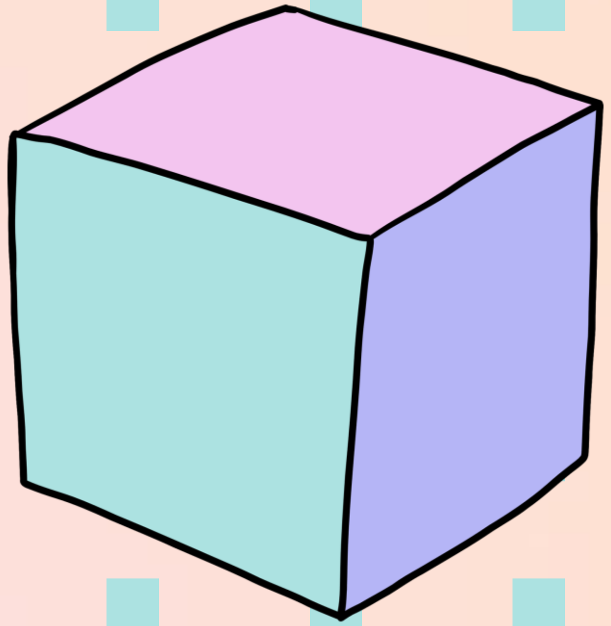
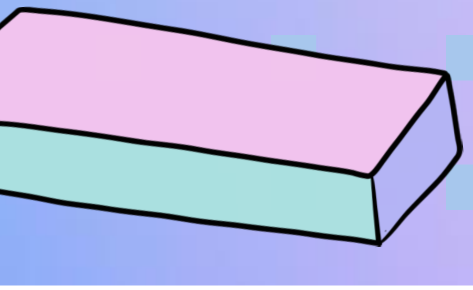
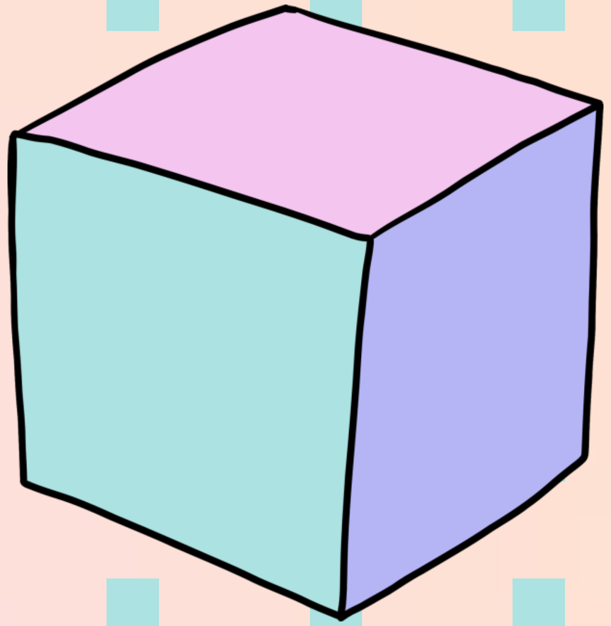


PROJECT PLANNING

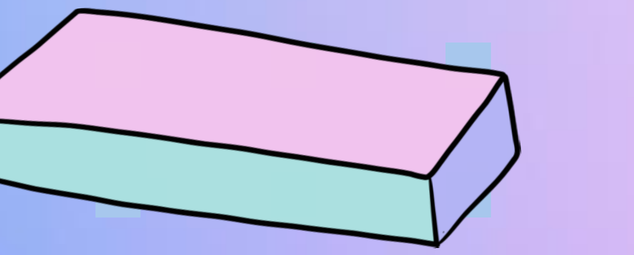
PRESENTATION





PaperMate

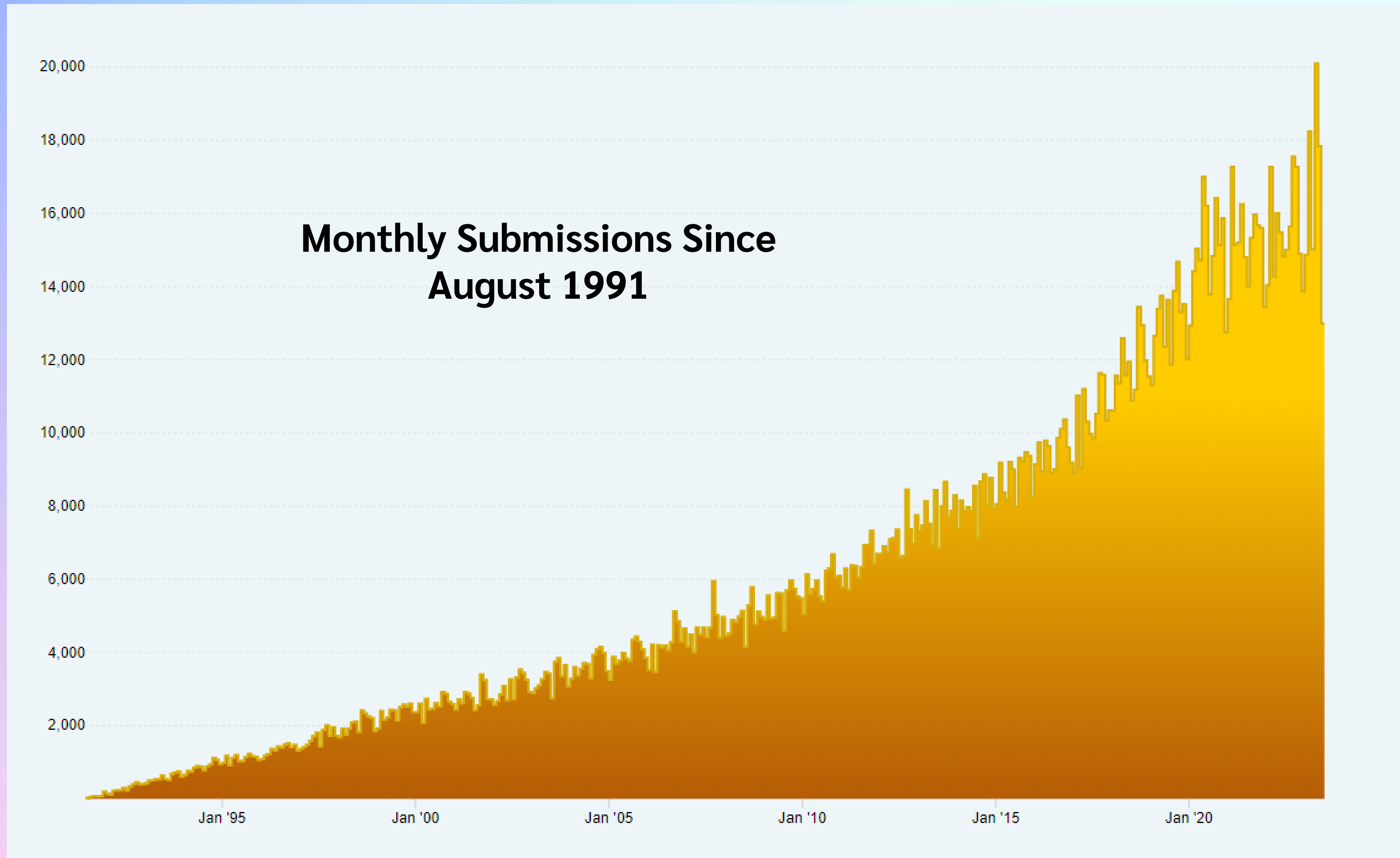
No More Endless Searching - PaperMate Delivers the Perfect Papers!



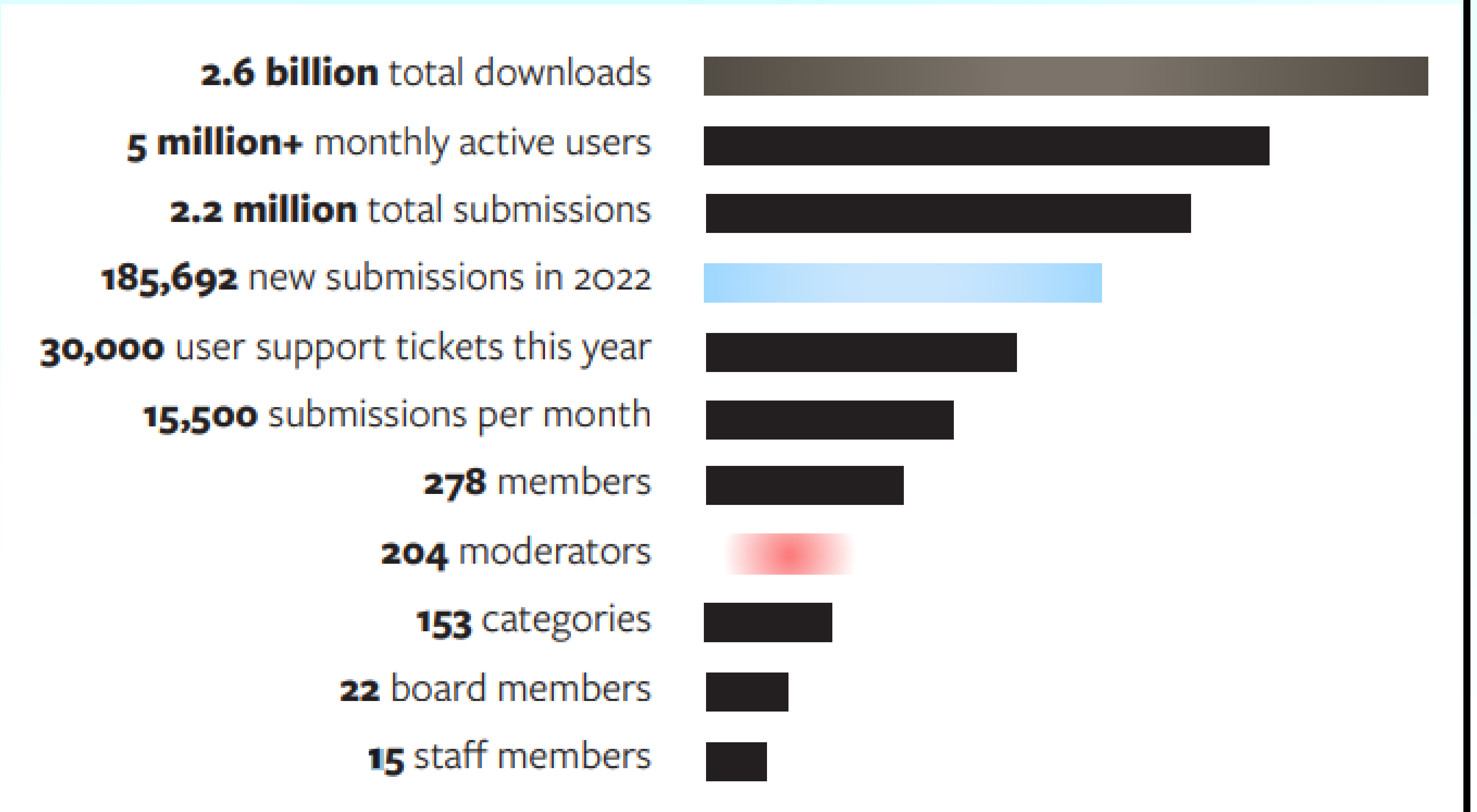
PROBLEM STATEMENT

A huge number of academic papers are coming out from a lot of conferences and journals these days. In these circumstances, It is hard for people who want to learn about it, such as researchers, students, or hobbyists, to find the papers that meet their needs and preferences. To ease this difficulty, I propose **PaperMate**, a tool that facilitates the discovery of relevant and useful machine learning papers , that may be interesting to her/him. *I have also been interested in machine learning , deep learning & NLP for a long time, and I wanted to keep up with the latest research and developments in this field. I realized that there is a need for a tool that can help people like me find machine learning papers that match their interests. That is why I decided to work on this project.*

YEARLY PAPER PUBLICATION ON ARXIV



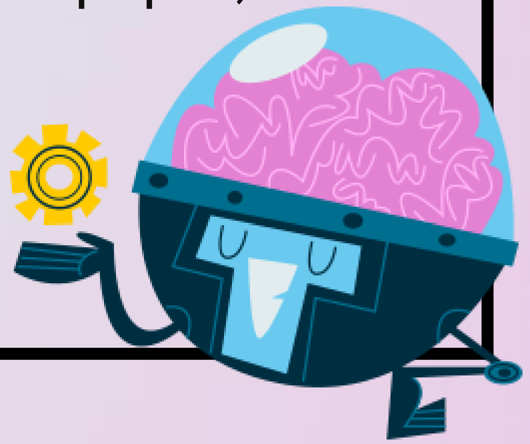
The number of new submissions received during each month since August 1991 (after 32.0 years). Hover over the graph to see the exact count for a given month.



arXiv in Numbers

IDENTIFYING MACHINE LEARNING PROBLEM

- The main machine learning problem that we want to tackle here is the recommendation of research papers. We can formulate this problem as a content-based recommendation system, where our aim is to recommend research papers that are suitable and helpful for a user's specific interests. To do this, we need to look at the content of the papers and the user's interests, and see how similar or related they are.
- We can use natural language processing techniques to analyze the text of the papers and the user's interests, and measure their similarity or relatedness. The output of the system is a list of research papers that match the user's interests, ranked by their similarity or relatedness. The system also provides some information about each paper, such as the title, the authors, the abstract, the keywords, the citations, and the related papers.



The Solution: **PaperMate**

PaperMate is a tool that helps people find machine learning papers that match their interests , It uses NLP to analyze the content of the papers and the user’s interests, and then measure how similar or related they are . It recommends the top 4 most similar papers for each interest, along with their abstract and links to the full text. It updates itself every month with new data from arXiv website . It helps users keep up with the latest research, learn new skills and knowledge, and find inspiration for their own projects .

Project Approach and Methodology

Project Approach and Methodology The project approach is to use NLP to recommend machine learning papers based on the user’s interests. **The project methodology has these steps:**

1. **Data collection:** I scrape machine learning papers from the arXiv website.
2. **Data preprocessing**: I clean and normalize the text data and make a user profile for each user.
3. **Feature extraction**: I use a TF-IDF vectorizer to change the text data into numerical features.
4. **Similarity computation:** I use a cosine similarity function to find the similarity between each paper and each user’s profile.
5. **Recommendation generation**: I use a collaborative filtering method to recommend the top 4 most similar papers for each user’s interest, with summaries and links.
6. **Integration with Hopsworks**: I use Hopsworks, a data platform for ML with a Feature Store and MLOp's capabilities. Hopsworks helps me store, manage, govern, and serve my features and models. I use Hopsworks’ feature store, model registry, and vector database.
7. **Evaluation**: I evaluate my system using different metrics and user feedback.

PLAN

1. ArXiv papers data is scrape from the arXiv website.

2. EDA, Data Processing, and Feature Engineering are used to produce the best text data for the embedding process.

3. Hugging Face Sentence Transformer is used to embed the text data.

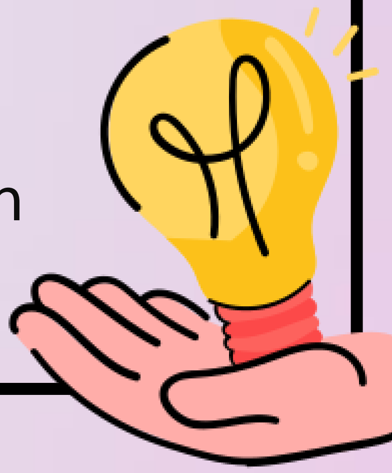
4. The similarity between papers is calculated using cosine similarity, once by comparing titles and once by comparing abstracts.

5. Once similarity scores are calculated, the top 4 most similar papers will be recommended to the user, along with their abstracts and links to the full text.

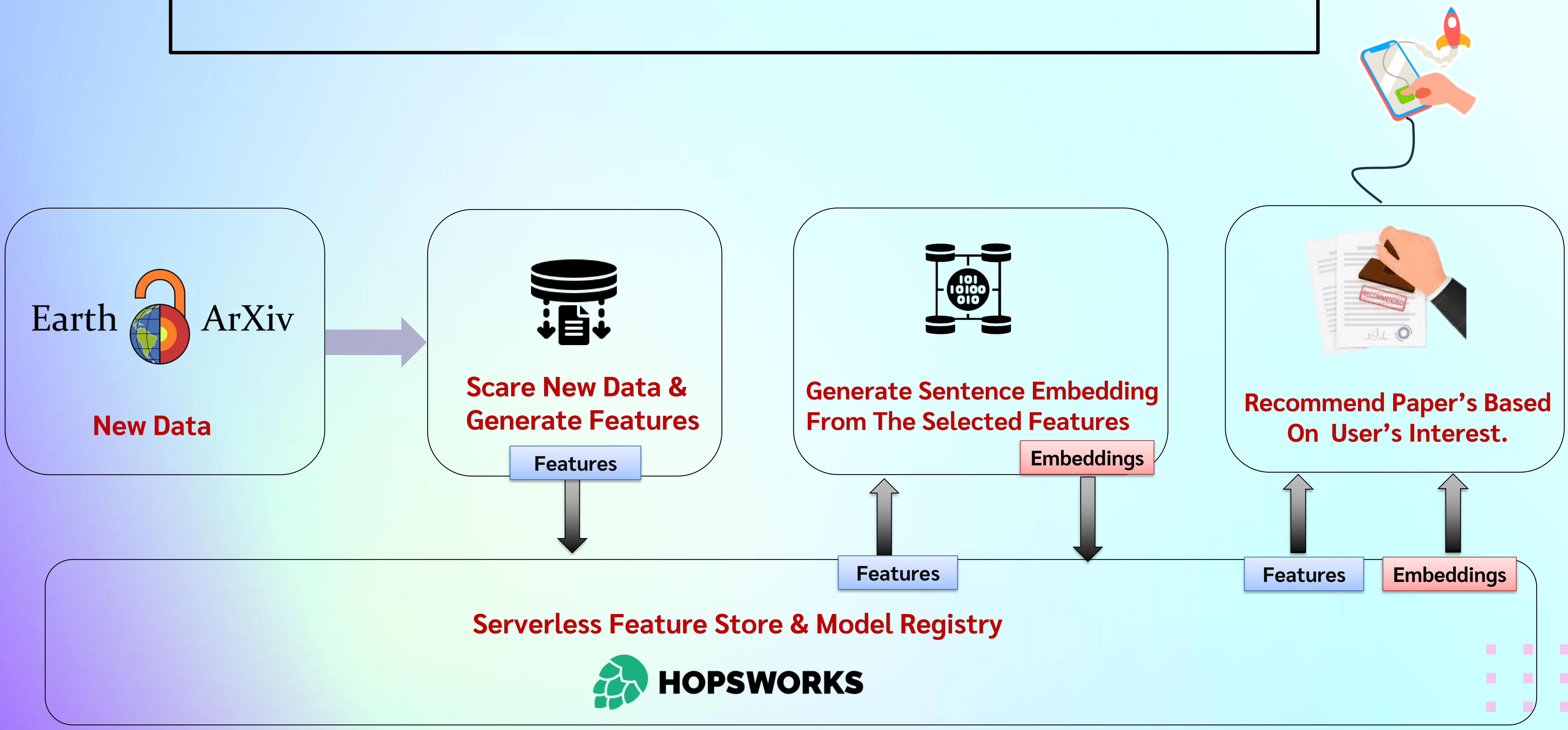
6. Pipelines will be set up to scrape new papers from the arXiv website every month and re-embed the new data collected when the pipeline is triggered.

7. Deploy the applocation.

8. Evaluation will be done using the Trubrics feedback component , enabling the collection of user feedback on paper recommendations.



STRUCTURE OVERVIEW



Earth ArXiv

New Data

Scare New Data & Generate Features

Features

Generate Sentence Embedding From The Selected Features

Embeddings

Recommend Paper’s Based On User’s Interest.

Features

Embeddings

Features

Serverless Feature Store & Model Registry

HOPSWORKS

Goals for the Week

1

1. Data Collection

2. Data Pre-Processing

2.1 EDA

2.2 Data Cleaning

2.3 Feature Engineering

Understand Docker

2

1. Integrate with Hopsworks.

2. Embed the text data.

3. Calculate Similarity.

4. Notebook Experiment.

Understand CI/CD pipelines & Git-hub actions & Heruku

3

1. Set the pipeline (monthly Extraction –arxiv data)

2. Complete th work of front-end.

3. Evaluation.

4. Deploy

Thank you for your time and attention