Multilingual
Semantic search
hackathon

We are team

Under Dawgs UD

Saran Sundar
Sudharshan

Suggested domains:

1. **Internal Knowledge base search:** An internal knowledge base is a company-made resource consisting of process documents and tools that members of the organization need to do their jobs properly

2. **Legal document search:** As the name suggests, this help us retrieve legal documents that outlines an agreement between two or more parties that has been signed by mutual assent and in all other respects can be relied upon in court

3. **Forum Search:** A forum is an online discussion board where people can ask questions, share their experiences, and discuss topics of mutual interest. Forums are an excellent way to create social connections and a sense of community

4. **Customer review:** An evaluation of a product or service made by someone who has purchased and used, or had experience with, a product or service

5. **Recommendations:** A filtering system that seeks to predict and show the items that a user would like to purchase or see

Our test case :

- As we are final year UG (UnderGrad Students) we have to read a lot of Research papers in order to make our own papers as we have no prior experience, people find it difficult to search papers, read them and some find it difficult in understanding the entire concept of the paper. So that's what lead us in devising the idea why not we provide a solution for this problem, how can we ensure that the person who reads the papers really understands them.

- Simply we provide the ideal features such as,

  1. Recommendation based on **Article Name**
  2. Recommendation based on **Article Content**
  3. **Article Translator**
  4. **Article Summarizer**

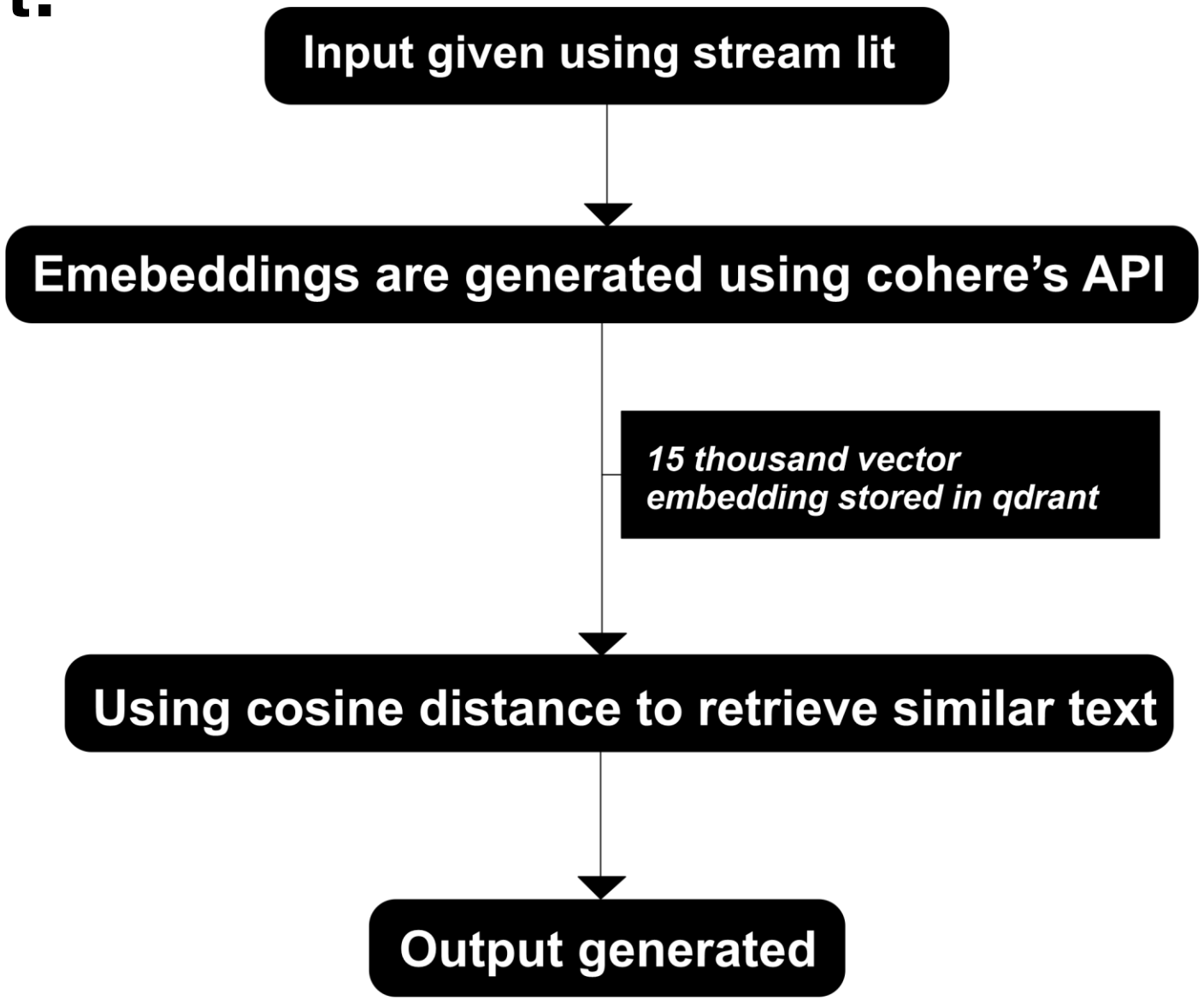
1. Recommendation based on **Article Name**: This search filter provides you recommendations based on the given article name, based on the vector similarity score based on cosine distance.

2. Recommendation based on **Article Content:**This search filter provides you recommendations based on the given article content

3. Article Translator: This feature help us to translate parts of article into users native language, so far we have included 8 languages as follows Tamil, Nepali, Indonesian, Thai, Spanish, Russian, Turkish and French.

4. Article Summarizer: This feature helps the reader to summarize the main points of the input content

Flow Chart:


Input given using stream lit
Emebeddings are generated using cohere's API
15 thousand vector embedding stored in qdrant
Using cosine distance to retrieve similar text
Output generated

Approach:

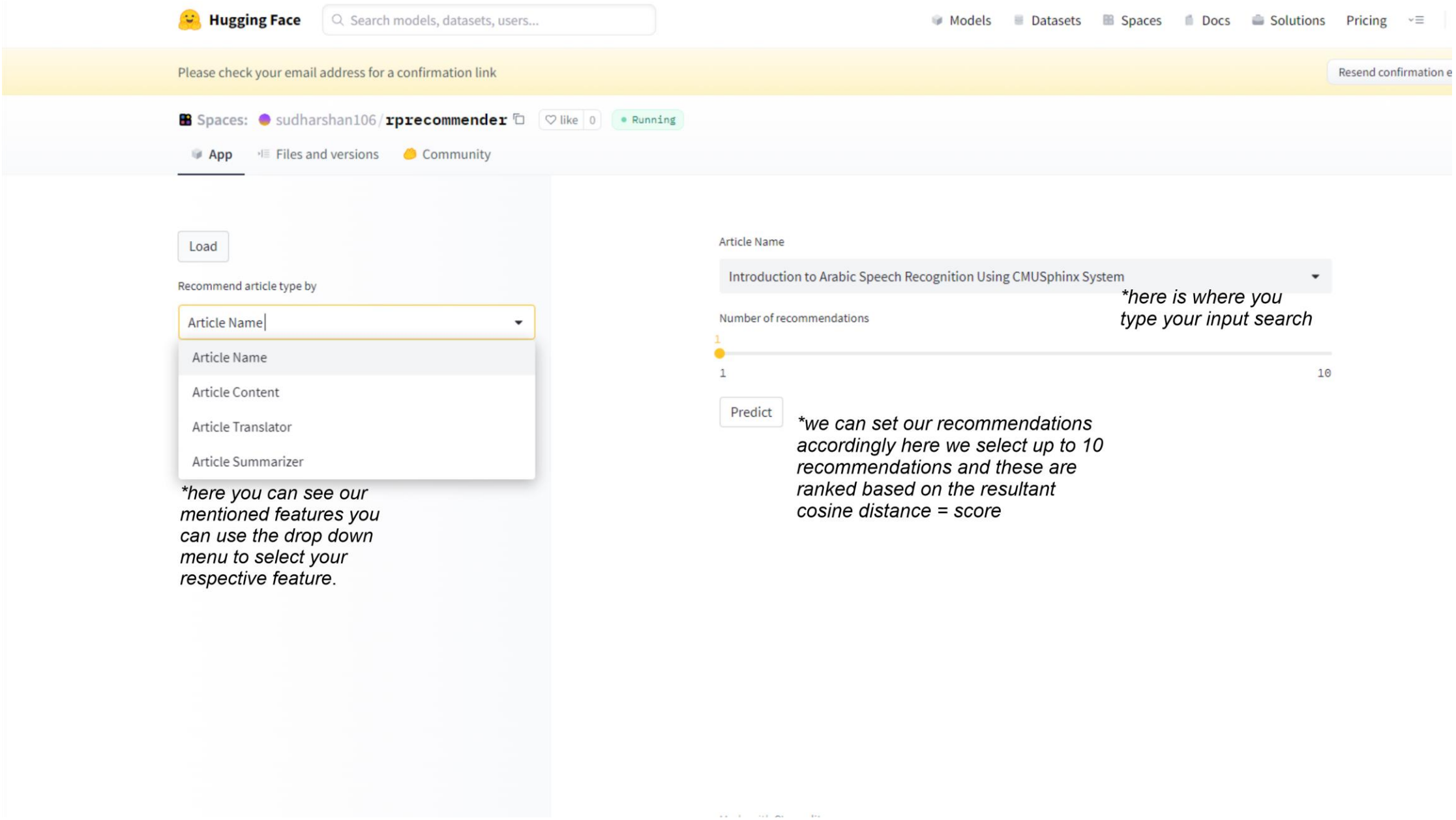
- The dataset we used is Arxiv's Open access data and we used 15000 records from it which contains ML and AI-oriented papers.

- Cohere's small model was used for creating embeddings which produced 1024 embeddings for each records. Then the 15000 X 1024 Vectors are uploaded to the created collections in Qdrant.

- Finally, the embeddings are generated for the abstract of chosen articles or the given prompt, and the Qdrant searches for similar texts in the collection and outputs the indices of it. The distance metric used to measure the similar vectors is Cosine distance.

- Besides the recommendation for articles we provided features such as research paper summarization and translating contents of given paragraphs from English to 8 different languages such as Tamil, Nepali, Indonesia, Thai, Spanish, Russian , Turkish, and French.
- The model we used for Language Translation is MBART Large-50-one-to-many for multilingual machine translation
- The Summarization part is done using cohere’s API.

Results:


Hugging Face
Search models, datasets, users...
Models
Datasets
Spaces
Docs
Solutions
Pricing
Please check your email address for a confirmation link
Resend confirmation e
Spaces: sudharshan106 / rprecommender
like 0
Running
App
Files and versions
Community
Load
Recommend article type by
Article Name
Article Name
Article Content
Article Translator
Article Summarizer
*here you can see our mentioned features you can use the drop down menu to select your respective feature.
Article Name
Introduction to Arabic Speech Recognition Using CMUSphinx System
*here is where you type your input search
Number of recommendations
1
1
10
Predict
*we can set our recommendations accordingly here we select up to 10 recommendations and these are ranked based on the resultant cosine distance = score