PONTIFICIA
UNIVERSIDAD
CATÓLICA
DE CHILE

A Tutorial on Wikimedia Visual Resources and its Application to Neural Visual Recommender Systems

Denis Parra, Antonio Ossa-Guerra, Manuel Cartagena,
Patricio Cerda-Mardini, Felipe del Río, Isidora Palma,
Diego Saez-Trumper, Miriam Redi

WIKIMEDIA
FOUNDATION

2001-2021

All Wikis

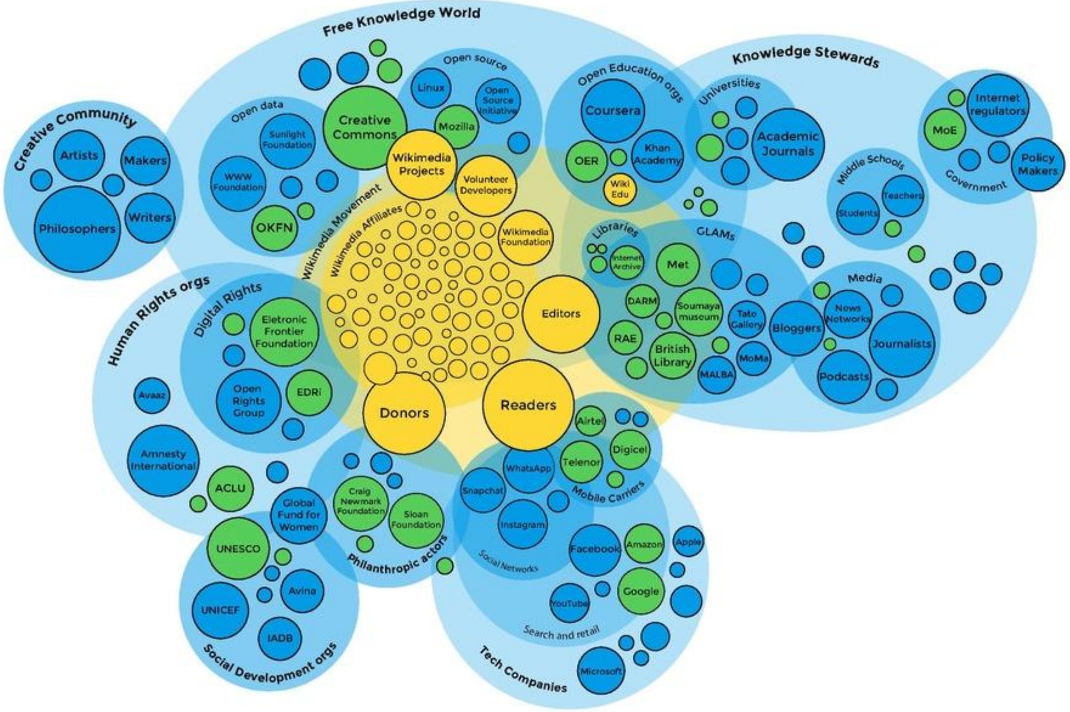
55+ million
articles

36+ million
edit/month

19+ billion
pageviews/month

Wikimedia Foundation

- Non-profit organization
- Operates Wikipedia and sister projects
- ~500 employees
  - **Research team have only 8 people!**



MARÍA CRUZ • ED BLAND • NICOLE EBBER • SHANNON KEITH • JAIME ANSTEE • GUILLAUME PAUMIER • SUZIE NUZZEL • CC-BY-SA 4.0

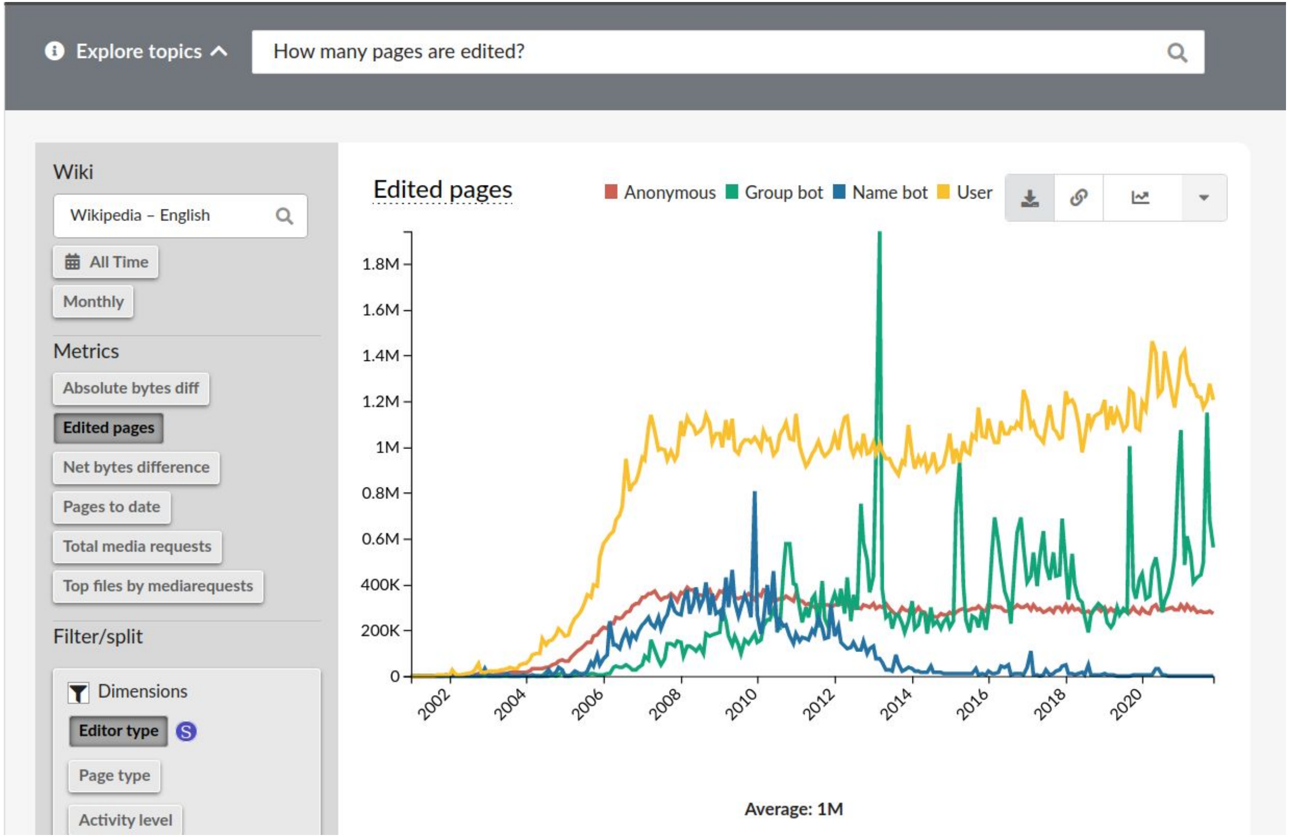
Minimalist user data collection

Revision histories are public data

No 3rd-party collection or sharing of user data

Private user data removed after 90 days

Statistics: stats.wikimedia.org



Explore topics ^

How many pages are edited?

Wiki

Wikipedia – English

All Time

Monthly

Metrics

Absolute bytes diff

Edited pages

Net bytes difference

Pages to date

Total media requests

Top files by mediarequests

Filter/split

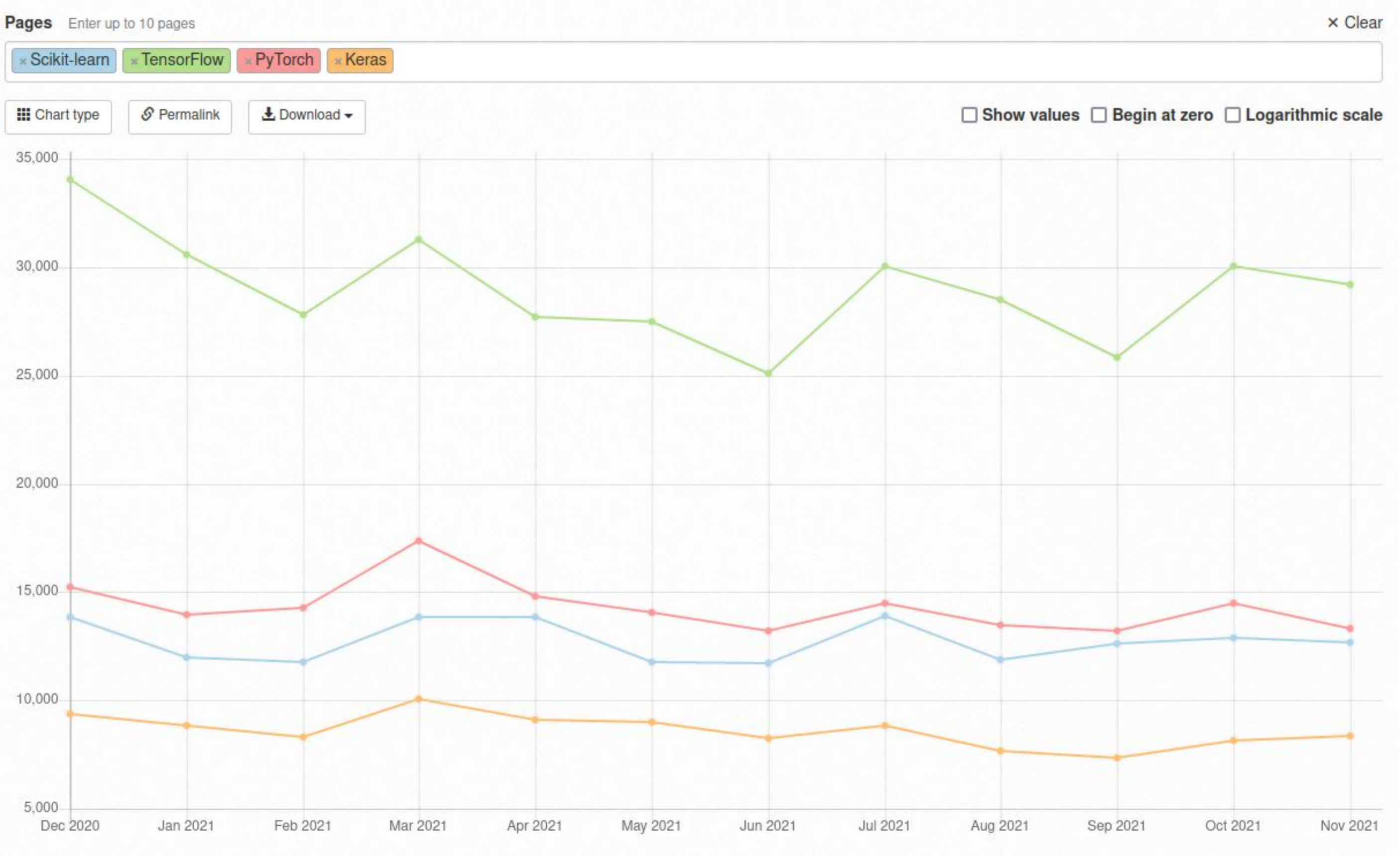
Dimensions

Editor type S

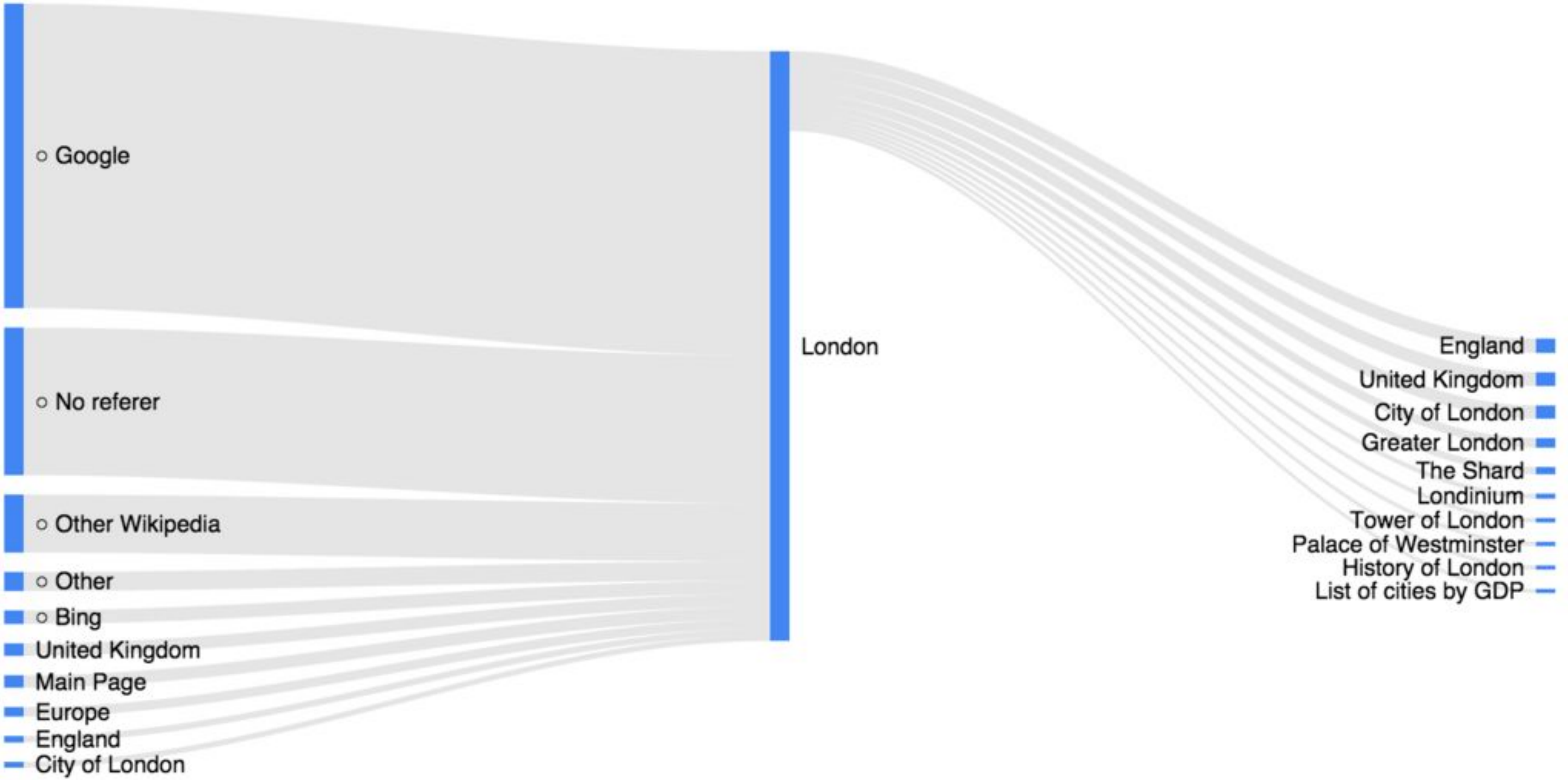
Page type

Activity level

Pageviews: pageviews.toolforge.org



Click Dataset (a.k.a Clickstream)



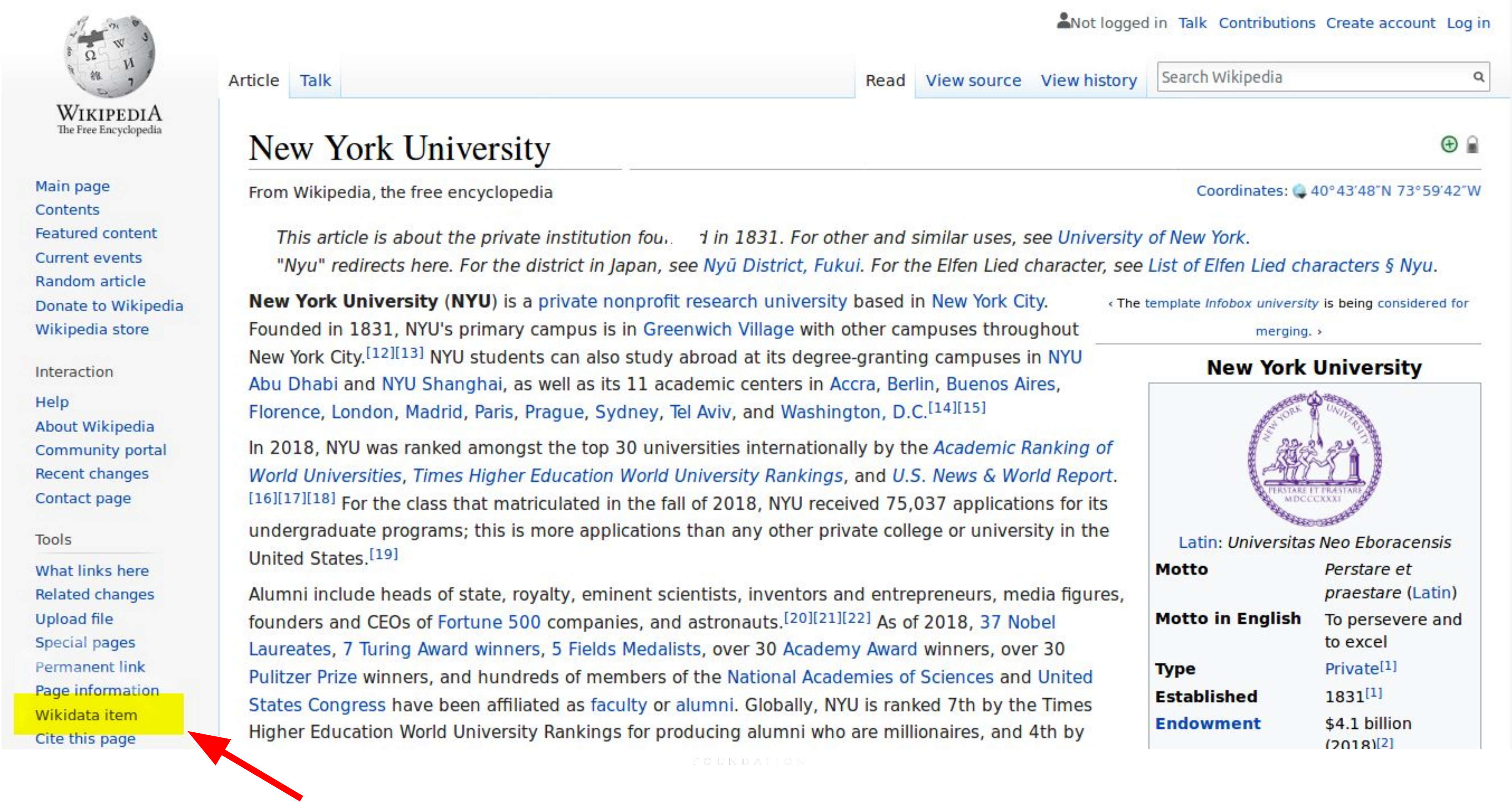
Wikidata

Wikidata is a free and open knowledge base that can be read and edited by both humans and machines.

Wikidata acts as central storage for the **structured data** of its Wikimedia sister projects including Wikipedia, Wikivoyage, Wikisource, and others.

In October 2018, there 23 M of edits in Wikidata, compared with around 4 M edits in English Wikipedia

- https://www.wikidata.org/
- https://query.wikidata.org/



Not logged in Talk Contributions Create account Log in

Article Talk

Read View source View history Search Wikipedia

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
Upload file
Special pages
Permanent link
Page information
Wikidata item
Cite this page

New York University

From Wikipedia, the free encyclopedia

Coordinates: 40°43′48″N 73°59′42″W

This article is about the private institution fou... d in 1831. For other and similar uses, see University of New York.

"Nyu" redirects here. For the district in Japan, see Nyū District, Fukui. For the Elfen Lied character, see List of Elfen Lied characters § Nyu.

‹ The template *Infobox university* is being considered for merging. ›

New York University (**NYU**) is a private nonprofit research university based in New York City. Founded in 1831, NYU's primary campus is in Greenwich Village with other campuses throughout New York City.[12][13] NYU students can also study abroad at its degree-granting campuses in NYU Abu Dhabi and NYU Shanghai, as well as its 11 academic centers in Accra, Berlin, Buenos Aires, Florence, London, Madrid, Paris, Prague, Sydney, Tel Aviv, and Washington, D.C.[14][15]

In 2018, NYU was ranked amongst the top 30 universities internationally by the *Academic Ranking of World Universities*, *Times Higher Education World University Rankings*, and *U.S. News & World Report*.[16][17][18] For the class that matriculated in the fall of 2018, NYU received 75,037 applications for its undergraduate programs; this is more applications than any other private college or university in the United States.[19]

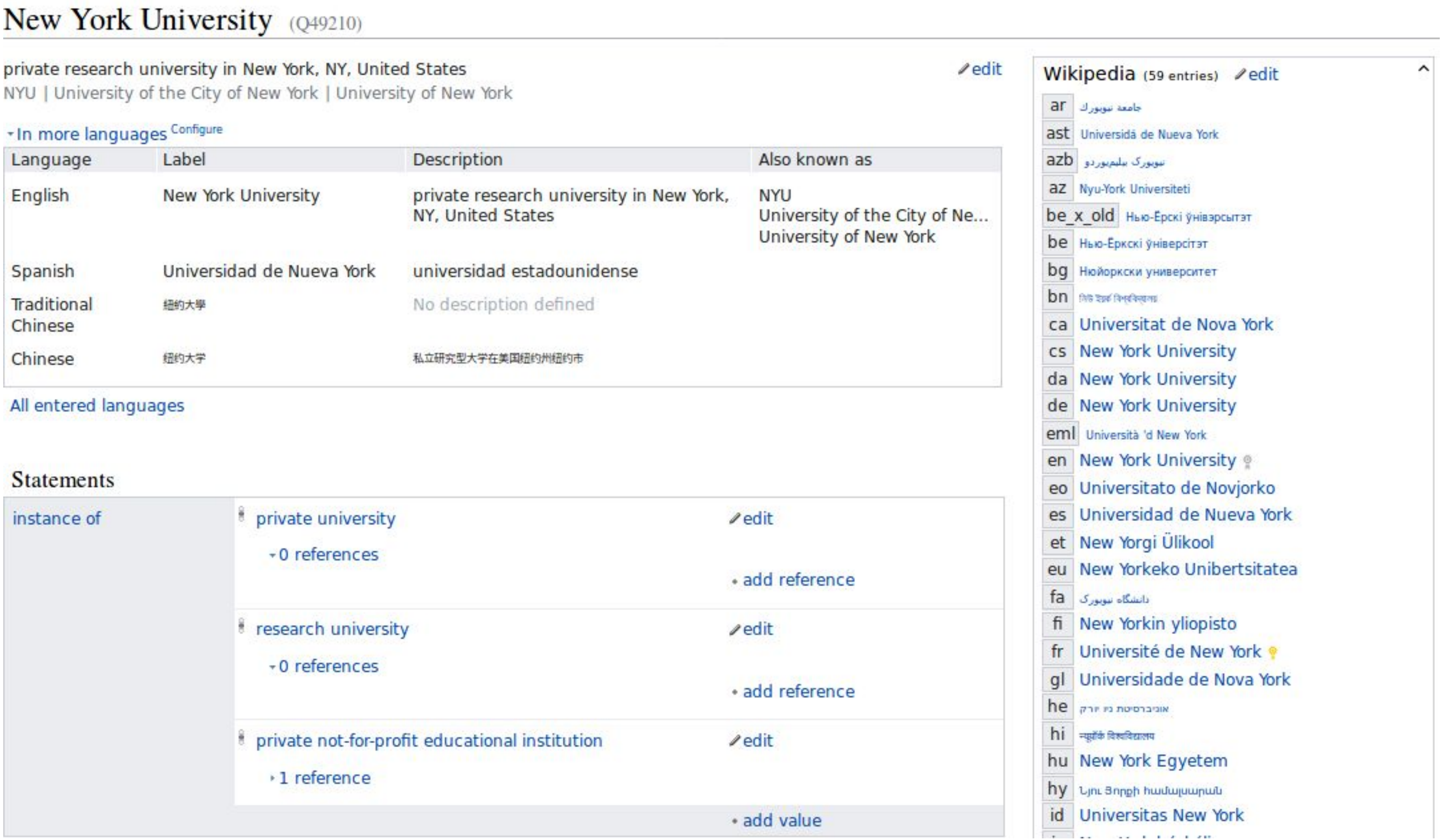
Alumni include heads of state, royalty, eminent scientists, inventors and entrepreneurs, media figures, founders and CEOs of Fortune 500 companies, and astronauts.[20][21][22] As of 2018, 37 Nobel Laureates, 7 Turing Award winners, 5 Fields Medalists, over 30 Academy Award winners, over 30 Pulitzer Prize winners, and hundreds of members of the National Academies of Sciences and United States Congress have been affiliated as faculty or alumni. Globally, NYU is ranked 7th by the Times Higher Education World University Rankings for producing alumni who are millionaires, and 4th by

New York University

Latin: *Universitas Neo Eboracensis*

Motto	*Perstare et praestare* (Latin)
Motto in English	To persevere and to excel
Type	Private[1]
Established	1831[1]
Endowment	$4.1 billion (2018)[2]

FOUNDATION



New York University (Q49210)

private research university in New York, NY, United States ✎edit

NYU | University of the City of New York | University of New York

▾In more languages Configure

Language	Label	Description	Also known as
English	New York University	private research university in New York, NY, United States	NYU University of the City of Ne... University of New York
Spanish	Universidad de Nueva York	universidad estadounidense	
Traditional Chinese	紐約大學	No description defined	
Chinese	纽约大学	私立研究型大学在美国纽约州纽约市	

All entered languages

Statements

instance of	private university ▾0 references	✎edit + add reference
	research university ▾0 references	✎edit + add reference
	private not-for-profit educational institution ▸1 reference	✎edit
		+ add value

Wikipedia (59 entries) ✎edit

ar	جامعة نيويورك
ast	Universidá de Nueva York
azb	نیویورک بیلیم‌یوردو
az	Nyu-York Universiteti
be_x_old	Нью-Ёрскі ўнівэрсытэт
be	Нью-Ёркскі ўніверсітэт
bg	Нюйоркски университет
bn	নিউ ইয়র্ক বিশ্ববিদ্যালয়
ca	Universitat de Nova York
cs	New York University
da	New York University
de	New York University
eml	Università 'd New York
en	New York University
eo	Universitato de Novjorko
es	Universidad de Nueva York
et	New Yorgi Ülikool
eu	New Yorkeko Unibertsitatea
fa	دانشگاه نیویورک
fi	New Yorkin yliopisto
fr	Université de New York
gl	Universidade de Nova York
he	אוניברסיטת ניו יורק
hi	न्यूयॉर्क विश्वविद्यालय
hu	New York Egyetem
hy	Նյու Յորքի համալսարան
id	Universitas New York

image

NYU07.JPG
2,592 × 1,952; 2.42 MB

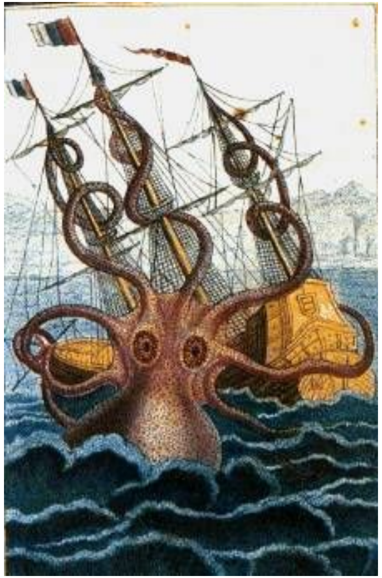
0 references

Data Sources

- Dumps
- Wikimedia API
- SQL Replicas / Quarry
- Event Stream
- Wikidata
- **Commons**

Wikipedia images: 5 Million on English Wikipedia only, most with categories

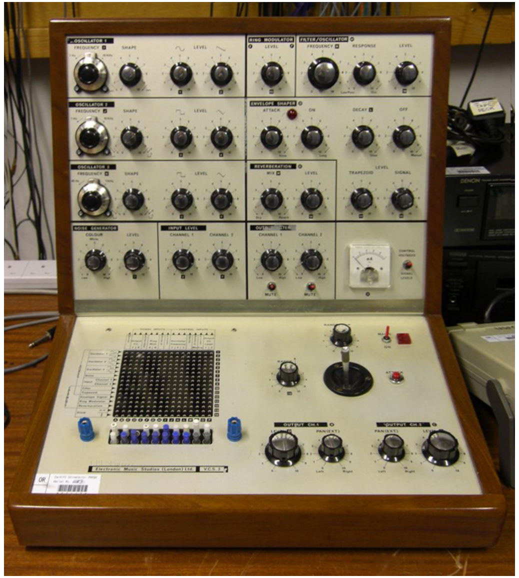
About 43% is captioned, and 10% alt-captioned



Example of good Alt Text: Coloured drawing of a huge octopus rising from the sea and attacking a sailing ship's three masts with its spiraling arms

Image: Colossal octopus by Pierre Denys de Montfort.jpg

Image by en:Pierre Denys de Montfort. uploaded by en:user:Salleman., Public Domain.



Example of bad Alt Text: Photograph

Caption: User's eye view of the "VCS 3";[# 1] top left, three main oscillators; bottom left, patch panel; bottom right, joystick. External keyboard not shown.

Note: it has printed logo: **"V.C.S. 3"**.

Image: EMS VCS 3.jpg

Image by The Standard Deviant - CC BY-SA 2.0,

(Almost) all images come from one source: **Wikimedia Commons**

Building Solutions: Generate lists of images from a specific class

Let's say we are want to generate a list of all pictures containing "Cats"

- **Manual** search: https://commons.wikimedia.org/wiki/Main_Page
- **API-search:** Use API to retrieve search results
https://commons.wikimedia.org/w/api.php?action=query&list=search&srnamespace=6&srsearch=cats
    - Search a specific Category:
https://commons.wikimedia.org/w/api.php?action=query&list=categorymembers&cmtitle=Category:Felis_silvestris_catus
    - Structured data search:
https://commons.wikimedia.org/w/api.php?action=query&list=search&srnamespace=6&srsearch=haswbstatement:P180=Q146

Building Solutions: Generate lists of images from a specific article

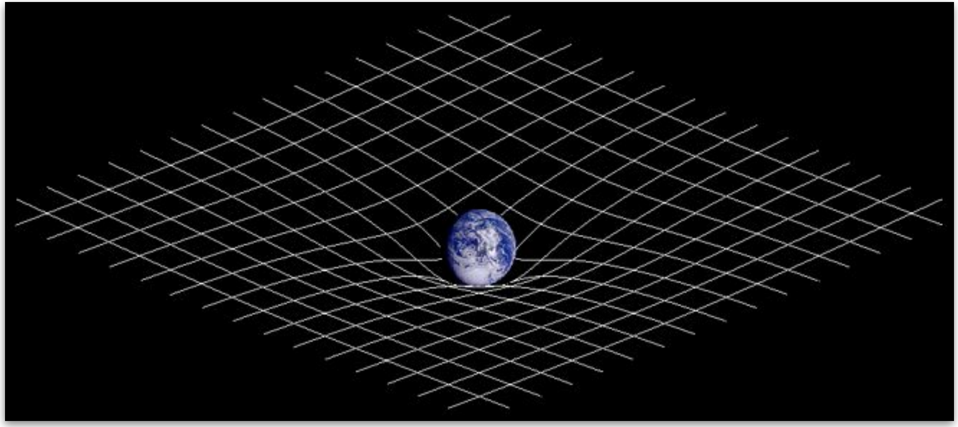
Let’s say we are want to generate a list of all pictures from Wikipedia pages related to “Cats”. We need to use **SQL Replicas**

- Let’s extract this list for one specific Wikipedia edition, using the **imagelinks table**:
  - https://it.wikipedia.org/wiki/Felis_silvestris_catus
  - https://quarry.wmflabs.org/query/38252

Image: image content data

ImageLinks: link between image and Wikipedias

Building Solutions: Image/Article/Caption Data



bn: সাধারণ আপেক্ষিকতা তত্ব অণুয়ায়ী সময় এবং কাল এর বক্রতা একটি দ্বি-মাত্রিক চিত্রের সাহায্যে উপস্থাপন করা হয়েছে।

ja: 一般相対性理論によって記述される、 2次元空間と時間の作る曲面。地球の質量によって空間が歪むとして記述して、重力を特殊相対性理論に取り入れる。実際の空間は 3次元であることに注意すべし。

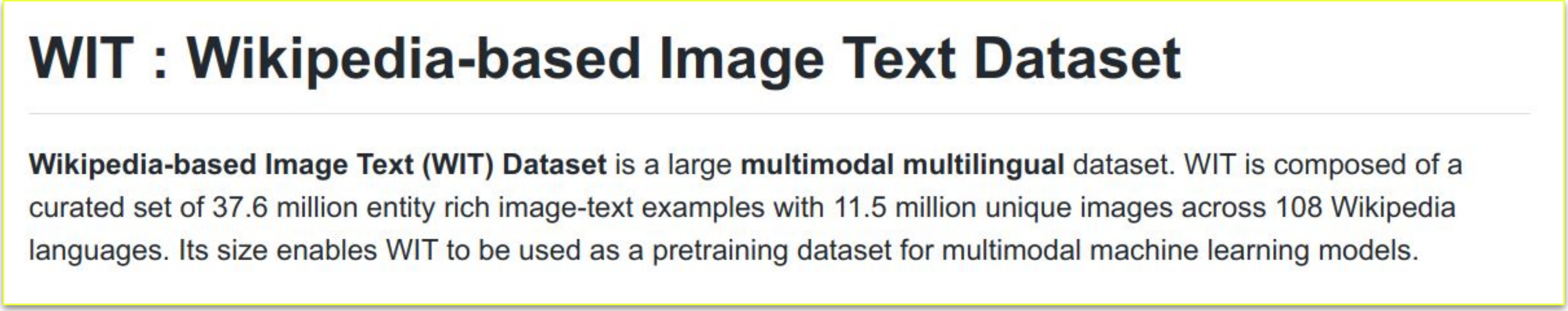
ko: 일반상대성이론에서 묘사된 시공의 곡률을 2차원으로 표현한 그림.

it: Una celebre illustrazione divulgativa della curvatura dello spaziotempo dovuta alla presenza di massa, rappresentata in questo caso dalla Terra.

en: Two-dimensional projection of a three-dimensional analogy of spacetime curvature described in general relativity

ckb: دەرهاوێشتەیەکی دووڕەهەندی لە چەمانەوەی کاتــجێ لە بۆشاییەکی سێڕەهەندیدا، کە لە تیۆریی ڕێژەیی ئاینشتایندا دێتە بەر باس.

my: နှိုင်းရသီအိုရီအရ သုံးဖက်မြင် အာကာသအချိန် ကွေးညွတ်ပုံအား နှစ်ဘက်အမြင်ဖြင့် ဖော်ပြထားပုံ



WIT : Wikipedia-based Image Text Dataset

Wikipedia-based Image Text (WIT) Dataset is a large **multimodal multilingual** dataset. WIT is composed of a curated set of 37.6 million entity rich image-text examples with 11.5 million unique images across 108 Wikipedia languages. Its size enables WIT to be used as a pretraining dataset for multimodal machine learning models.

Downloading image files

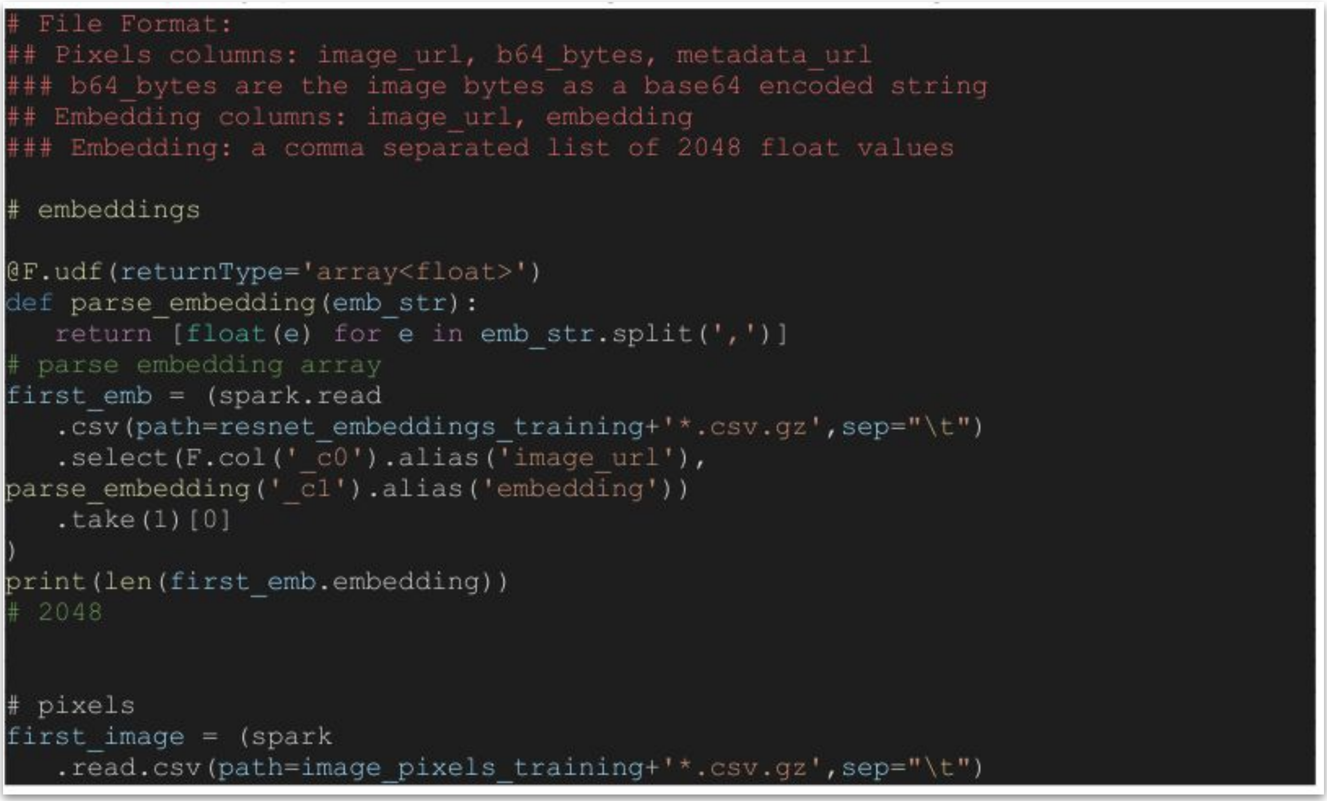
If we want to analyze images, we need to download the files, so that we can process the pixel data. But here we have only image file names!

- Use the Commons Downloader - a Python Library which also allows to download images for a specific Commons category: https://pypi.org/project/CommonsDownloader/

- Other tools for image download: https://commons.wikimedia.org/wiki/Commons:Tools#Download_media

Building Solutions: Pixel Data

https://analytics.wikimedia.org/published/datasets/one-off/caption_competition/



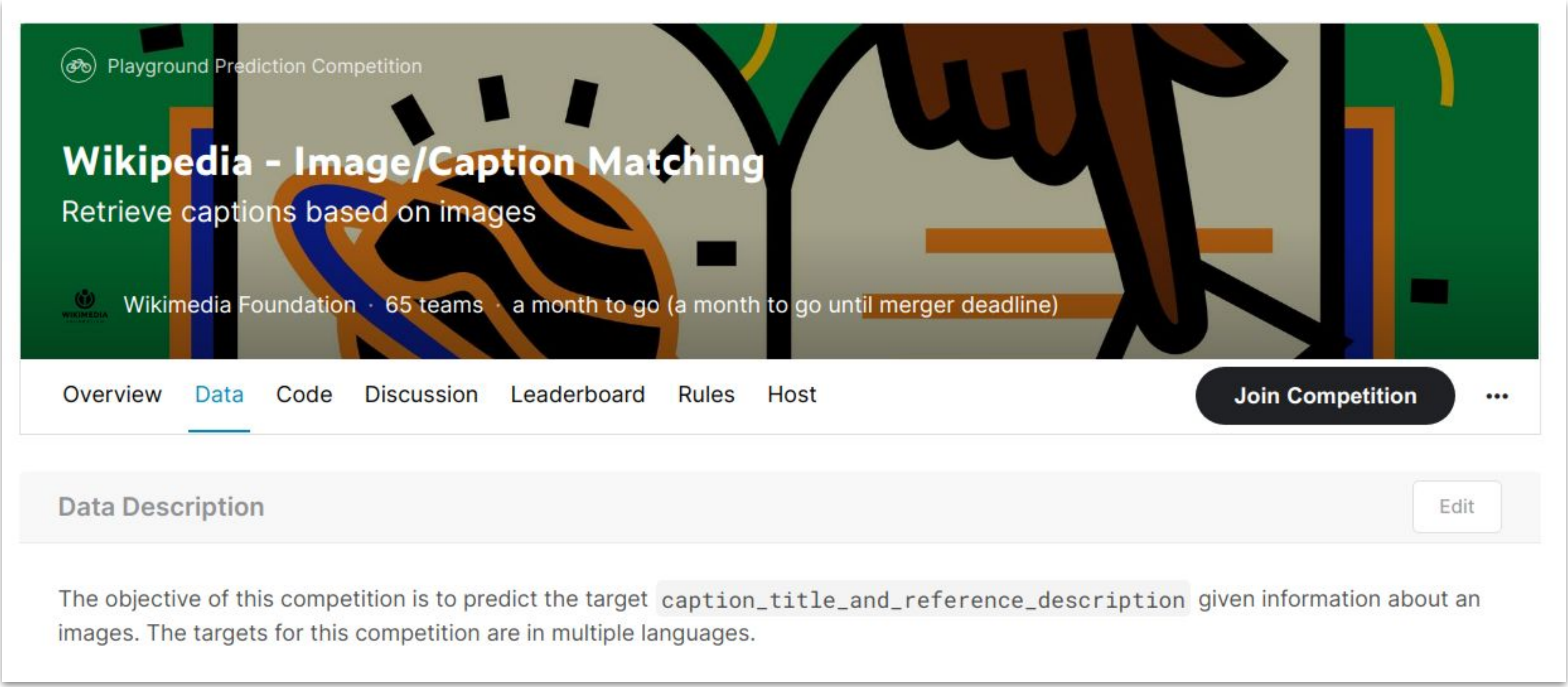
```
# File Format:
## Pixels columns: image_url, b64_bytes, metadata_url
### b64_bytes are the image bytes as a base64 encoded string
## Embedding columns: image_url, embedding
### Embedding: a comma separated list of 2048 float values

# embeddings

@F.udf(returnType='array<float>')
def parse_embedding(emb_str):
   return [float(e) for e in emb_str.split(',')]
# parse embedding array
first_emb = (spark.read
   .csv(path=resnet_embeddings_training+'*.csv.gz',sep="\t")
   .select(F.col('_c0').alias('image_url'),
parse_embedding('_c1').alias('embedding'))
   .take(1)[0]
)
print(len(first_emb.embedding))
# 2048


# pixels
first_image = (spark
   .read.csv(path=image_pixels_training+'*.csv.gz',sep="\t")
```

Building Solutions: Competitions



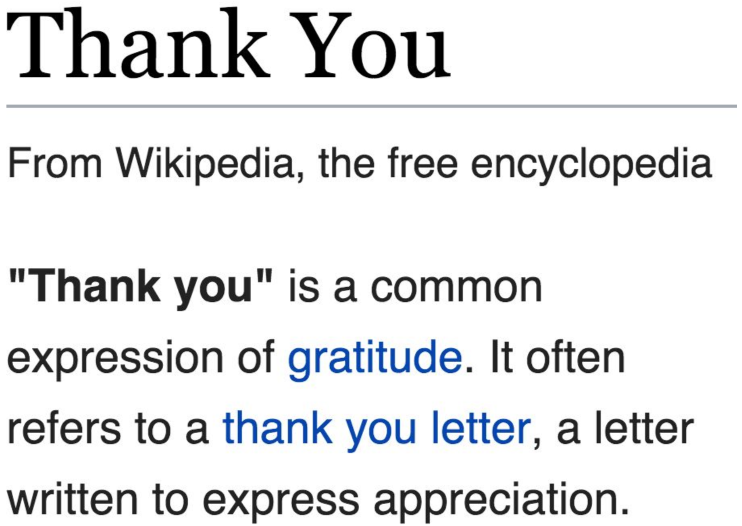
Wikipedia - Image/Caption Matching

Retrieve captions based on images

Wikimedia Foundation · 65 teams · a month to go (a month to go until merger deadline)

Data Description

The objective of this competition is to predict the target `caption_title_and_reference_description` given information about an images. The targets for this competition are in multiple languages.



Thank You

From Wikipedia, the free encyclopedia

"Thank you" is a common expression of gratitude. It often refers to a thank you letter, a letter written to express appreciation.

✉ {miriam,diego}@wikimedia.org

http://research.wikimedia.org