USC Student Health | Part of K

https://ParseMeIfYouCan/CSCI-544/Poster-Presentation

SemSE

Zero-shot multilingual semantic search pipeline

USC University of Southern California

Motivation | Model Design | Dataset Creation | Analysis and Results | बीमा

Search Results for "बीमा" Query

NOTHING FOUND

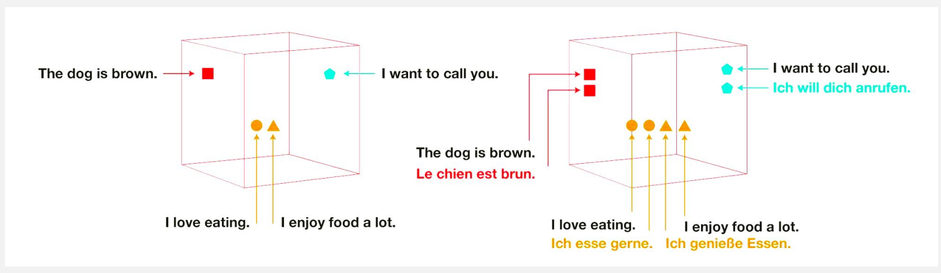
Sorry, but nothing matched your search terms. Please try again with some different keywords. Or try using SemSE API

Motivation / प्रेरणा / motivación

- ➔ Keyword based search does not always help the end-users as documents may not always contain the exact search phrase but may contain its semantic equivalent.
- ➔ Most semantic search engines require training on large datasets to produce good results. This is a time-consuming and resource-intensive process.
- ➔ We propose an unsupervised pipeline that automates the entire process of creating a semantic search engine for any website using a single API. Given the website URL, our automated pipeline will be able to create a multilingual semantic search engine without human intervention with just a few clicks!
- ➔ Novelty: Along with the API, we also present a novel dataset, which is a first of its kind - multilingual student health (medical) domain dataset. This dataset can be used for further research purposes as a benchmarking dataset.

Design / डिजाइन / diseño

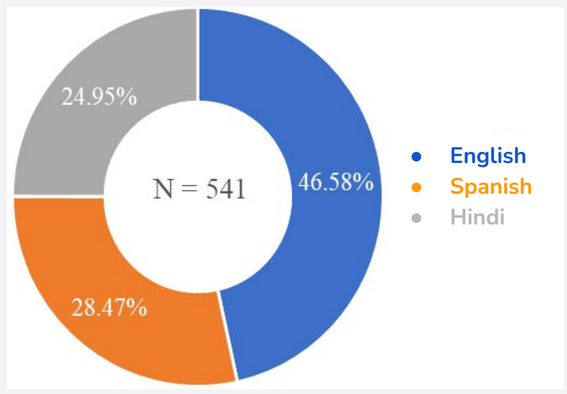
- ➔ LASER by Facebook - Trained on 93 different languages in 23 different alphabets. Embeds all the language into a single shared embedding space and is a ZERO-SHOT approach, allows us to directly use it for encoding text, without the need of fine-tuning!



- ➔ We created a corpus of 50475 lines of text from over 1250 webpages found on USC's Student Health website.
- ➔ The processed web content was indexed using FAISS for fast retrieval and indexing.
- ➔ Cosine similarity was chosen to be the metric of choice after researching various similarity measures for comparison.
- ➔ The application provides an endpoint which takes in the user query and returns the webpage url with the most similar content, and the similarity score for the user to see - all in real-time.

Dataset / डाटासेट / conjunto de datos

- ➔ A total of 650+ queries were crowdsourced using a google form that was distributed to a wide variety of end-users.
- ➔ After preprocessing and removing queries that do not have answers in the health website, a total of 541 queries remained with a split of ~47:25:28 (English:Hindi:Spanish) as shown in the pie chart below.

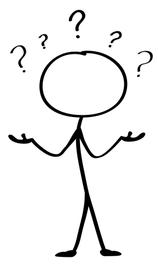


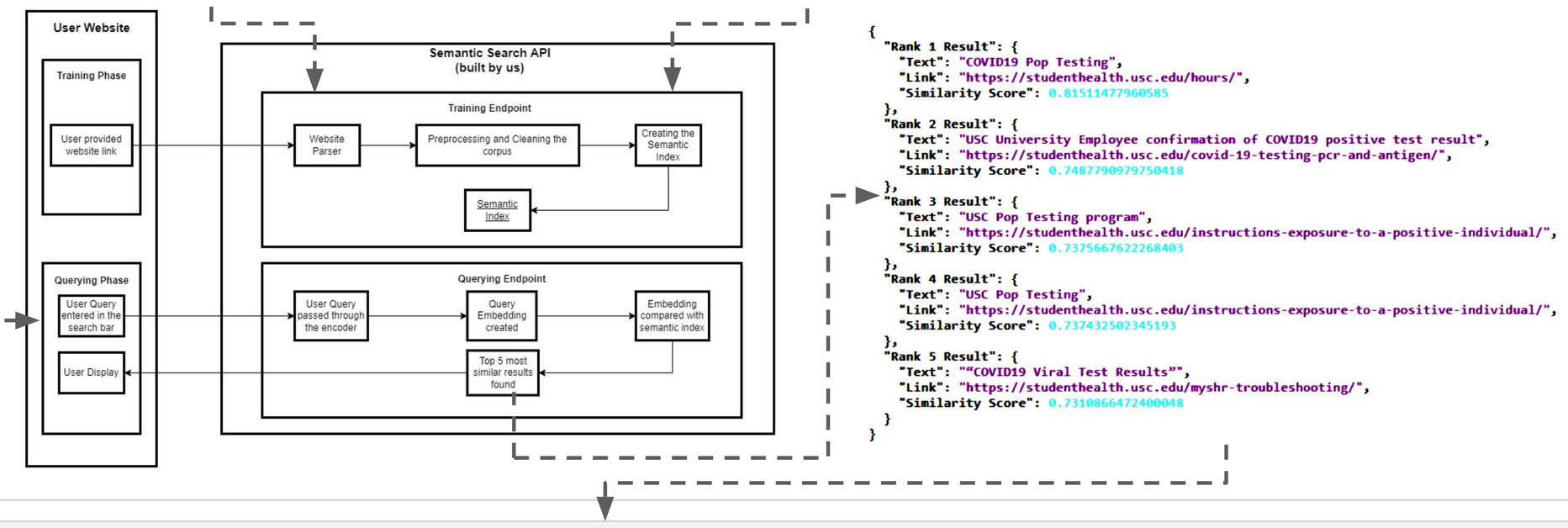
- ➔ To reduce the bias & complexity of annotating, BERT was used to cluster similar queries.
- ➔ The dataset was then manually annotated to indicate the top-3 preferred url results from USC Student Health website for each query group.

USC Covid19 pop testing

यूएससी कोविड -19 पॉप परीक्षण

Prueba pop de USC Covid19






```
{
  "Rank 1 Result": {
    "Text": "COVID19 Pop Testing",
    "Link": "https://studenthealth.usc.edu/hours/",
    "Similarity Score": 0.81511477960585
  },
  "Rank 2 Result": {
    "Text": "USC University Employee confirmation of COVID19 positive test result",
    "Link": "https://studenthealth.usc.edu/covid-19-testing-pcr-and-antigen/",
    "Similarity Score": 0.7487790979750418
  },
  "Rank 3 Result": {
    "Text": "USC Pop Testing program",
    "Link": "https://studenthealth.usc.edu/instructions-exposure-to-a-positive-individual/",
    "Similarity Score": 0.7375667622268403
  },
  "Rank 4 Result": {
    "Text": "USC Pop Testing",
    "Link": "https://studenthealth.usc.edu/instructions-exposure-to-a-positive-individual/",
    "Similarity Score": 0.737432502345193
  },
  "Rank 5 Result": {
    "Text": "“COVID19 Viral Test Results”",
    "Link": "https://studenthealth.usc.edu/myshr-troubleshooting/",
    "Similarity Score": 0.7310866472400048
  }
}
```


Analysis and Results / विश्लेषण और परिणाम / análisis y resultados

- ➔ Models were benchmarked on the novel multilingual health queries dataset.
- ➔ SemSE's zero shot performance beats the current system of traditional keyword matching by a significant margin.
- ➔ SemSE's zero shot performance beats zero shot & fine-tuned mBERT performance for all 3 languages.
- ➔ SemSE's zero shot performance beats zero shot SOTA XLM-R but fails to beat fine-tuned XLM-R.