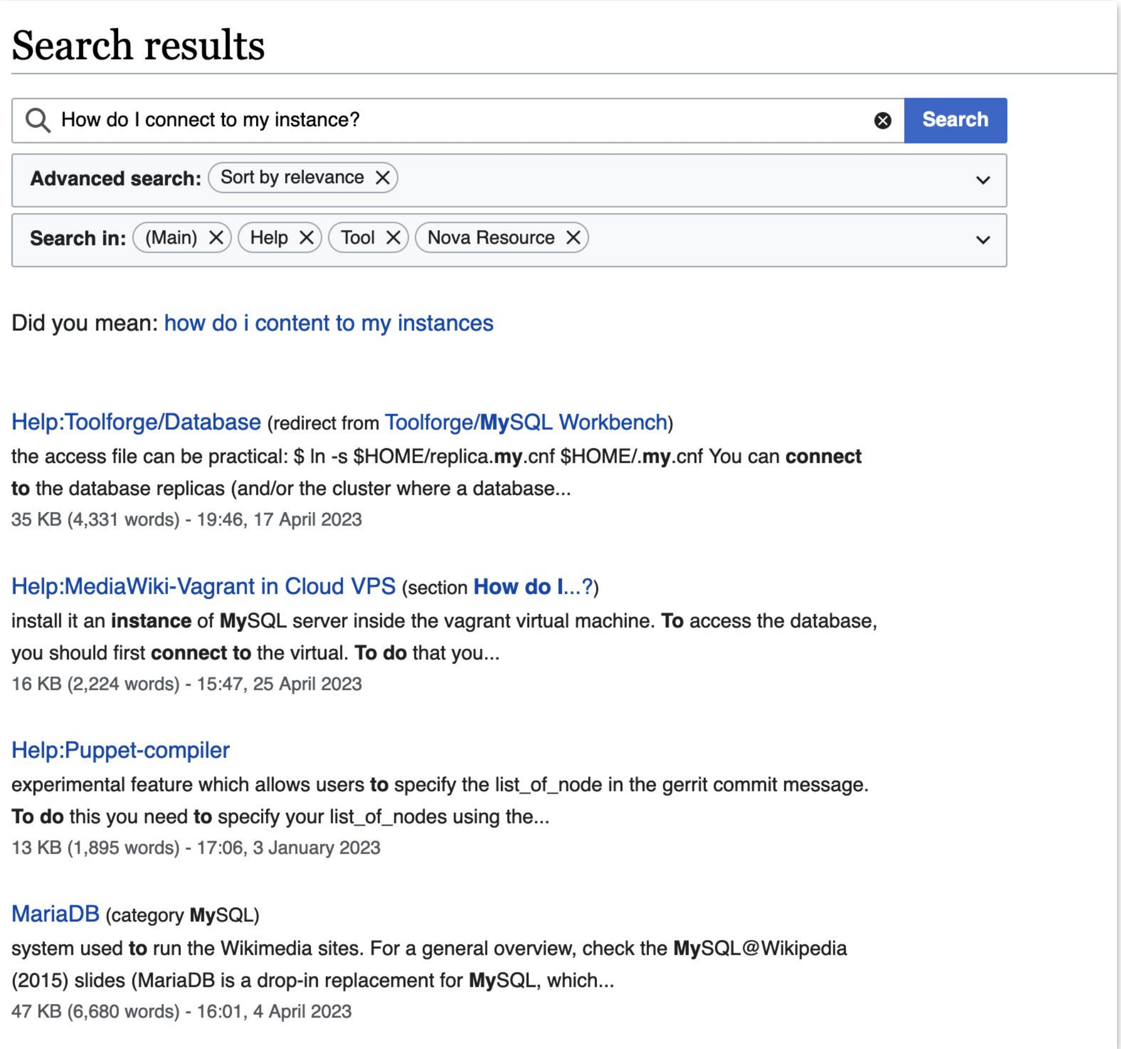
WIKIMEDIA
HACKATHON

# ML on Cloud Services

Isaac Johnson: https://meta.wikimedia.org/wiki/User:Isaac_(WMF)
Slavina Stefanova: https://meta.wikimedia.org/wiki/User:SStefanova_(WMF)
21 May 2023 – Wikimedia Hackathon

# Wikitech Search (example)

Existing search does not work very well for natural-language queries because Wikitech has very complicated/diverse pages and natural-language queries often don't have good keyword overlap with them.



## Search results

How do I connect to my instance?

**Advanced search:** Sort by relevance ✕

**Search in:** (Main) ✕ Help ✕ Tool ✕ Nova Resource ✕

Did you mean: how do i content to my instances

Help:Toolforge/Database (redirect from Toolforge/**My**SQL Workbench)
the access file can be practical: $ ln -s $HOME/replica.**my**.cnf $HOME/.**my**.cnf You can **connect to** the database replicas (and/or the cluster where a database...
35 KB (4,331 words) - 19:46, 17 April 2023

Help:MediaWiki-Vagrant in Cloud VPS (section **How do I**...?)
install it an **instance** of **My**SQL server inside the vagrant virtual machine. **To** access the database, you should first **connect to** the virtual. **To do** that you...
16 KB (2,224 words) - 15:47, 25 April 2023

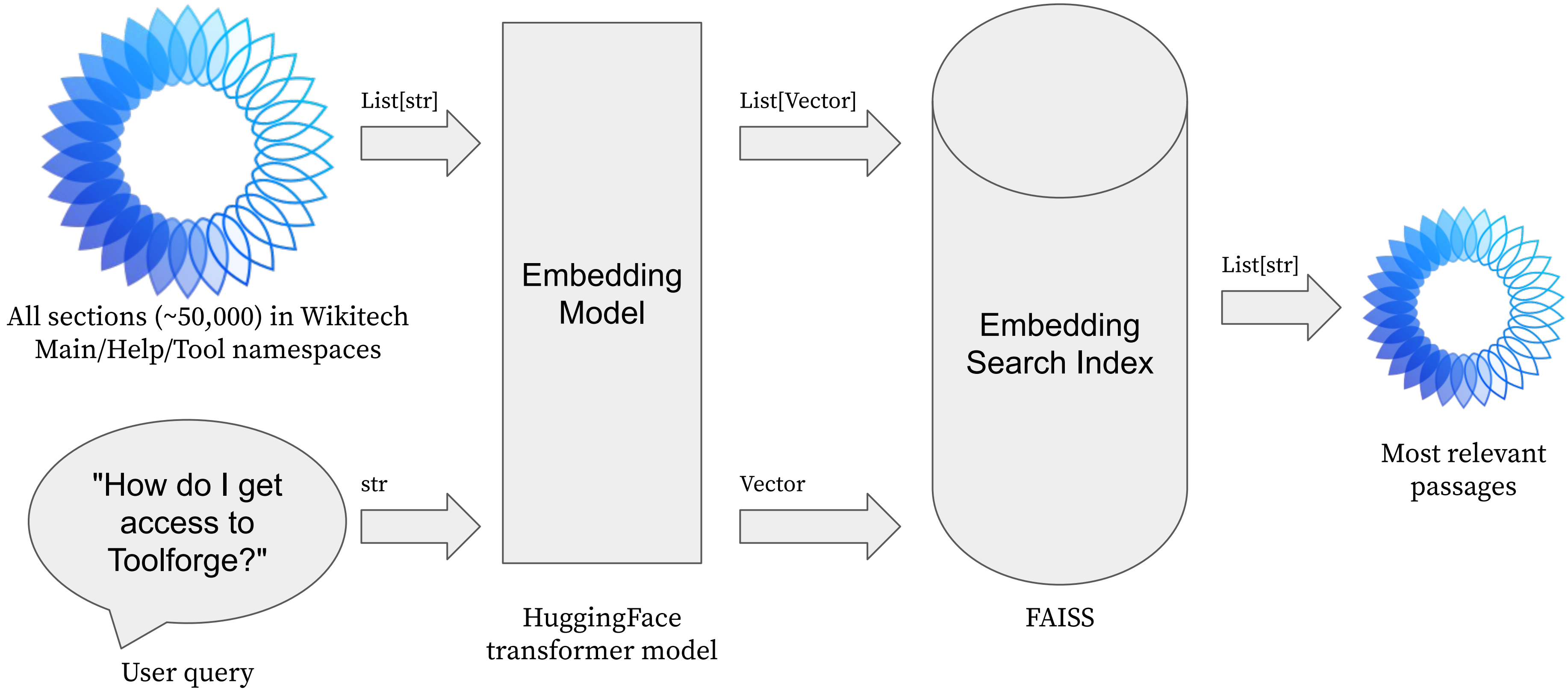
Help:Puppet-compiler
experimental feature which allows users **to** specify the list_of_node in the gerrit commit message. **To do** this you need **to** specify your list_of_nodes using the...
13 KB (1,895 words) - 17:06, 3 January 2023

MariaDB (category **My**SQL)
system used **to** run the Wikimedia sites. For a general overview, check the **My**SQL@Wikipedia (2015) slides (MariaDB is a drop-in replacement for **My**SQL, which...
47 KB (6,680 words) - 16:01, 4 April 2023

# Demo!

https://search-wikitech.wmcloud.org/docs

# Demo: Wikitech Search



All sections (~50,000) in Wikitech Main/Help/Tool namespaces

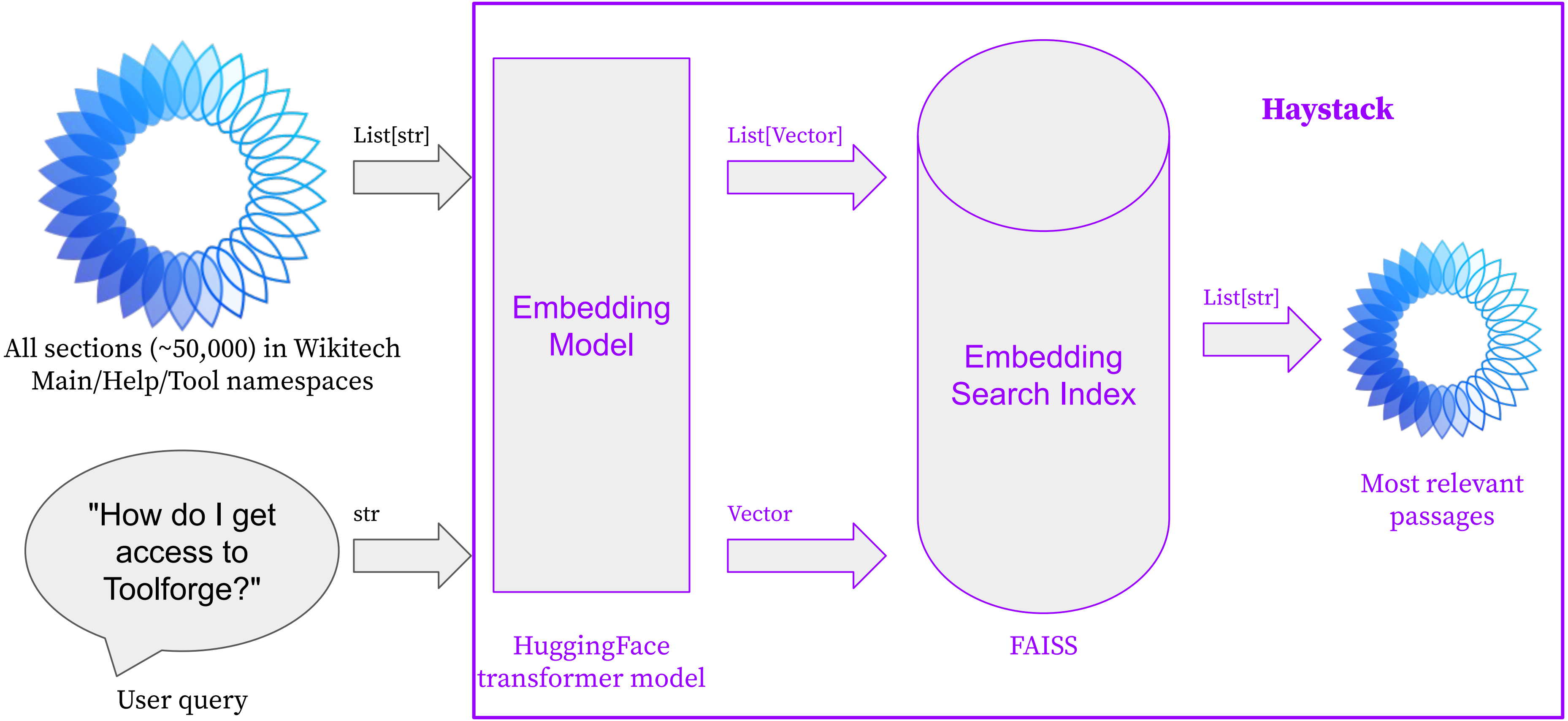
User query

HuggingFace transformer model

FAISS

Most relevant passages

# Demo: Wikitech Search



All sections (~50,000) in Wikitech Main/Help/Tool namespaces

List[str]

"How do I get access to Toolforge?"

User query

str

Haystack

Embedding Model

HuggingFace transformer model

List[Vector]

Vector

Embedding Search Index

FAISS

List[str]

Most relevant passages

# Demo: Wikitech Search

- NLP Framework (Python Haystack)
  - ML: Transformers (PyTorch)
    - Many alternatives but Transformers is the de facto standard for most NLP applications and PyTorch has the best support of any of the ML back-end libraries (you could also choose Tensorflow or FLAX)
  - Database: FAISS
    - Many alternatives such as Elastic – check out Haystack for other options
- API: FastAPI
  - Many alternatives – Flask etc. Mostly interchangeable and what you're familiar with

# Learnings – Open Source + AI

- Challenges with GPUs → take care when downloading PyTorch dependencies to not include NVidia packages (proprietary). Instead:
  - `pip3 install torch torchvision torchaudio --index-url https://download.pytorch.org/whl/cpu`
  - pip-licenses is your friend here if you're not sure
- Look carefully at what models you're using – at least four relevant components:
  - Self-hosting:
    - Model weights:
      - Growing number of openly-licensed models
      - Debate around RAIL licenses; Alpaca as tricky example
    - Model serving code:
      - de facto standard is HuggingFace's `transformers` (Apache 2.0)
  - Full ecosystem:
    - Model training code – generally trivial but ideally open
    - Model training data – often not public and rarely open (generally a lot of fair-use exceptions are used for training ML models)

# Learnings – File Size and Permissions

- Caches
  - HuggingFace by default puts all datasets/models/etc. into a single `~/.cache/huggingface` directory
  - Torch will put model files into `~/.cache/torch`
  - These can be set to other directories via your OS environments or, in some cases, when invoking models
  - Many of these libraries also have extensive dependencies to cover the many modalities etc. that will go into your virtual environment and `~/.cache/pip`
- These caches can cause odd file permission errors, bloat your image, or over-fill drives if you're not aware of that. It can be set explicitly to be another folder as well.

# Example – transformers

| | ❌ Suggested installation from HuggingFace | ✅ Open-source/size friendly install |
|---|---|---|
| Command | `$ pip install transformers[torch]` | `$ pip3 install torch --index-url https://download.pytorch.org/whl/cpu`<br>`$ pip3 install transformers` |
| Virtual environment size | 4.4G | 976M |
| Cache size | 2.2G | 231M |
| # of packages | 39 | 23 |
| # of proprietary packages | 11 | 0 |

# Learnings – Threading

- Threading
    - PyTorch has its own threading which in the past has caused issues with certain web app configurations
    - Our docker container solution seems to solve this but if you're having issues going from a localhost API to webapp with stacks like nginx+uwsgi or nginx+gunicorn, try switching to a single worker for uwsgi/gunicorn

# Learnings – Model Choice

- Choosing a model
  - Beyond open-source, how to find an appropriate model for what you want to achieve?
  - Considerations:
    - Objective (is it doing what you want?)
    - Coverage (how many languages does it support?)
    - Size (will it fit into RAM?)
    - Performance (are the results useful?)
    - Latency (how slow is inference?)
    - Optimize-able (can it be optimized for inference on CPUs?)
- Examples:
  - Ideal case: https://www.sbert.net/docs/pretrained_models.html#model-overview
  - Usually: https://huggingface.co/models?pipeline_tag=sentence-similarity&sort=downloads

# Thank you! Feedback? Questions?

Contact:

- User:Isaac_(WMF)
- User:SStefanova_(WMF)

Documentation:

- Demo: https://search-wikitech.wmcloud.org/docs
- Code: https://github.com/blancadesal/wikitech-search/
- Generating the Search Index: https://public-paws.wmcloud.org/User:Isaac_(WMF)/hackathon-2023/wikitech-natural-language-search.ipynb

# Attribution

- Slide 2:
  - Screenshot of Wikitech search results: CC BY-SA 3.0
  - URL: https://wikitech.wikimedia.org/w/index.php?go=Go&search=How+do+I+connect+to+my+instance%3F&title=Special%3ASearch&ns0=1&ns12=1&ns116=1&ns498=1
- Slide 4:
  - Wikitech logo: By Serhio MagpieKrinkle - File:Wikitech-2021-logo-blue.svg, CC BY-SA 4.0, https://commons.wikimedia.org/w/index.php?curid=104659586