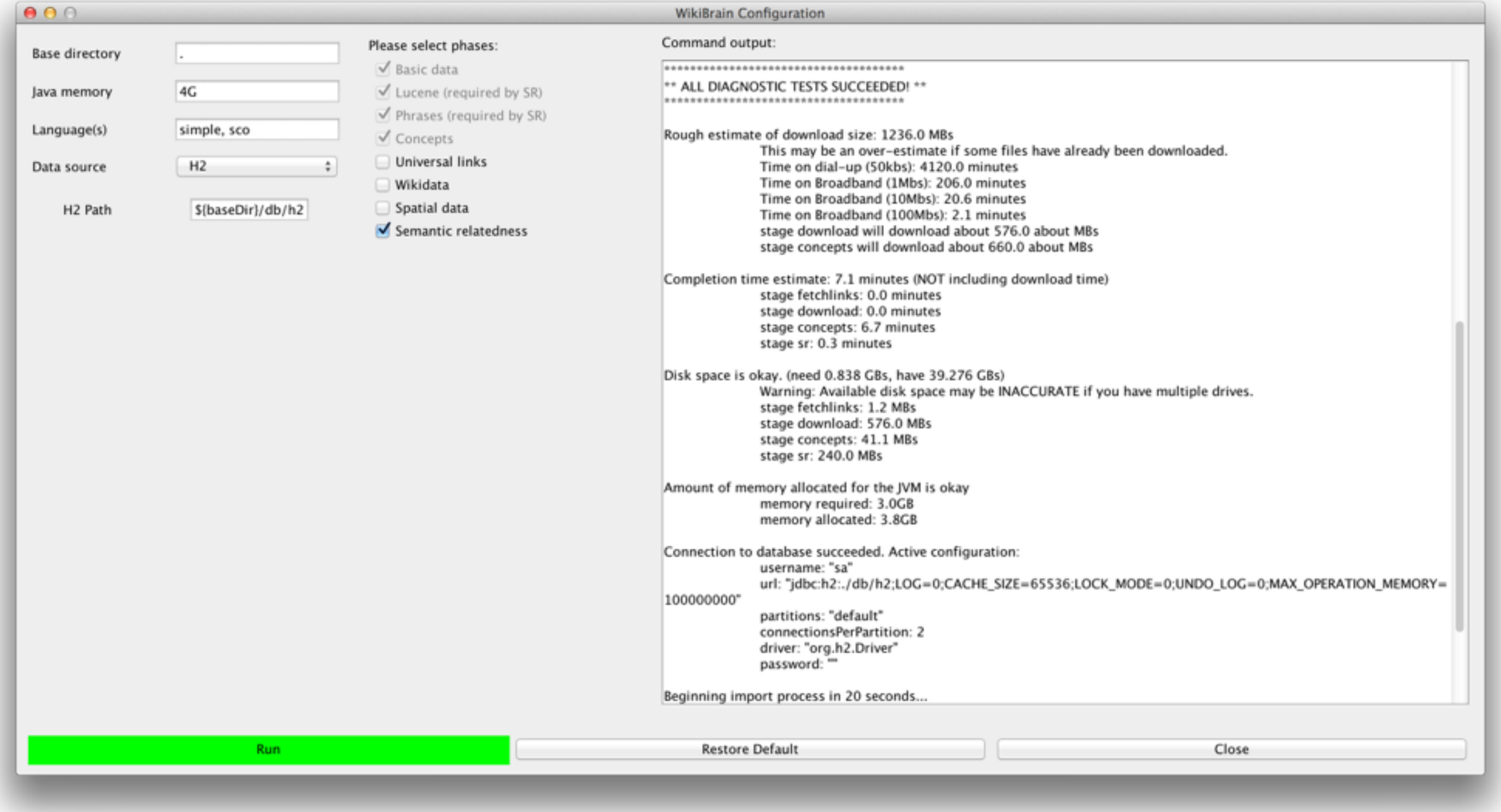
Maven dependency:

```
<dependency>
    <groupId>org.wikibrainapi</gr
    <artifactId>wikibrain</artifa
    <version>0.3.1</version>
</dependency>
```

WikiBrain is maintained by Shilad Sen, Brent Hecht, and many others.

This page uses the GitHub Pages Architect theme by Jason Long.

Developers:
- IDE setup
- Release checklist
- Travis CI status



Base directory: .

Java memory: 4G

Language(s): simple, sco

Data source: H2

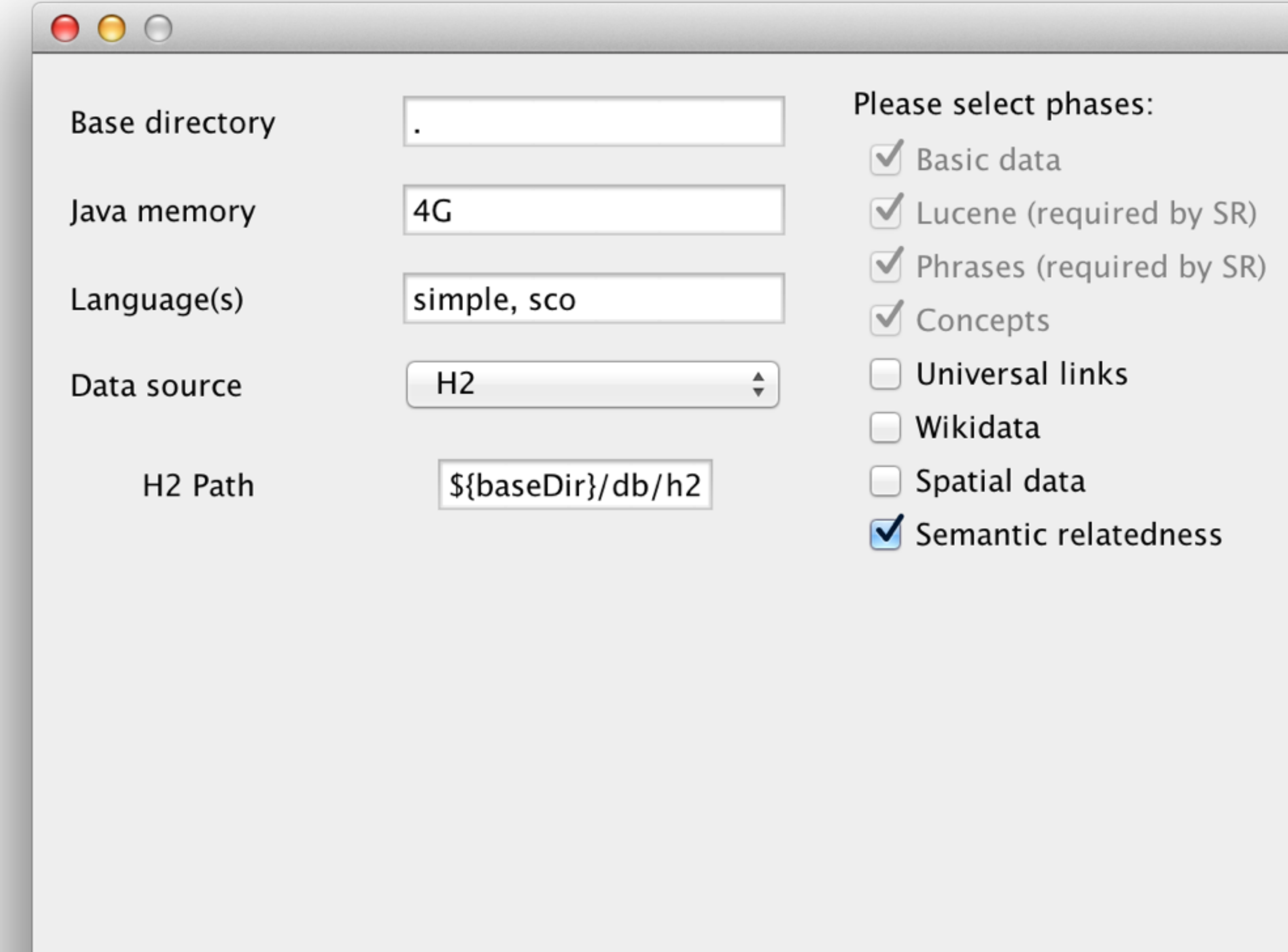
H2 Path: ${baseDir}/db/h2

Please select phases:

- ☑ Basic data
- ☑ Lucene (required by SR)
- ☑ Phrases (required by SR)
- ☑ Concepts
- ☐ Universal links
- ☐ Wikidata
- ☐ Spatial data
- ☑ Semantic relatedness

Command output:

```
****************************************
** ALL DIAGNOSTIC TESTS SUCCEEDED! **
****************************************

Rough estimate of download size: 1236.0 MBs
                This may be an over-estimate if some files have already been downloaded.
                Time on dial-up (50kbs): 4120.0 minutes
                Time on Broadband (1Mbs): 206.0 minutes
                Time on Broadband (10Mbs): 20.6 minutes
                Time on Broadband (100Mbs): 2.1 minutes
                stage download will download about 576.0 about MBs
                stage concepts will download about 660.0 about MBs

Completion time estimate: 7.1 minutes (NOT including download time)
                stage fetchlinks: 0.0 minutes
                stage download: 0.0 minutes
                stage concepts: 6.7 minutes
                stage sr: 0.3 minutes

Disk space is okay. (need 0.838 GBs, have 39.276 GBs)
                Warning: Available disk space may be INACCURATE if you have multiple drives.
                stage fetchlinks: 1.2 MBs
                stage download: 576.0 MBs
                stage concepts: 41.1 MBs
                stage sr: 240.0 MBs

Amount of memory allocated for the JVM is okay
                memory required: 3.0GB
                memory allocated: 3.8GB

Connection to database succeeded. Active configuration:
                username: "sa"
                url: "jdbc:h2:./db/h2;LOG=0;CACHE_SIZE=65536;LOCK_MODE=0;UNDO_LOG=0;MAX_OPERATION_MEMORY=
100000000"
                partitions: "default"
                connectionsPerPartition: 2
                driver: "org.h2.Driver"
                password: ""

Beginning import process in 20 seconds...
```



Base directory .

Java memory 4G

Language(s) simple, sco

Data source H2

H2 Path ${baseDir}/db/h2

Please select phases:

☑ Basic data

☑ Lucene (required by SR)

☑ Phrases (required by SR)

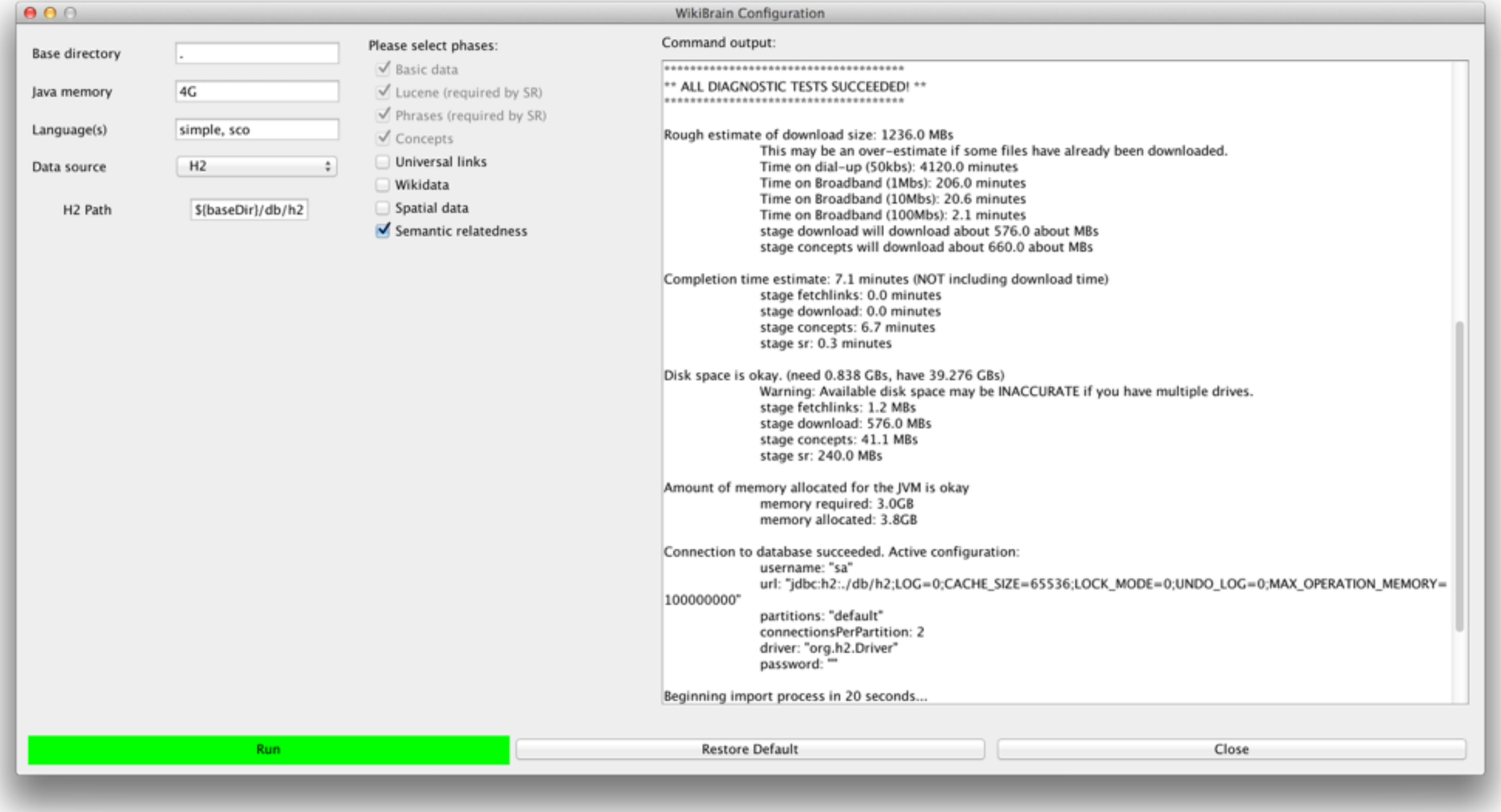
☑ Concepts

☐ Universal links

☐ Wikidata

☐ Spatial data

☑ Semantic relatedness



Base directory .

Java memory 4G

Language(s) simple, sco

Data source H2

H2 Path ${baseDir}/db/h2

Please select phases:

- ☑ Basic data
- ☑ Lucene (required by SR)
- ☑ Phrases (required by SR)
- ☑ Concepts
- ☐ Universal links
- ☐ Wikidata
- ☐ Spatial data
- ☑ Semantic relatedness

Command output:

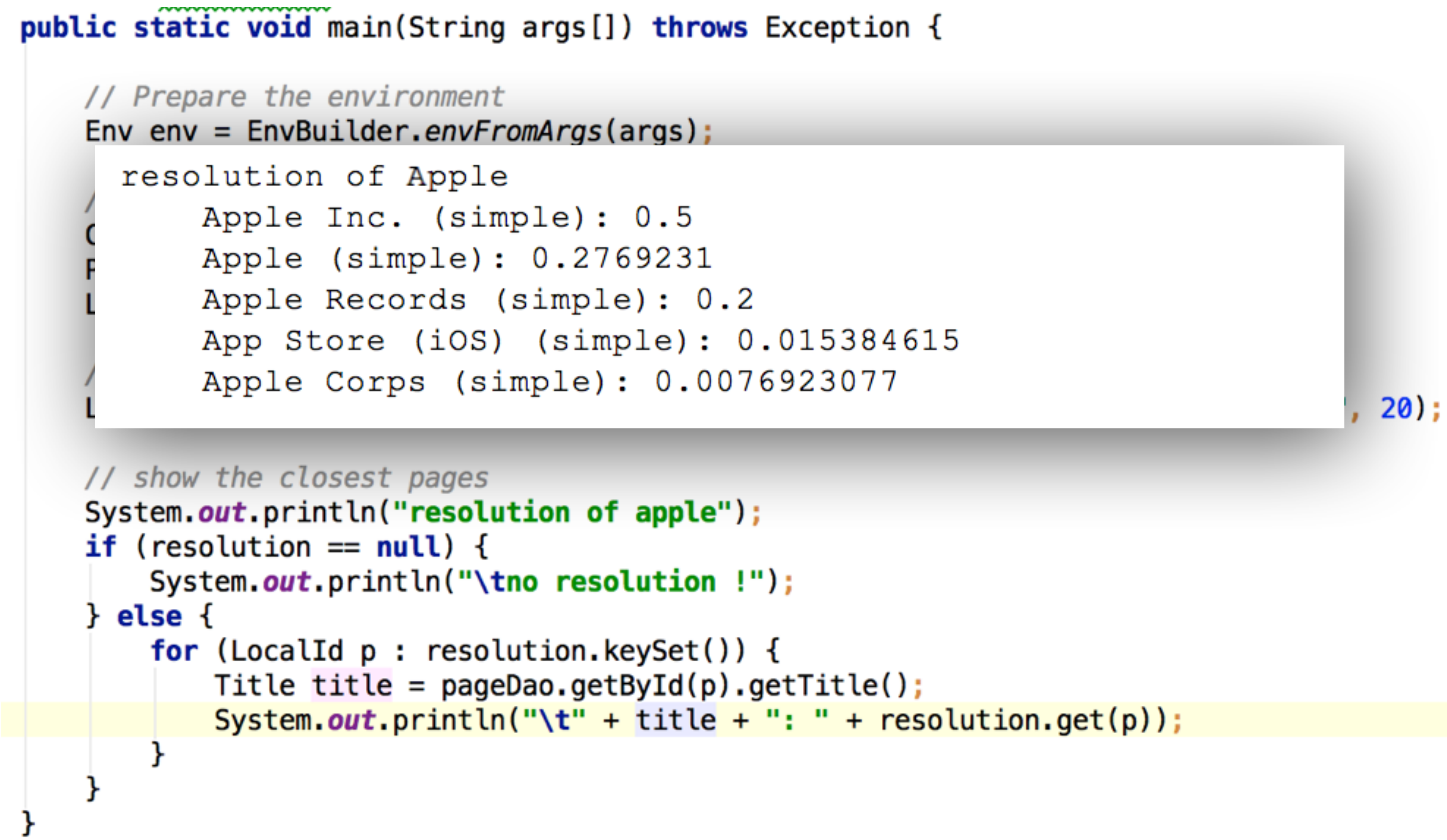
```
*************************************
** ALL DIAGNOSTIC TESTS SUCCEEDED! **
*************************************

Rough estimate of download size: 1236.0 MBs
                This may be an over-estimate if some files have already been downloaded.
                Time on dial-up (50kbs): 4120.0 minutes
                Time on Broadband (1Mbs): 206.0 minutes
                Time on Broadband (10Mbs): 20.6 minutes
                Time on Broadband (100Mbs): 2.1 minutes
                stage download will download about 576.0 about MBs
                stage concepts will download about 660.0 about MBs

Completion time estimate: 7.1 minutes (NOT including download time)
                stage fetchlinks: 0.0 minutes
                stage download: 0.0 minutes
                stage concepts: 6.7 minutes
                stage sr: 0.3 minutes

Disk space is okay. (need 0.838 GBs, have 39.276 GBs)
                Warning: Available disk space may be INACCURATE if you have multiple drives.
                stage fetchlinks: 1.2 MBs
                stage download: 576.0 MBs
                stage concepts: 41.1 MBs
                stage sr: 240.0 MBs

Amount of memory allocated for the JVM is okay
                memory required: 3.0GB
                memory allocated: 3.8GB

Connection to database succeeded. Active configuration:
                username: "sa"
                url: "jdbc:h2:./db/h2;LOG=0;CACHE_SIZE=65536;LOCK_MODE=0;UNDO_LOG=0;MAX_OPERATION_MEMORY=
100000000"
                partitions: "default"
                connectionsPerPartition: 2
                driver: "org.h2.Driver"
                password: ""

Beginning import process in 20 seconds...
```

Import times for core data:

language	**# articles**	**# links**	**runtime***
Simple English	102K	6M	8 min
German	1.9M	96M	210 min
English	4.6M	470M	640 min
25 largest	25M	1,670M	3163 min

*Additional time required for SR, geospatial, wikidata.

Quickstart.java



```
public static void main(String args[]) throws Exception {

    // Prepare the environment
    Env env = EnvBuilder.envFromArgs(args);
    /
    C
    F
    L

    /
    L                                                                  , 20);
```

```
resolution of Apple
    Apple Inc. (simple): 0.5
    Apple (simple): 0.2769231
    Apple Records (simple): 0.2
    App Store (iOS) (simple): 0.015384615
    Apple Corps (simple): 0.0076923077
```

```
    // show the closest pages
    System.out.println("resolution of apple");
    if (resolution == null) {
        System.out.println("\tno resolution !");
    } else {
        for (LocalId p : resolution.keySet()) {
            Title title = pageDao.getById(p).getTitle();
            System.out.println("\t" + title + ": " + resolution.get(p));
        }
    }
}
```

WikiBrain features

Core data structures

Graphs: link, category, redirect.

Article text: wikitext and plaintext.

Full text search using Lucene.

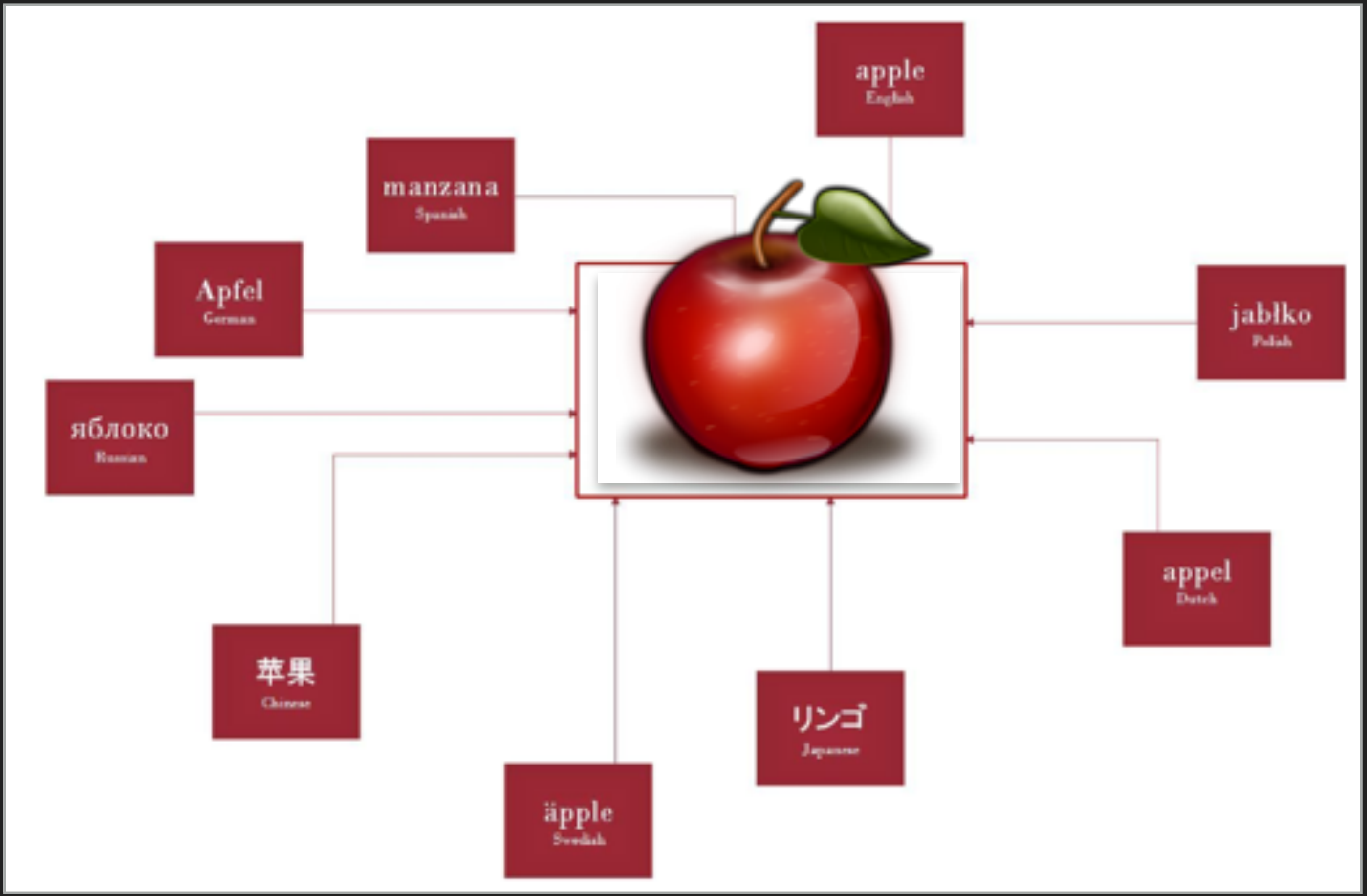
Highly optimized disk and memory caches.

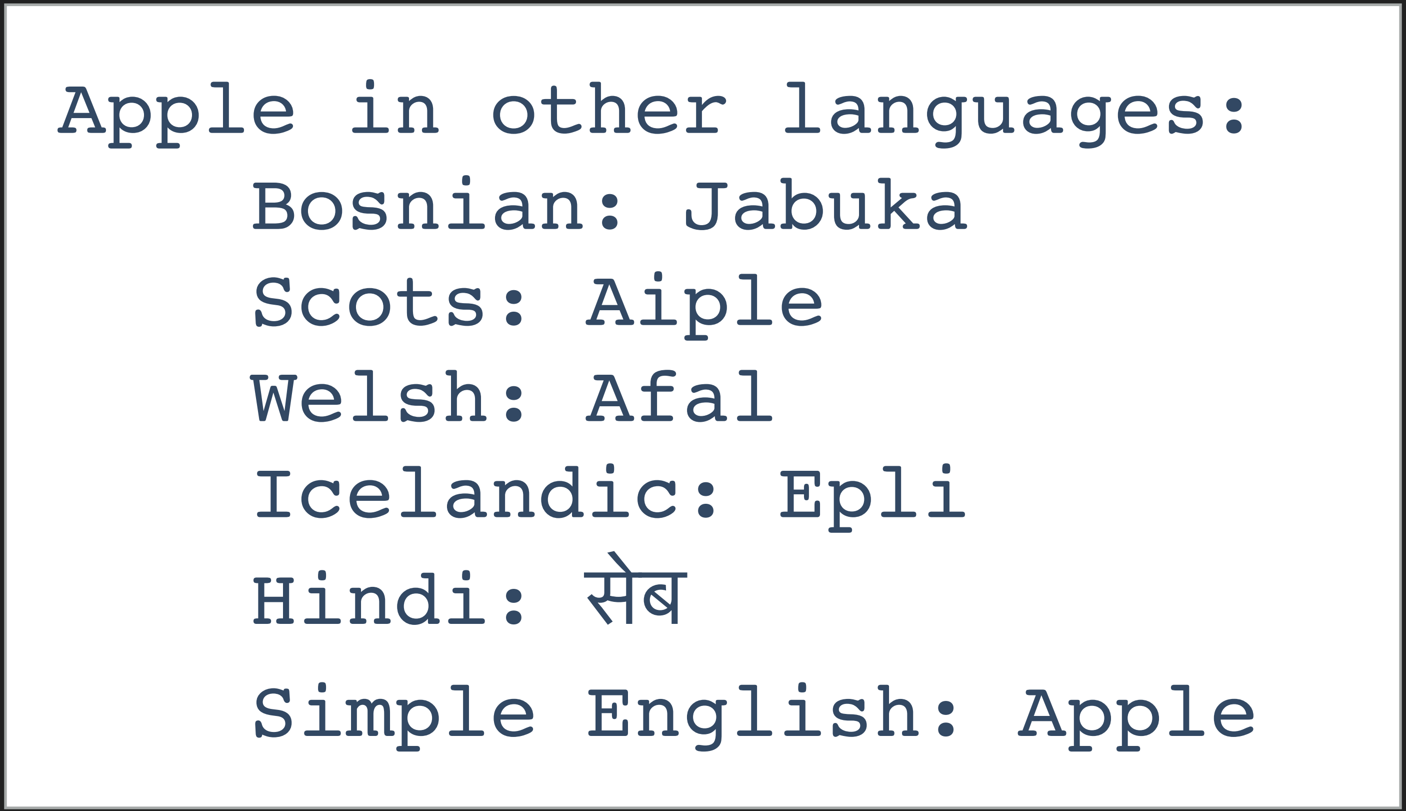
Multilingual

All languages**

Concept alignment

Universal links





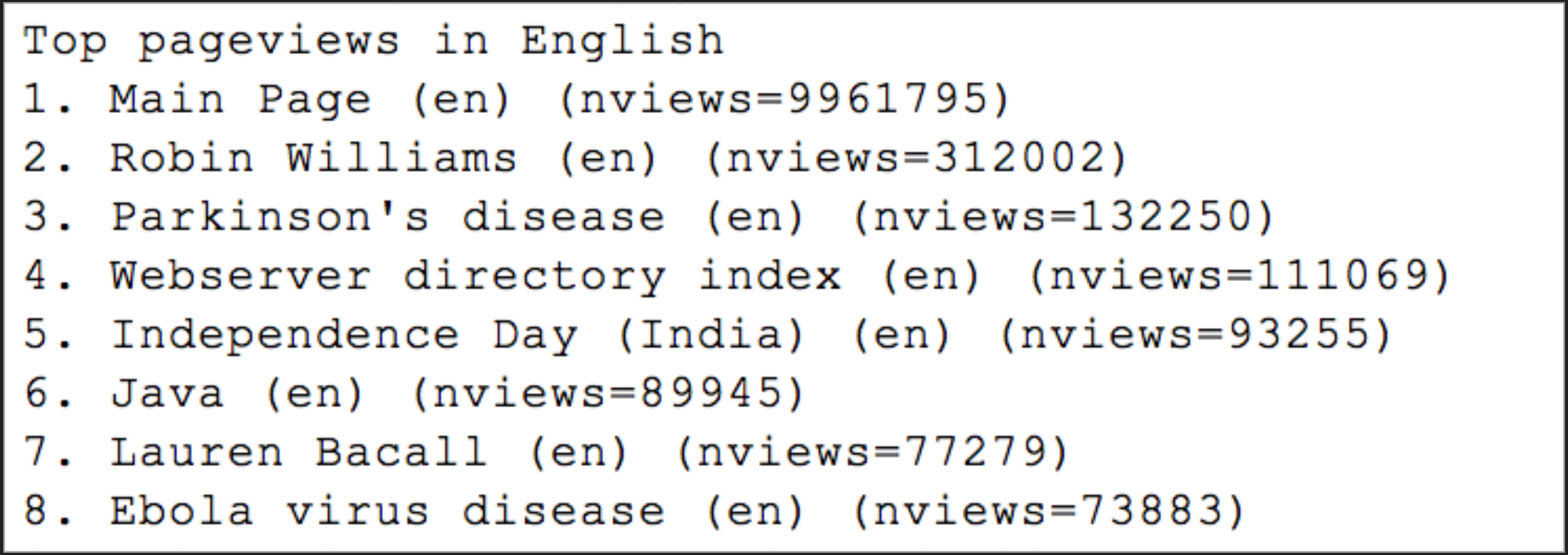
Apple in other languages:
Bosnian: Jabuka
Scots: Aiple
Welsh: Afal
Icelandic: Epli
Hindi: सेब
Simple English: Apple

apple image: CC public domain, http://www.freestockphotos.biz/stockphoto/16787

Pageview module

Num of page views for a requested date range.

On August 14, 2014:

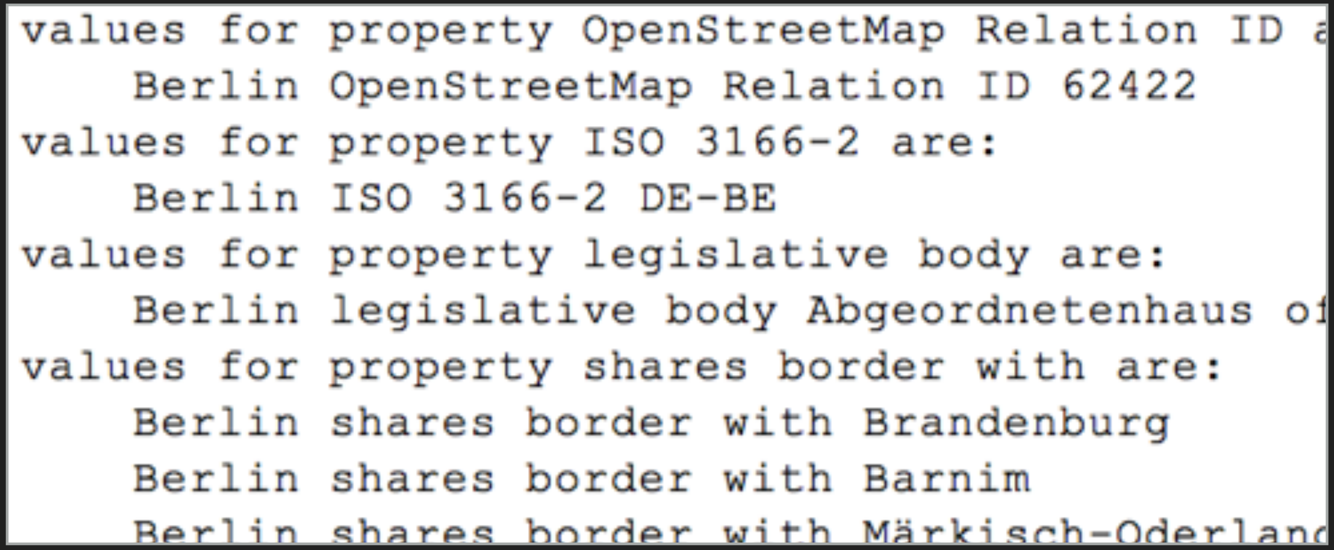


```
Top pageviews in English
1. Main Page (en) (nviews=9961795)
2. Robin Williams (en) (nviews=312002)
3. Parkinson's disease (en) (nviews=132250)
4. Webserver directory index (en) (nviews=111069)
5. Independence Day (India) (en) (nviews=93255)
6. Java (en) (nviews=89945)
7. Lauren Bacall (en) (nviews=77279)
8. Ebola virus disease (en) (nviews=73883)
```

Wikidata module

Over 40M statements about 15M concepts.

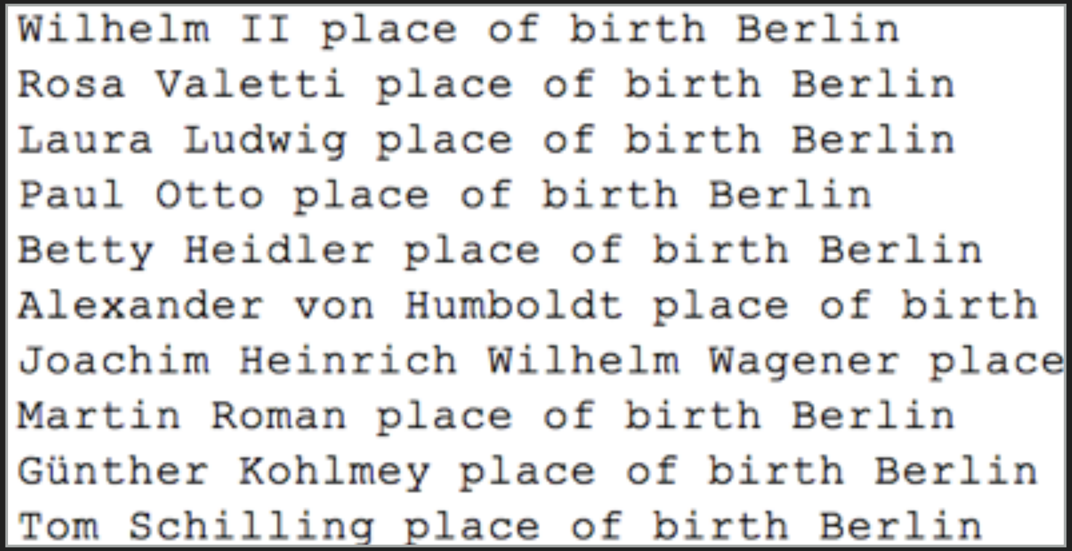
What does WikiBrain know about Berlin?



```
values for property OpenStreetMap Relation ID a
    Berlin OpenStreetMap Relation ID 62422
values for property ISO 3166-2 are:
    Berlin ISO 3166-2 DE-BE
values for property legislative body are:
    Berlin legislative body Abgeordnetenhaus of
values for property shares border with are:
    Berlin shares border with Brandenburg
    Berlin shares border with Barnim
    Berlin shares border with Märkisch-Oderlan
```

+ 100 more statements

Who was born in Berlin?



```
Wilhelm II place of birth Berlin
Rosa Valetti place of birth Berlin
Laura Ludwig place of birth Berlin
Paul Otto place of birth Berlin
Betty Heidler place of birth Berlin
Alexander von Humboldt place of birth
Joachim Heinrich Wilhelm Wagener place
Martin Roman place of birth Berlin
Günther Kohlmey place of birth Berlin
Tom Schilling place of birth Berlin
```

+ 2000 more people in EN

Semantic relatedness module

similarity(x,y)

mostSimilar(x)

cosimilarity(x_1,x_2,…x_n)

```
mostSimilar(“Berlin”):
  1. Munich
  2. Hamburg
  3. Vienna

mostSimilar(“Berlin”, <all movie ids>):
  1. The Wall (1962 film)
  2. The Tunnel (2001 film)
  3. The Road to the Wall
```

Named Entity Recognition

```
resolve(phrase)

resolve(phrase, context)

wikify(text)
```

```
resolve(“Apple”):
  1. Apple Inc. (simple): 0.50
  2. Apple (simple): 0.28
  3. Apple Records (simple): 0.20

wikify(“Wikipedia is a free-access…”):
   Wikipedia is a free-access, free-content Internet
   encyclopedia, supported and hosted by the non-
   profit Wikimedia Foundation.
```

Geospatial module

Built on PostGIS and OpenGeo.

Layers connected to articles:

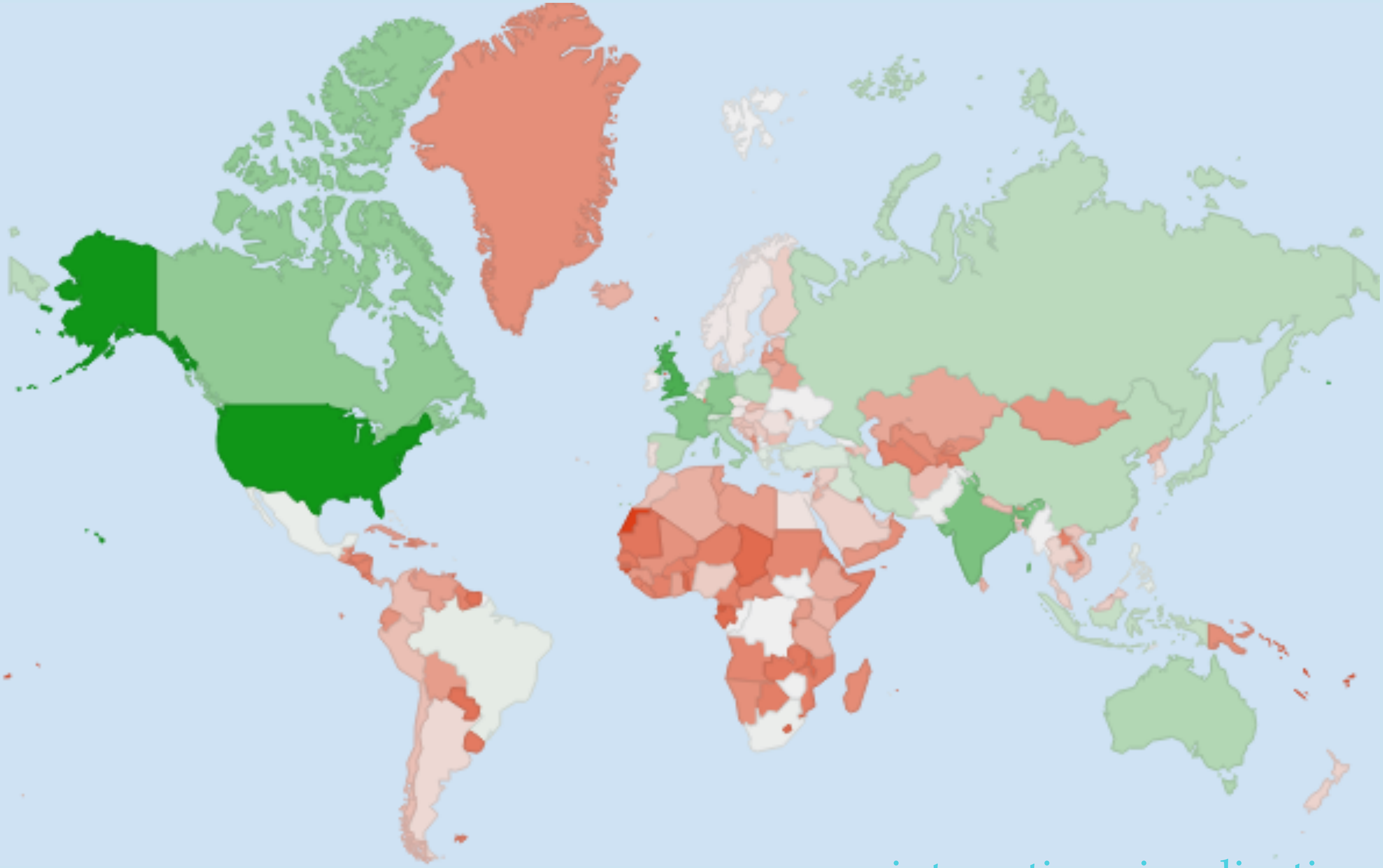
- Wikidata (coordinate points).
- Country (polygons from NaturalEarth).
- State (polygons from NaturalEarth).

Q: How many kms separate Berlin and Alaska?

Q: How many countries separate Berlin and Shanghai?

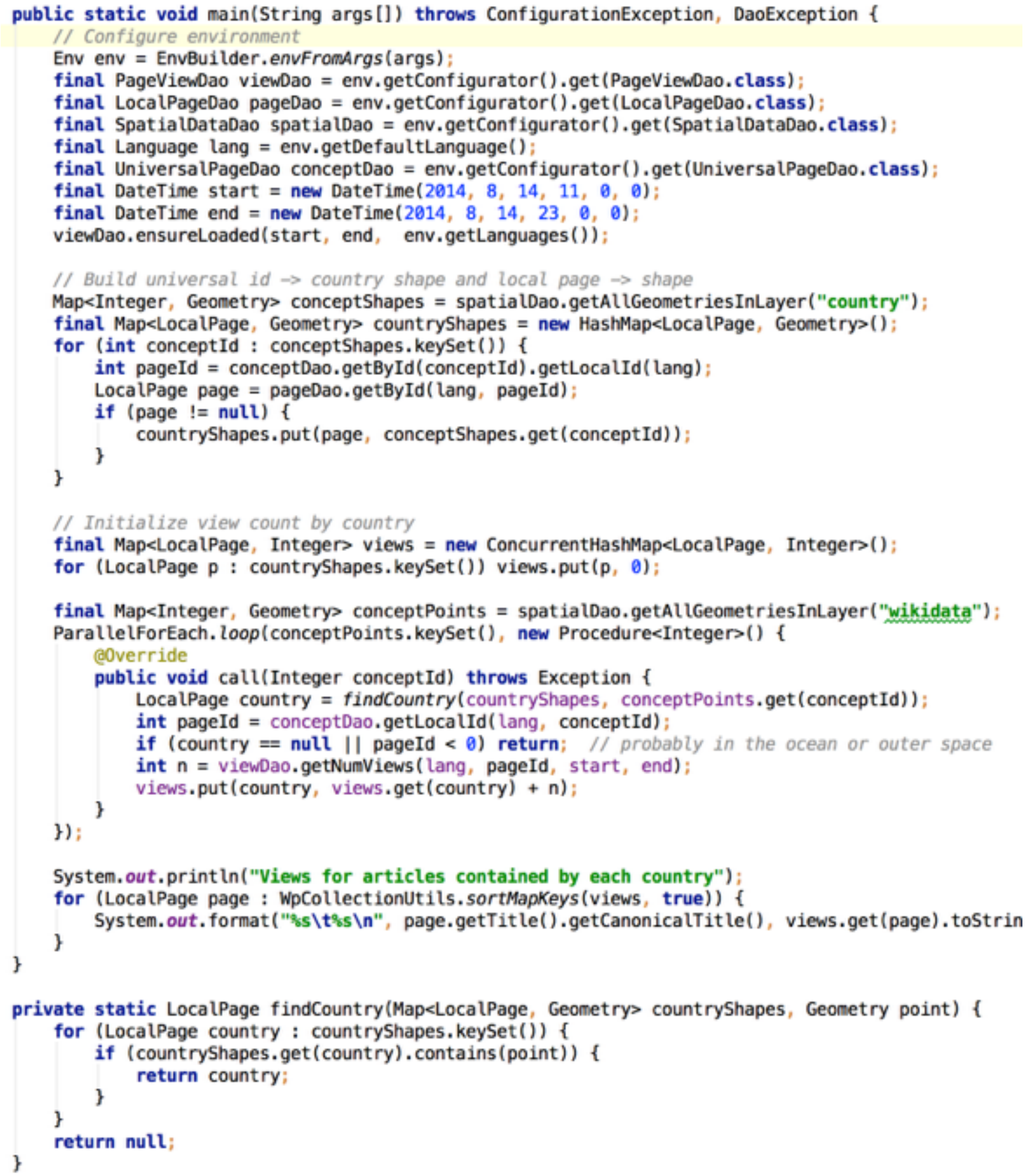
Q: What articles are about places in Minnesota?

Spatial article views by country (EN)



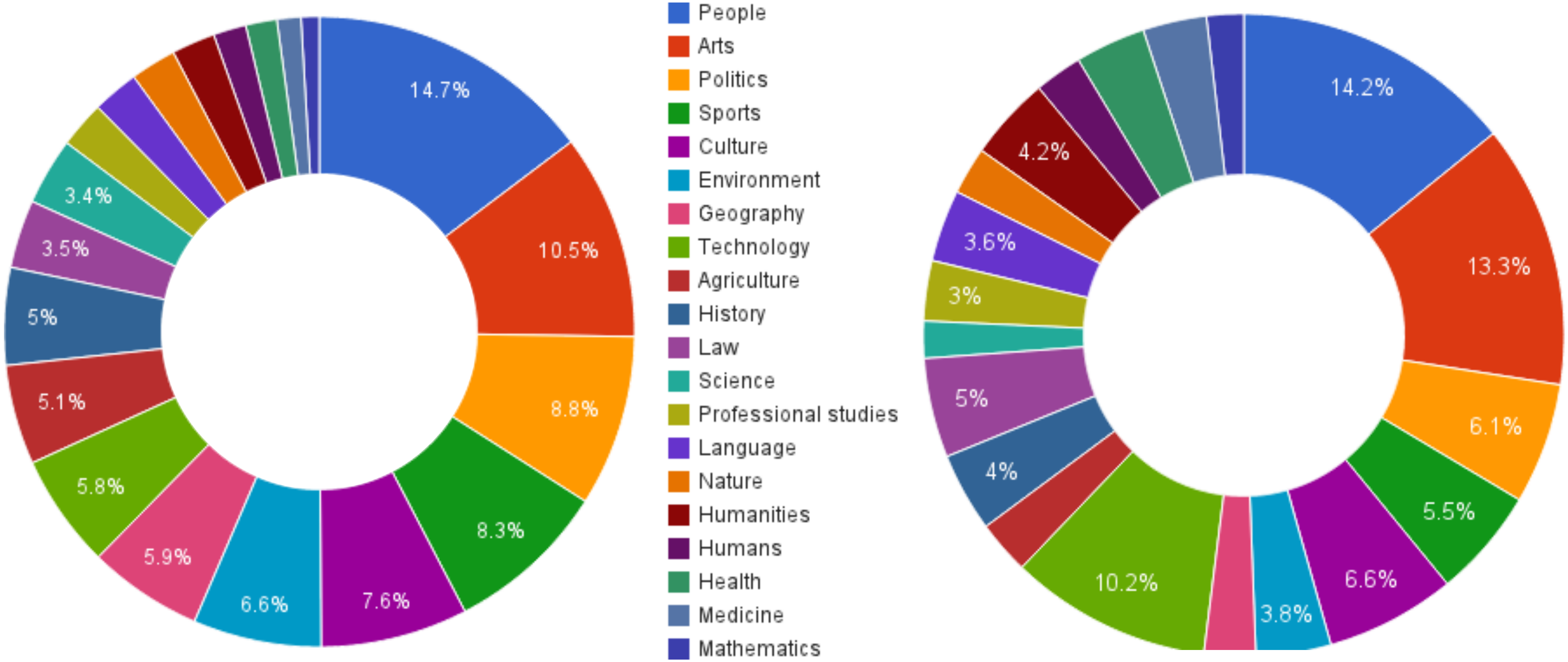
interactive visualization

CountryPageViews.java



```
public static void main(String args[]) throws ConfigurationException, DaoException {
    // Configure environment
    Env env = EnvBuilder.envFromArgs(args);
    final PageViewDao viewDao = env.getConfigurator().get(PageViewDao.class);
    final LocalPageDao pageDao = env.getConfigurator().get(LocalPageDao.class);
    final SpatialDataDao spatialDao = env.getConfigurator().get(SpatialDataDao.class);
    final Language lang = env.getDefaultLanguage();
    final UniversalPageDao conceptDao = env.getConfigurator().get(UniversalPageDao.class);
    final DateTime start = new DateTime(2014, 8, 14, 11, 0, 0);
    final DateTime end = new DateTime(2014, 8, 14, 23, 0, 0);
    viewDao.ensureLoaded(start, end,  env.getLanguages());

    // Build universal id -> country shape and local page -> shape
    Map<Integer, Geometry> conceptShapes = spatialDao.getAllGeometriesInLayer("country");
    final Map<LocalPage, Geometry> countryShapes = new HashMap<LocalPage, Geometry>();
    for (int conceptId : conceptShapes.keySet()) {
        int pageId = conceptDao.getById(conceptId).getLocalId(lang);
        LocalPage page = pageDao.getById(lang, pageId);
        if (page != null) {
            countryShapes.put(page, conceptShapes.get(conceptId));
        }
    }

    // Initialize view count by country
    final Map<LocalPage, Integer> views = new ConcurrentHashMap<LocalPage, Integer>();
    for (LocalPage p : countryShapes.keySet()) views.put(p, 0);

    final Map<Integer, Geometry> conceptPoints = spatialDao.getAllGeometriesInLayer("wikidata");
    ParallelForEach.loop(conceptPoints.keySet(), new Procedure<Integer>() {
        @Override
        public void call(Integer conceptId) throws Exception {
            LocalPage country = findCountry(countryShapes, conceptPoints.get(conceptId));
            int pageId = conceptDao.getLocalId(lang, conceptId);
            if (country == null || pageId < 0) return;  // probably in the ocean or outer space
            int n = viewDao.getNumViews(lang, pageId, start, end);
            views.put(country, views.get(country) + n);
        }
    });

    System.out.println("Views for articles contained by each country");
    for (LocalPage page : WpCollectionUtils.sortMapKeys(views, true)) {
        System.out.format("%s\t%s\n", page.getTitle().getCanonicalTitle(), views.get(page).toStrin
    }
}

private static LocalPage findCountry(Map<LocalPage, Geometry> countryShapes, Geometry point) {
    for (LocalPage country : countryShapes.keySet()) {
        if (countryShapes.get(country).contains(point)) {
            return country;
        }
    }
    return null;
}
```

Supply vs demand of categories

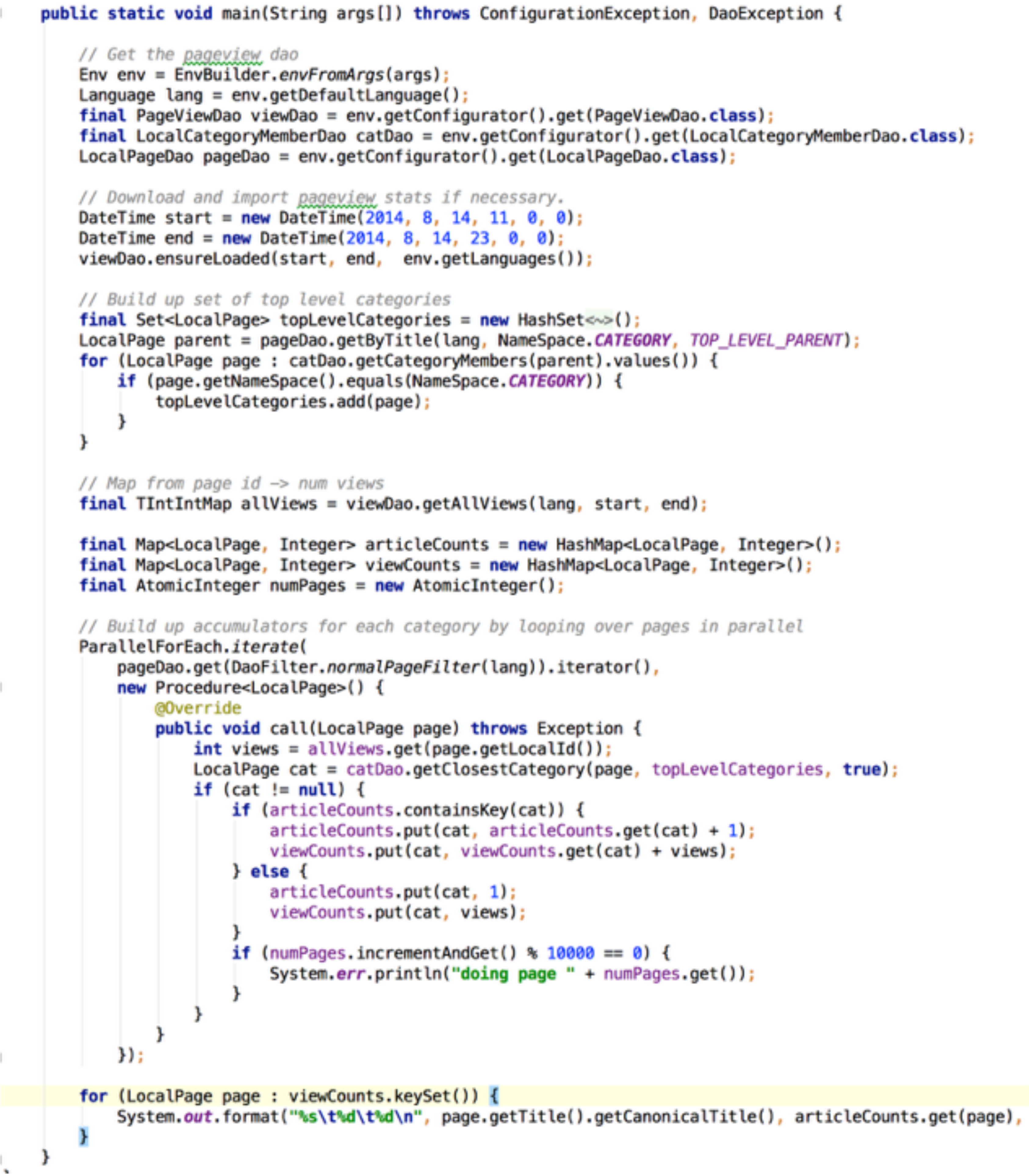


Supply:
# articles per category

Demand:
# views per category

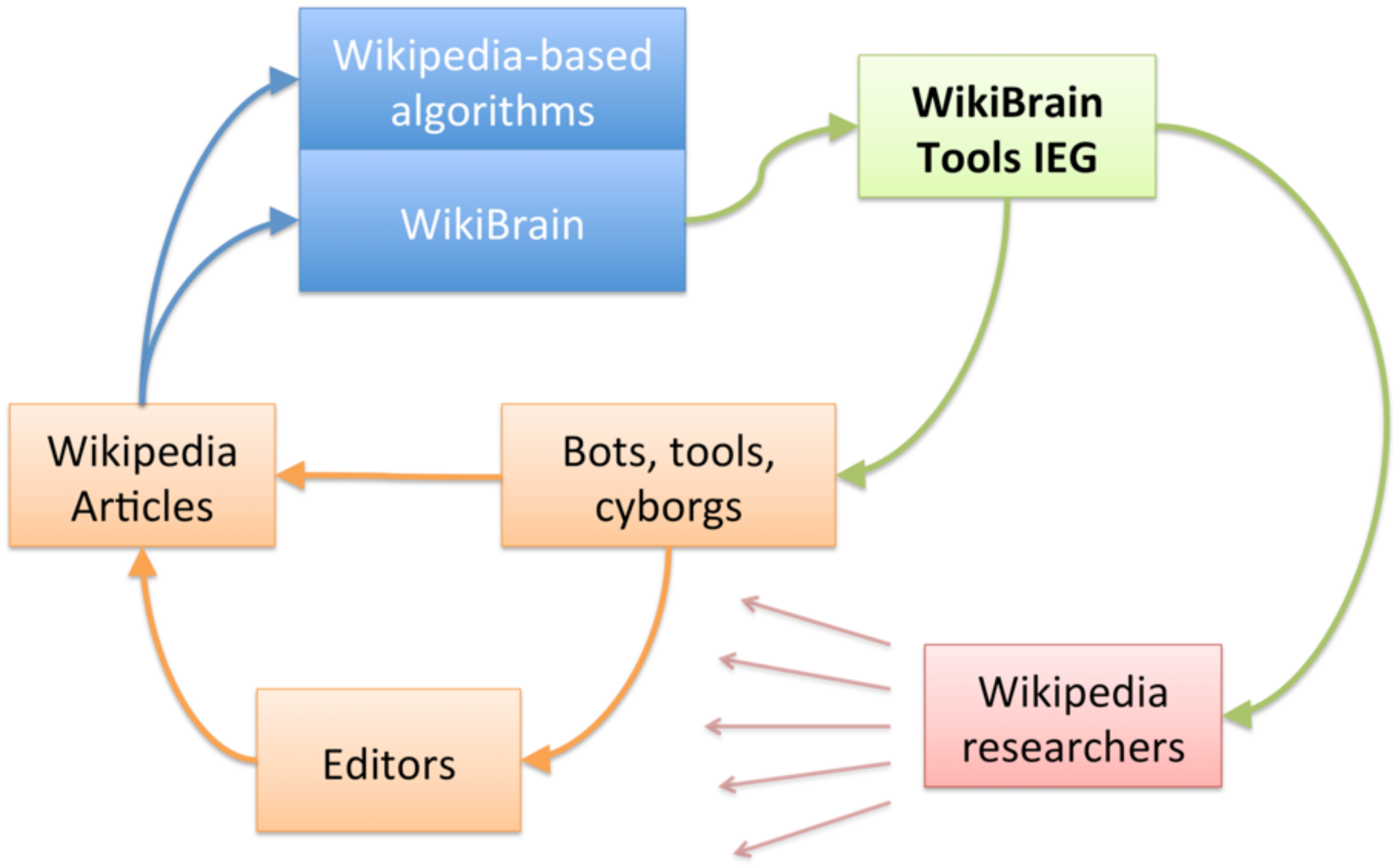
Inspired by Kittur, Chi, and Suh. "What's in Wikipedia?: mapping topics and conflict using socially annotated category structure." *CHI*, 2009.

CategoryViews.java

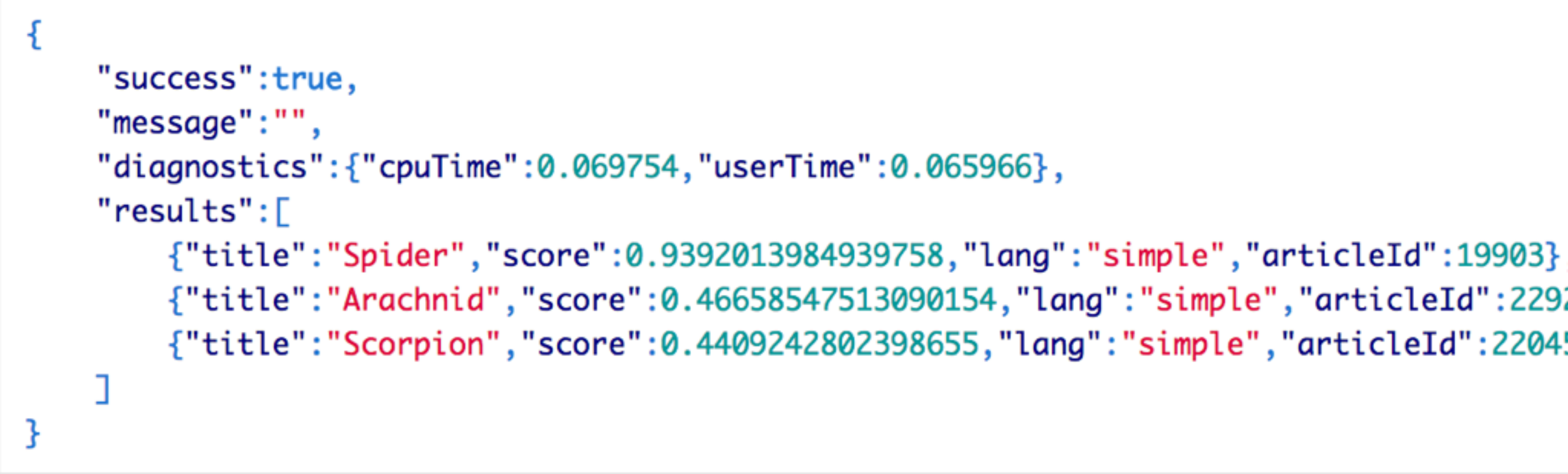


```java
public static void main(String args[]) throws ConfigurationException, DaoException {

    // Get the pageview dao
    Env env = EnvBuilder.envFromArgs(args);
    Language lang = env.getDefaultLanguage();
    final PageViewDao viewDao = env.getConfigurator().get(PageViewDao.class);
    final LocalCategoryMemberDao catDao = env.getConfigurator().get(LocalCategoryMemberDao.class);
    LocalPageDao pageDao = env.getConfigurator().get(LocalPageDao.class);

    // Download and import pageview stats if necessary.
    DateTime start = new DateTime(2014, 8, 14, 11, 0, 0);
    DateTime end = new DateTime(2014, 8, 14, 23, 0, 0);
    viewDao.ensureLoaded(start, end,  env.getLanguages());

    // Build up set of top level categories
    final Set<LocalPage> topLevelCategories = new HashSet<~>();
    LocalPage parent = pageDao.getByTitle(lang, NameSpace.CATEGORY, TOP_LEVEL_PARENT);
    for (LocalPage page : catDao.getCategoryMembers(parent).values()) {
        if (page.getNameSpace().equals(NameSpace.CATEGORY)) {
            topLevelCategories.add(page);
        }
    }

    // Map from page id -> num views
    final TIntIntMap allViews = viewDao.getAllViews(lang, start, end);

    final Map<LocalPage, Integer> articleCounts = new HashMap<LocalPage, Integer>();
    final Map<LocalPage, Integer> viewCounts = new HashMap<LocalPage, Integer>();
    final AtomicInteger numPages = new AtomicInteger();

    // Build up accumulators for each category by looping over pages in parallel
    ParallelForEach.iterate(
        pageDao.get(DaoFilter.normalPageFilter(lang)).iterator(),
        new Procedure<LocalPage>() {
            @Override
            public void call(LocalPage page) throws Exception {
                int views = allViews.get(page.getLocalId());
                LocalPage cat = catDao.getClosestCategory(page, topLevelCategories, true);
                if (cat != null) {
                    if (articleCounts.containsKey(cat)) {
                        articleCounts.put(cat, articleCounts.get(cat) + 1);
                        viewCounts.put(cat, viewCounts.get(cat) + views);
                    } else {
                        articleCounts.put(cat, 1);
                        viewCounts.put(cat, views);
                    }
                    if (numPages.incrementAndGet() % 10000 == 0) {
                        System.err.println("doing page " + numPages.get());
                    }
                }
            }
        });

    for (LocalPage page : viewCounts.keySet()) {
        System.out.format("%s\t%d\t%d\n", page.getTitle().getCanonicalTitle(), articleCounts.get(page),
    }
}
```

WikiBrain Tools IEG

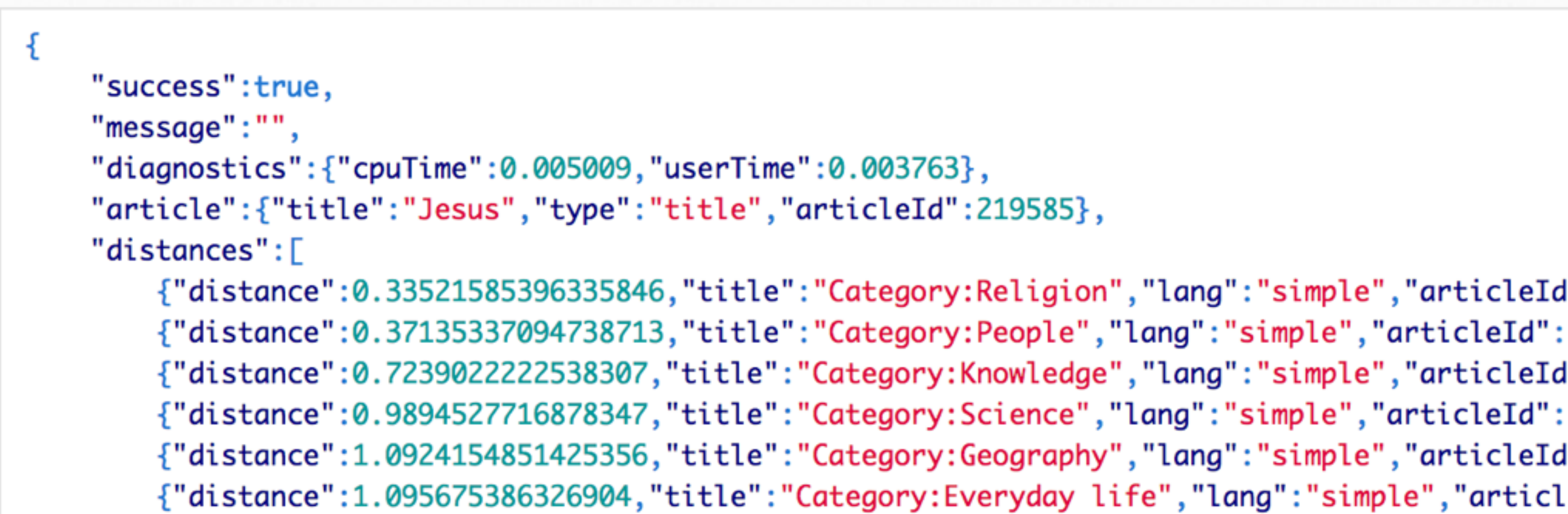


mostSimilar?lang=simple&phrase=spider&n=3



```
{
    "success":true,
    "message":"",
    "diagnostics":{"cpuTime":0.069754,"userTime":0.065966},
    "results":[
        {"title":"Spider","score":0.9392013984939758,"lang":"simple","articleId":19903},
        {"title":"Arachnid","score":0.46658547513090154,"lang":"simple","articleId":2292
        {"title":"Scorpion","score":0.4409242802398655,"lang":"simple","articleId":22045
    ]
}
```

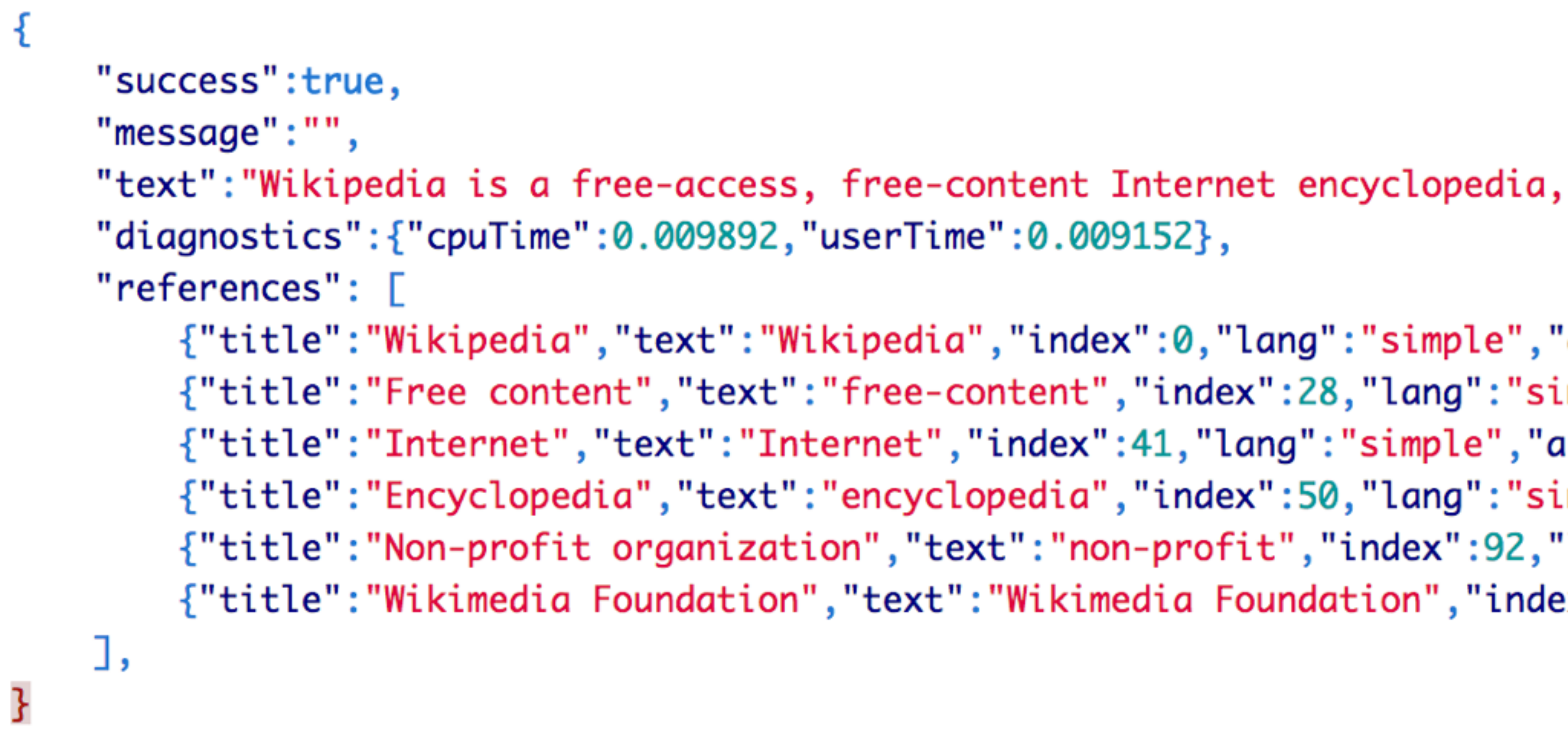
categoriesForArticle?title=Jesus&lang=simple



```
{
    "success":true,
    "message":"",
    "diagnostics":{"cpuTime":0.005009,"userTime":0.003763},
    "article":{"title":"Jesus","type":"title","articleId":219585},
    "distances":[
        {"distance":0.33521585396335846,"title":"Category:Religion","lang":"simple","articleId
        {"distance":0.37135337094738713,"title":"Category:People","lang":"simple","articleId":
        {"distance":0.723902222538307,"title":"Category:Knowledge","lang":"simple","articleId
        {"distance":0.9894527716878347,"title":"Category:Science","lang":"simple","articleId":
        {"distance":1.0924154851425356,"title":"Category:Geography","lang":"simple","articleId
        {"distance":1.095675386326904,"title":"Category:Everyday life","lang":"simple","articl
```

Wikification

> *Wikipedia is a free-access, free-content Internet encyclopedia, supported and hosted by the non-profit Wikimedia Foundation.*

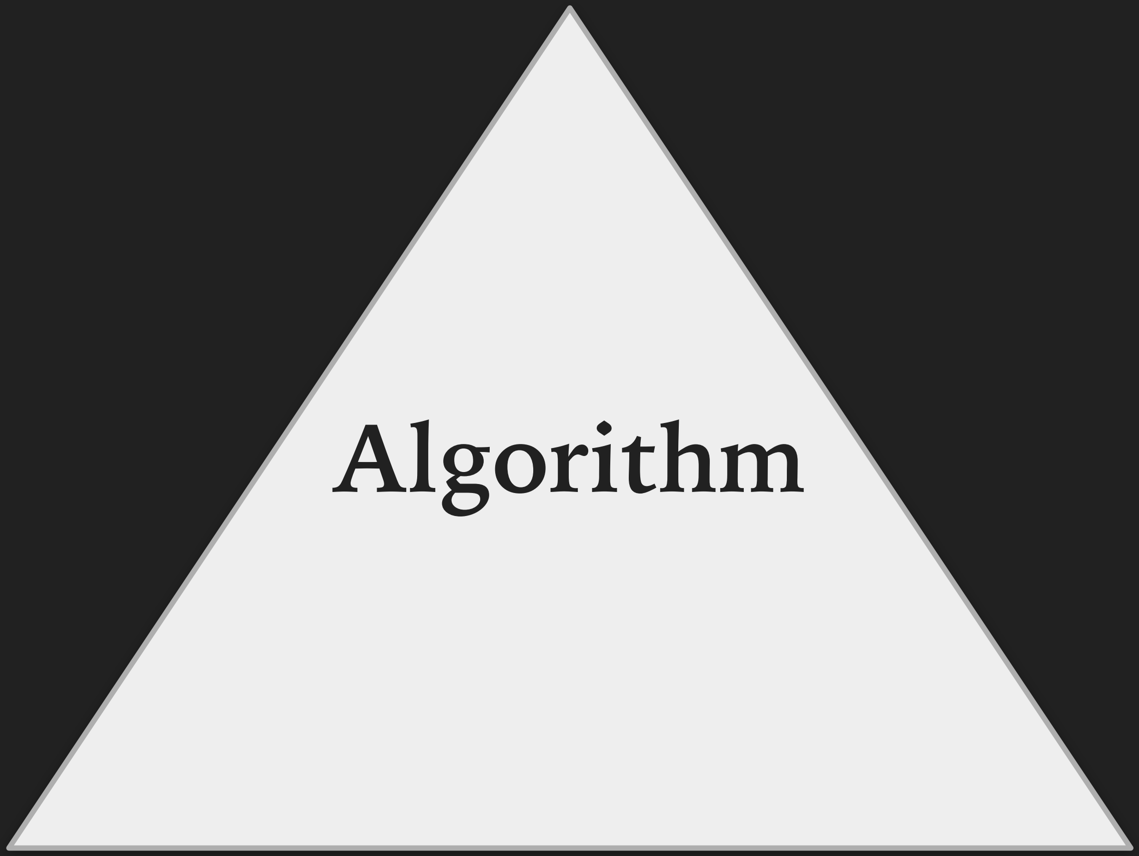


```
{
    "success":true,
    "message":"",
    "text":"Wikipedia is a free-access, free-content Internet encyclopedia,
    "diagnostics":{"cpuTime":0.009892,"userTime":0.009152},
    "references": [
        {"title":"Wikipedia","text":"Wikipedia","index":0,"lang":"simple","
        {"title":"Free content","text":"free-content","index":28,"lang":"si
        {"title":"Internet","text":"Internet","index":41,"lang":"simple","a
        {"title":"Encyclopedia","text":"encyclopedia","index":50,"lang":"si
        {"title":"Non-profit organization","text":"non-profit","index":92,"
        {"title":"Wikimedia Foundation","text":"Wikimedia Foundation","inde
    ],
}
```

WikiBrain developers:
Alan Morales Blanco, Margaret Giesel, Rebecca Gold, Becca Harper, Brent Hecht, Ben Hillman, Sam Horlbeck, Aaron Jiang, Matthew Lesicko, Toby Li, Yulun Li, Huy Mai, Ben Mathers, Sam Naden, Jesse Russell, Shilad Sen, Laura Sousa Vonessen, Zixiao Wang, and Ari Weilland

WikiBrain Case Studies

Cultural Alignment in Algorithms



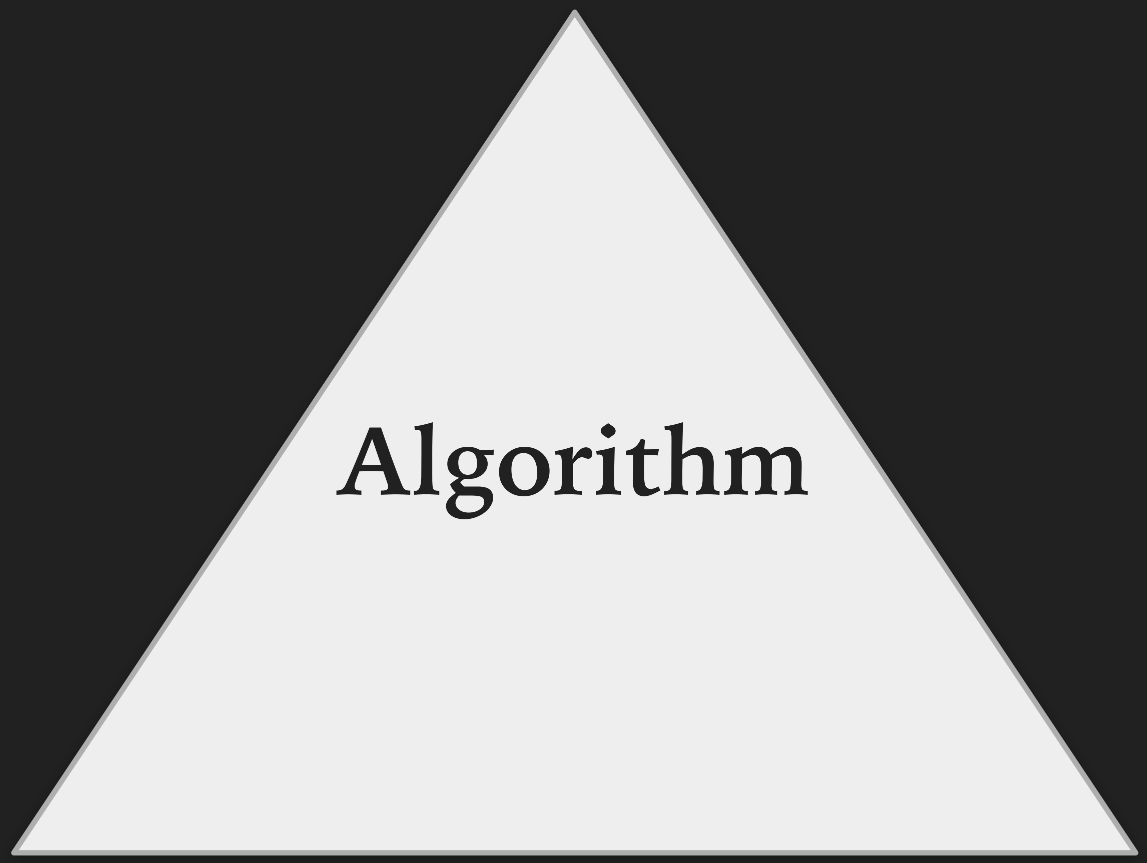
Knowledge Base
(Wikipedia Editors)

Gold Standard
(Human-Labeled Data)

Application Audience
(End-Users)

Sen, Giesel, Gold, Hillmann, Lesicko, Nadan, Russell, Wang, Hecht. Turkers, scholars, Arafat and peace: Cultural communities and algorithmic gold standards. CSCW 2015

Cultural Alignment in Algorithms



Psychology Articles

Gold Standard by Pyschologists

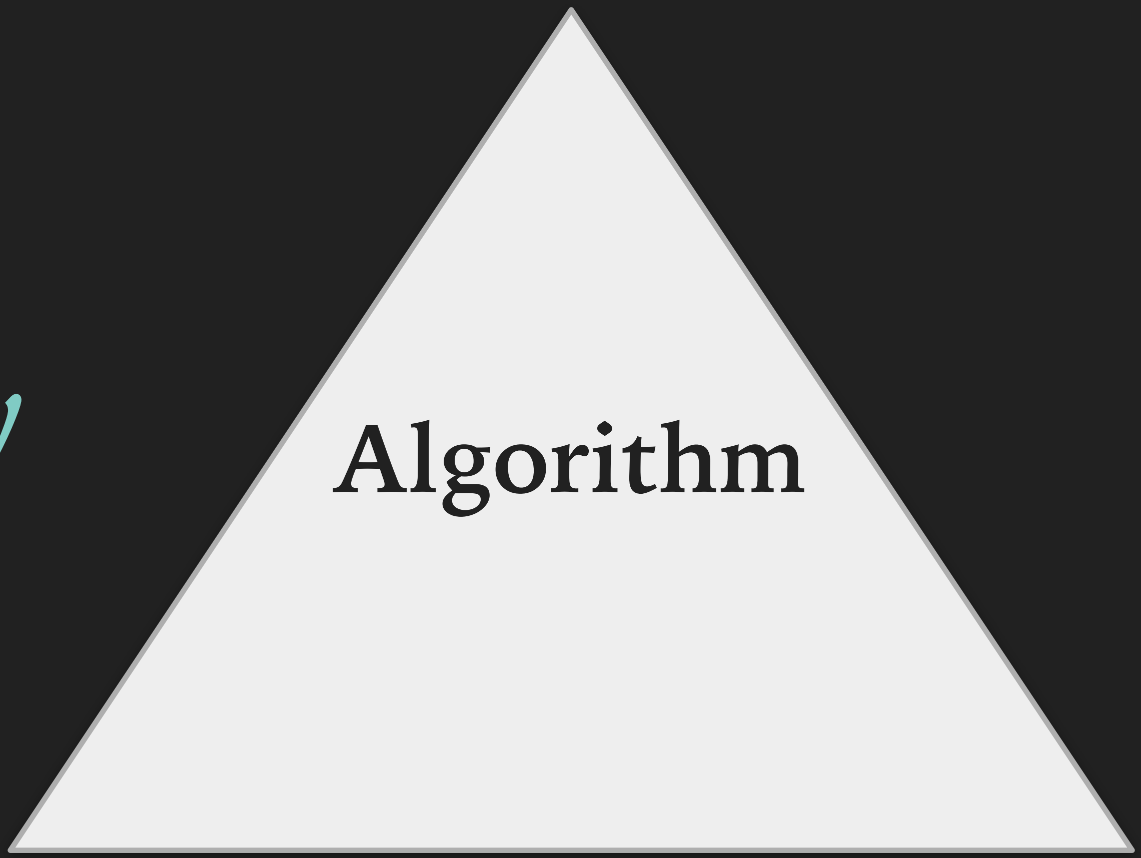
General Population

Sen, Giesel, Gold, Hillmann, Lesicko, Nadan, Russell, Wang, Hecht. Turkers, scholars, Arafat and peace: Cultural communities and algorithmic gold standards. CSCW 2015

Cultural Alignment in Algorithms

emotion

personality



Psychology Articles

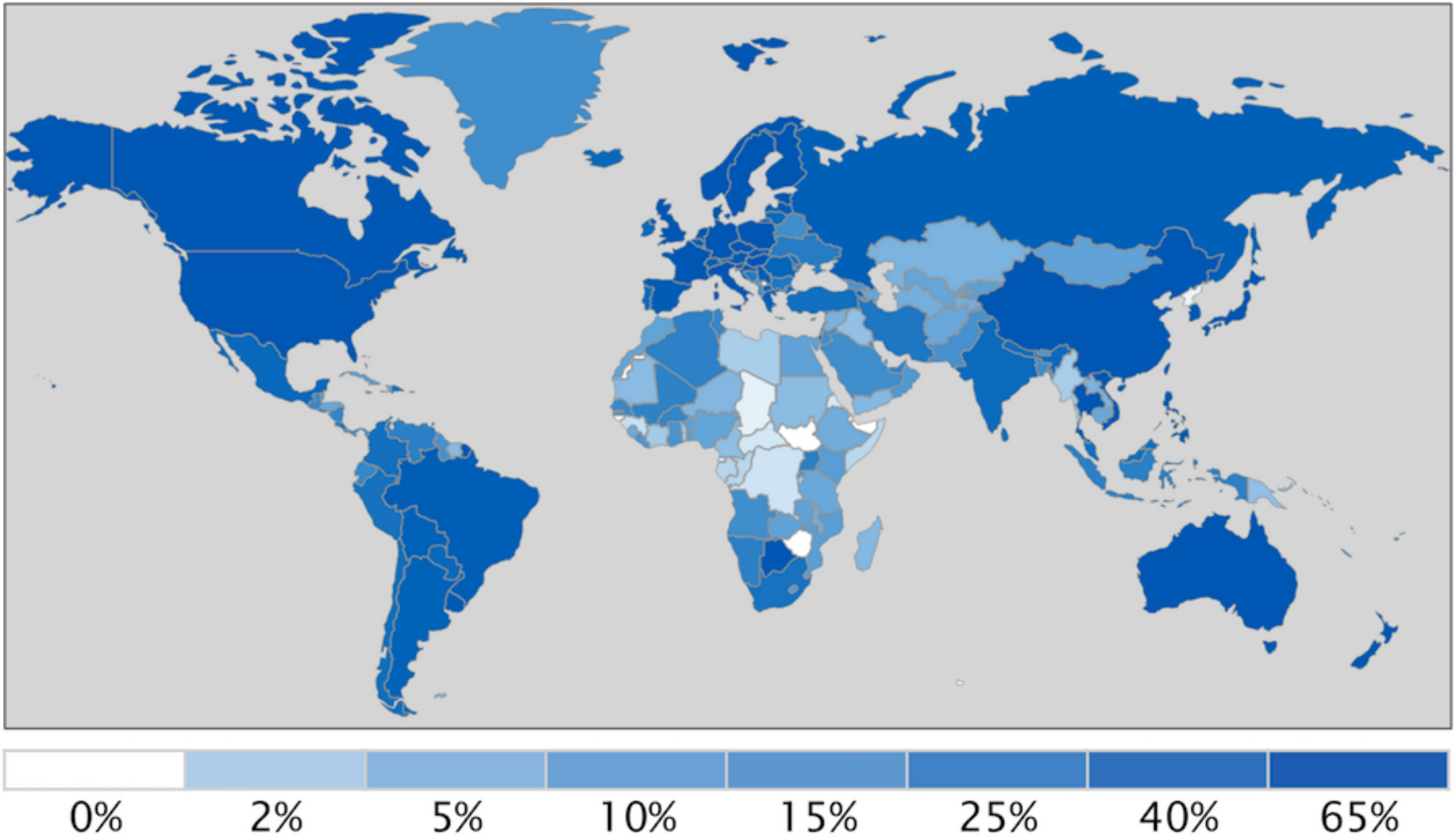
Algorithm

Gold Standard by Pyschologists

General Population

Sen, Giesel, Gold, Hillmann, Lesicko, Nadan, Russell, Wang, Hecht. Turkers, scholars, Arafat and peace: Cultural communities and algorithmic gold standards. CSCW 2015

Localness of Sources Cited in Spatial Articles



http://shilad.com/localness

Sen, Ford, Musicant, Graham, Keyes, Hecht. Barriers to the localness of volunteered geographic information. CHI 2015

$$\boldsymbol{SR}\left(\begin{matrix}\text{SF}\\ \text{Minneapolis}\end{matrix}\right) = \mathbf{0.6}$$

$$\boldsymbol{SR}\left(\begin{matrix}\text{SF}\\ \text{Sahara}\end{matrix}\right) = \mathbf{0.2}$$

$$\textbf{GESR(A,B)} = \begin{matrix} -0.019 * \textbf{states(A,B)} + \\ -0.173 * \textbf{ordinal(A,B)} + \\ +2.598 * \textbf{general-SR(A,B)} \end{matrix} \; +$$

		scale(B)		
		POI	**state**	**country**
scale(A)	**country**	+1.68 -0.27	n/a -1.25	n/a -0.65
	state	+0.77 -0.16	n/a -0.12	
	POI	n/a -0.09		

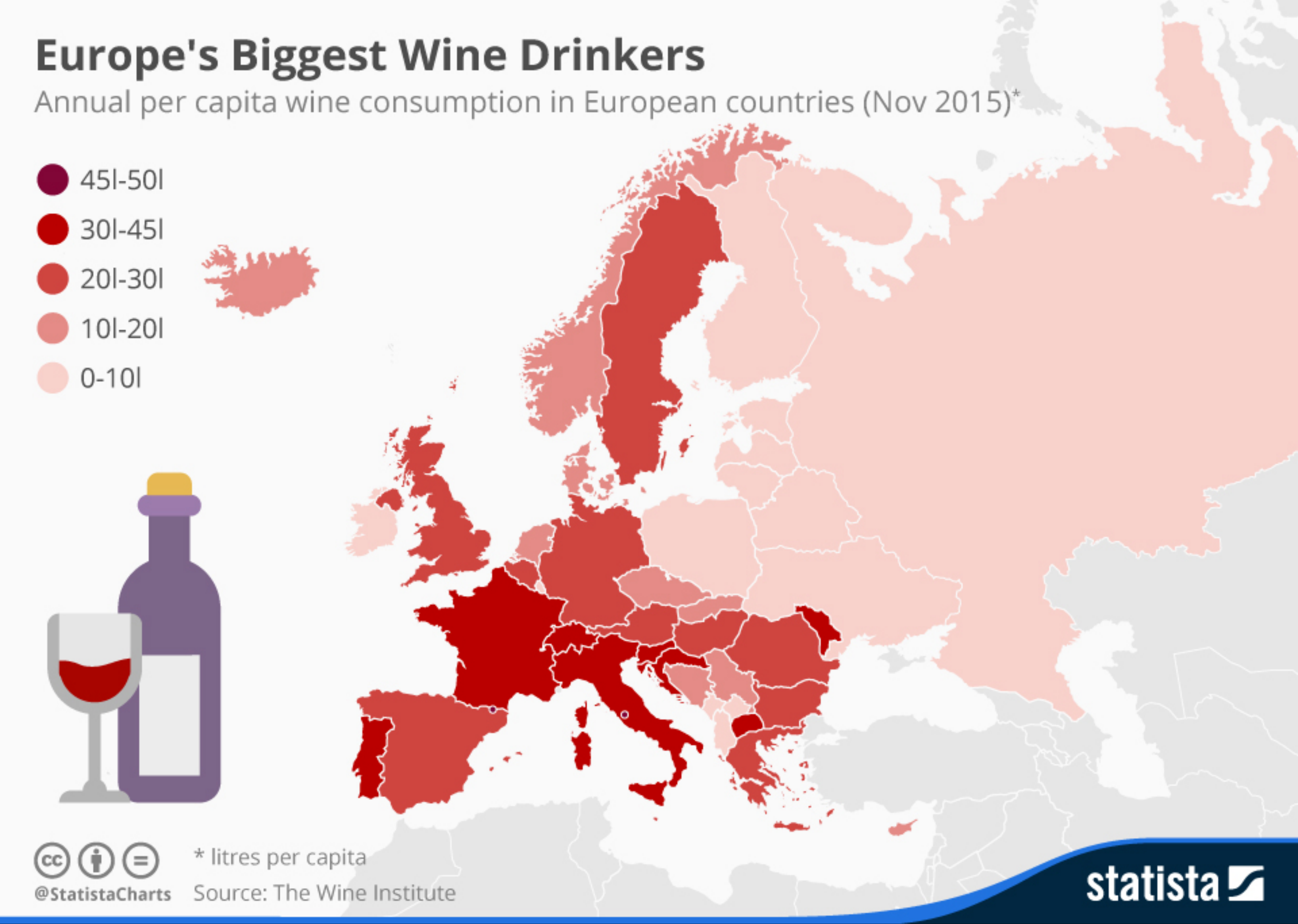
Towards domain-specific semantic relatedness : A case study from geography. Sen; Johnson; Harper; Mai; Olsen; Mathers; Vonessen; Wright; Hecht. IJCAI, 2015.

Cartograph

http://cartograph.info

Sen, Swoap, Li, Boatman, Dippenaar, Gold, Ngo, Pujol, Jackson, Hecht. "Cartograph: Unlocking Spatial Visualization Through Semantic Enhancement." IUI, 2017.

Thematic Cartography



Europe's Biggest Wine Drinkers

Annual per capita wine consumption in European countries (Nov 2015)*

* litres per capita

Source: The Wine Institute

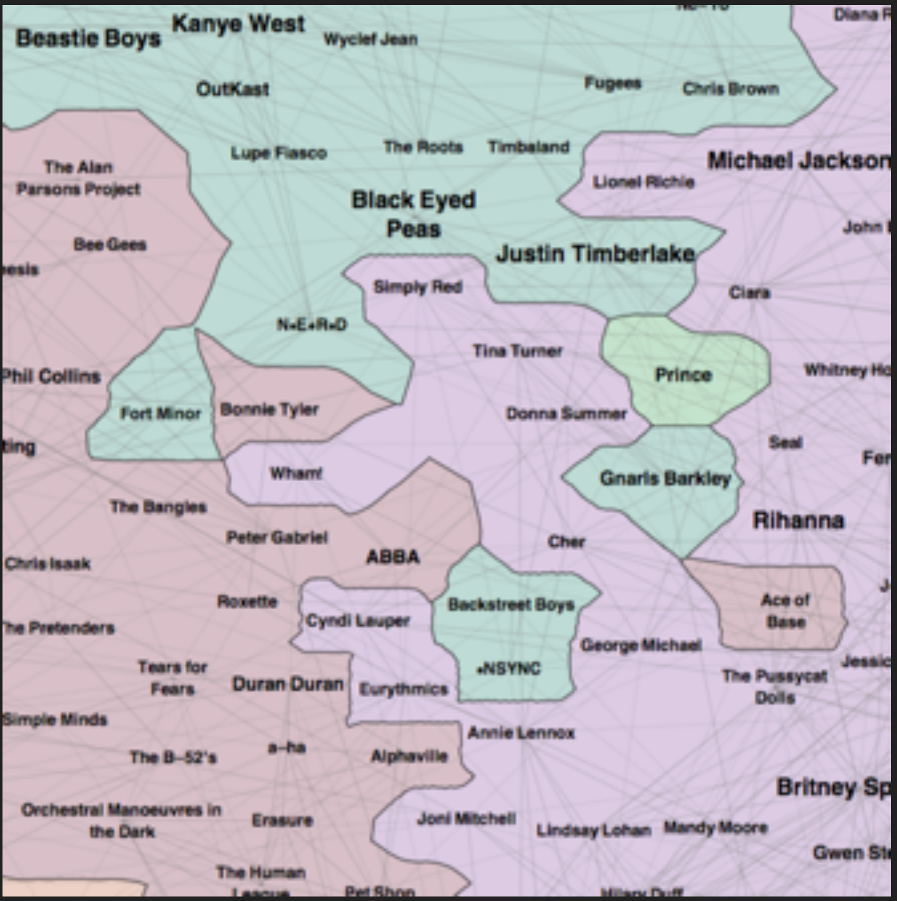
https://www.statista.com/chart/4837/europes-biggest-wine-drinkers/

Tobler’s First Law of Geography:

Everything is related to everything else, but near things are more related than distant things.

Spatialization

Prior Spatialization Systems

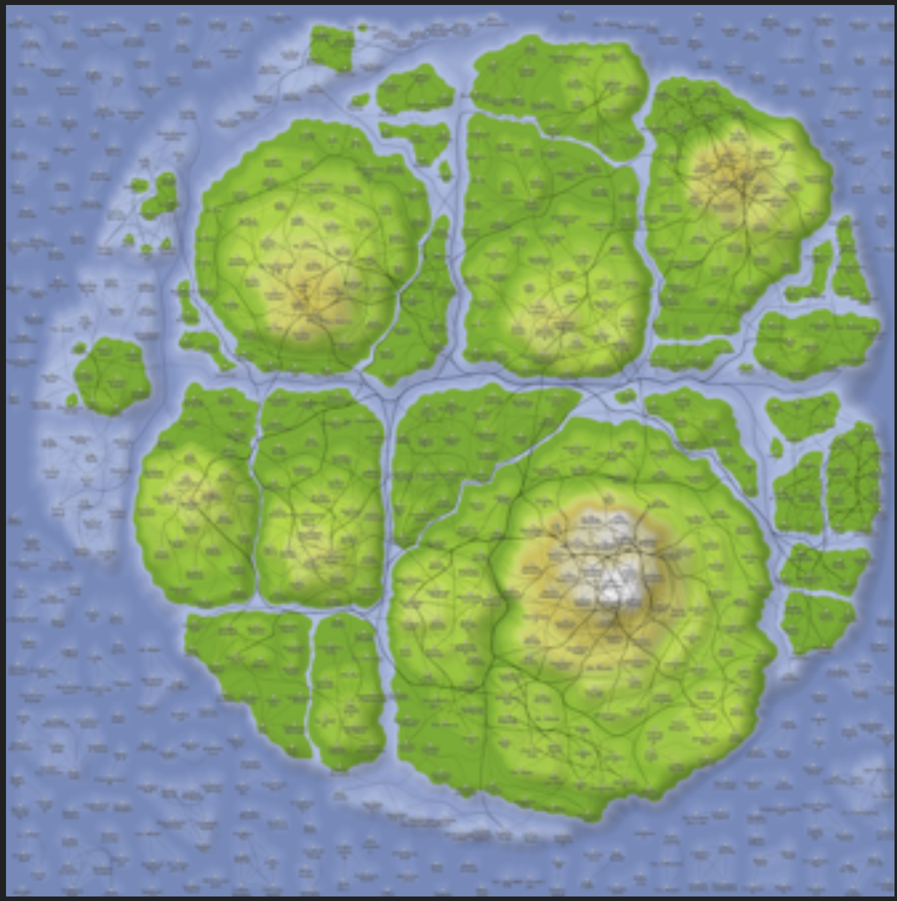


GMap

Hu et al., 2010

Atlasify

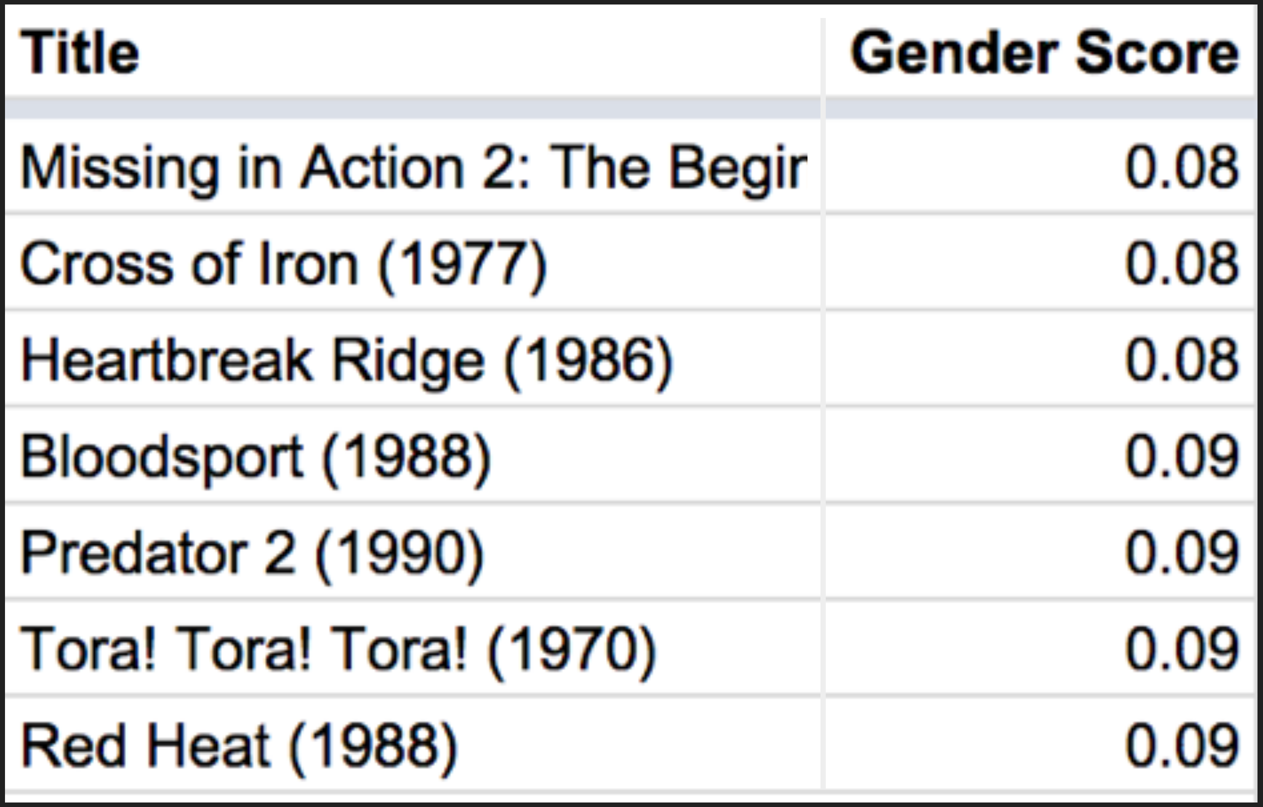
Hecht et al., 2012



Gronemann and Jünger, 2013

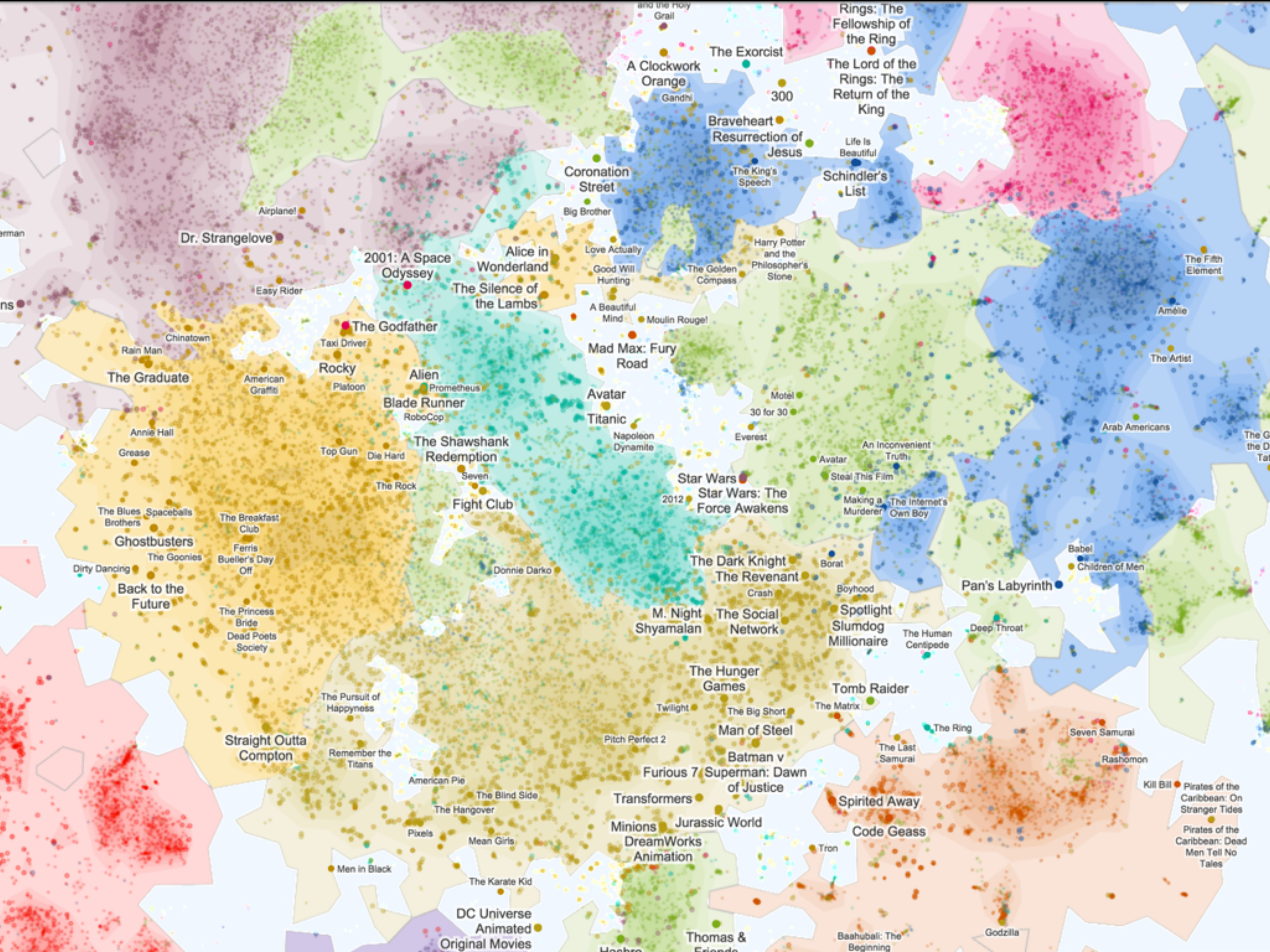
…and **many** others!

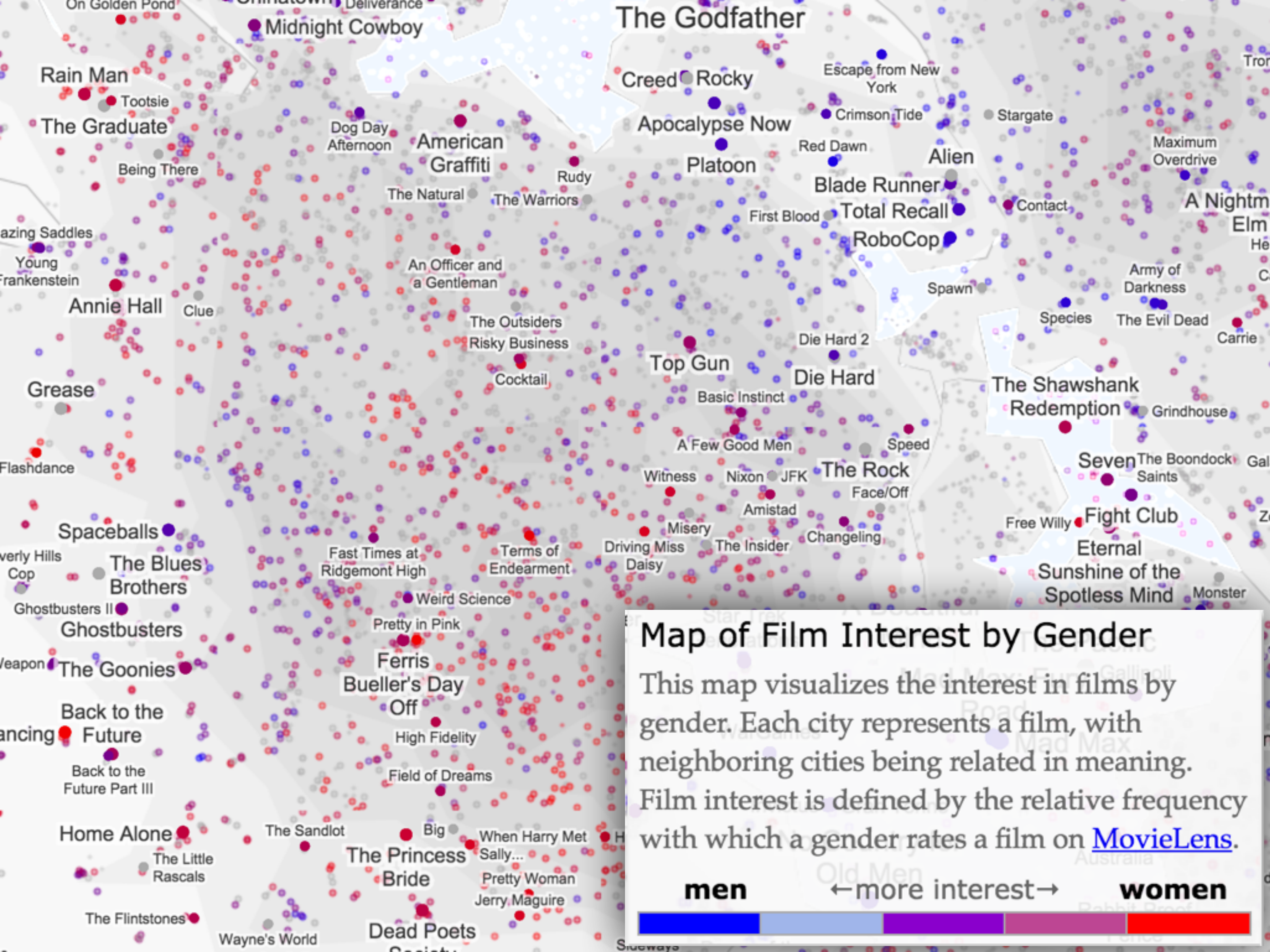
Spatialization in Cartograph



Title	Gender Score
Missing in Action 2: The Begir	0.08
Cross of Iron (1977)	0.08
Heartbreak Ridge (1986)	0.08
Bloodsport (1988)	0.09
Predator 2 (1990)	0.09
Tora! Tora! Tora! (1970)	0.09
Red Heat (1988)	0.09

+ Wikipedia =


and the Holy Grail
Rings: The Fellowship of the Ring
The Exorcist
A Clockwork Orange
Gandhi
300
The Lord of the Rings: The Return of the King
Braveheart
Resurrection of Jesus
Life Is Beautiful
Coronation Street
The King's Speech
Schindler's List
Big Brother
Airplane!
Dr. Strangelove
erman
Love Actually
Harry Potter and the Philosopher's Stone
The Fifth Element
2001: A Space Odyssey
Alice in Wonderland
Good Will Hunting
The Golden Compass
Easy Rider
The Silence of the Lambs
ns
A Beautiful Mind
Moulin Rouge!
Amélie
The Godfather
Chinatown
Taxi Driver
Rain Man
Mad Max: Fury Road
The Artist
Rocky
The Graduate
American Graffiti
Alien
Prometheus
Platoon
Blade Runner
RoboCop
Avatar
Motel
30 for 30
Titanic
Arab Americans
Annie Hall
Napoleon Dynamite
Everest
The G the D Tat
Grease
The Shawshank Redemption
Top Gun
Die Hard
An Inconvenient Truth
Avatar
Seven
Star Wars
Steal This Film
The Rock
Making a Murderer
The Internet's Own Boy
Fight Club
2012
Star Wars: The Force Awakens
The Blues Brothers
Spaceballs
The Breakfast Club
Ghostbusters
Ferris Bueller's Day Off
The Goonies
Babel
Dirty Dancing
Donnie Darko
The Dark Knight
Borat
Children of Men
The Revenant
Back to the Future
Crash
Boyhood
Pan's Labyrinth
The Princess Bride
M. Night Shyamalan
The Social Network
Spotlight
Slumdog Millionaire
Dead Poets Society
The Human Centipede
Deep Throat
The Hunger Games
Tomb Raider
The Pursuit of Happyness
Twilight
The Big Short
The Matrix
The Ring
Seven Samurai
Straight Outta Compton
Man of Steel
The Last Samurai
Pitch Perfect 2
Rashomon
Remember the Titans
Batman v Superman: Dawn of Justice
Furious 7
Kill Bill
Pirates of the Caribbean: On Stranger Tides
American Pie
Transformers
Spirited Away
The Blind Side
The Hangover
Pirates of the Caribbean: Dead Men Tell No Tales
Minions
Jurassic World
Code Geass
Pixels
Mean Girls
DreamWorks Animation
Tron
Men in Black
The Karate Kid
DC Universe Animated Original Movies
Godzilla
Thomas & Friends
Baahubali: The Beginning
Hasbro



Map of Film Interest by Gender

This map visualizes the interest in films by gender. Each city represents a film, with neighboring cities being related in meaning. Film interest is defined by the relative frequency with which a gender rates a film on MovieLens.

http://cartograph.info

Cartograph Innovations:

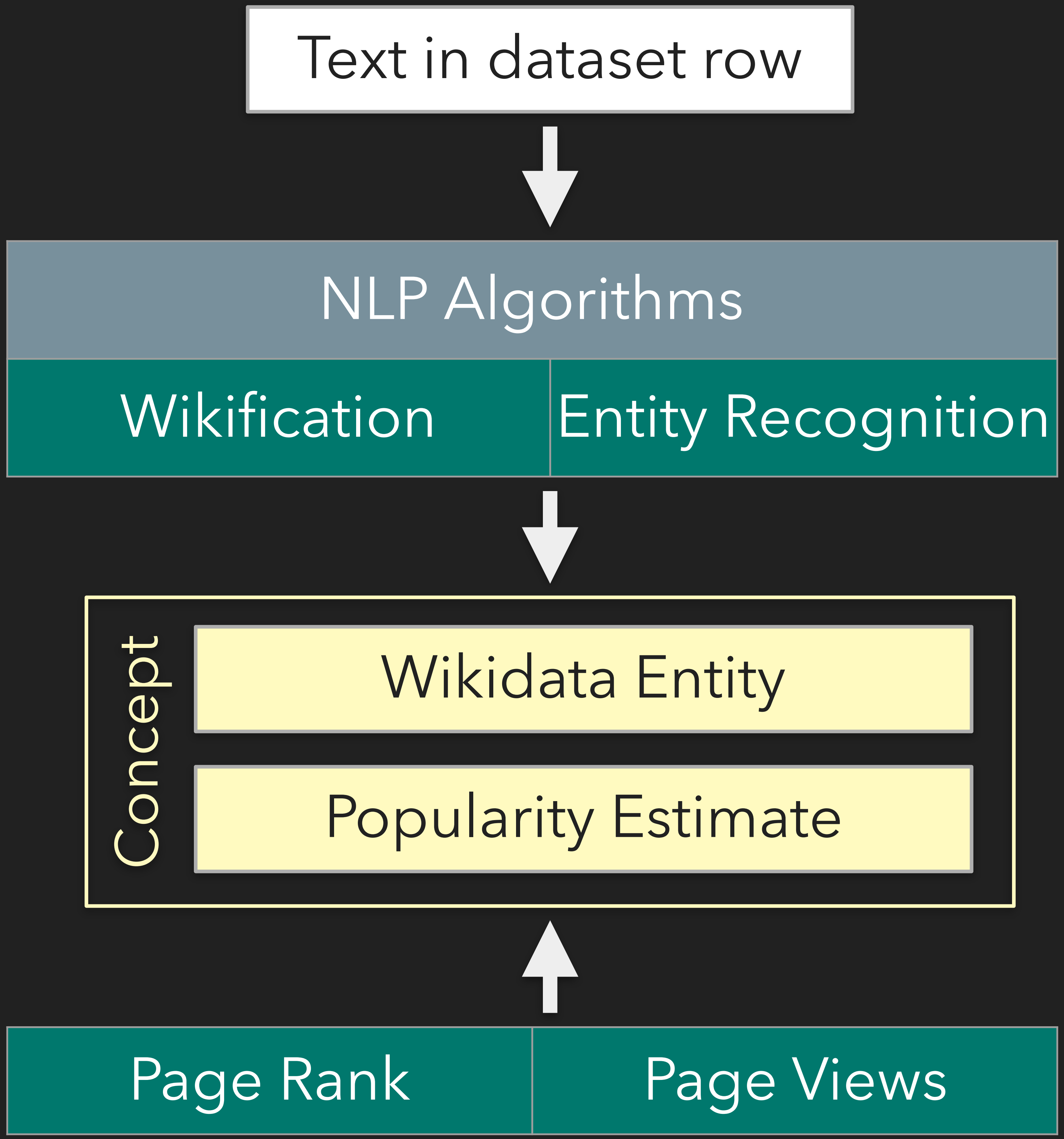
1. Taps vast world knowledge encoded in Wikipedia.
2. Leverages recent advances in NLP algorithms.
3. Maps delivered via cutting-edge web technologies.

Thematic cartography that is **scalable**, **interactive**, **applicable** to almost any dataset.

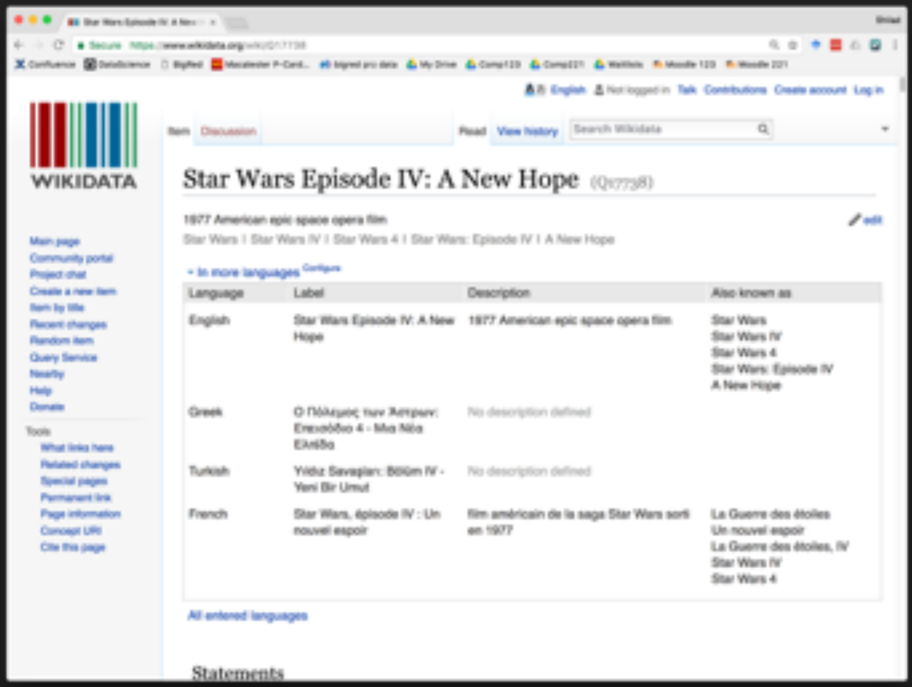
Cartograph Pipeline

1. Concept definition
2. X, Y embedding
3. Thematic layers
4. Map delivery

Step 1: Concept definition



"Star Wars"



Star Wars Episode IV: A New Hope (Q17738)

1977 American epic space opera film

Star Wars | Star Wars IV | Star Wars 4 | Star Wars: Episode IV | A New Hope

In more languages Configure

Language	Label	Description	Also known as
English	Star Wars Episode IV: A New Hope	1977 American epic space opera film	Star Wars Star Wars IV Star Wars 4 Star Wars: Episode IV A New Hope
Greek	Ο Πόλεμος των Άστρων: Επεισόδιο 4 - Μια Νέα Ελπίδα	No description defined	
Turkish	Yıldız Savaşları: Bölüm IV - Yeni Bir Umut	No description defined	
French	Star Wars, épisode IV : Un nouvel espoir	film américain de la saga Star Wars sorti en 1977	La Guerre des étoiles Un nouvel espoir La Guerre des étoiles, IV Star Wars IV Star Wars 4

All entered languages

Statements

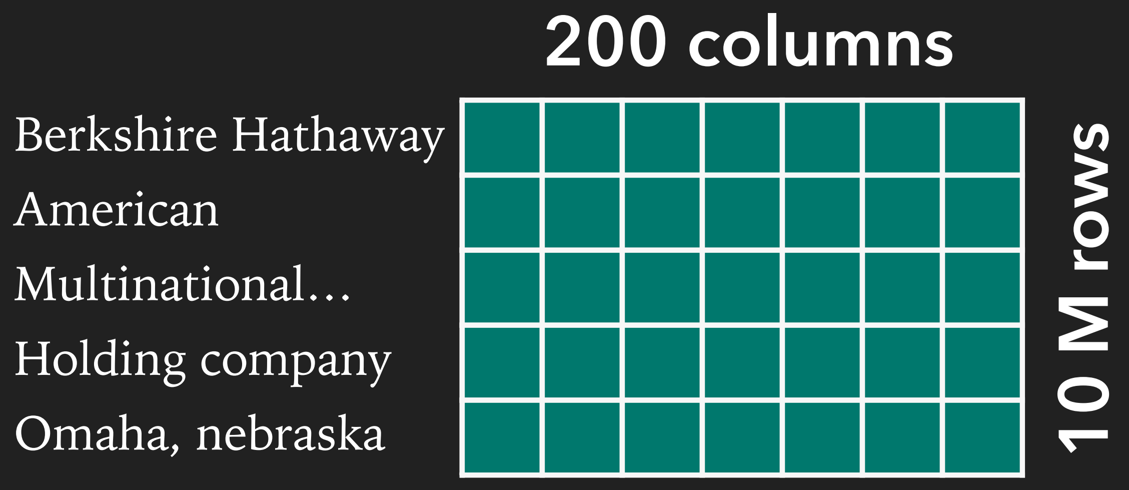
article in 69 langs

Importance: #5 of 72K

Step 2: X, Y embedding

Berkshire Hathaway Inc. is an American multinational conglomerate holding company headquartered in Omaha, Nebraska...

Word2Vec



200 columns

Berkshire Hathaway

American

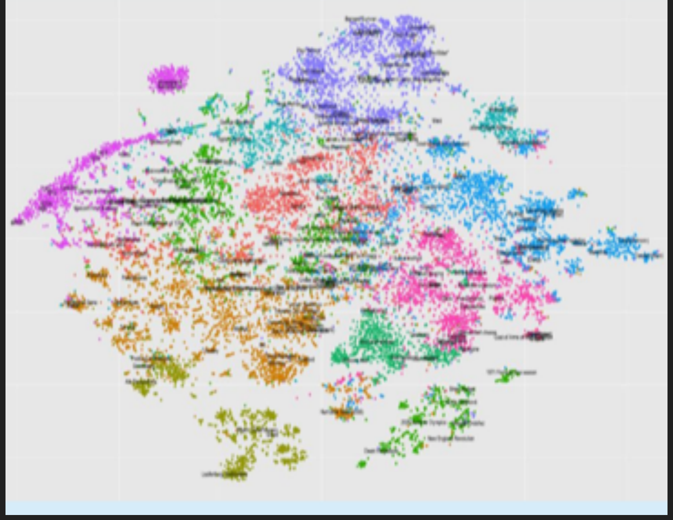
Multinational...

Holding company

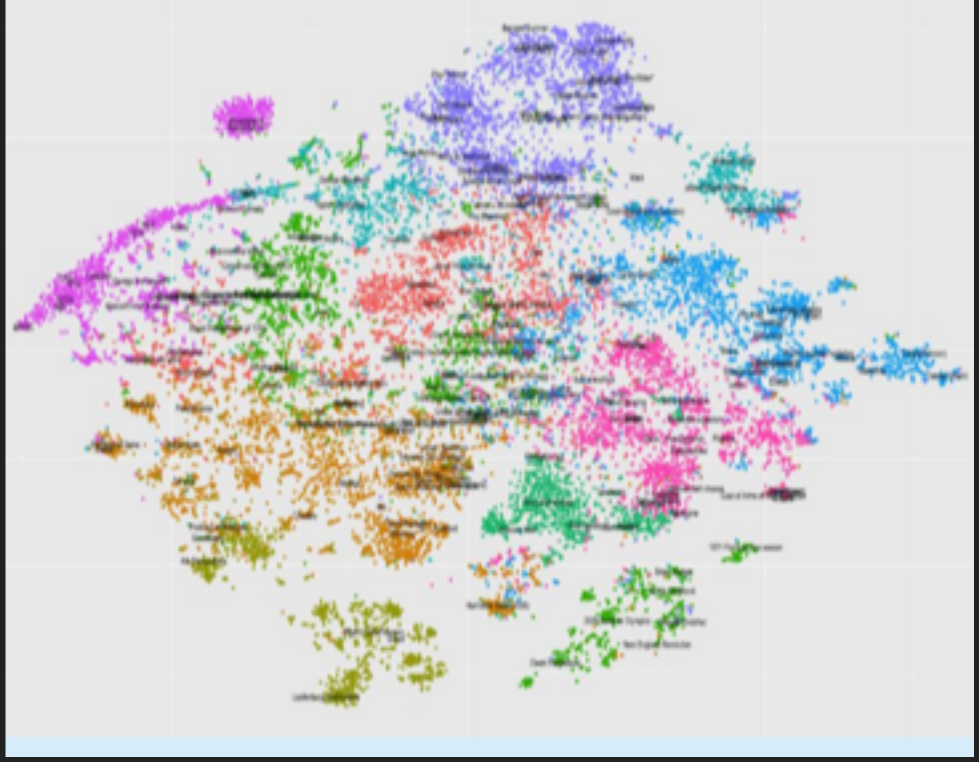
Omaha, nebraska

10 M rows

t-SNE on 50K sample points

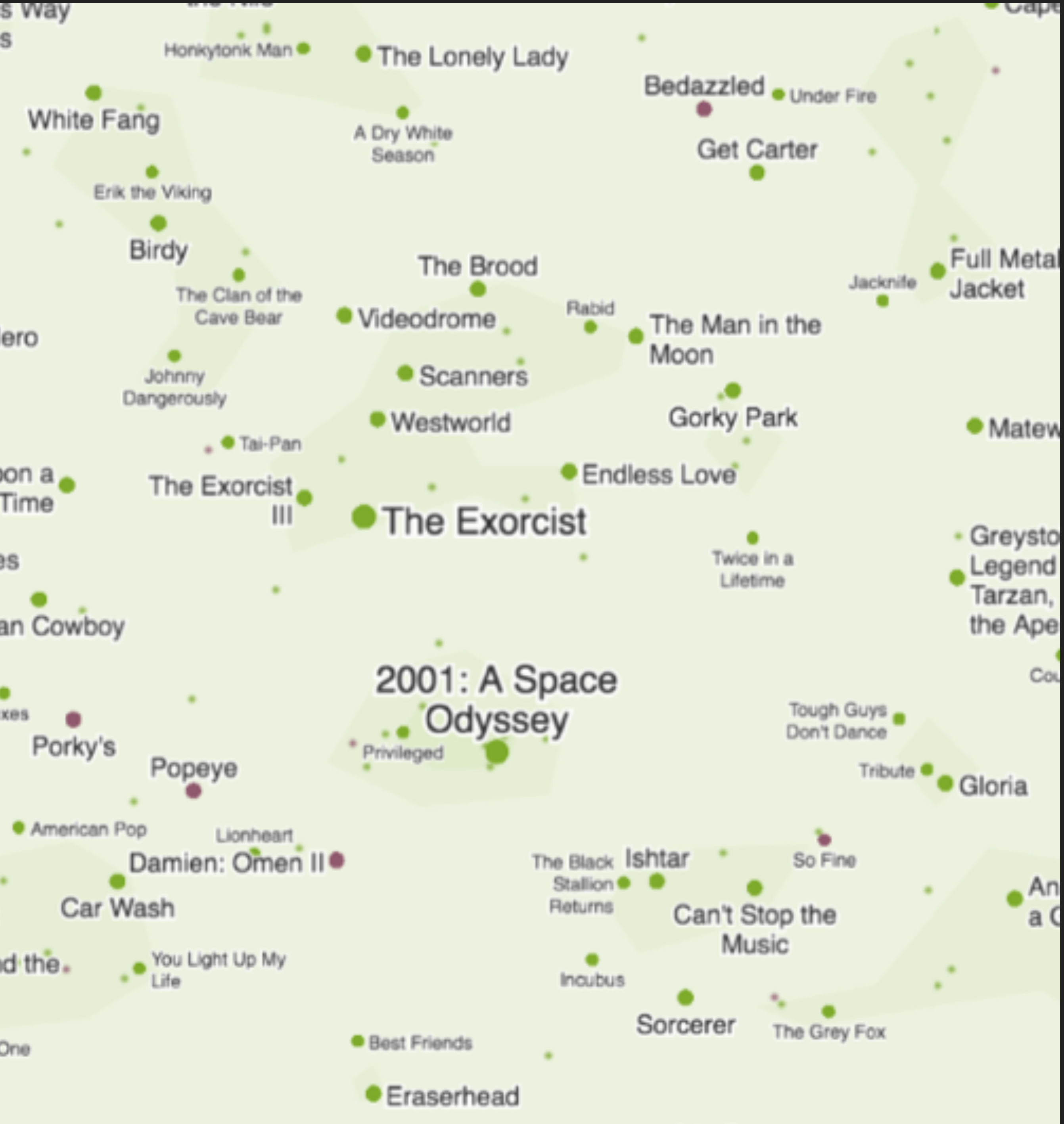


interpolation for out of sample points

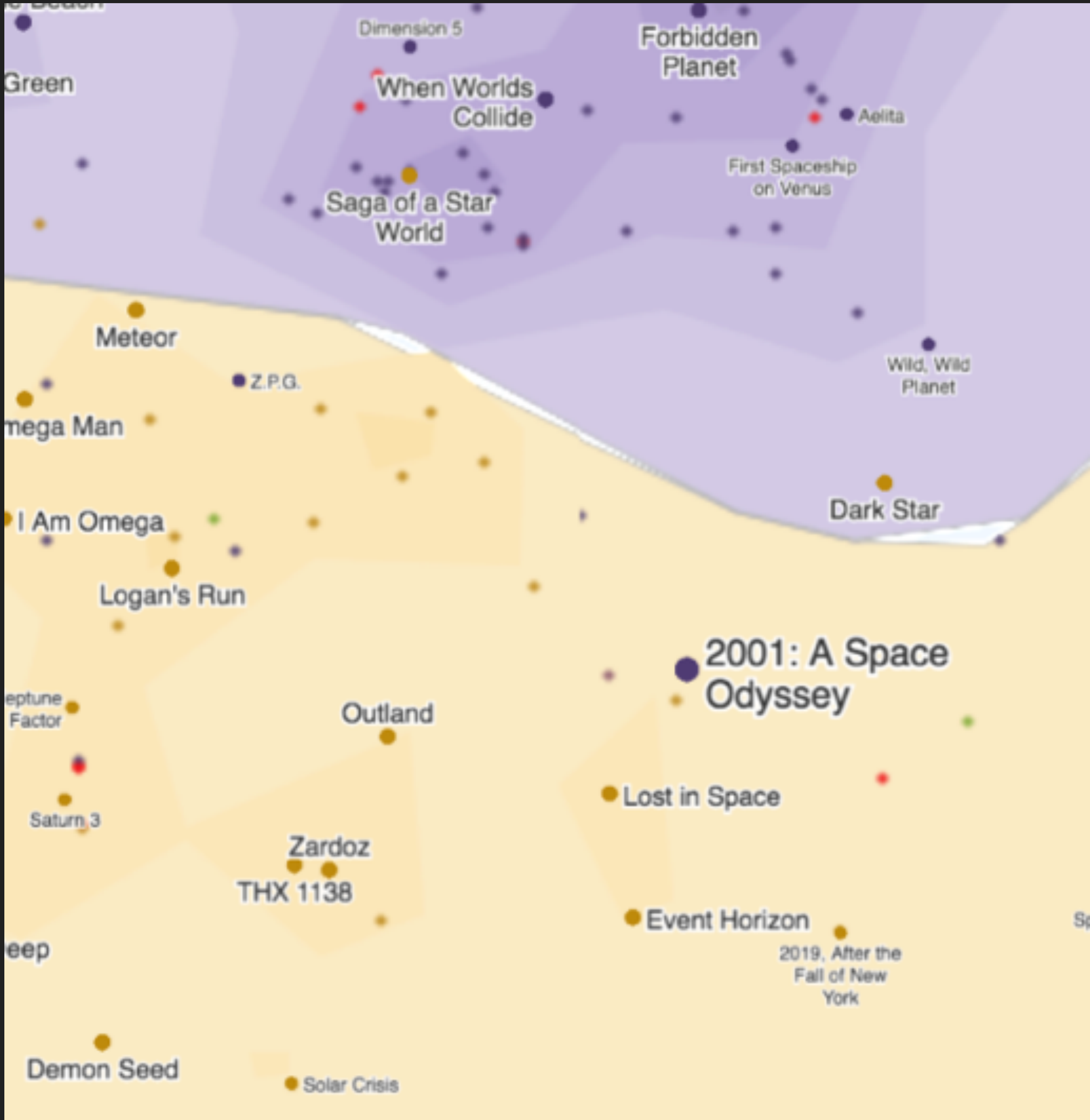


Step 2a: A Word2Vec Sentence is...

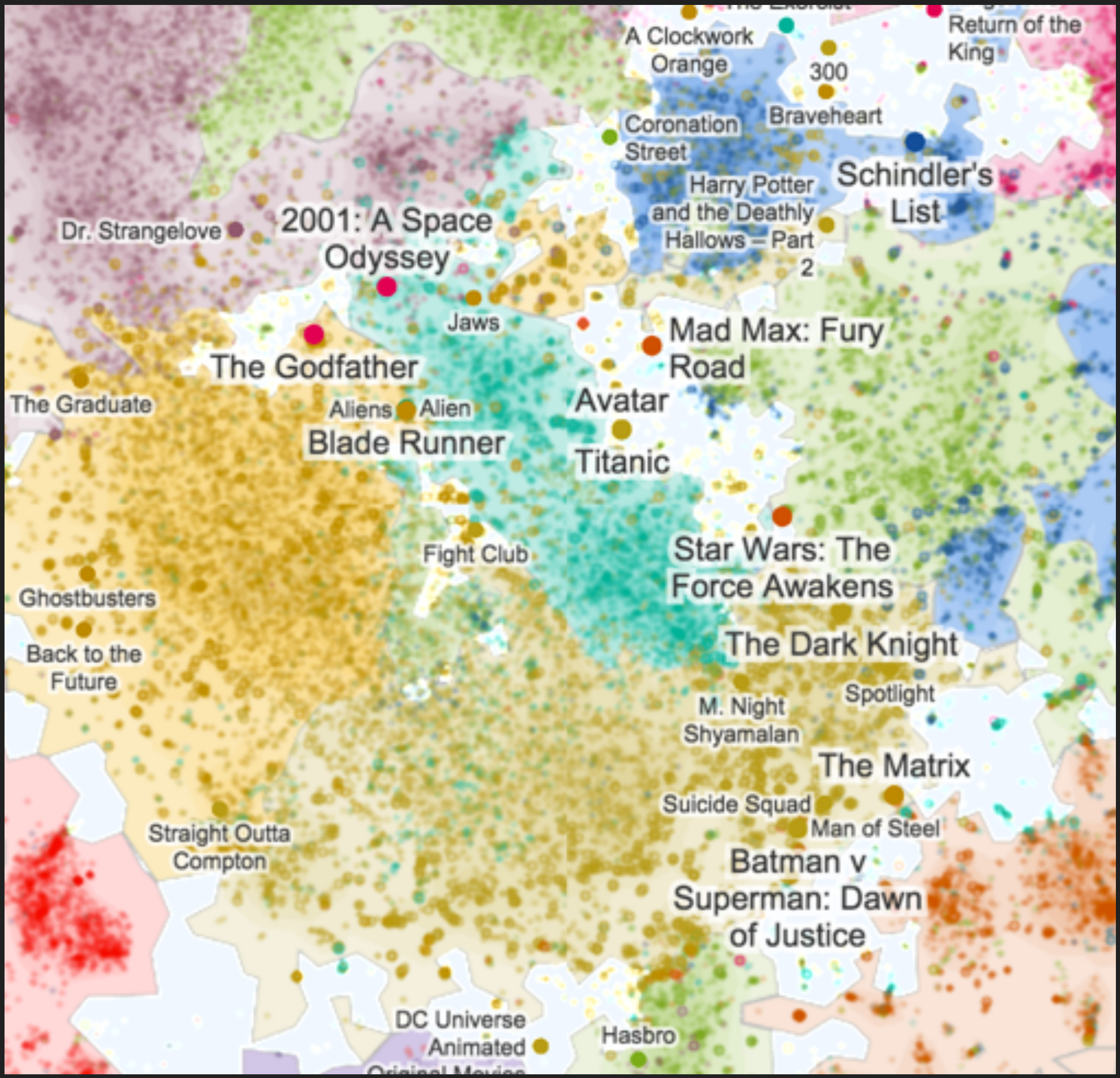
Wikipedia article text



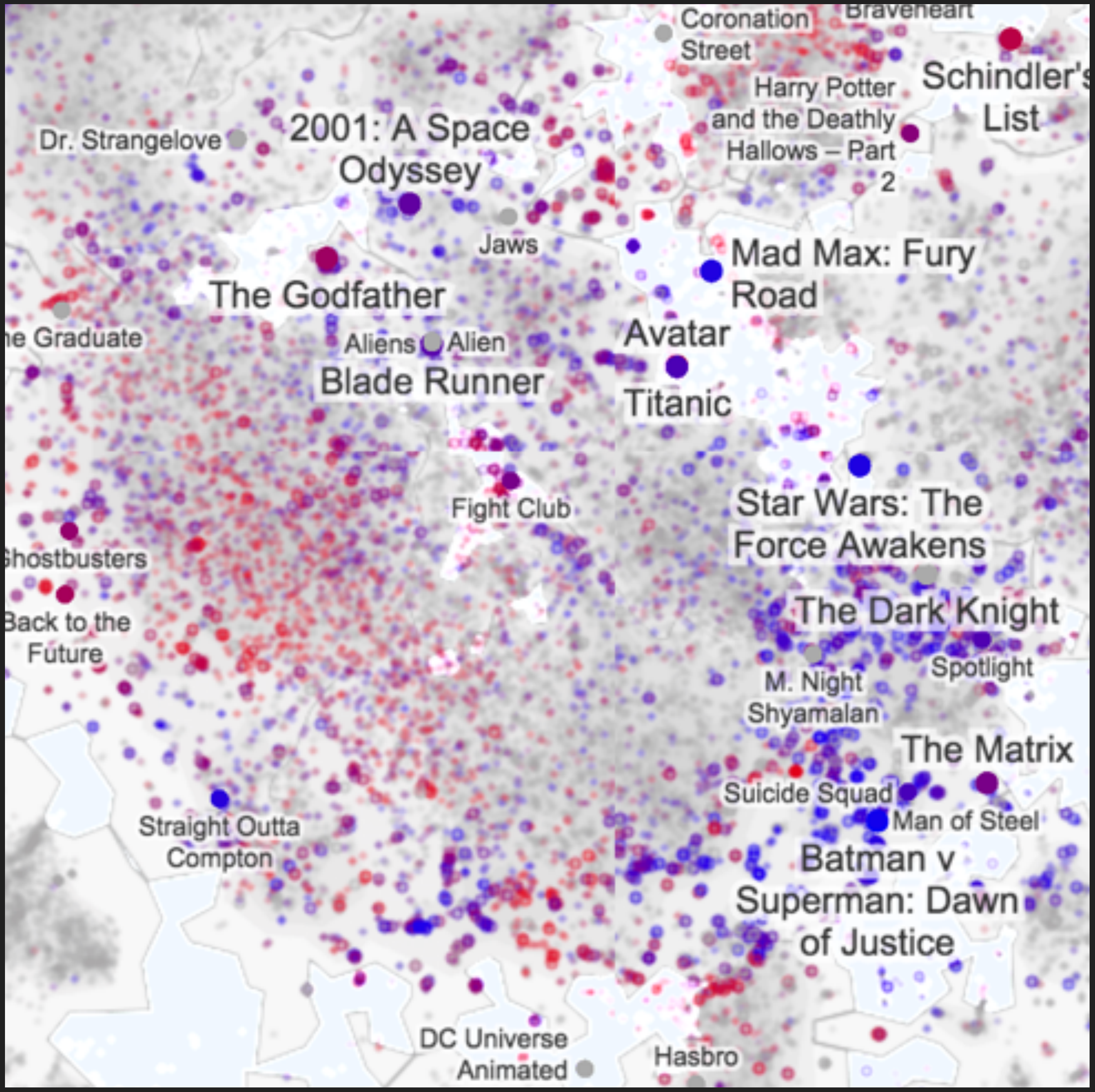
User session article clicks



Step 3: Thematic Layers



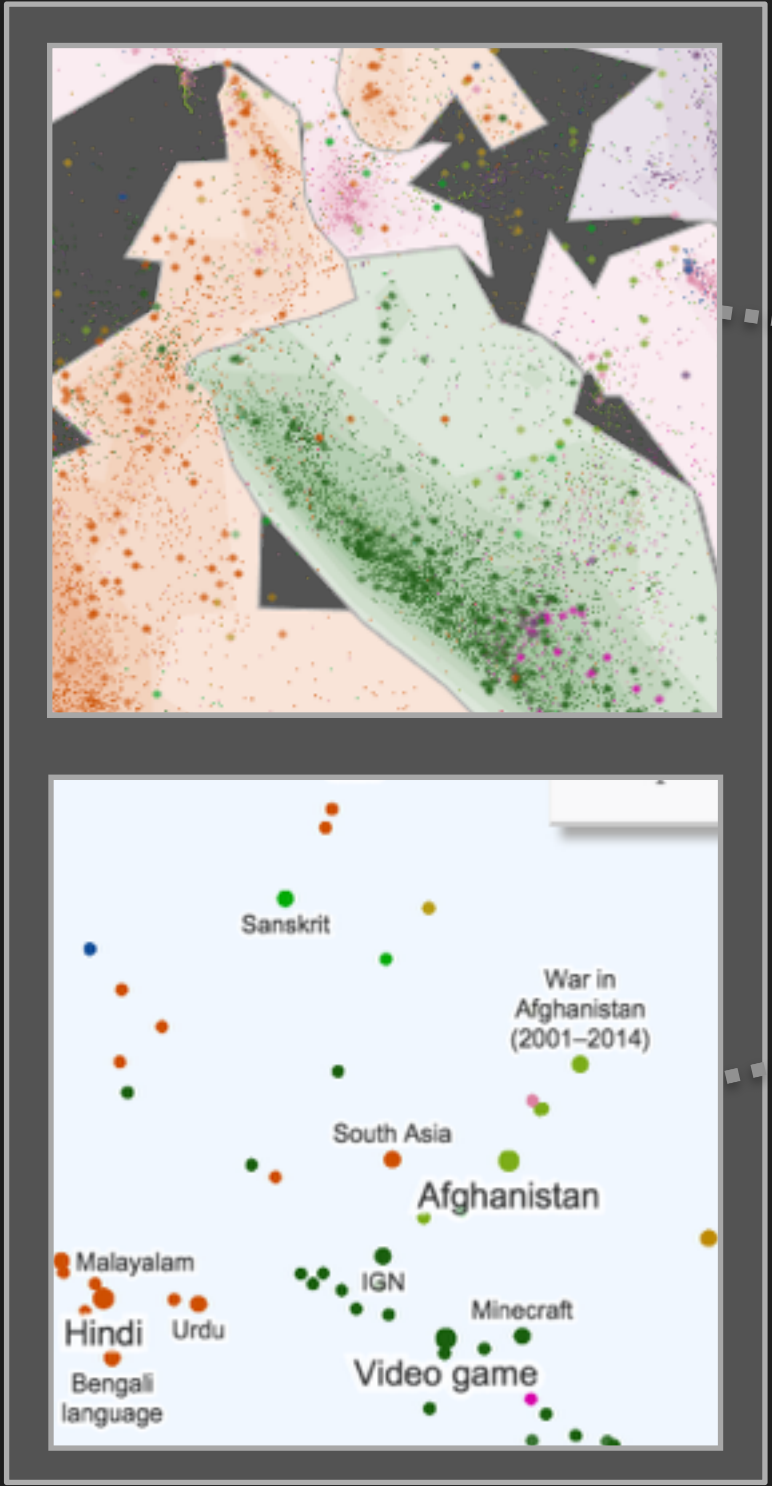
Thematic Cluster
(Categorical Layer)



Relative Gender Interest
(Proportional Layer)

Step 4: Web Delivery

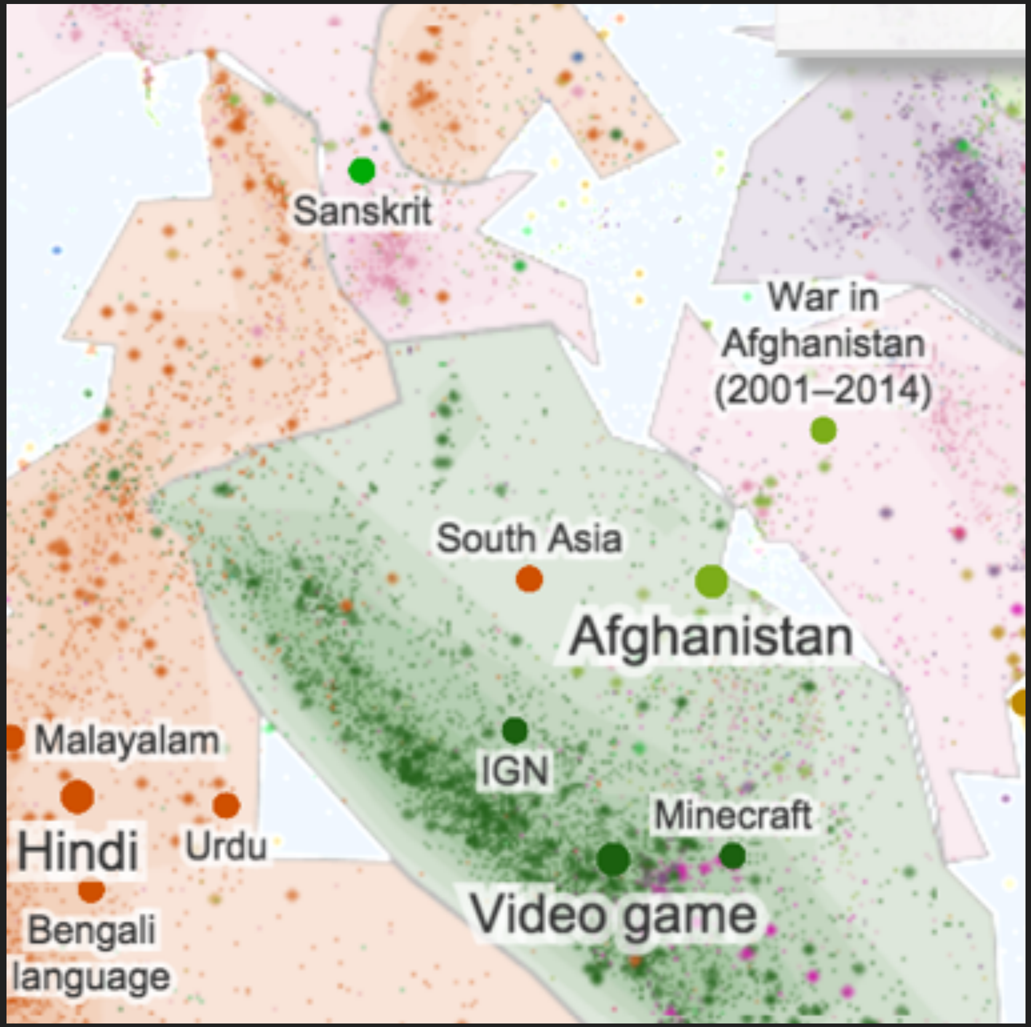
Custom Tile Server



Raster Background Tiles

Vector Foreground Data Tiles

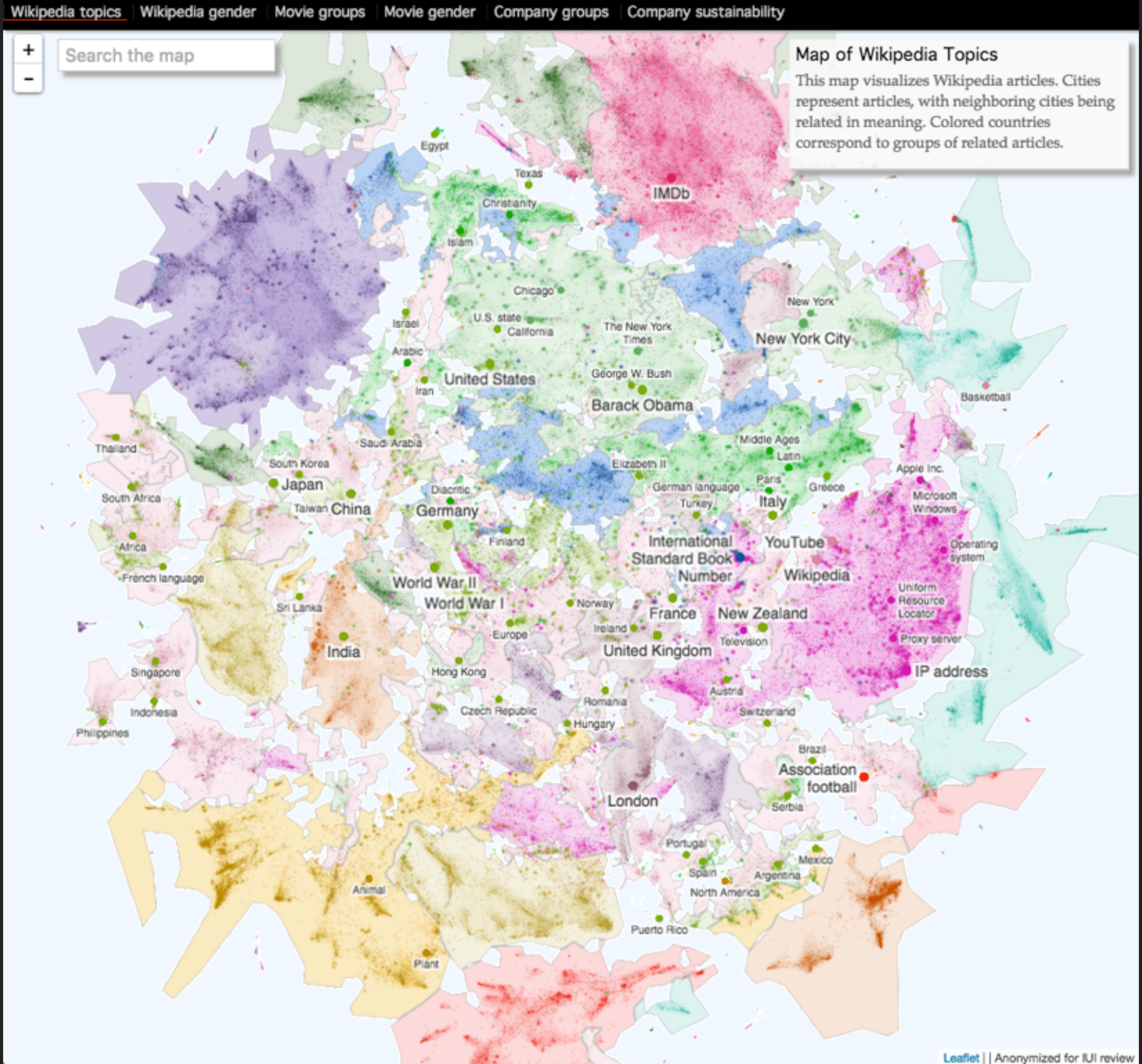
Browser (WebGL)



Case Studies

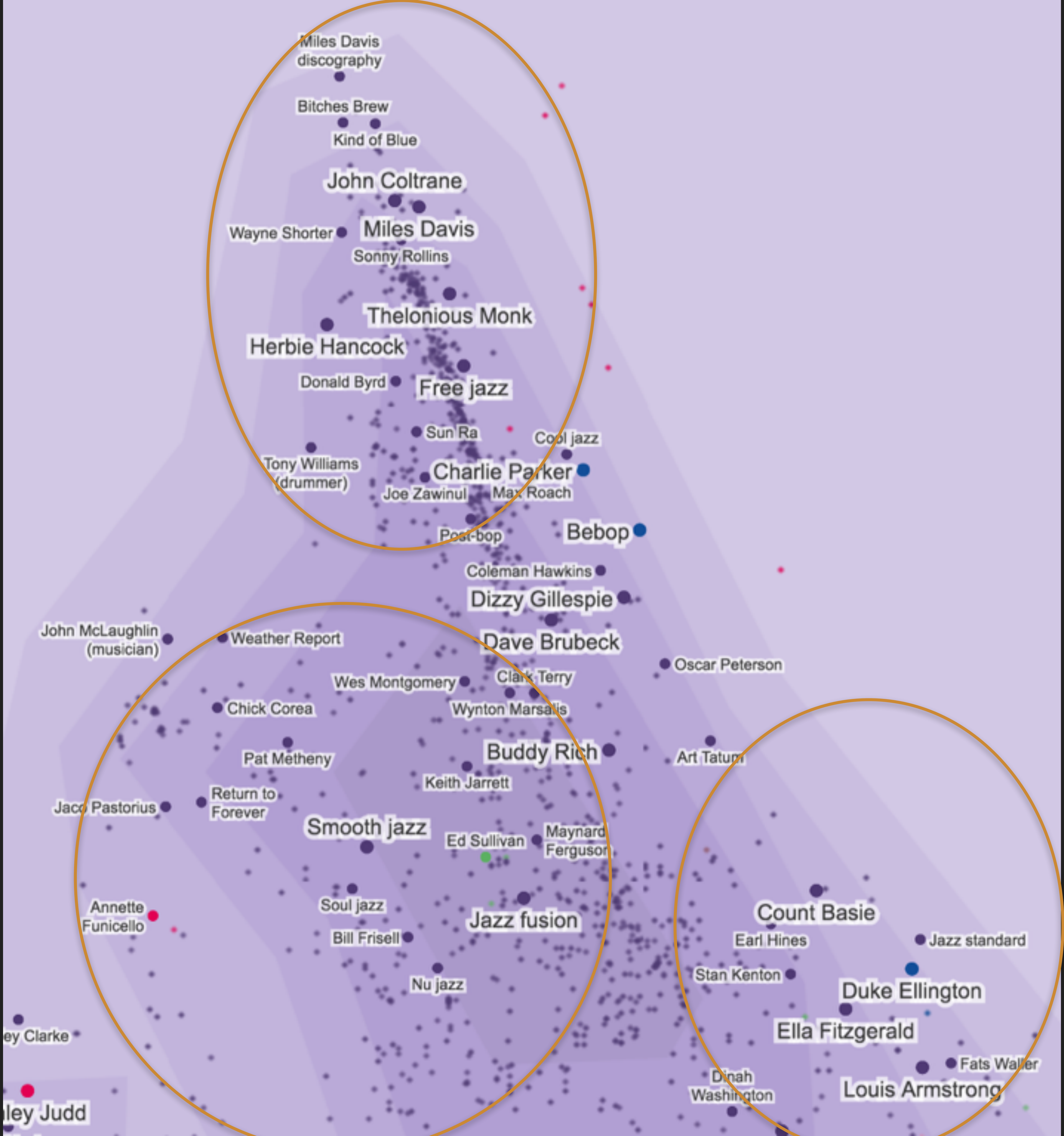
Case Study: Map of Wikipedia

1.4 million Wikipedia articles with sufficient page views



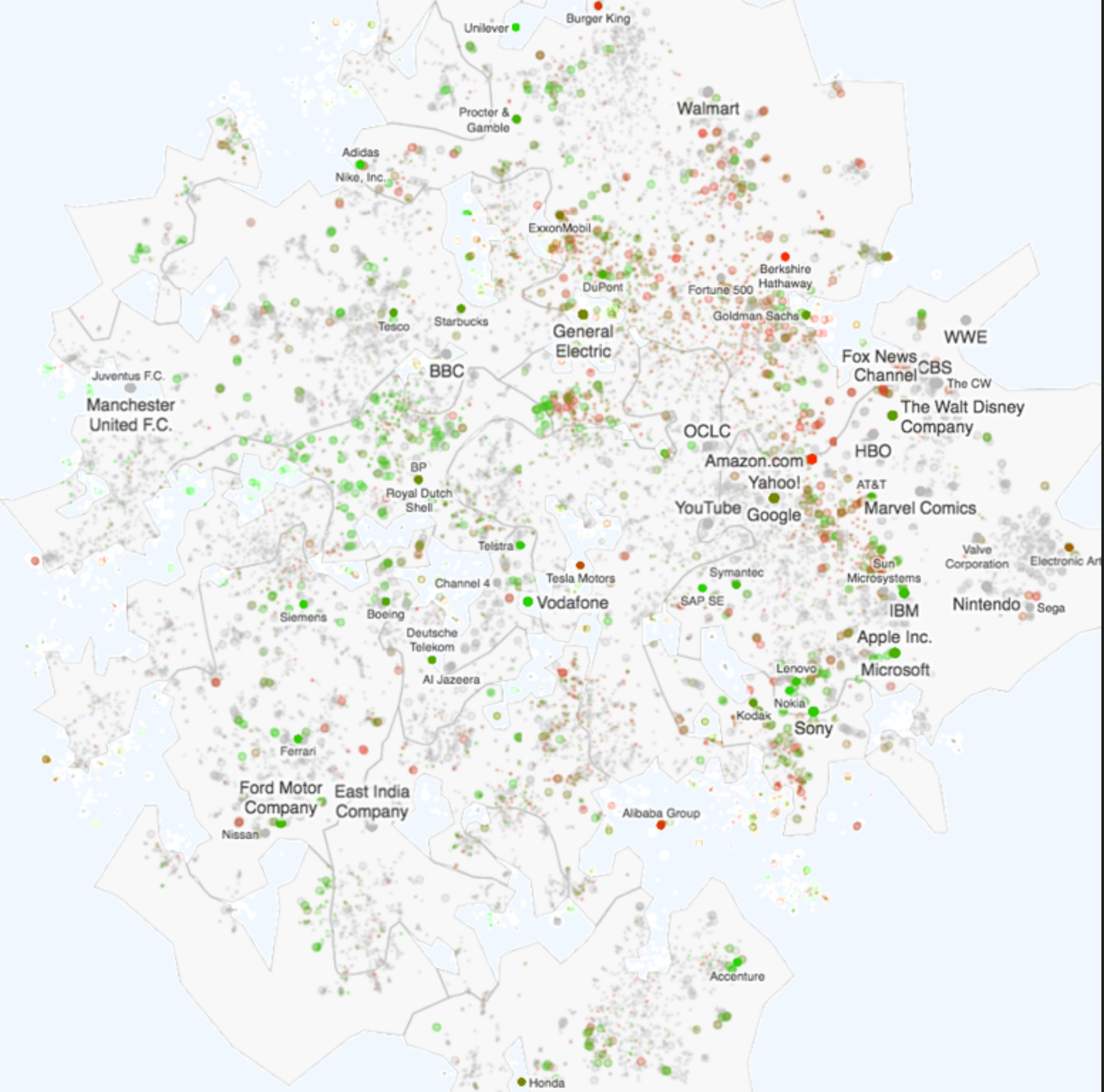
Map of Wikipedia Topics

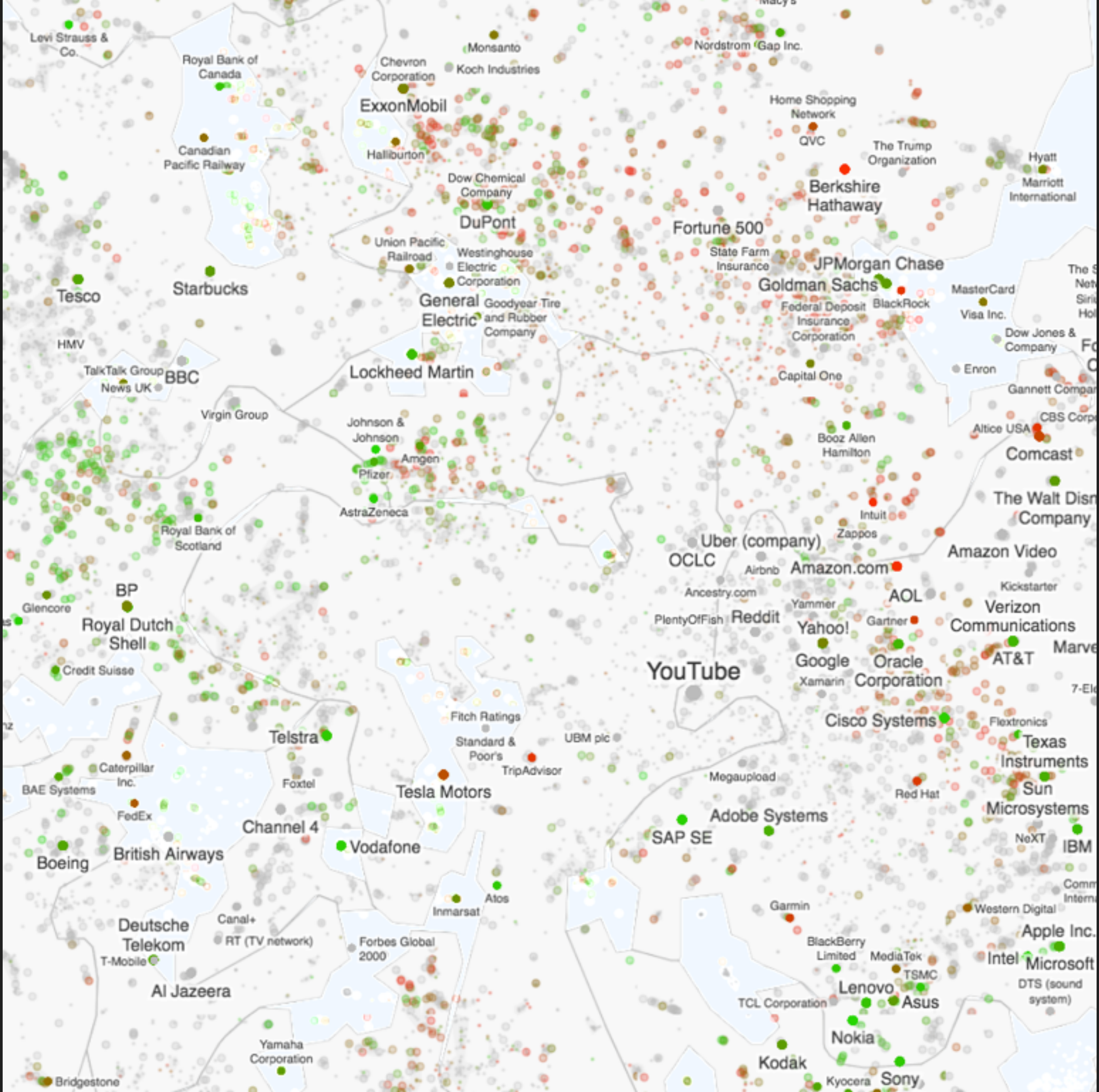
This map visualizes Wikipedia articles. Cities represent articles, with neighboring cities being related in meaning. Colored countries correspond to groups of related articles.


Miles Davis discography
Bitches Brew
Kind of Blue
John Coltrane
Wayne Shorter
Miles Davis
Sonny Rollins
Thelonious Monk
Herbie Hancock
Donald Byrd
Free jazz
Sun Ra
Cool jazz
Tony Williams (drummer)
Charlie Parker
Joe Zawinul
Max Roach
Post-bop
Bebop
Coleman Hawkins
Dizzy Gillespie
John McLaughlin (musician)
Weather Report
Dave Brubeck
Oscar Peterson
Wes Montgomery
Clark Terry
Chick Corea
Wynton Marsalis
Pat Metheny
Buddy Rich
Art Tatum
Keith Jarrett
Jaco Pastorius
Return to Forever
Smooth jazz
Ed Sullivan
Maynard Ferguson
Annette Funicello
Soul jazz
Jazz fusion
Count Basie
Bill Frisell
Earl Hines
Jazz standard
Stan Kenton
Nu jazz
Duke Ellington
ey Clarke
Ella Fitzgerald
Fats Waller
Dinah Washington
Louis Armstrong
ley Judd

Case Study: Corporate Sustainability

Data from https://www.csrhub.com/


Burger King
Unilever
Procter & Gamble
Walmart
Adidas
Nike, Inc.
ExxonMobil
Berkshire Hathaway
DuPont
Fortune 500
Goldman Sachs
Tesco
Starbucks
General Electric
WWE
Fox News Channel
CBS
The CW
Juventus F.C.
BBC
Manchester United F.C.
The Walt Disney Company
OCLC
HBO
Amazon.com
BP
Yahoo!
AT&T
Royal Dutch Shell
YouTube
Google
Marvel Comics
Telstra
Valve Corporation
Electronic Ar
Sun Microsystems
Tesla Motors
Symantec
Channel 4
Vodafone
SAP SE
Nintendo
Sega
IBM
Siemens
Boeing
Deutsche Telekom
Apple Inc.
Microsoft
Lenovo
Al Jazeera
Nokia
Kodak
Sony
Ferrari
Ford Motor Company
East India Company
Alibaba Group
Nissan
Accenture
Honda

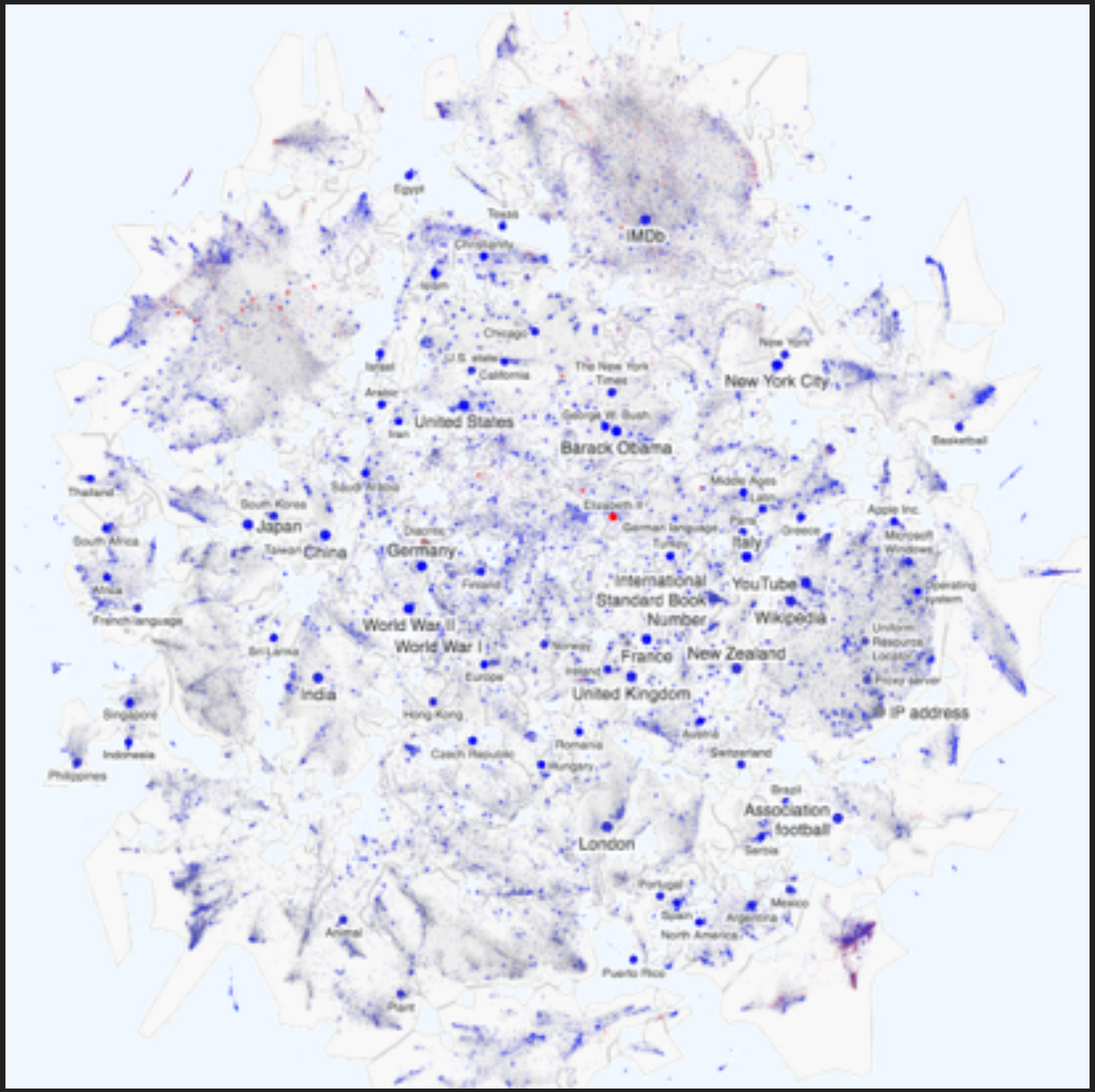

Macy's
Levi Strauss & Co.
Royal Bank of Canada
Chevron Corporation
Monsanto
Koch Industries
Nordstrom
Gap Inc.
ExxonMobil
Home Shopping Network
QVC
Canadian Pacific Railway
Halliburton
The Trump Organization
Hyatt
Dow Chemical Company
Berkshire Hathaway
Marriott International
DuPont
Fortune 500
Union Pacific Railroad
Westinghouse Electric Corporation
State Farm Insurance
JPMorgan Chase
Tesco
Starbucks
Goldman Sachs
MasterCard
General Electric
Goodyear Tire and Rubber Company
Federal Deposit Insurance Corporation
BlackRock
Visa Inc.
Dow Jones & Company
HMV
TalkTalk Group
News UK
BBC
Lockheed Martin
Capital One
Enron
Gannett Compa
CBS Corp
Virgin Group
Johnson & Johnson
Altice USA
Booz Allen Hamilton
Comcast
Amgen
Pfizer
AstraZeneca
Intuit
The Walt Disn Company
Royal Bank of Scotland
Zappos
Uber (company)
Amazon Video
OCLC
Airbnb
Amazon.com
Kickstarter
Ancestry.com
AOL
BP
Glencore
Yammer
Verizon Communications
Royal Dutch Shell
PlentyOfFish
Reddit
Yahoo!
Gartner
Marve
YouTube
Google
Oracle Corporation
AT&T
Credit Suisse
Xamarin
7-Ele
Fitch Ratings
Cisco Systems
Flextronics
Telstra
Standard & Poor's
UBM plc
Texas Instruments
Caterpillar Inc.
TripAdvisor
BAE Systems
Foxtel
Tesla Motors
Megaupload
Red Hat
Sun Microsystems
FedEx
Adobe Systems
Channel 4
SAP SE
NeXT
IBM
Boeing
British Airways
Vodafone
Comm Interna
Atos
Inmarsat
Garmin
Western Digital
Canal+
Deutsche Telekom
RT (TV network)
Forbes Global 2000
BlackBerry Limited
Apple Inc.
MediaTek
Intel
Microsoft
T-Mobile
TSMC
Lenovo
DTS (sound system)
Al Jazeera
TCL Corporation
Asus
Nokia
Yamaha Corporation
Kodak
Bridgestone
Kyocera
Sony
The S Netw Siri Hol
Fo C

Exploratory Study: Feedback from Users

Goals

Learn how domain experts interpret the cartographic embedding

Learn how to design better spatialization tools to support exploratory analytical tasks

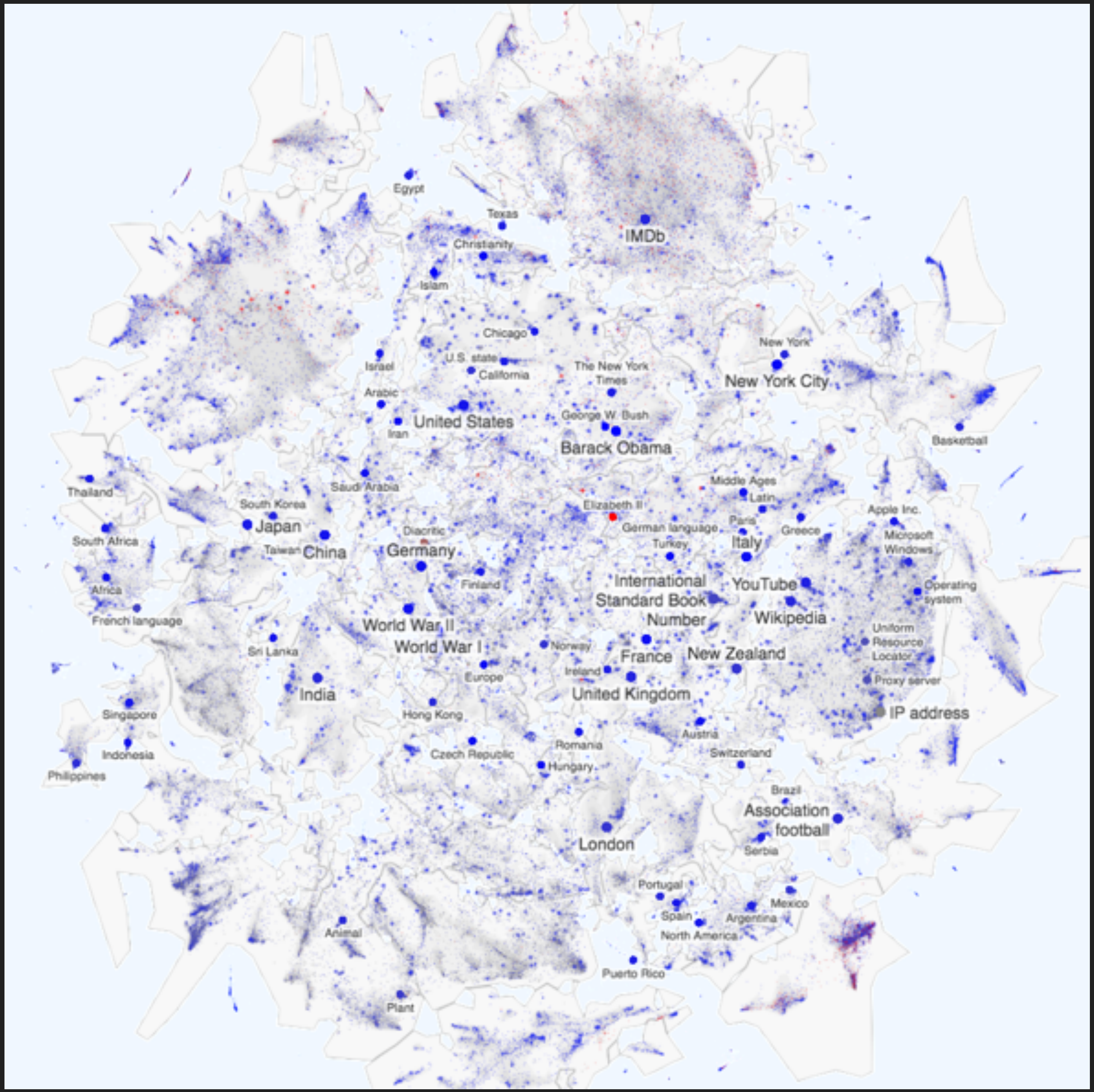


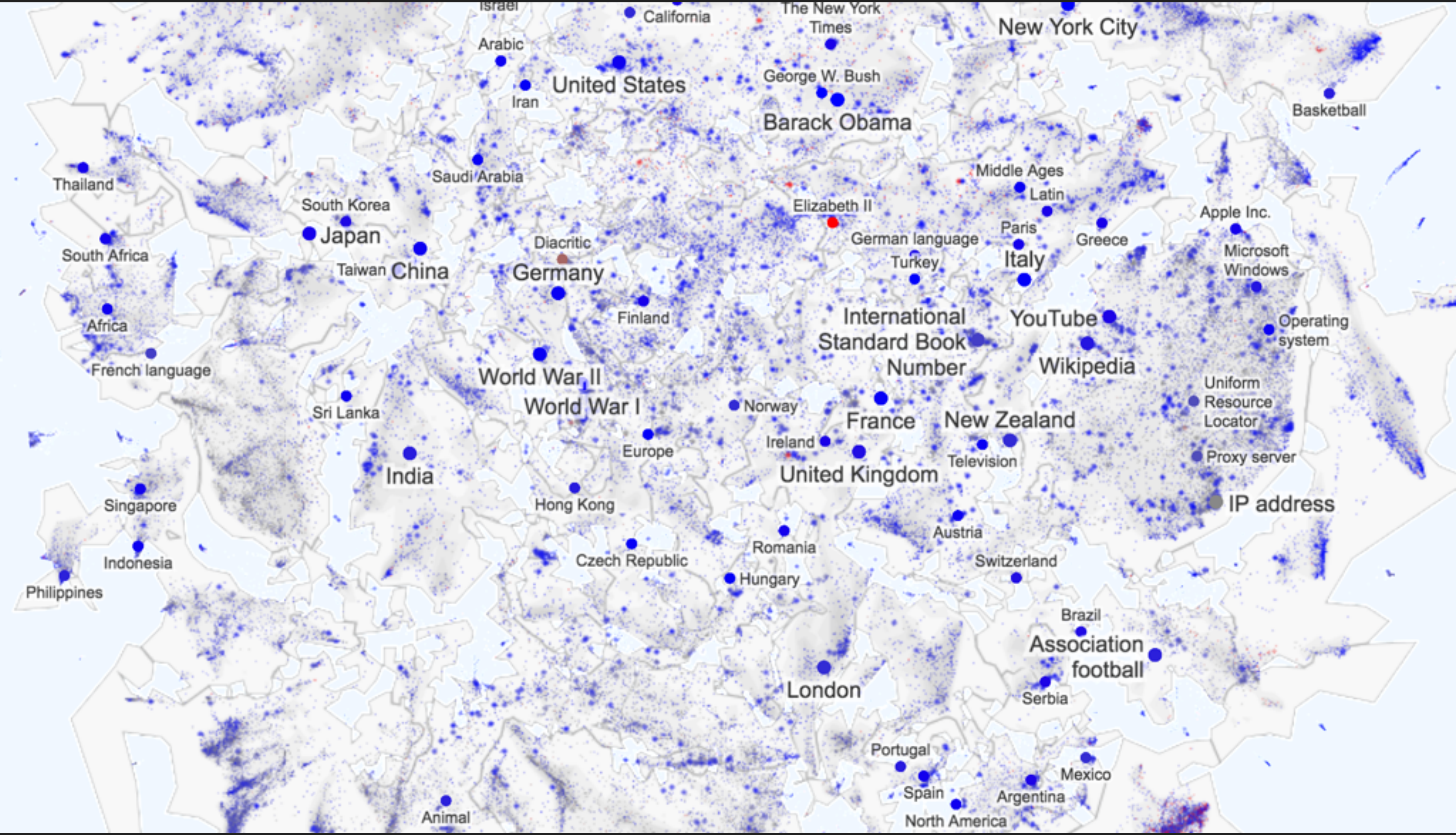
Dataset

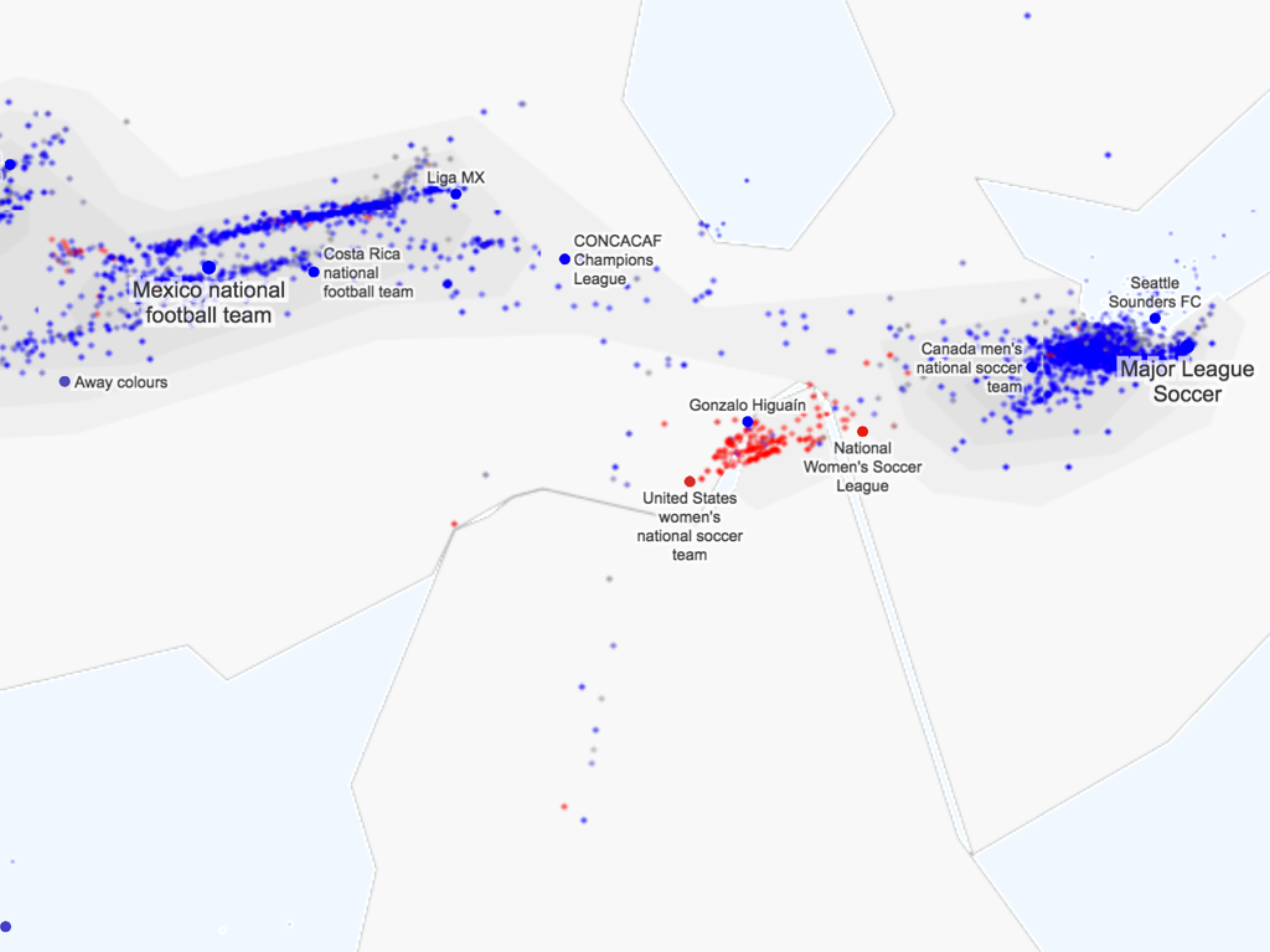
Gender focus of articles

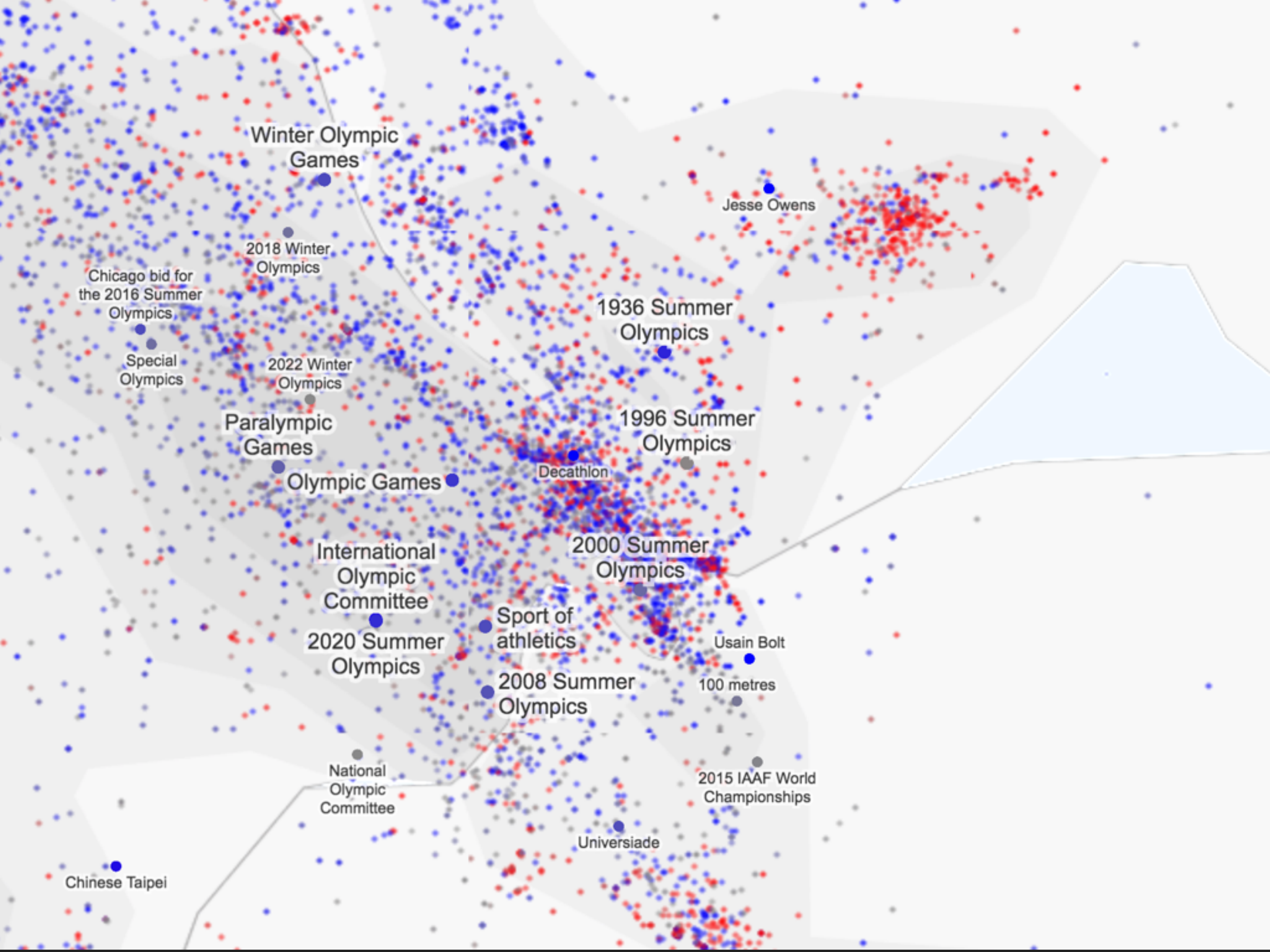
Ratio of links to men articles / women articles

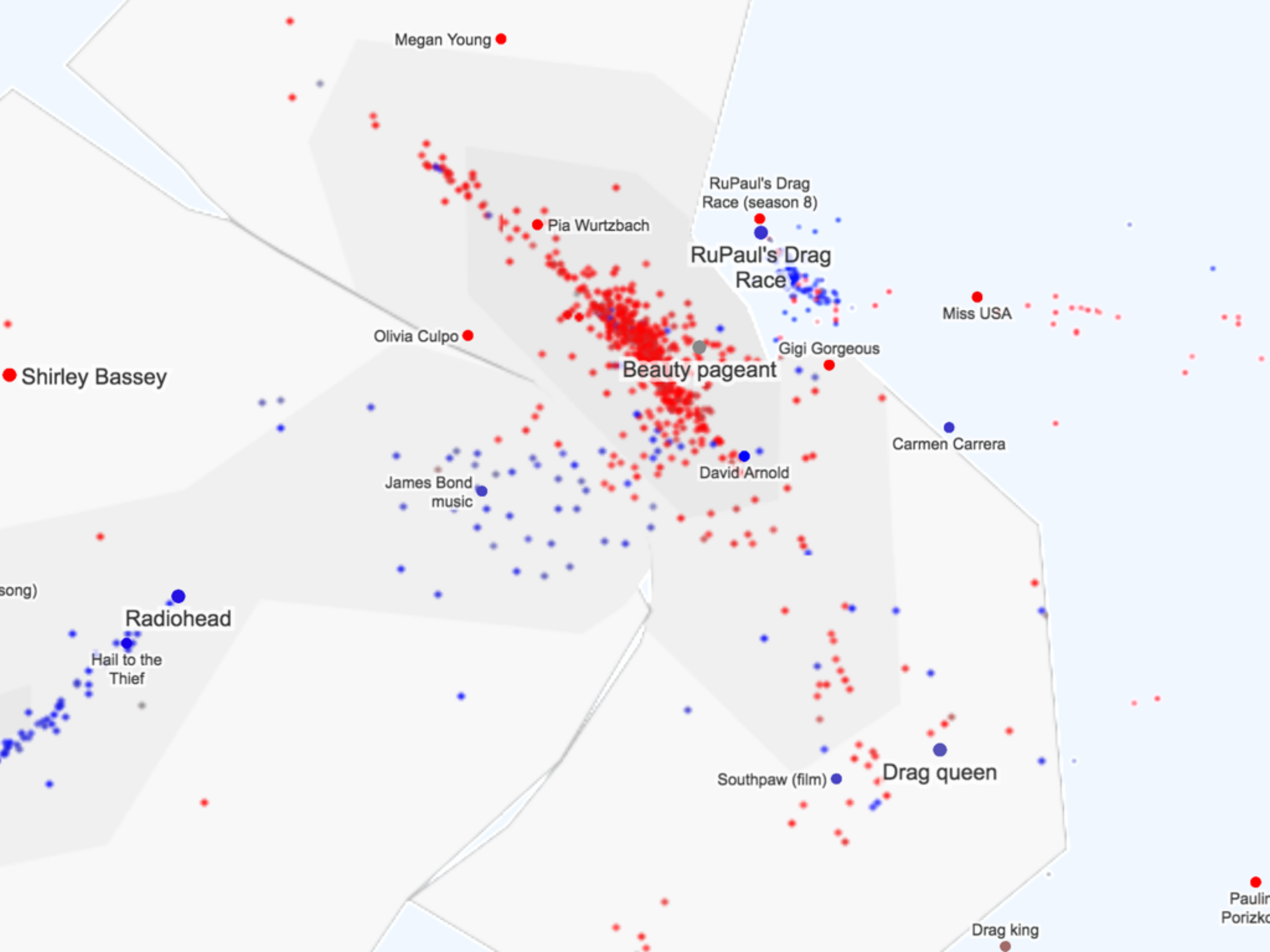
Gender of article came from WikiData

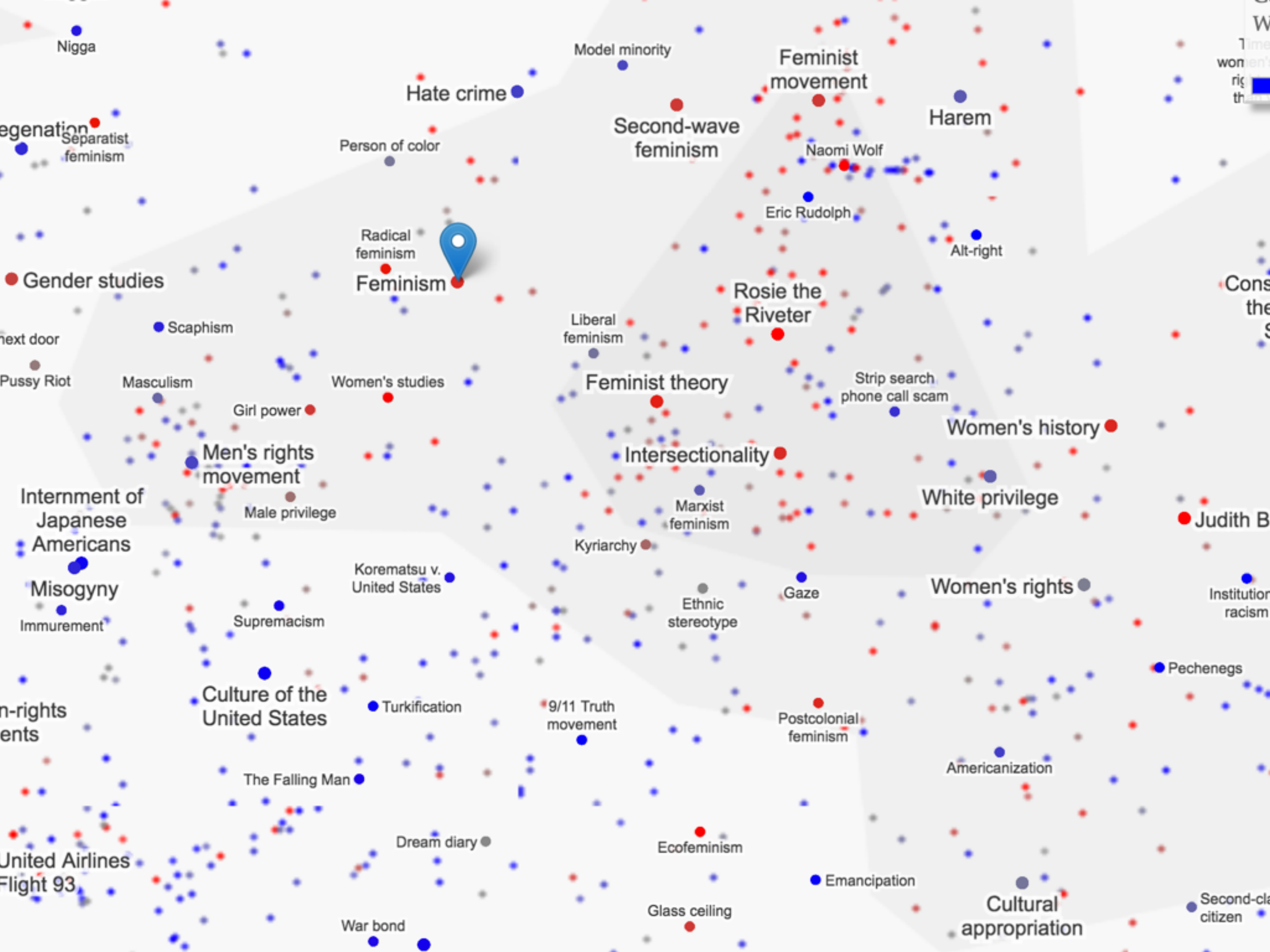


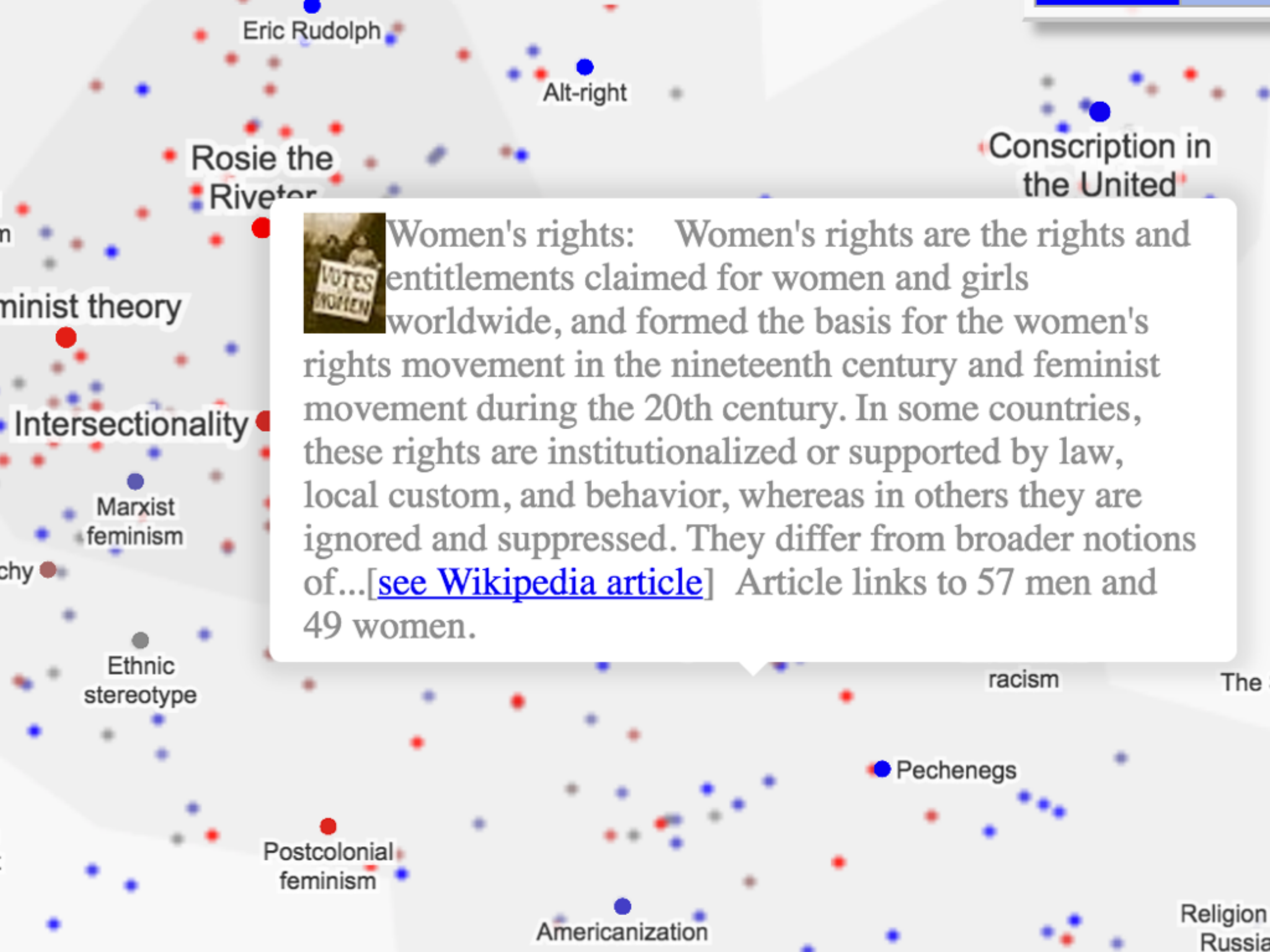

Israel
California
The New York Times
New York City
Arabic
United States
George W. Bush
Iran
Barack Obama
Basketball
Thailand
Saudi Arabia
Middle Ages
Latin
Elizabeth II
South Korea
Apple Inc.
Japan
Paris
German language
Greece
South Africa
Diacritic
Microsoft Windows
Taiwan
China
Turkey
Italy
Germany
Africa
Finland
International Standard Book Number
YouTube
Operating system
French language
Wikipedia
World War II
Uniform Resource Locator
World War I
Sri Lanka
Norway
France
New Zealand
Europe
Ireland
Television
Proxy server
India
United Kingdom
Singapore
Hong Kong
IP address
Austria
Romania
Indonesia
Czech Republic
Switzerland
Philippines
Hungary
Brazil
Association football
Serbia
London
Portugal
Mexico
Spain
Argentina
Animal
North America


Liga MX
CONCACAF Champions League
Costa Rica national football team
Mexico national football team
Away colours
Seattle Sounders FC
Canada men's national soccer team
Major League Soccer
Gonzalo Higuaín
National Women's Soccer League
United States women's national soccer team

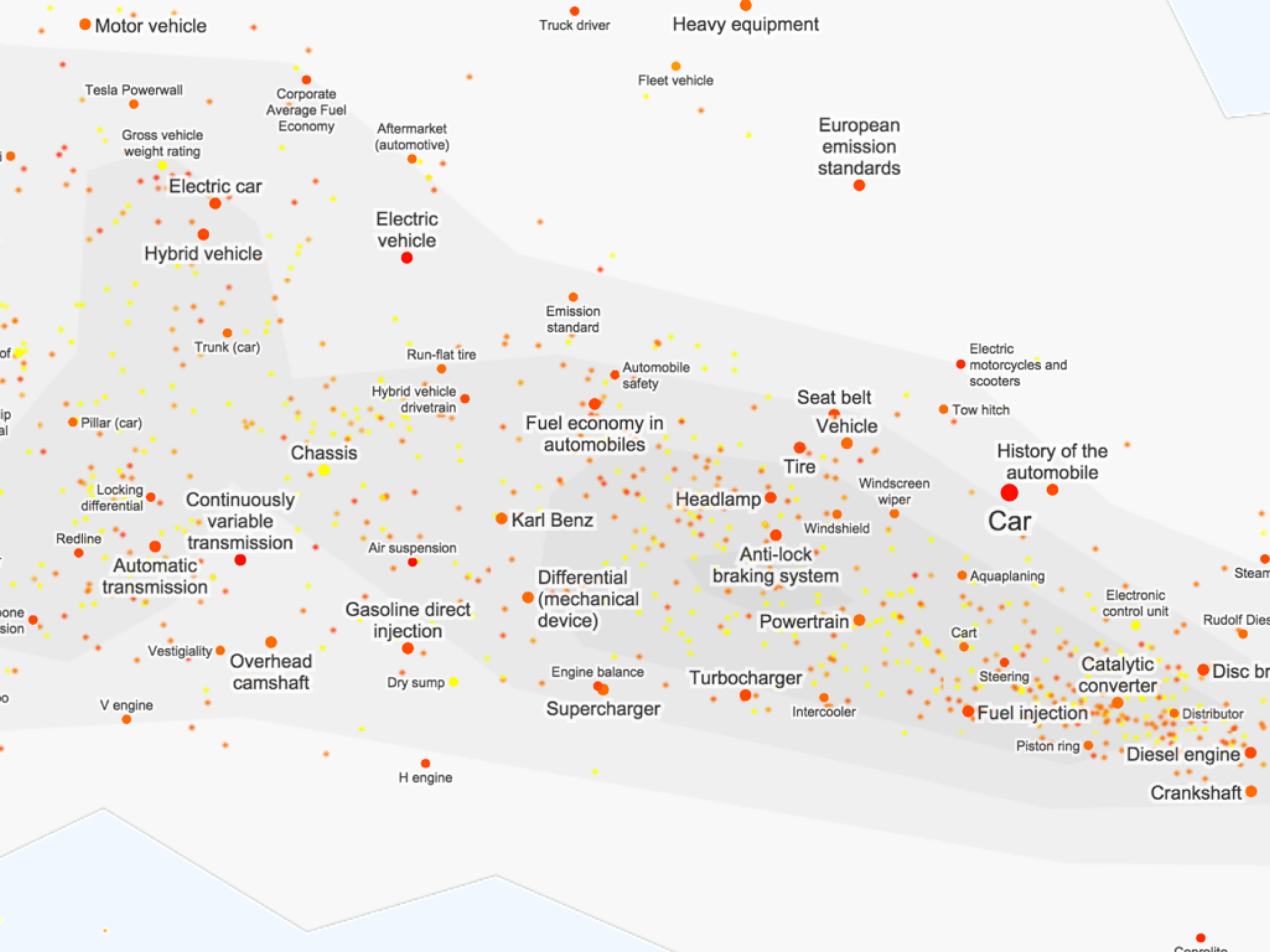

Winter Olympic Games
Jesse Owens
2018 Winter Olympics
Chicago bid for the 2016 Summer Olympics
1936 Summer Olympics
Special Olympics
2022 Winter Olympics
Paralympic Games
1996 Summer Olympics
Decathlon
Olympic Games
2000 Summer Olympics
International Olympic Committee
Sport of athletics
Usain Bolt
2020 Summer Olympics
2008 Summer Olympics
100 metres
National Olympic Committee
2015 IAAF World Championships
Universiade
Chinese Taipei


Megan Young
RuPaul's Drag Race (season 8)
Pia Wurtzbach
RuPaul's Drag Race
Miss USA
Olivia Culpo
Gigi Gorgeous
Shirley Bassey
Beauty pageant
Carmen Carrera
James Bond music
David Arnold
song)
Radiohead
Hail to the Thief
Southpaw (film)
Drag queen
Paulir Porizkc
Drag king


Nigga
Model minority
Feminist movement
Hate crime
Harem
Second-wave feminism
egenation
Separatist feminism
Person of color
Naomi Wolf
Eric Rudolph
Radical feminism
Alt-right
Gender studies
Feminism
Rosie the Riveter
Cons
the
Liberal feminism
Scaphism
next door
Pussy Riot
Masculism
Women's studies
Feminist theory
Strip search phone call scam
Girl power
Women's history
Men's rights movement
Intersectionality
Internment of Japanese Americans
Male privilege
White privilege
Marxist feminism
Judith B
Kyriarchy
Korematsu v. United States
Misogyny
Ethnic stereotype
Gaze
Women's rights
Institution racism
Immurement
Supremacism
Pechenegs
Culture of the United States
Turkification
9/11 Truth movement
Postcolonial feminism
n-rights
ents
Americanization
The Falling Man
Dream diary
Ecofeminism
United Airlines Flight 93
Emancipation
Glass ceiling
Cultural appropriation
Second-cla citizen
War bond



Women's rights: Women's rights are the rights and entitlements claimed for women and girls worldwide, and formed the basis for the women's rights movement in the nineteenth century and feminist movement during the 20th century. In some countries, these rights are institutionalized or supported by law, local custom, and behavior, whereas in others they are ignored and suppressed. They differ from broader notions of...[see Wikipedia article] Article links to 57 men and 49 women.

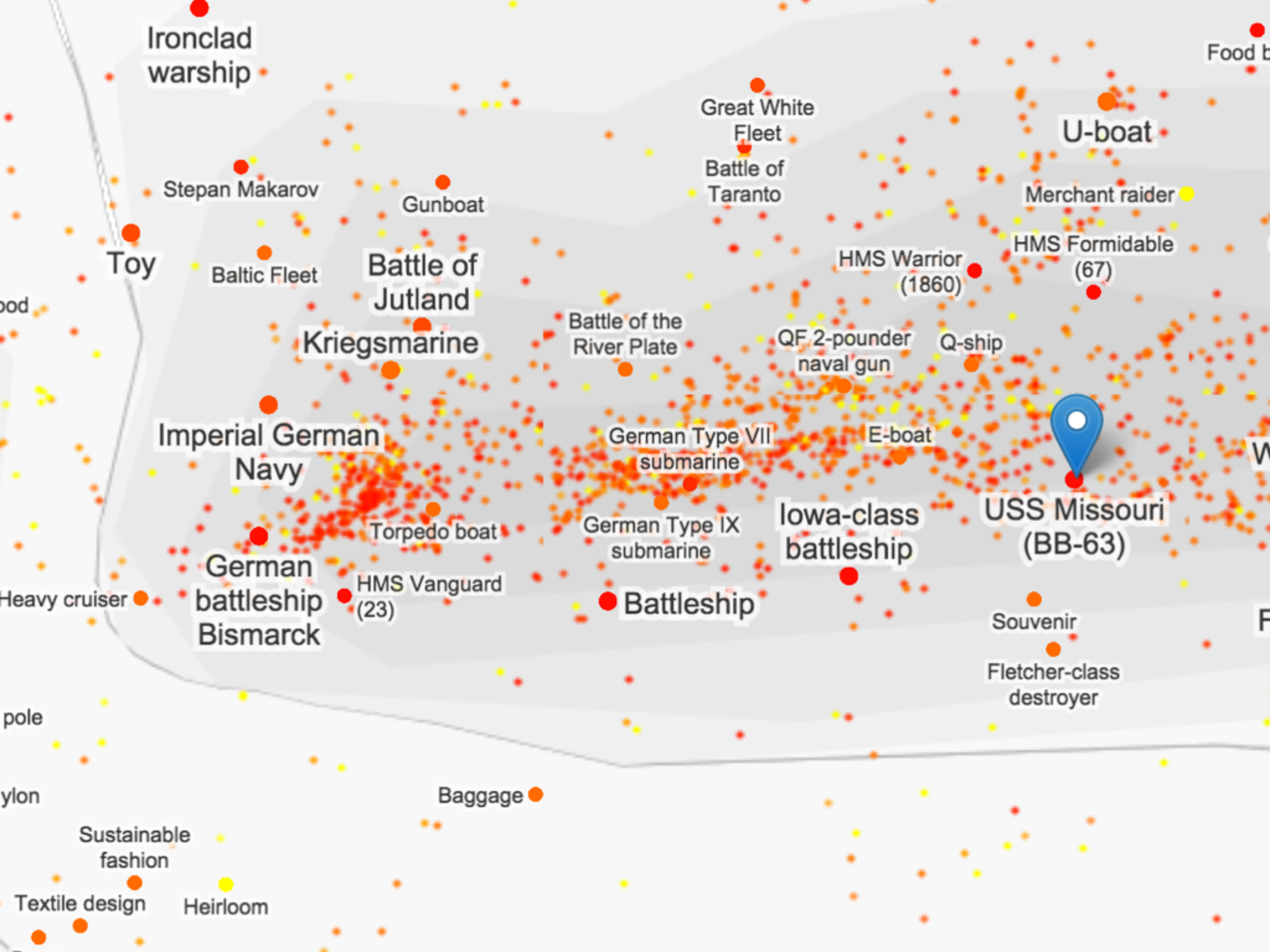

Motor vehicle
Truck driver
Heavy equipment
Fleet vehicle
Tesla Powerwall
Corporate Average Fuel Economy
Gross vehicle weight rating
Aftermarket (automotive)
European emission standards
Electric car
Electric vehicle
Hybrid vehicle
Emission standard
Trunk (car)
Run-flat tire
Automobile safety
Electric motorcycles and scooters
Hybrid vehicle drivetrain
Seat belt
Tow hitch
Pillar (car)
Fuel economy in automobiles
Vehicle
Chassis
Tire
History of the automobile
Windscreen wiper
Locking differential
Continuously variable transmission
Headlamp
Car
Karl Benz
Windshield
Redline
Air suspension
Anti-lock braking system
Automatic transmission
Differential (mechanical device)
Aquaplaning
Steam
Electronic control unit
Gasoline direct injection
Powertrain
Rudolf Dies
Cart
Vestigiality
Overhead camshaft
Catalytic converter
Disc br
Engine balance
Steering
Dry sump
Turbocharger
Supercharger
Intercooler
Fuel injection
Distributor
V engine
Piston ring
Diesel engine
Crankshaft
H engine

Article Quality Visualization

Quality estimates from ORES (Halfaker)

High Quality 🟥 🟧 🟨 Low Quality

https://figshare.com/articles/Monthly_Wikipedia_article_quality_predictions/3859800


Ironclad warship
Food b
Great White Fleet
U-boat
Battle of Taranto
Stepan Makarov
Gunboat
Merchant raider
Toy
Baltic Fleet
Battle of Jutland
HMS Warrior (1860)
HMS Formidable (67)
ood
Battle of the River Plate
Kriegsmarine
QF 2-pounder naval gun
Q-ship
Imperial German Navy
German Type VII submarine
E-boat
W
Torpedo boat
German Type IX submarine
Iowa-class battleship
USS Missouri (BB-63)
German battleship Bismarck
Heavy cruiser
HMS Vanguard (23)
Battleship
Souvenir
F
Fletcher-class destroyer
pole
ylon
Baggage
Sustainable fashion
Textile design
Heirloom

User Study Results

Overall positive feedback

Easy, fun

Some confusion about article placement

Next Steps

1. Stand up WikiBrain API in labs
2. Regular releases of navigation embeddings
3. Release Cartograph
4. Cartograph enhancements

Thank You!

Research collaborators, Wikipedians, Aaron Halfaker

http://shilad.com

http://wikibrainapi.org

http://cartograph.info

NSF

MACALESTER COLLEGE

Northwestern University

WIKIMEDIA FOUNDATION