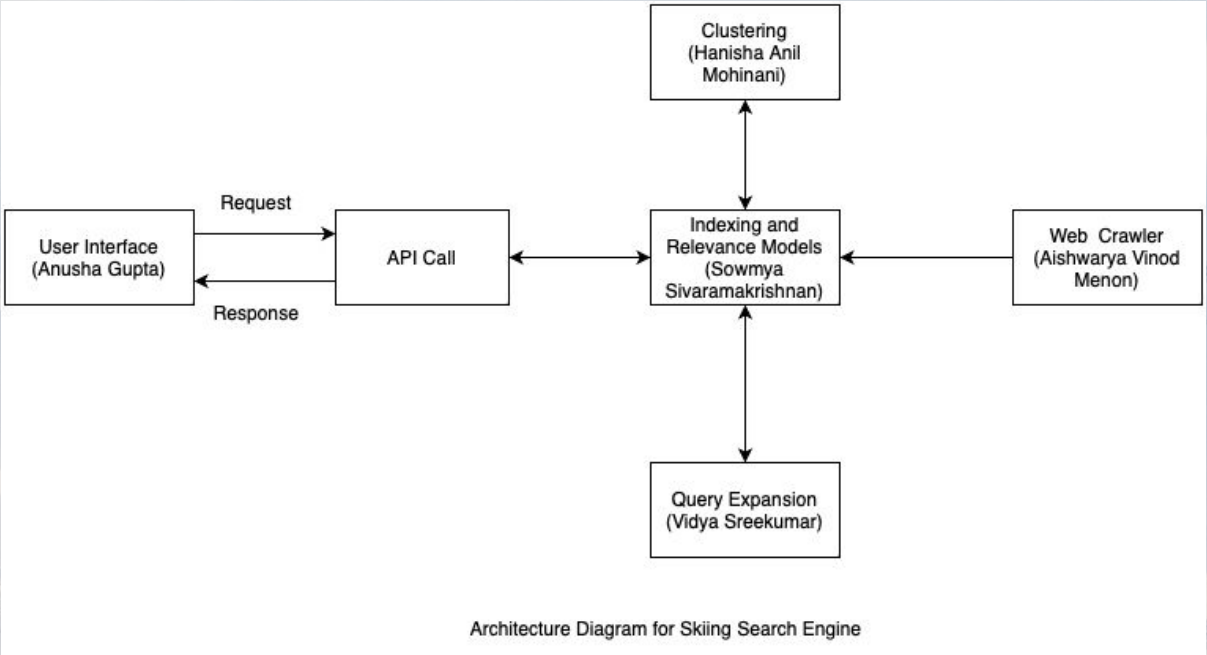
# Skiing Search Engine

## Team 5


**Aishwarya Vinod Menon(AXV220062)**
**Sowmya Sivaramakrishnan(SXS230043)**
**Anusha Gupta(AXG230026)**
**Hanisha Anil Mohinani(HXM220089)**
**Vidya Sreekumar(VXS220066)**

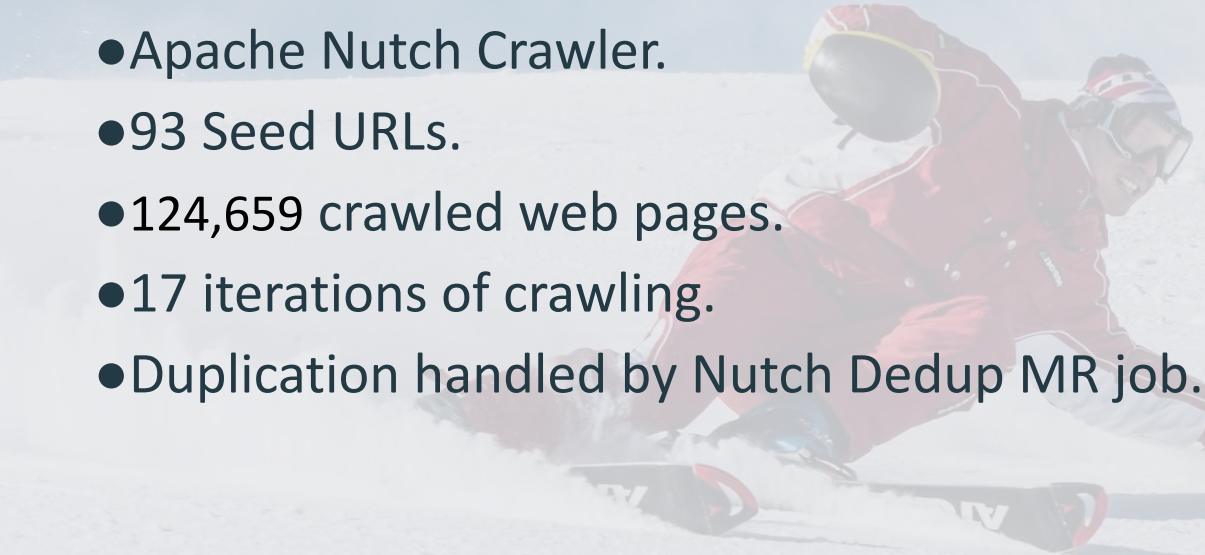

# Architecture Diagram of the System




Architecture Diagram for Skiing Search Engine

# CRAWLING (Aishwarya Vinod Menon AXV220062)

- Apache Nutch Crawler.
- 93 Seed URLs.
- 124,659 crawled web pages.
- 17 iterations of crawling.
- Duplication handled by Nutch Dedup MR job.



# INDEXING AND RELEVANCE MODELS

# (Sowmya Sivaramakrishnan SXS230043)

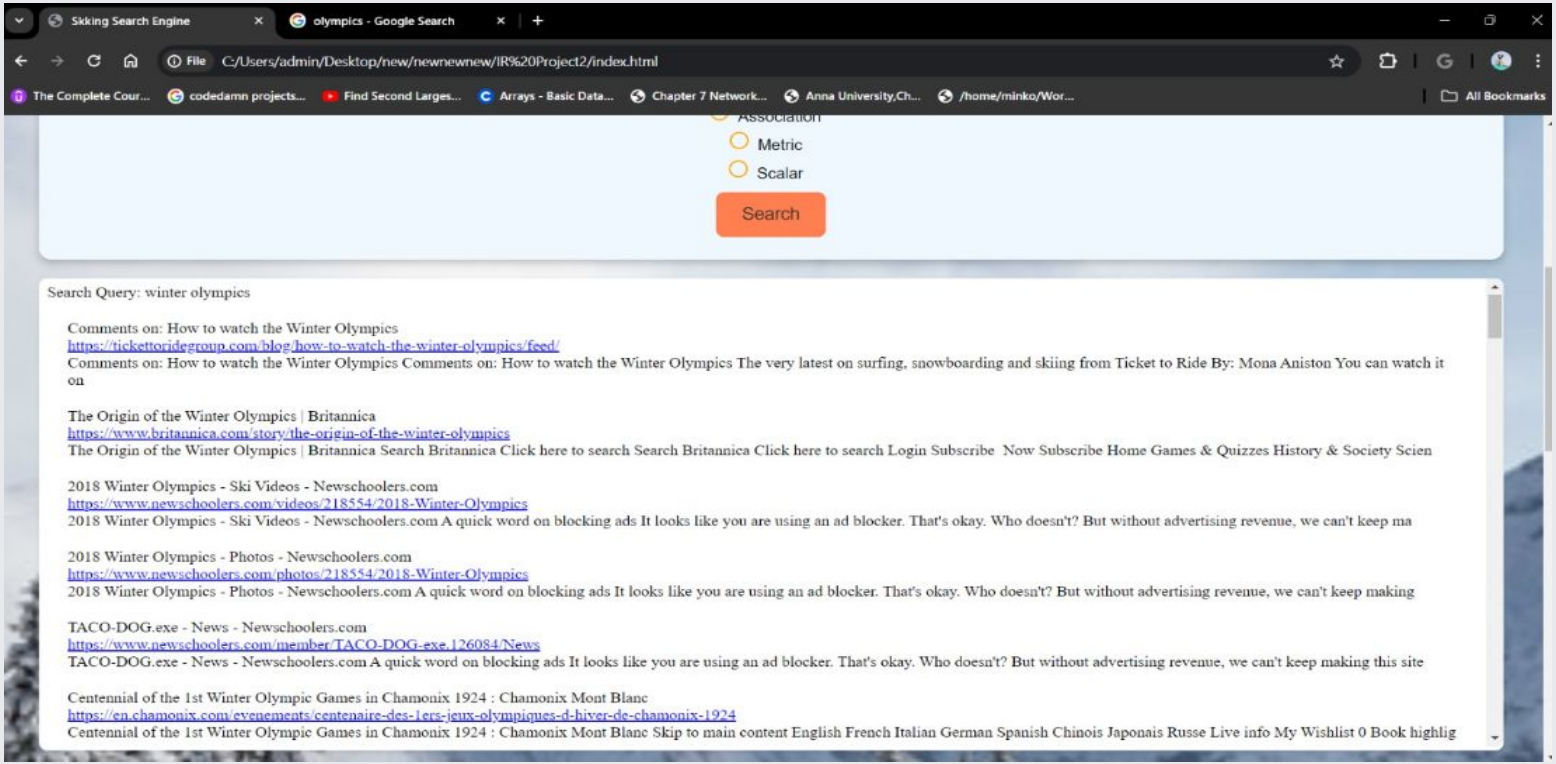


- Apache Solr is used for indexing.
- Vector Space model implemented while indexing in Solr.
- Webgraph created using nutch.
- Page Rank algorithm generated using nutch Link Rank command.
- HITS algorithm implemented using python library (networkx).
- 50 queries were used for testing the relevance models.

# WEB UI and BACKEND (Anusha Gupta AXG230026)

- Used HTML, CSS and Javascript for the front end and Flask for the backend.
- Shows results from Skiing Search Engine, Google and Bing.
- Backend connects with Solr API, preprocess query, gets Solr results,and calls relevant ranking model with Solr results as parameters.
- Tested 10 queries for all the models.

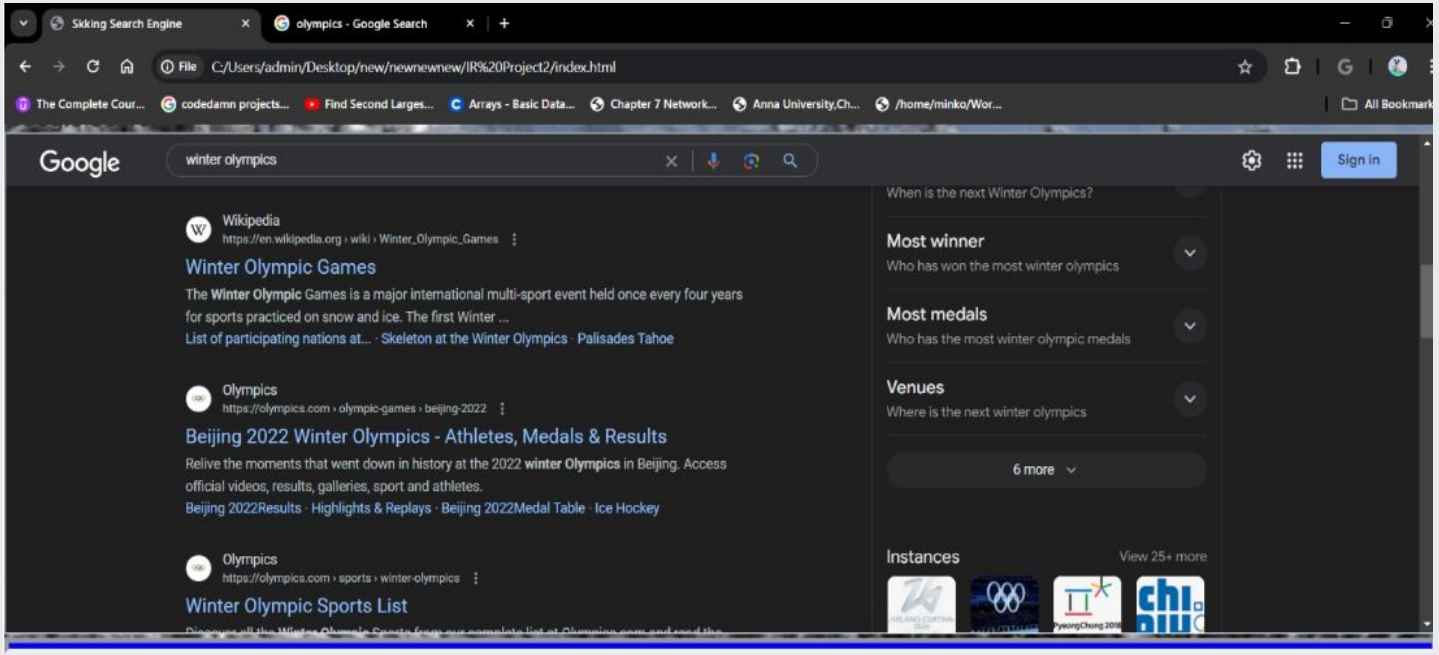
# Winter Olympics with Page Rank




Skiing Search Engine Results

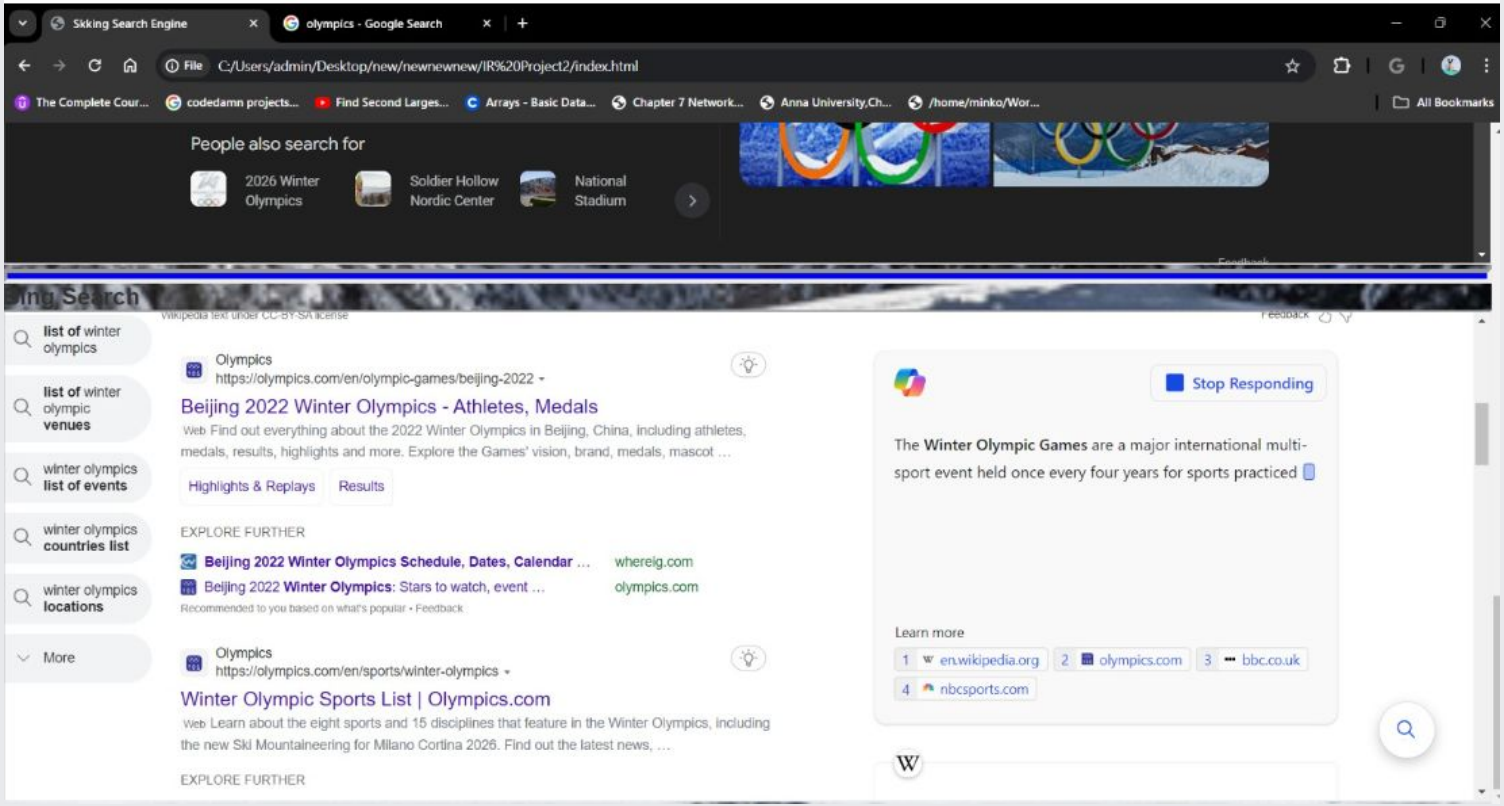
# Winter Olympics with Page Rank

I




Google Search Results

# Winter Olympics with Page Rank



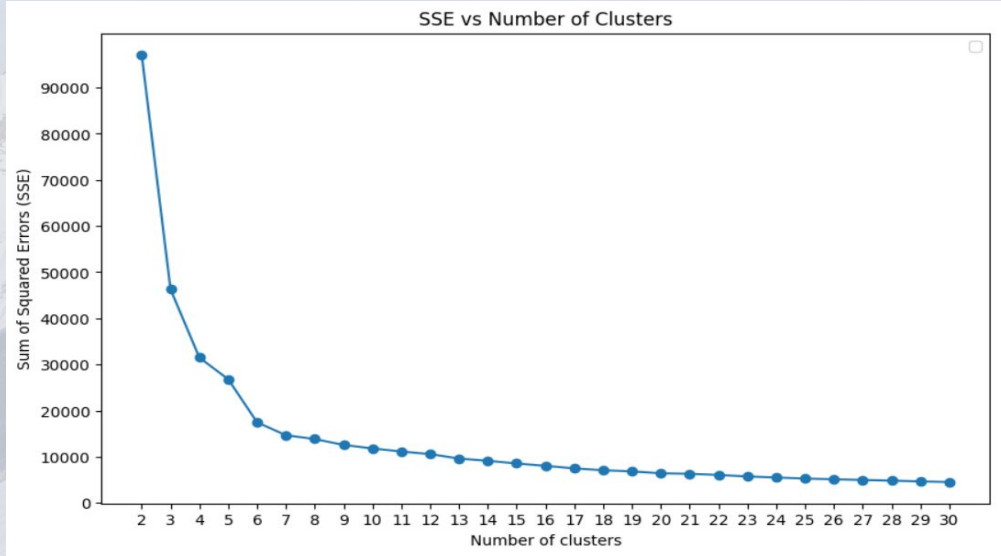

Bing Search Results

# Clustering (Hanisha Anil Mohinani HXM220089)

Solr Results -> TF-IDF Vectorization -> Removing stop words from content and title

## 1. Flat Clustering

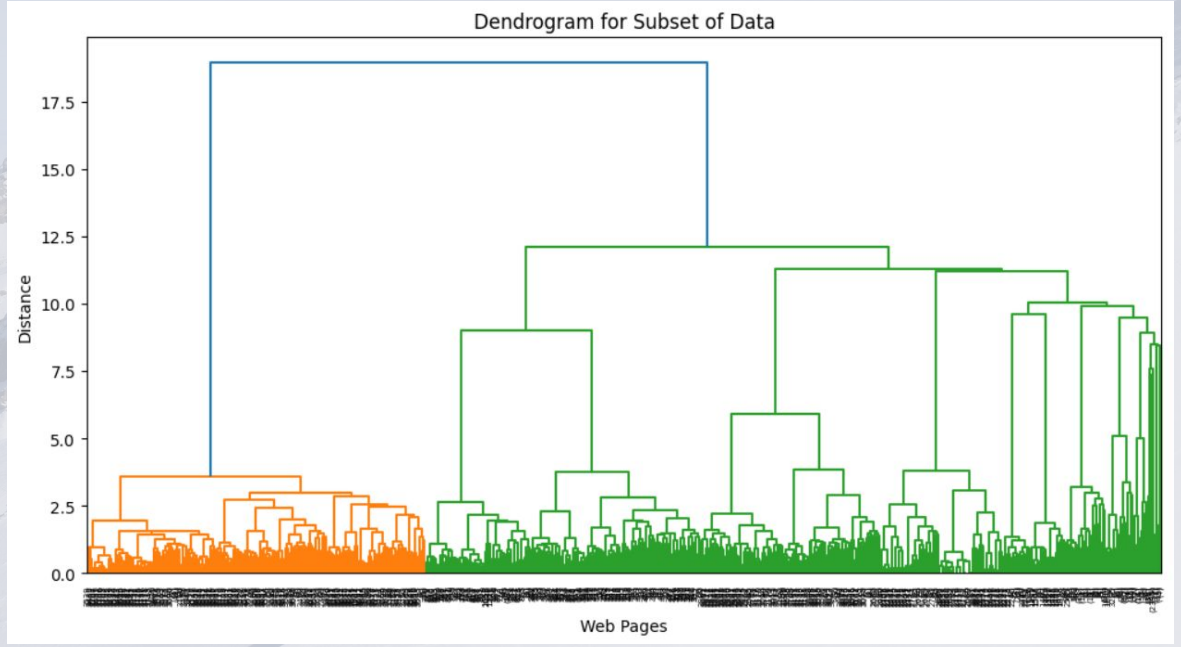
- K-means algorithm
- Elbow method to find number of clusters (Sum of Squared Errors vs Number of Clusters)
- Number of clusters = 7



# Clustering (Hanisha Anil Mohinani HXM220089)

## 2. Agglomerative Clustering

- Implemented using single(minimum distance between the two clusters) and average( average distance between the two clusters) linkage methods
- Plot dendrogram to find number of clusters
- Number of clusters = 8



# Query Expansion (Vidya Sreekumar VXS220066)

- Implemented Pseudo relevance feedback using Association, Metric and Scalar clustering- based query expansion, adding 3 words to the original query to form the modified query.
- Tested it with over  50 queries and found the expanded queries.
- Implemented the Rocchio Algorithm for relevance feedback.
- Generated Roccio based expanded query for 20 queries based on the relevant and irrelevant document set.

# Demo

Queries for demo:

1. Skiing Resort
2. Winter Olympics
3. Skiing Gear
4. Ski Terrain
5. Avalanche Safety


THANK YOU!