***The University of Texas at Dallas***

***CS 6322 Information Retrieval Spring 2024***

***Class Project Proposal***

*Project TITLE: Search Engine for Skiing*

*Group: No 5*

*Students:*

*Aishwarya Vinod Menon,*
*axv220062@utdallas.edu*

*Anusha Gupta,*
*axg230026@utdallas.edu*

*Hanisha Anil Mohinani,*
*hxm220089@utdallas.edu*

*Sowmya Sivaramakrishnan,*
*sxs230043@utdallas.edu*

*Vidya Sreekumar,*
*vxs220066@utdallas.edu*

Aishwarya Vinod Menon (AXV220062), Sowmya Sivaramakrishnan (SXS230043), Anusha Gupta (AXG230026), Hanisha Anil Mohinani (HXM220089), Vidya Sreekumar (VXS22066)

## ***1. The Problem***

To develop a search engine for skiing with data of over 1,00,000 crawled web pages. The crawling was performed and a webgraph was obtained, the data from which gets indexed. Two relevance models namely page rank and HITS are created for ranking the search results after which flat clustering and agglomerative hierarchical clustering is performed. These improve the search results and then query expansion is performed.

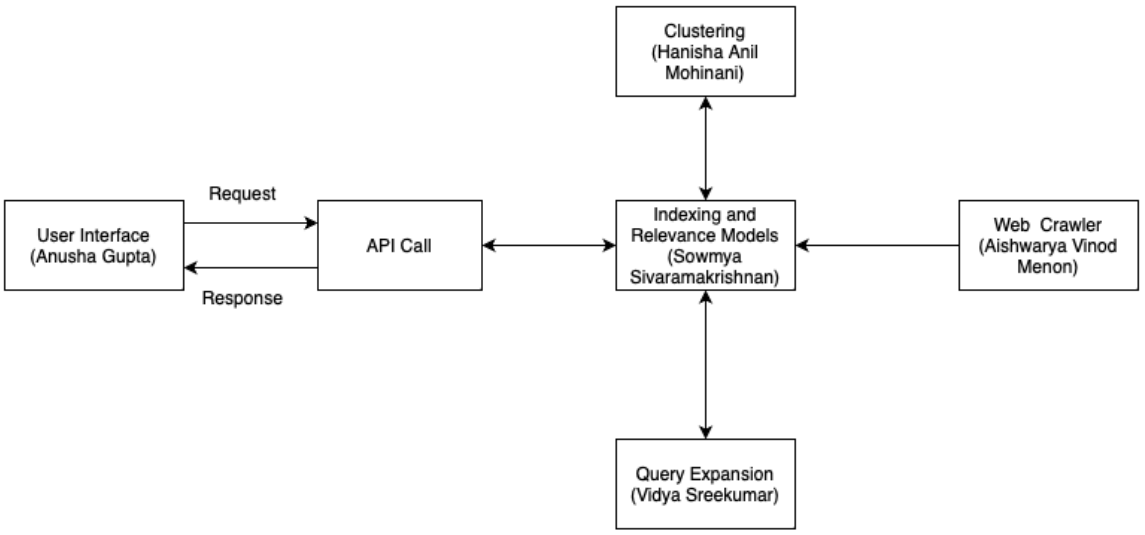
To summarize, we developed an advanced search engine which can be queried to return data about skiing, skiing gear, resorts for skiing, ski festivals, skiing locations etc. using multiple types of relevance models,

clustering techniques and query expansion methodologies, the effectiveness of which are evaluated in the experimental results.

We have divided the development into 5 parts based on different aspects of the project:

- Aishwarya Vinod Menon (Crawling)
- Sowmya Sivaramakrishnan (Indexing and relevance models)
- Anusha Gupta (User Interface)
- Hanisha Anil Mohinani (Clustering)
- Vidya Sreekumar (Query Expansion)

## Architecture




Architecture Diagram for Skiing Search Engine

Aishwarya Vinod Menon (AXV220062), Sowmya Sivaramakrishnan (SXS230043), Anusha Gupta (AXG230026), Hanisha Anil Mohinani (HXM220089), Vidya Sreekumar (VXS22066)

## Learnings

- Understanding Information Retrieval Concepts: Developing a search engine requires a solid grasp of information retrieval concepts such as indexing, querying, relevance ranking, and evaluation metrics. Through the process of building a search engine we gained a lot of practical insights into how these concepts are applied in real-world scenarios.
- Hands-on Experience with Search Engine Technologies: Building a search engine involves working with various technologies and tools commonly used in the field, such as web crawlers (eg: Nutch), indexing algorithms like page rank and HITS algorithms, query processing techniques, and relevance ranking algorithms. Experience with these technologies is invaluable for us.
- Overcoming Challenges and Trade-offs: Building a search engine involves making numerous design decisions and trade-offs to balance factors such as search quality, efficiency, scalability, and user experience. We learnt how to overcome the challenges involved in designing and implementing effective search systems, including dealing with large-scale data, optimizing performance, handling query intent ambiguity, and addressing user feedback.

## Experience

Working as a project team to develop the search engine was difficult but incredibly gratifying. Collaborating on a project of this nature allowed us to explore the intricacies involved in information retrieval, ranging from developing web indexing algorithms to putting relevance categorization and UI elements into practice. Together, we overcame several technological obstacles, such managing big datasets, optimizing search performance, and debugging intricate code. Nevertheless, overcoming these challenges gave our group a strong feeling of collaboration and improved our problem-solving abilities. Furthermore, it was immensely satisfying to witness the search engine's slow development from a notion to a tool that can find pertinent information on skiing.

## Challenges

Developing a search engine with 170,642 indexed web pages and incorporating multiple clustering techniques and relevance models presents several significant challenges.

First and foremost is the sheer scale of the data. Processing and indexing such a large number of web pages requires a solid infrastructure and efficient algorithms to ensure timely search responses. Managing the storage and retrieval of these pages

while maintaining retrieval performance can strain computing resources and require careful optimization.

Second, the integration of multiple clustering techniques increases system complexity. Each clustering method may have its own parameters, assumptions, and computational requirements. Coordinating these techniques to effectively organize search results can be difficult, especially when considering different types of ski-related content such as gear, resorts, techniques and news.

Additionally, incorporating different relevance models adds complexity. Relevance models determine how search results are ranked and presented to users based on their query intent. Developing and refining these models to precisely match user preferences and information needs for ski-related queries requires extensive testing and evaluation.In addition, ensuring the accuracy and relevance of search results is a major challenge. The quality, relevance and reliability of ski-related content can vary greatly from site to site. Filtering out irrelevant or low-quality content and promoting authoritative and useful resources requires complex ranking algorithms and constant monitoring and adjustment.

## ***2. Crawling***

Web crawling is the process of automatically collecting data from websites. To do this we use web crawlers or spiders which search the internet and get the relevant websites related to the one we are crawling.

First, we provide a list of seed URLs to the web crawlers and then the crawlers visit the hyperlinks mentioned in those websites. Web crawlers continue this process iteratively until we have visited all the seed URLs.

Web crawlers can collect text, images, hyperlinks, videos and metadata from these sites and we can use these to build Search Engine, SEO optimization, Price Monitoring and Market Research.

An open source framework called Apache Nutch is used to create search engines and index the web. This targets a collection of seed URLs that serve as the foundation for research and starts the indexing process. Nutch uses its search functions to effectively get the related web pages by sending HTTP requests to these URLs. During the search process, the framework looks at these pages' content structures and uses a variety of plugins to extract useful data including text, links, and metadata.

A total of 93 seed URLs were gathered by visiting relevant skiing sites and pasted into a text file and passed to Nutch, from which we were able to generate 124659 webpages.

### Seed URLs

https://www.sunvalley.com/
https://www.brettonwoods.com/
https://en.wikipedia.org/wiki/Skiing
https://www.skicanada.org/ready/what-is-skiing/
https://www.britannica.com/sports/skiing
https://www.skimag.com/
https://www.onthesnow.com/
https://www.theskibum.com/ski-snow/ski-equipment
https://www.sunandski.com/c/snow-ski-equipment
https://shop.rockymountainskiandboard.com/
https://www.blackdiamondequipment.com/en_US/shop/ski-snowboard/
https://www.amazon.com/ski-gear/b?ie=UTF8&node=2342470011
https://freeskier.com/
https://thesnowmag.com/
https://www.fall-line.co.uk/

https://magazines.feedspot.com/ski_magazines/
https://alltracksacademy.com/blog/the-top-5-skiers-in-the-world/
https://snowbrains.com/the-10-greatest-alpine-skiers-of-all-time/
https://www.sportsmanskihaus.com/
https://www.skitalk.com/forums/
https://snowheads.com/ski-forum/
https://www.newschoolers.com/forum
https://www.tetongravity.com/forums/forumdisplay.php/3-Ski-Snowboard
https://www.the-house.com/ski
https://justacoloradogal.com/skier-slang-decoded/
https://www.nbs.org/
https://en.wikipedia.org/wiki/Alpine_skiing
https://tickettoridegroup.com/blog/why-skiers-snowboarders-make-the-best-mates/
https://planetski.eu/
https://www.yardbarker.com/skiing
https://www.snowindustrynews.com/
https://www.eurosport.com/alpine-skiing/
https://apnews.com/hub/skiing
https://www.rei.com/learn/expert-advice/how-to-ski.htmlhttps://www.sunvalley.com/
https://www.brettonwoods.com/
https://en.wikipedia.org/wiki/Skiing
https://www.skicanada.org/ready/what-is-skiing/
https://www.britannica.com/sports/skiing
https://www.skimag.com/
https://www.onthesnow.com/
https://www.theskibum.com/ski-snow/ski-equipment
https://www.sunandski.com/c/snow-ski-equipment
https://shop.rockymountainskiandboard.com/
https://www.blackdiamondequipment.com/en_US/shop/ski-snowboard/
https://www.amazon.com/ski-gear/b?ie=UTF8&node=2342470011
https://freeskier.com/
https://thesnowmag.com/
https://www.fall-line.co.uk/
https://magazines.feedspot.com/ski_magazines/
https://alltracksacademy.com/blog/the-top-5-skiers-in-the-world/
https://snowbrains.com/the-10-greatest-alpine-skiers-of-all-time/
https://www.sportsmanskihaus.com/
https://www.skitalk.com/forums/
https://snowheads.com/ski-forum/
https://www.newschoolers.com/forum
https://www.tetongravity.com/forums/forumdisplay.php/3-Ski-Snowboard
https://www.the-house.com/ski
https://justacoloradogal.com/skier-slang-decoded/
https://www.nbs.org/

https://en.wikipedia.org/wiki/Alpine_skiing
https://tickettoridegroup.com/blog/why-skiers-snowboarders-make-the-best-mates/
https://planetski.eu/
https://www.yardbarker.com/skiing
https://www.snowindustrynews.com/
https://www.eurosport.com/alpine-skiing/
https://apnews.com/hub/skiing
https://www.rei.com/learn/expert-advice/how-to-ski.html
https://www.sundayriver.com/learning-to-ski-and-snowboard
https://thepointsguy.com/guide/right-age-start-kids-skiing/
https://www.googleadservices.com/pagead/aclk?sa=L&ai=DChcSEwjwmtKzwcqFAxXYONQBHX8ZCA4YABAAGgJvYQ&ase=2&gclid=CjwKCAjw5v2wBhBrEiwAXDDoJRukByqSP-Si7g9QM3tiMUTQBGNU0NpWQGj4lpR8vMJakPx70ju6RBoCrpEQAvD_BwE&ohost=www.google.com&cid=CAESVuD2fZ7SDkADMhPr5uRWRnSDbTIu1NOKEClsjKABYAP84eCyeizXRhrcC9yX-bS4CZ0sdLyiSUqecWEymh3diV6d8Mct7F3RULNCltFKL-QQ_kfMIuaa&sig=AOD64_29QpNhaonG5owqrNU4h9yspuhwAg&q&nis=4&adurl&ved=2ahUKEwjrxcmzwcqFAxVj6skDHQqhCSgQ0Qx6BAgJEAE
https://unofficialnetworks.com/category/news-3/
https://fasterskier.com/
https://www.skiutah.com/blog/categories/news
https://www.denverpost.com/tag/skiing/
https://www.parkrecord.com/news/final-week-of-ski-season-begins-in-park-city/
https://www.forbes.com/sites/wendysachs/2019/01/14/women-are-transforming-the-ski-industry-on-the-slopes-and-in-the-boardroom/
https://www.ridestore.com/mag/female-skiers-to-follow-on-instagram/
https://www.theskidiva.com/forums/index.php?threads/skiings-gender-gap-widens-with-age.28264/
https://welove2ski.com/how-to-ski/top-olympic-male-skiers/
https://en.wikipedia.org/wiki/List_of_FIS_Alpine_Ski_World_Cup_men%27s_race_winners
https://www.ranker.com/list/famous-male-alpine-skiers/reference
https://www.ranker.com/list/famous-male-alpine-skiers/reference
https://www.onthesnow.com/
https://travel.usnews.com/rankings/best-ski-destinations/
https://www.forbes.com/uk/advisor/travel-insurance/worlds-best-ski-resorts/
https://www.travelandleisure.com/best-ski-resorts-in-the-world-8406637
https://www.lakeplacid.com/do/outdoors/winterspring/skiing-riding
https://www.denverpost.com/tag/ski-accidents/
https://www.lonestarskiers.com/
https://www.metroplexskiclub.com/
https://www.texasskirangers.org/
https://www.dallasskiclub.org/content.aspx?page_id=22&club_id=592979&module_id=177125

https://en.wikipedia.org/wiki/Cross-country_skiing
https://www.curated.com/journal/20000/a-guide-to-the-different-kinds-of-skis
https://buckmans.com/blog/the-complete-guide-to-different-types-of-skis-bpid_445.aspx
https://www.mountainheaven.co.uk/blog/an-introduction-to-different-types-of-skis/
https://www.onthesnow.com/news/which-ski-is-right-for-you/
https://www.rei.com/learn/expert-advice/alpine-ski-bindings.html
https://www.undercovertourist.com/blog/best-time-ski/
https://www.thorindustries.com/stories/skiing-north-america-year-round

## Nutch Crawling Process

The following steps are followed for crawling:

1. Collecting seed URL list and inject it to CrawlDB database.
2. Produce a fetch list from the CrawlDB database and upload this fetch list in a new segment directory.
3. On this new segment directory, fetch the data from the seed URLs.
4. Parse the contents of the seed URLs.
5. Update CrawlDB with the new set of URLs.
6. Repeat steps 2-5 for N number of iterations to get the desired number of web pages crawled. (For our project, we crawled 124,659 webpages).

## Components of Nutch Data

CrawlDB contains information about all the URLs, including whether it was fetched or not.
LinkDB contains the links of the known seed URLs which includes the source URL and anchor text of the link.
Segments are a set of URLs fetched together as a unit. It consists of

- crawl_generate - Set of URLs that needs to be fetched.
- crawl_fetch - fetch status of each URL
- Content - raw content got each URL
- parse_text - parsed text of each URL
- parse_data - outlinks and metadata parsed from each URL
- crawl_parse - outlink URLs which are used to update the crawlDB

Below is the Nutch command we used to automate the crawling process to get 124,659 webpages as a result iteratively for 17 iterations.

bin/crawl -i -D solr.server.url=http://localhost:8983/solr/nutch -s urls crawl 17

## Duplicate Results Handling

The CrawlDB component of the Nutch can easily detect if the URL is already parsed during the update CrawlDB command execution after each iteration. For the duplicate results which have similar content but different URL is handled by Nutch script which internally runs the “dedup” MapReduce job which marks these results in the CrawlDB. These marked results are later deleted in the Solr during the indexing

process. Shortest URL, highest score or most recently fetched are some of the heuristics we can use to choose the item which is not marked as duplicate.

## Collaboration with Indexing and Relevance Models

The process involved setting up the crawling infrastructure, which included the installation and configuration of Nutch and Solr, as they are interconnected components. The functionality of Nutch's automated script, linked with Solr for indexing, was comprehensively discussed. Input of seed URLs and the corresponding output were reviewed collaboratively. Remote access to the server facilitated joint exploration and understanding of the automation script provided by Nutch, which manages the indexing process through Solr.

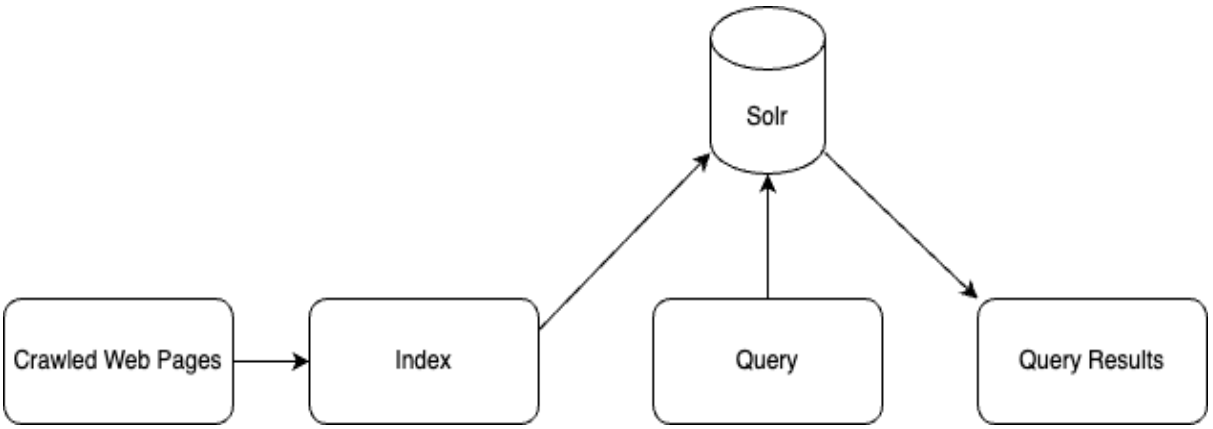
## ***3. Indexing and Relevance***

After performing the web crawling with apache nutch, we then proceeded to index the crawled pages using Apache Solr. We configured nutch to internally use solr for indexing and created the index using the following command:

**bin/nutch index crawl/crawldb/ -linkdb crawl/linkdb -dir crawl/segments/ -filter -normalize -deleteGone**

In the above command, bin/nutch points to the binary of apache nutch and its an indication that the nutch tool is being utilised to run the command. This is followed by the keyword index which is essentially the command instructing nutch to index the crawled data. We then provide the paths to crawldb- a database storing information about crawled web pages, such as their URLs, metadata, and status and to linkdb- which contains data about the links between pages. The directory where the segments to be indexed are stored is added next in the command. The -filter option indicates that you want to apply any configured URL filters during the indexing process. URL filters can be used to include or exclude certain URLs based on patterns or criteria. -normalize tells Nutch to normalize URLs during the indexing process. URL normalization ensures that URLs are standardized and consistent, which can help improve indexing and reduce duplication of content. Lastly, -deleteGone option instructs Nutch to delete any documents from the index that are marked as "gone" in the crawldb. "Gone" typically refers to web pages that were previously crawled but are no longer accessible or valid.

The diagram below represents flow in the process of assembling an index:




Next we move onto creating web graphs. Creating a web graph in this context helps in understanding and analyzing the structure and organization of content on the websites better. It also facilitates detection of updates made to the pages by drawing comparisons with previously obtained link structure. Web crawler was used to crawl the web pages and extract their urls which are then used by us to create a web

graph. The web pages are each represented as a node of the web graph and the edges connecting these nodes represent the links between them.

The next course of action was to determine the relevance or importance of these web pages for which we explored three different models such as the vector space relevance model, page rank and HITS algorithm.

The first one was the Vector Space Relevance model which is provided by Solr itself. The VSM employed by Solr for information retrieval is pivotal in efficiently assessing the relevance of documents to user queries. Within this model, documents and queries are transformed into vectors within a high-dimensional space, with each dimension corresponding to a unique term in the document collection. Central to this approach is the Term Frequency-Inverse Document Frequency (TF-IDF) weighting scheme, which evaluates the importance of a term within a document relative to its frequency across the entire collection. Documents are represented as vectors, with each dimension reflecting the TF-IDF weight of the corresponding term, while queries are similarly transformed into vector representations based on their constituent terms and TF-IDF weights.

Once documents and queries are represented as vectors, Apache Solr calculates relevance scores using cosine similarity, measuring the cosine of the angle between the query vector and the document vectors. This comparison determines the degree of similarity between the query and each document, with higher cosine similarity scores indicating greater relevance. By ranking documents based on these similarity scores, Solr identifies the most relevant documents to return as search results. Through this Vector Space Model, Apache Solr leverages TF-IDF weighting and cosine similarity to efficiently and effectively retrieve information, facilitating accurate and meaningful search experiences for users across diverse applications.

The next approach was to use the PageRank algorithm for determining the relevance. This algorithm assigns a numerical weight to each element of a hyperlinked set of documents, with the purpose of measuring its relative importance within the set. PageRank interprets links between pages as endorsements or votes for the linked page's importance, with more influential pages receiving higher scores. This iterative algorithm considers the entire link structure of the web, providing a nuanced understanding of relevance beyond individual page content. The built-in nutch webgraph command was employed to create the webgraph from crawled URLs.

**bin/nutch webgraph -filter -normalize -segmentDir crawl/segments/ -webgraphdb crawl/**

The PageRank algorithm runs repeatedly until the point of convergence and this was done with the following command:

**bin/nutch linkrank -webgraphdb crawl/**

With a damping factor of 0.80, we got the page rank algorithm to assign scores to the web pages which was then updated in the crawled database and indexed using Solr. The score update was done with the help of the following command:

**bin/nutch scoreupdater -crawldb crawl/crawldb -webgraphdb crawl/**

This is followed by the indexing as mentioned above which is done with the following command:

**bin/nutch index crawl/crawldb/ -linkdb crawl/linkdb -dir crawl/segments/ -filter -normalize -deleteGone**

The last algorithm we used was the HITS algorithm for creating a relevance model. This works on determining the relevance on the basis of the authority and hub scores. We wrote our own code to generate these scores based on the inlinks and outlinks we'd obtained earlier for all the web pages. We first identified the most relevant web pages which become our roots, these were identified based on the queries provided by the user. Pages linking to these roots are called hubs while the pages being linked to by the roots are called authorities. We collected a root set of web pages based on page ranks and this was then converted into a base set by adding the inlinks and outlinks. Then we calculated the HIT scores and sorted the max authority scores.

To summarize, different models were used to find the relevance. While each one had its own set of strengths and limitations, we used them specific to the context of our project. The purpose of these relevance models was to give accurate results when the user queries our engine.

## Web Graph & Statistics

We created a web graph from the crawled documents. The following command was used to create the web graph from all the crawled web pages:

## bin/nutch webgraph -segmentDir crawl/segments/ -webgraphdb crawl/webgraphdb

This command processes multiple segments of crawled data together and generates a web graph in different components to a given directory. The three components created are:

• InLink Database: List of URLs and all their inlinks.

• OutLink Database: List of URLs and their outlinks.
• Node Database: List of URLs with meta information including the number of inlinks and outlinks.

**Statistics**

Total Number of Links: 124659
Total Number of Nodes: 123668
Max number of incoming links: 44
Max number of outgoing links: 2

## Web graph information connected to the index

While creating the index in Solr, we have run the LinkRank program of Nutch to perform an iterative link analysis with the web graph. LinkRank is a PageRank-like link analysis program that converges to stable global scores for each URL. The LinkRank program starts with a common score for all URLs. It then creates a global score for each URL based on the number of incoming links and the scores for those links and the number of outgoing links from the page.

## Highest Hub and Authority Scores

Highest hub score- 0.0027258360565092206
URL with highest hub score- https://www.skiinfo.no/
Highest authority score- 0.0027258360565107728
URL with highest authority score- https://www.skiinfo.no/

## Topic Based Page Ranking

For the purpose of testing topic based page ranking, we chose the topic:
Skiing Jumps
The results from our search engine are as follows:
1. Skiing - Freestyle, Tricks, Jumps | Britannica
   https://www.britannica.com/sports/skiing/Freestyle-skiing
   Score- 0.21385658
2. Ski jumps and balls | SkiTalk | Ski reviews, Ski Selector
   https://www.skitalk.com/threads/ski-jumps-and-balls.33349/
   Score- 0.21375658
3. Jacob_W - Videos - Newschoolers.com
   https://www.newschoolers.com/member/Jacob-W.5247/Videos
   Score- 0.16821946
4. U.S. ski team's 5 essential ski training exercises - OnTheSnow
   https://www.onthesnow.com/news/us-ski-teams-training-exercises/
   Score- 0.15607622

5. SRAM jumps into the ebike powertrain game | SkiTalk | Ski reviews, Ski Selector
https://www.skitalk.com/threads/sram-jumps-into-the-ebike-powertrain-game.33579/
Score- 0.15000461

## Testing with Queries

The relevance models were tested using 60 queries which we obtained by looking into the most frequently searched queries with respect to skiing on google. Upon obtaining the search results we manually inspected the results by comparing the top results obtained for each query with their page rank or hits scores, depending on the model being tested, and the search results were both relevant to the query and ordered according to their relevance scores.

## Collaboration with UI and test relevance models results

We generated several queries to test the various relevance models implemented in this search engine. I have generated 60 queries and tested them on the UI by differentiating the results of PageRank and HITS relevance model results. Upon obtaining the search results we manually inspected the results by comparing the top results obtained for each query with their page rank or hits scores, depending on the model being tested, and the search results were both relevant to the query and ordered according to their relevance scores. The results were judged in some ways such as the HITS relevance model giving the most relevant results in comparison of page rank.

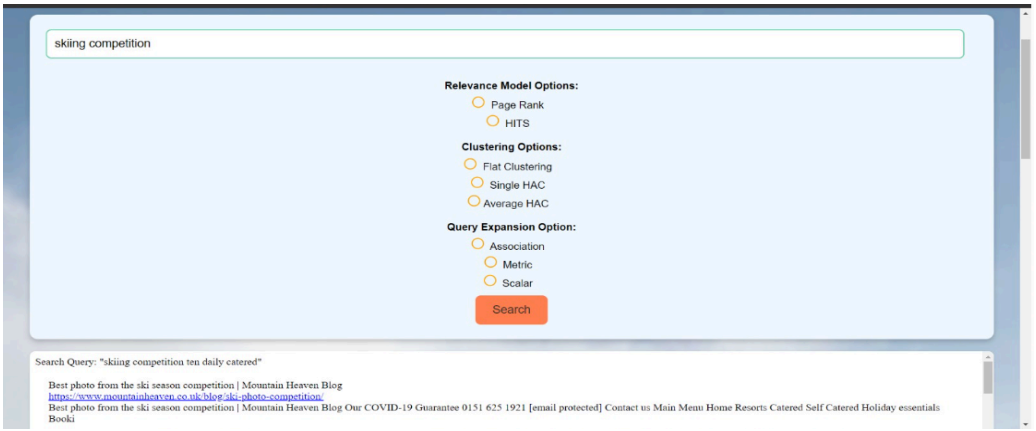
## Collaboration with Clustering

We collaborated with clustering to improve the results of the relevance models I have used. For all the above queries mentioned, clustering gives better results in comparison of page rank and HITS relevance models.

# ***4. User Interface and Comparison with Google and Bing:***

## Design of the Interface

I have used HTML, CSS and Javascript to create the front end of the search engine. Additionally, I have used the Flask API to connect it to the backend and then call relevant functions to display the results.




Front end consists of a search bar, ranking algorithm options,submit button, frames to display results of Skiing Search Engine, Google and Bing.

To search a query, the user has to do 3 steps:

- Enter the query in the search bar
- Select one of the following options to rank the result: Page Rank, Hits [Relevance Model]; Flat, Single HAC clustering, Average HAC clustering [Clustering Model Options]; Association, Metric and Scalar [Query Expansion Model options].
- Click on the Submit button.

When the user clicks on submit, a GET API is called with the searched query and selected ranking option as parameters to the flask backend app. The backend app connects to Solr to find relevant results. The results are then sent to the selected ranking model and then ranked accordingly. The ranked results in JSON format are returned and displayed in the front end. For the same query, Google and Bing results are also displayed in their respective frames.

I have also designed a backend app with Python and Flask that does the following:

1. It receives query and rank type parameters from the front end.

2. Preprocess the user query by removing stopwords and punctuations and sends it to Solr API.

3. Receives data from Solr API and calls appropriate ranking model to rank the Solr results.

4. JSONify the results and send it to the front end.

**Accessing the Relevance Model**

Sowmya was responsible for indexing and relevance models PageRank and HITS. Sowmya provided me PageRank and HITS scores for crawled URLs. In the backend, I have written code to sort the Solr results based on scores provided by Sowmya for PageRank and HITS relevance models.

**Queries Tested**

I tested 10 common queries related to skiing with Sowmya and checked if sorted results are being returned as per the scores she provided. Additionally, I tested 10 most searched skiing queries as per Google to check if all the ranking models are working correctly and correct data is being sent via API.

**Accessing the Clustering Model:**

Hanisha was responsible for clustering. She created functions for flat, Single HAC and Average HAC and I integrated it with the Flask application. The functions are called whenever the user chooses clustering model options. Accordingly results are returned and displayed on the front end. We both tested about 15 queries if it was returning relevant results.

**Comparison to Google and Bing:**

As compared to Google and Bing, our results are not that good. The primary reason for this is that Google and Bing have more computing resources which has allowed them to crawl much more webpages as compared to ours. Plus, they have gathered the user data since decades and improved upon the result. They also use realtime, state of the art algorithms and have high computational power to process the data giving far more accurate results.

**Clustering component:**

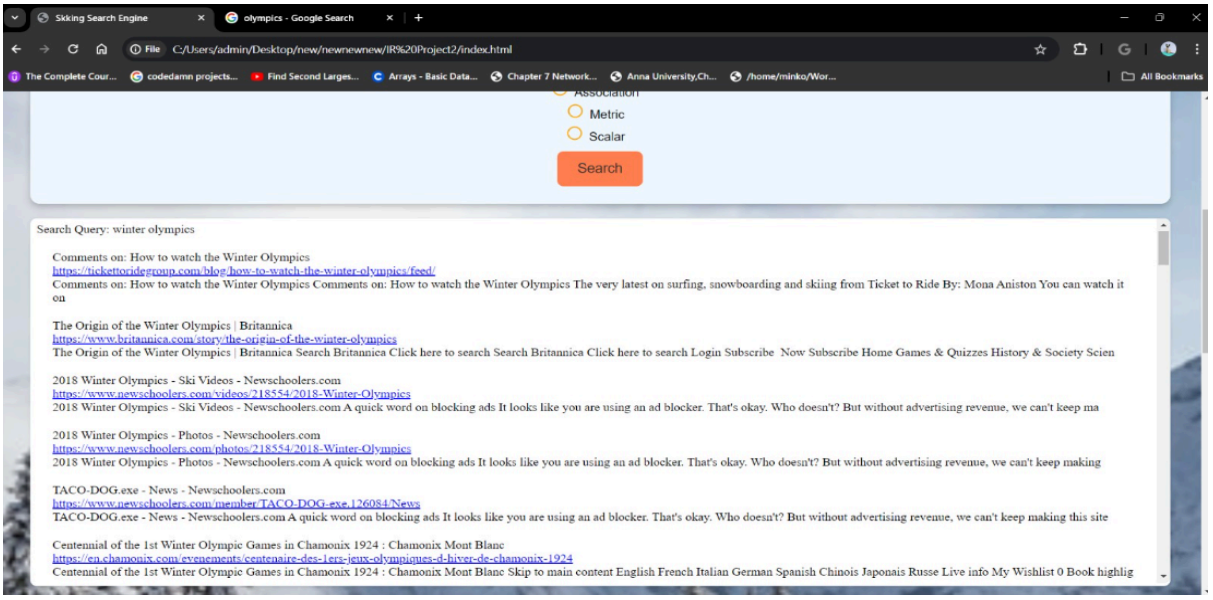
When the user chooses a clustering option, backend code calls a clustering algorithm and returns appropriate results to the front end. The clustering algorithm rearranges the result based on clusters containing top results.

**Selecting the queries:**

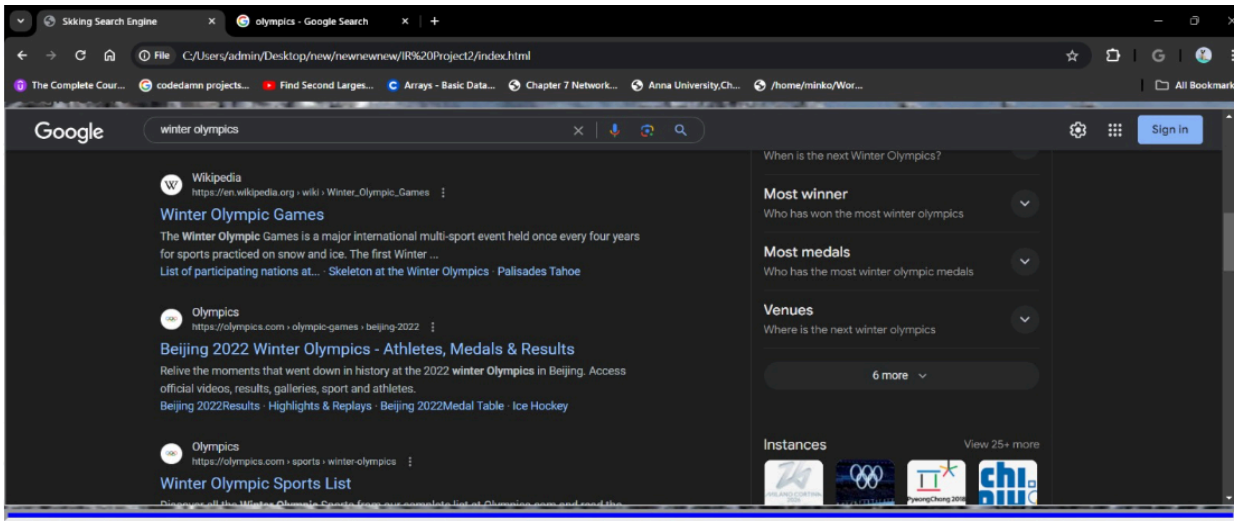
I googled most searched skiing related queries and from those chose the ones which were giving more accurate results in our search engine. We also checked if the results were being improved after clustering and query expansion.

**Query Results:**

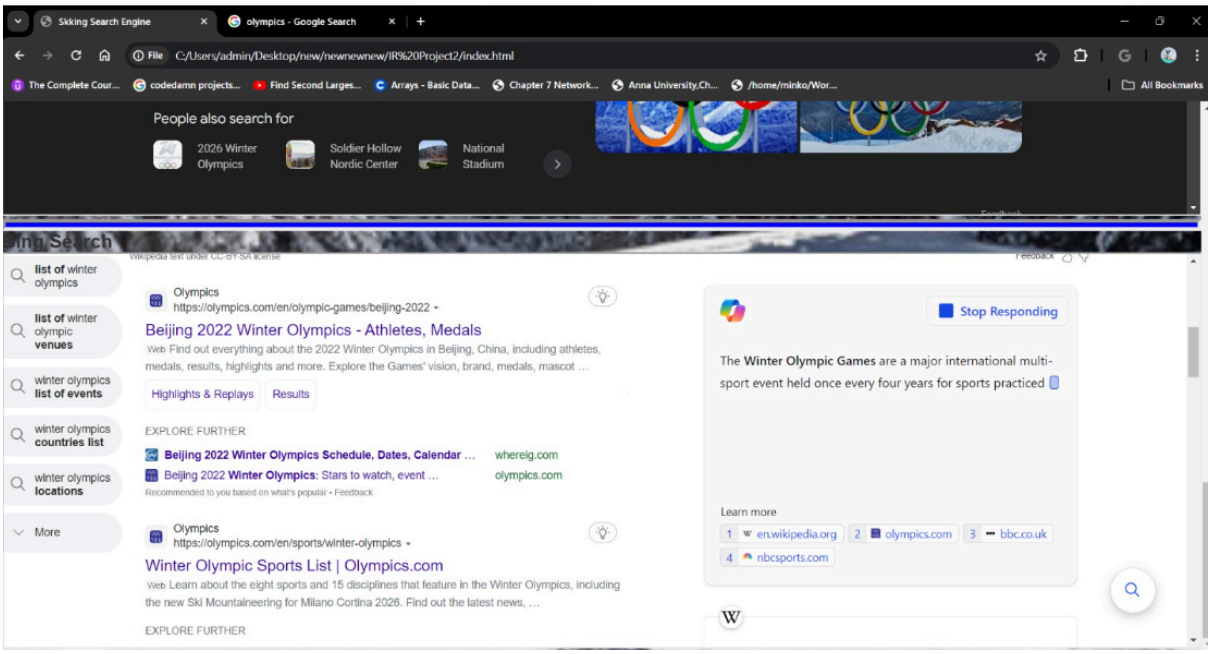
Query 1: Winter Olympics with Page Rank




Skiing Search Engine Result

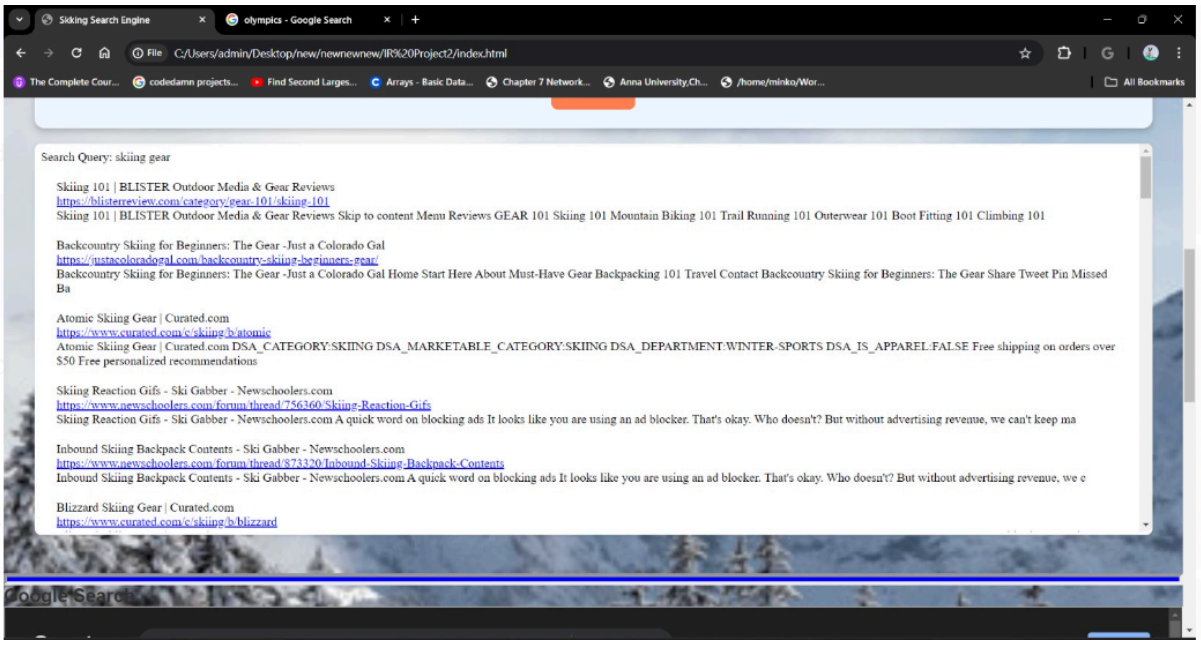



Google Results

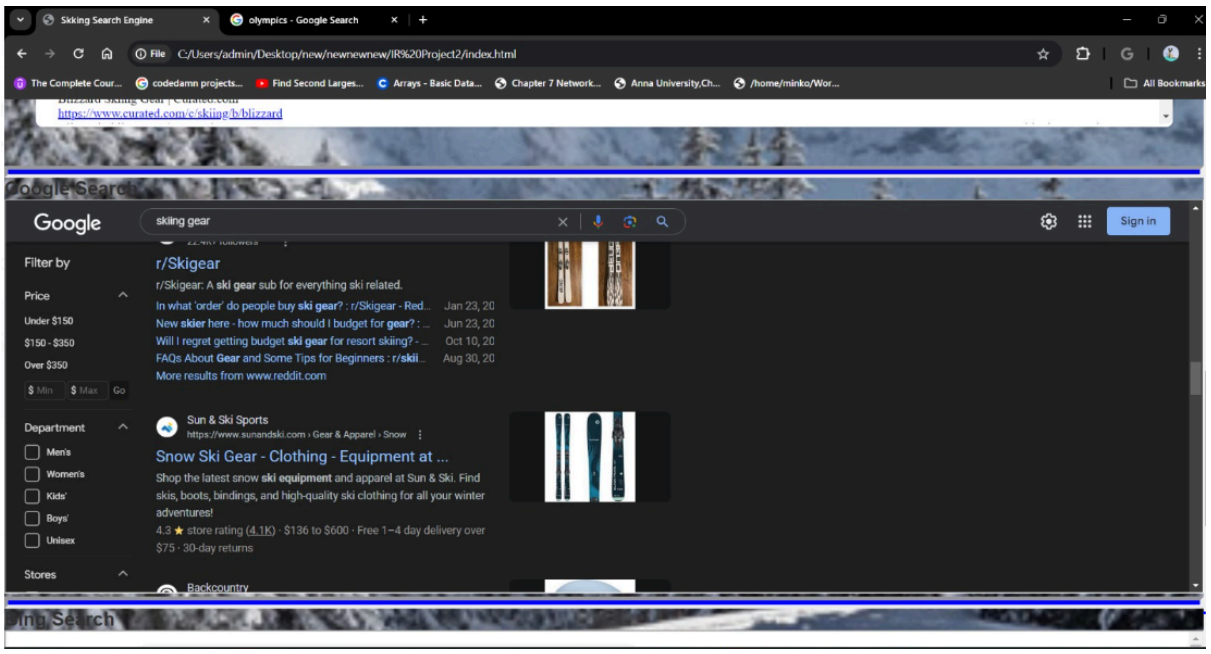



Bing Results

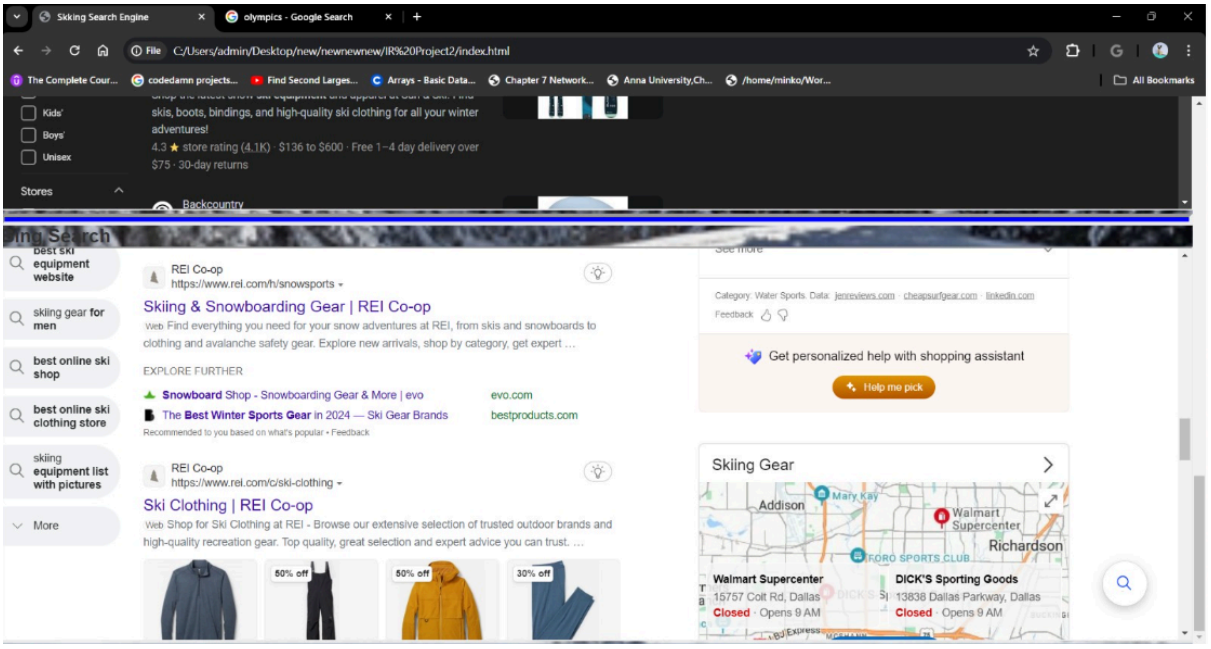
## Query 2: Skiing Gear with Flat Clustering




Skiing Search Engine Results

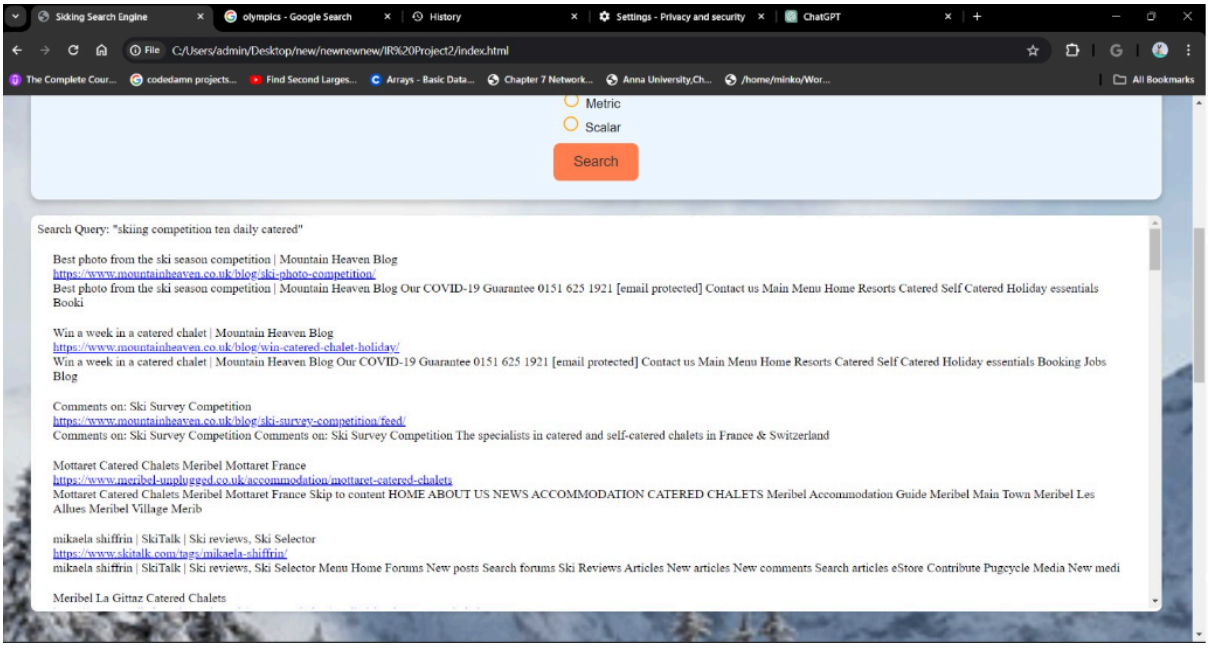



Google Results

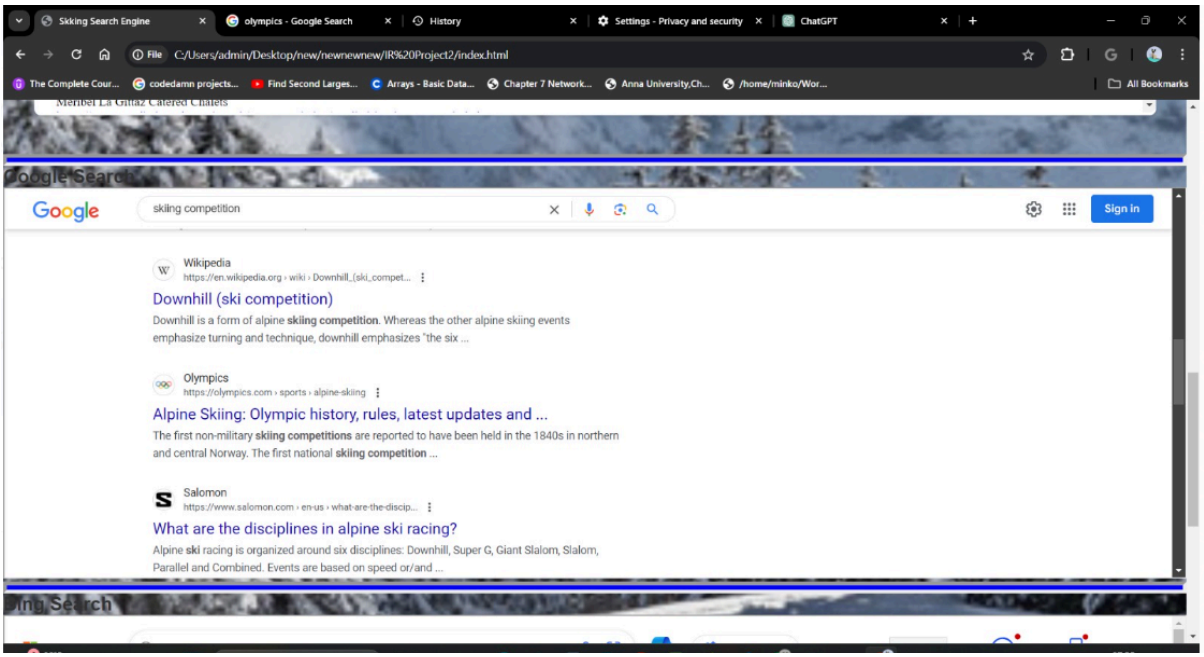



Bing Results

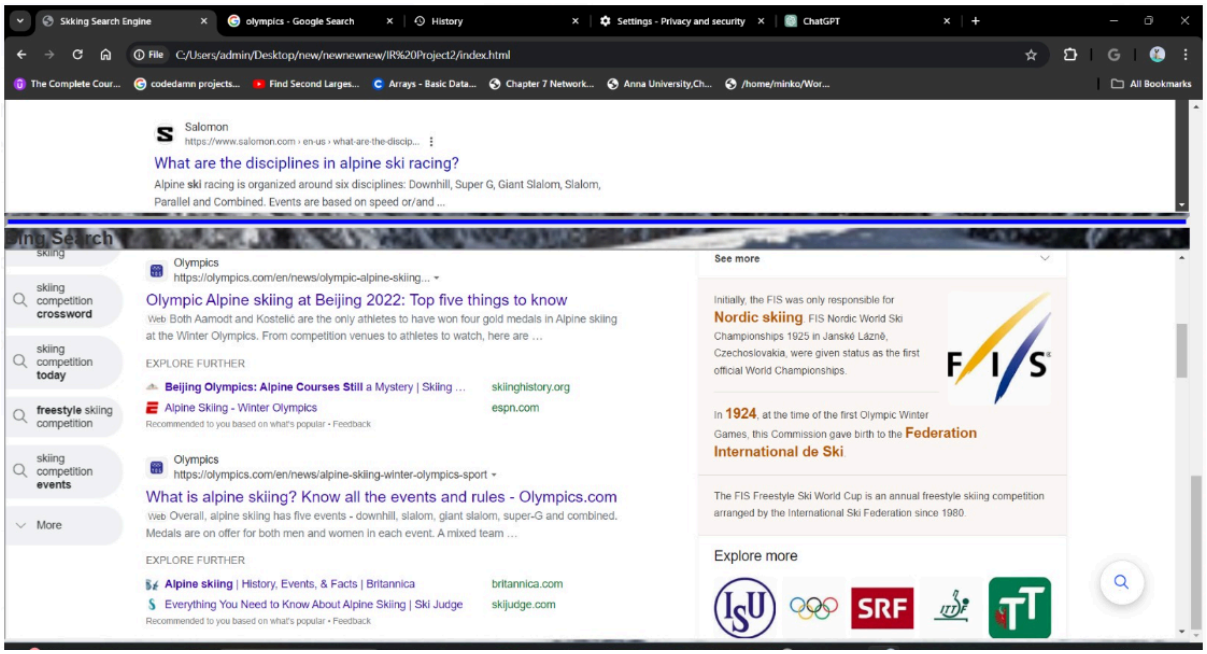
Query 3: Skiing Competition with Metric Query Expansion




Skiing Search Engine Results




Google Results




Bing Results

# ***5. Clustering***

Clustering is the process of arranging data items according to how similar they are to one another. The similarity is calculated based on metrics such as Euclidean distance, Cosine similarity, Manhattan distance, etc. The points with the highest similarity score are then grouped together.

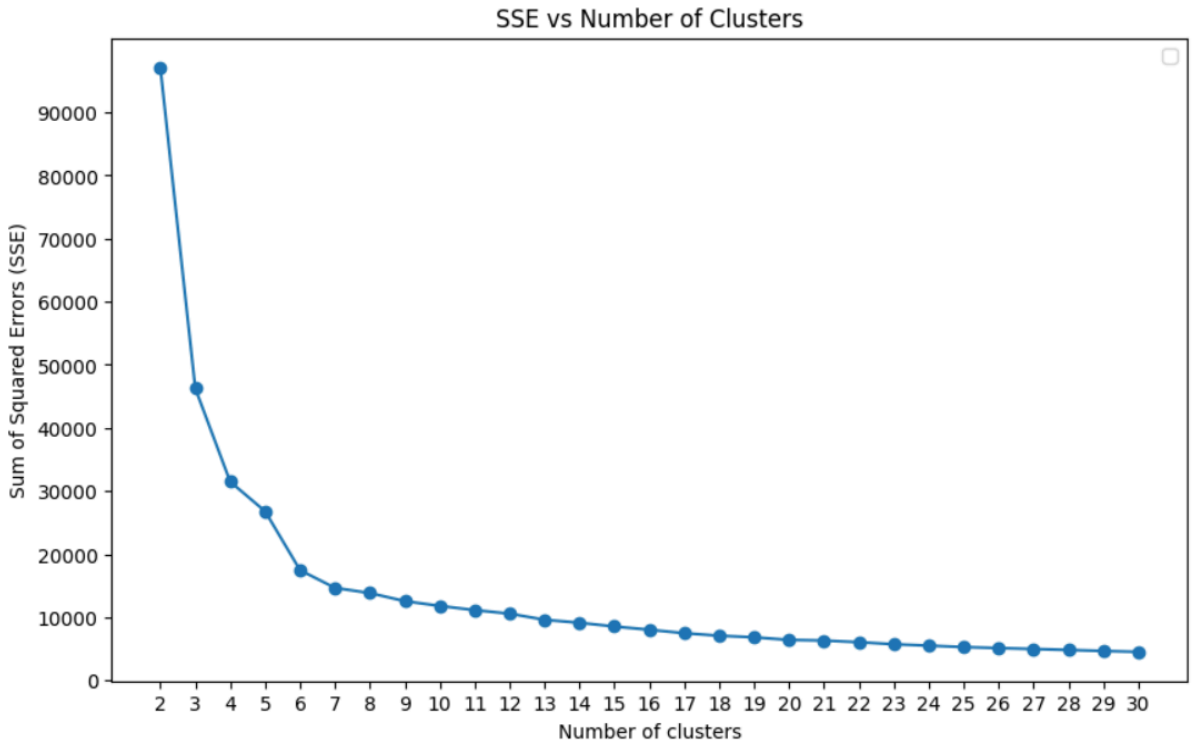
After crawling we received 124,659 webpages. Some pre-processing steps were performed before clustering as given below:

1. The webpages obtained after crawling was in json format
2. The essential features like Title, URL and content were extracted from the file and converted to a pandas dataframe
3. URLs with empty titles or content were removed. Stopwords were also removed
4. The data frame was vectorized using TfidfVectorizer using 10000 features
5. To convert a query to vector notation the tfidf object is saved using pickle

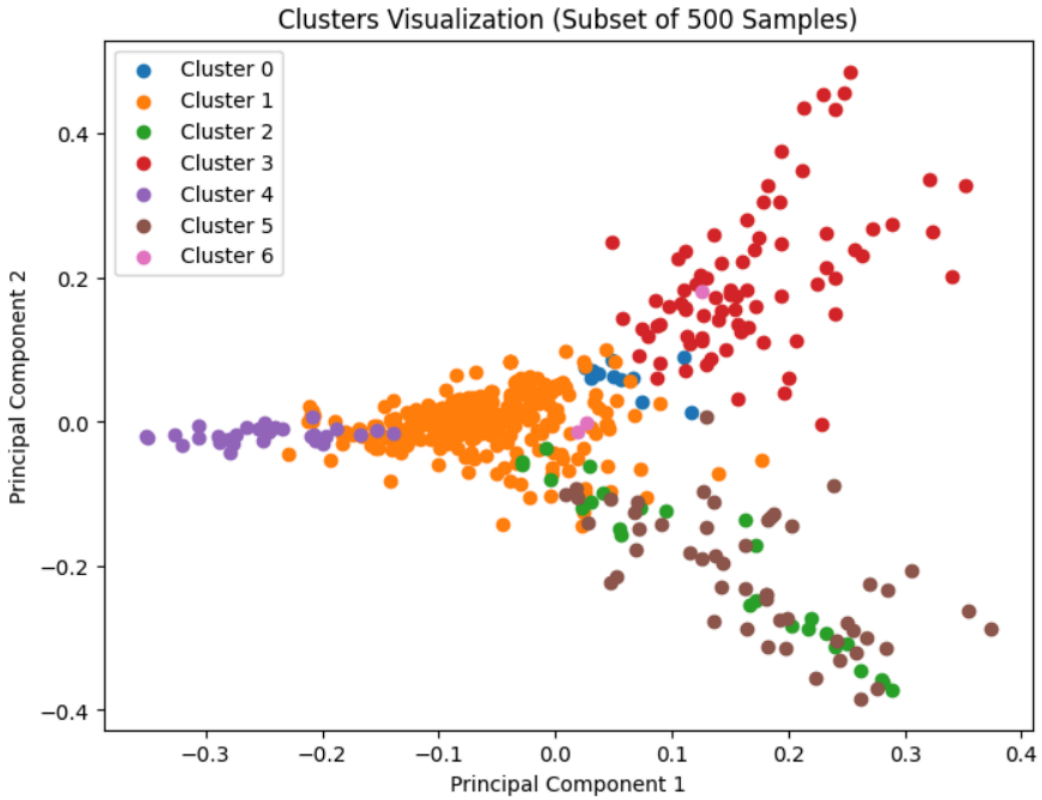
## Flat Clustering

Flat clustering was performed using the k-means algorithm. To find the number of clusters required for the k-means algorithm, we used the elbow method. The value of k(cluster number range) was chosen from 2 to 30. The algorithm was applied to each value of k and we calculated the sum of squared distances(SSE) between every data point and its corresponding cluster center. The line chart given below shows the plot of SSE against the number of clusters. Increasing the number clusters decreases the SSE as the distance between points reduces when the number of clusters are increased.

From the line chart we can take k = 7 as the SSE decreases at a slower rate at that point.



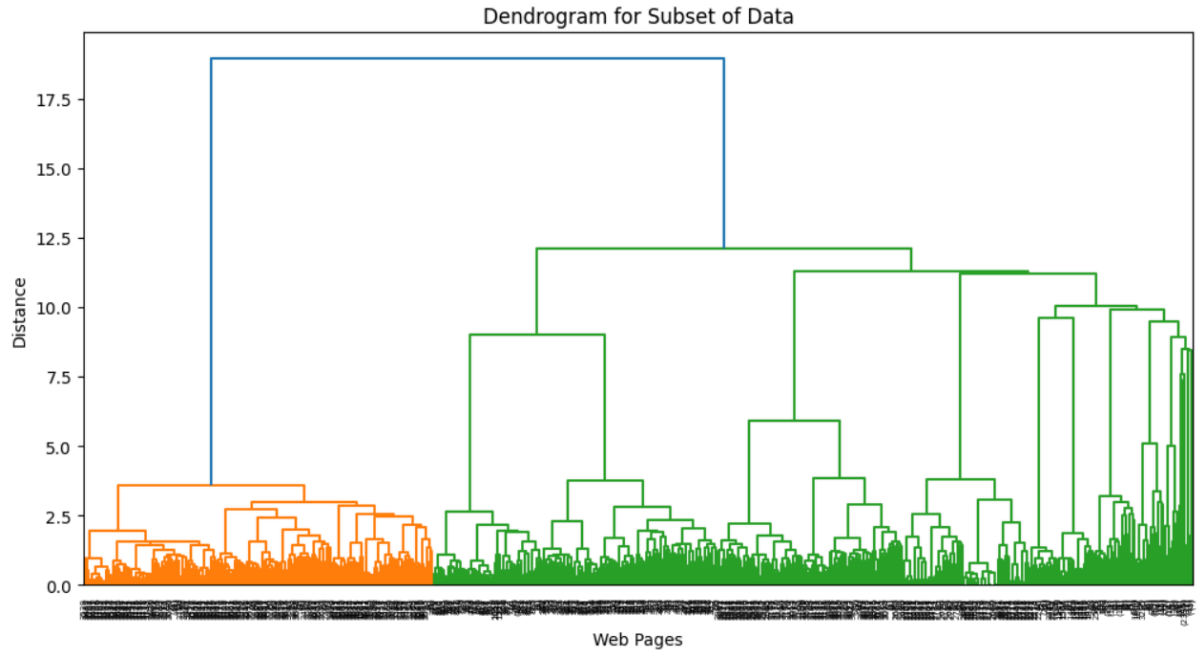

We reduced the dimensionality using Principal Component Analysis(PCA) and plotted 500 samples of the data.



After clustering, a cluster label was assigned to every URL and saved to a file in the format (URL, cluster number). For every cluster, the centroid information was also saved to a file in the format(cluster number, centroid)

## Agglomerative Clustering

The same preprocessing steps were performed as that in flat clustering. To determine the number of clusters the hierarchical clustering dendrogram was plotted on the dataset.




Find the highest vertical difference between nodes in the dendrogram, and then pass a horizontal line across the middle. The ideal number of clusters is the number of vertical lines that cross it. In our case we find the number of clusters = 8

Agglomerative clustering was done in batches due to memory constraint. After performing hierarchical clustering with each linkage method (single linkage and average linkage), the centroids of the clusters are calculated. We compare the centroids of each pair of clusters - If the distance between the centroids of two clusters is less than the specified threshold distance of 0.7, the clusters are considered similar enough to be merged. The merging process updates the cluster labels accordingly, ensuring that data points belonging to the merged clusters are assigned the same label.

The above process is repeated until no further clusters can be merged based on the threshold distance and the total number of clusters is 8.

After clustering, a cluster label was assigned to every URL and saved to a file in the format (URL, cluster number). For every cluster, the centroid information was also saved to a file in the format(cluster number, centroid). This was done for both single and average methods in hierarchical clustering.

## Integration with front end using Flask API

When the flask app is initialized it reads the 2 files - (URL - Cluster Number) and (Cluster Number - Centroid) and is saved to a dictionary.

Format of results obtained after clustering:

- Input query - All queries are of string type that the user inputs
- Input solr results - Contains the solr results obtained based on the given query
- Output - returns a list of URL objects with its corresponding title, content and other relevant fields

When the user selects a clustering method, the corresponding clustering function with the type of clustering is called. The clustering methods would then find the relevant URLs that match the query that the user had entered and return it to the front-end in JSON format. This would be parsed and results will be displayed in the front-end

## Integrate Clustering with Relevance Model

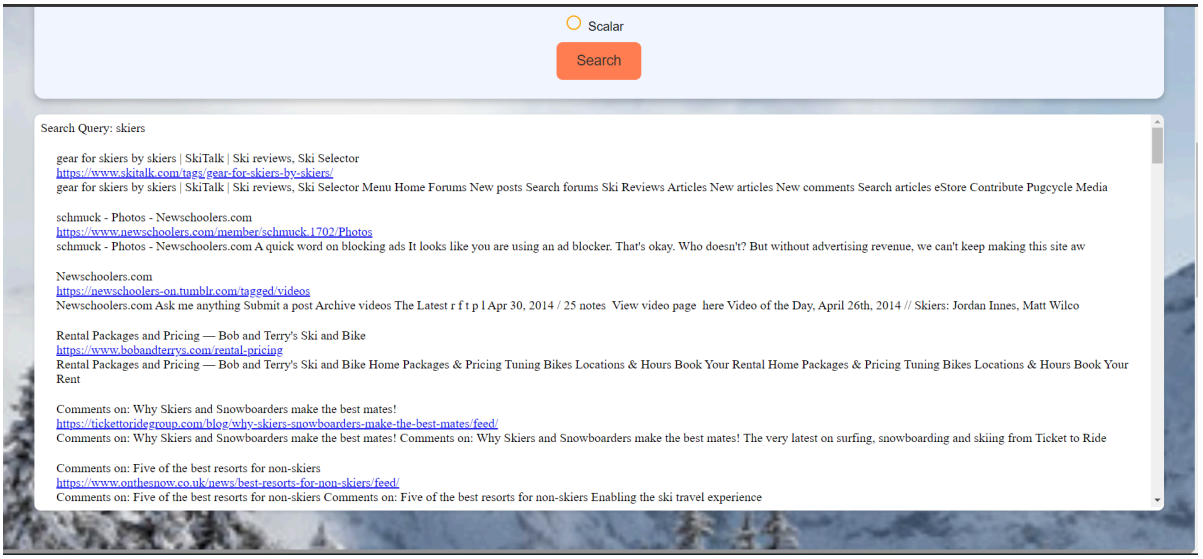
After converting crawled web pages to vector space using Apache SOLR, we integrated related web sites using the hierarchical algometric clustering approach and k-means.The centroid of the cluster was utilized to apply pertinent information. We collect a cluster id, keep track of a list of URLs around the centroid, and utilize that list as our output. To identify the relevant clusters, we utilize the vector space relevance approach for the given query and cluster centroids. We utilize the Page-Rank technique to identify the most important web pages from each cluster after relevant clusters have been extracted. The relevance model is used to identify and elevate the most relevant clusters.

## Improving the Search Results using Clustering

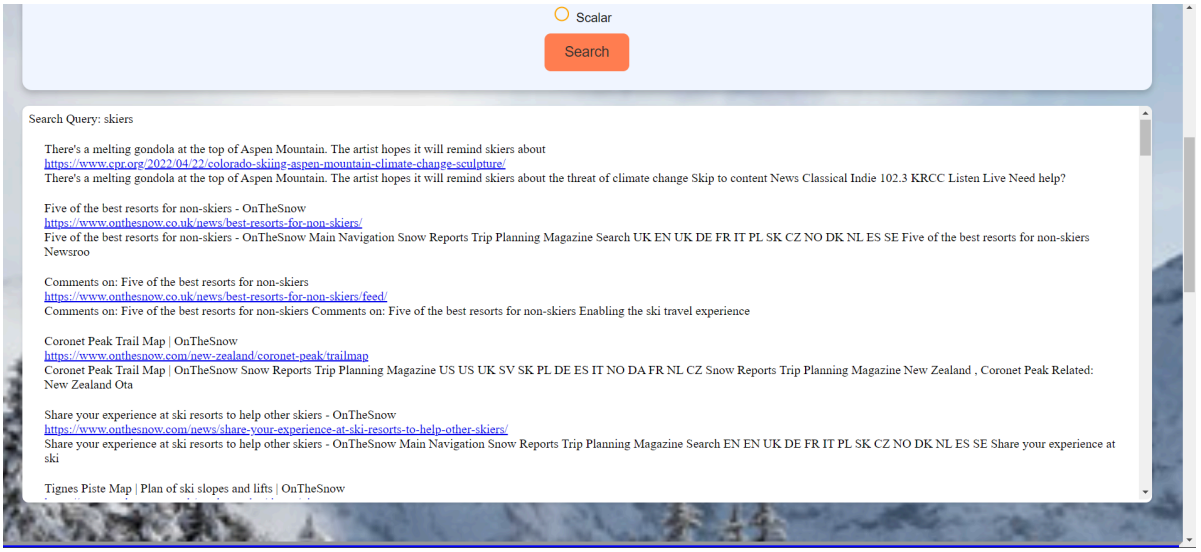
- The TfidfVectorizer is used to convert the query entered by the user to a vector form (numerical vector representation).
- The euclidean distance between the query vector and the centroids of all the clusters are calculated.
- The minimum distance gives us the centroid of the cluster that is closest to the query vector.

- Thus all the URLs that belong to that cluster will be returned in JSON format

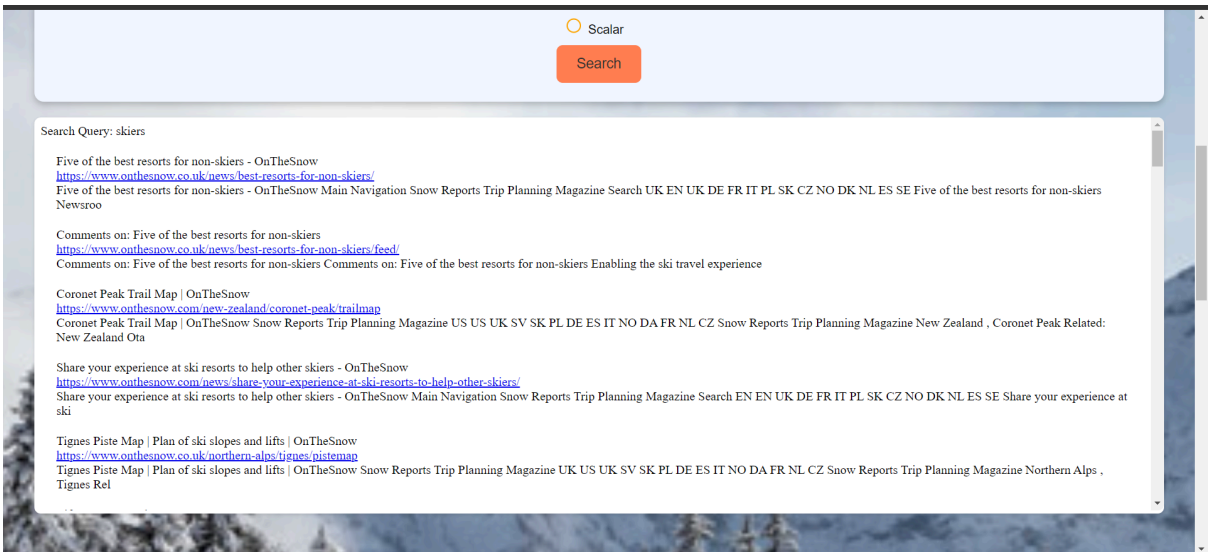
## Results of Flat Clustering




## Results of Single Hierarchical Clustering



Results of Average Hierarchical Clustering




## Testing Clustering with Queries

60 queries were used to test the clustering methods. The queries were selected based on top google searches according to the topic ‘skiing’. The results were judged manually.

## Query list

Skiing Equipment Checklist
Skiing Slang Phrases
Skiing Accident Statistics
Winter Olympics Skiing
Skiing Health Benefits
Skiing Competition Types
Skier Demographics Analysis
Skiing Events Calendar
Skiing Gear Guide
Skiing History Timeline
Skiing Skill Levels
Skiing Locations Worldwide
Skiing News Updates
Advanced Skiing Techniques
Dallas Skiing Trails
Ski Vacation Packages
Water Skiing Basics
Skiing Costs Breakdown
Freestyle Skiing Tricks

Skiing Medal Winners
Skiing Safety Tips
Skiing Legends Stories
Skiing Competition Rules
Skiing Benefits Overview
Skiing Gear Innovations
Skiing Levels Explained
Skiing Events Schedule
Dallas Ski Resorts
Water Skiing Equipment
Skiing Vacation Deals
Freestyle Skiing Competitions
Skiing Accident Prevention
Olympic Skiers' Profiles
Skiing Gear Maintenance
Skiing Levels Progression
Skiing Events Near Me
Water Skiing Techniques
Skiing Cost Factors
Freestyle Skiing Jumps
Skiing Medalists History
Skiing Gear Reviews
Skiing Levels Comparison
Skiing Events Today
Water Skiing Safety Tips
Skiing Cost Estimate
Freestyle Skiing Tricks List
Skiing Legends Legacy
Skiing Gear Recommendations
Skiing Levels Assessment
Skiing Events Calendar
Water Skiing Locations
Skiing Cost Calculator
Freestyle Skiing Styles
Skiing Legends Stories
Skiing Gear Upgrades
Skiing Levels Evaluation
Skiing Events This Weekend
Water Skiing Gear Guide
Skiing Cost Breakdown
Freestyle Skiing Techniques
Skiing Legends Achievements
Skiing Gear Comparison
Skiing Levels Progress

Skiing Events Nearby
Water Skiing Basics Guide
Skiing Cost Analysis
Freestyle Skiing Competitors
Skiing Legends Impact
Skiing Gear Selection
Skiing Levels Mastery

**Query:** Skiing Gear

**Solr Results**

1. https://www.curated.com/c/skiing/b/atomic
2. https://www.curated.com/c/skiing/b/blizzard
3. https://www.curated.com/c/skiing/b/fischer
4. https://www.curated.com/c/skiing/b/head
5. https://www.curated.com/c/skiing/b/k2

**Flat Clustering Results**

1. https://www.skicooper.com/snowsports-safety/
2. https://justacoloradogal.com/backcountry-skiing-beginners-gear/
3. https://www.newschoolers.com/forum/thread/938284/Most-annoying-term-in-skiing
4. https://www.theskidiva.com/
5. https://www.newschoolers.com/forum/thread/756360/Skiing-Reaction-Gifs

**Single Hierarchical Clustering Results**

1. https://www.onthesnow.co.uk/news/ski-packing-list/
2. https://www.onthesnow.co.uk/news/the-best-sustainable-ski-gear/
3. https://www.onthesnow.co.uk/news/how-to-find-the-best-skiing-gloves/
4. https://www.onthesnow.com/news/gear/page/2/
5. https://www.tetongravity.com/community/login

**Average Hierarchical Clustering Results**

1. https://www.onthesnow.co.uk/news/ski-packing-list/
2. https://www.onthesnow.co.uk/news/the-best-sustainable-ski-gear/
3. https://www.onthesnow.co.uk/news/how-to-find-the-best-skiing-gloves/
4. https://www.onthesnow.com/news/gear/page/2/
5. https://www.tetongravity.com/community/login

**Query:** Skiing resort

**Solr Results**

1. https://www.smuggs.com/pages/welcome/
2. https://www.smuggs.com/pages/winter/
3. https://www.smuggs.com/pages/winter/activities/cat-trax.php
4. https://www.smuggs.com/pages/winter/activities/iceSkating.php
5. https://www.smuggs.com/pages/winter/activities/tubing.php

**Flat Clustering Results**

1. https://www.smuggs.com/pages/welcome/
2. https://www.smuggs.com/pages/winter/activities/iceSkating.php
3. https://www.smuggs.com/pages/winter/
4. https://www.smuggs.com/pages/winter/amenities/massage.php
5. https://www.smuggs.com/pages/winter/lodging/resortMap.php

**Single Hierarchical Clustering Results**

1. https://www.onthesnow.co.uk/upper-austria/hinterstoder/skireport
2. https://www.onthesnow.co.uk/tyrol/imst/skireport
3. https://www.onthesnow.com/australia/dinner-plain/ski-resort
4. https://www.onthesnow.com/australia/dinner-plain/skireport
5. https://www.onthesnow.co.uk/aosta-valley/chamois/skireport

**Average Hierarchical Clustering Results**

1. https://www.onthesnow.co.uk/aosta-valley/chamois/skireport
2. https://www.onthesnow.co.uk/carinthia/gerlitzen/skireport
3. https://www.onthesnow.com/michigan/crystal-mountain/reviews
4. https://www.onthesnow.co.uk/upper-austria/hinterstoder/skireport
5. https://www.onthesnow.co.uk/tyrol/imst/skireport

**Query:** Skiing Safety

**Solr Results**

1. https://kinderlift.com/
2. https://www.skitalk.com/tags/avalanche-safety/
3. https://www.mountainheaven.co.uk/blog/the-skiers-responsibility-code/
4. https://www.skitalk.com/threads/daughter-is-ready-to-move-off-of-bunny-slopes-but-too-small-for-lifts.33020/

**Flat Clustering Results**

1. https://kinderlift.com/childrens-ski-vests/feed/
2. https://cochranskiarea.com/safety/

3. https://www.tahoedonner.com/amenities/
4. https://www.curated.com/c/winter-sports-backcountry-safety-equipment
5. https://www.skicooper.com/snowsports-safety/

**Single Hierarchical Clustering Results**

1. https://www.onthesnow.co.uk/news/family-freeriding/
2. https://www.perisher.com.au/reports-cams/cams
3. https://www.skitalk.com/forums/general-skiing.5/index.rss
4. https://www.perisher.com.au/
5. https://www.mtpeter.com/

**Average Hierarchical Clustering Results**

1. https://www.skitalk.com/forums/general-skiing.5/index.rss
2. https://www.sugarbush.com/
3. https://www.onthesnow.co.uk/carinthia/falkert/skireport
4. https://kinderlift.com/
5. https://www.palisadestahoe.com/footer/safety

# ***6. Query Expansion and Relevance FeedBack***

Query Expansion is a technique used to enhance the search results using various methods like:

1. Rocchio Algorithm (Relevance Feedback method)
2. Pseudo Relevance Feedback method:
    a. Association Cluster based Query Expansion
    b. Metric Cluster based Query Expansion
    c. Scalar Cluster based Query expansion

## 1. Rocchio Algorithm (Relevance Feedback method)

It is used to improve the results of the search engine by refining original query by using a set of relevant documents.

$$q' = \alpha.q + \beta.\frac{1}{|Dr|}\sum_{dj \epsilon Dr} dj - \gamma\frac{1}{|Dnr|}\sum_{dj \epsilon Dnr} dj$$

$Where,$
$q - original\ query$
$q' - modified\ query$
$Dr - set\ of\ relevant\ vectors$
$Dnr - set\ of\ non - relevant/irrelevant\ vectors$
$\alpha, \beta, \gamma - weights$

The values of $\alpha$=1, $\beta$=0.85, $\gamma$=0.1 were set to separate relevant documents from the irrelevant documents.

It was tested against 20 queries. The queries were chosen based on how closely they are related to the topic(skiing) . While a few queries were directly related to skiing, the other words were not so closely associated with skiing.

The top few relevant and irrelevant pages for different queries are given below

| Query | Relevant Pages | Irrelevant Pages |
|---|---|---|
| skiing | • https://thesnowmag.com/category/heli-skiing-articles/heli-alaska/feed/<br>• https://thesnowmag.com/category/heli-skiing-articles/heli-canada/feed/<br>• https://www.onthesnow.co.uk/news/year-round-skiing/<br>• http://www.chamonixskiguide.com/<br>• https://thesnowmag.com/category/heli-skiing-articles/feed/ | • https://www.fis-ski.com/alpine-skiing<br>• https://thesnowmag.com/category/heli-skiing-articles/heli-new-zealand/<br>• https://thesnowma<br>• g.com/category/custom-skis/<br>• https://thesnowmag.com/category/park-city/<br>• https://thesnowmag.com/category/heli-skiing-articles/heli-united-states/feed/ |
| snowboarding | • http://www.mechanicsofsport.com/snowboarding/equipment.html<br>• http://www.mechanicsofsport.com/advertising.html<br>• https://www.onthesnow.co.uk/news/best-snowboarding-boots-how-to-find-the-right-pair/<br>• https://www.onthesnow.co.uk/news/best-snowboarding-resorts/feed/<br>• https://www.onthesnow.com/news/travel-guide-skiing-and-snowboarding-in-canada/feed/<br>• https://www.curated.com/journal/1178001/ski-size-chart-how-to-size-skis | • https://www.onthesnow.co.uk/news/featured-gear/feed/<br>• https://www.curated.com/c/ski-bindings<br>• https://www.curated.com/c/ski-boot-accessories<br>• https://www.curated.com/c/ski-boots<br>• https://www.curated.com/c/ski-poles<br>• https://www.curated.com/c/ski-skins<br>• https://www.curated.com/c/ski-snowboard-helmets<br>• https://www.curated.com/c/ski-snowboard-sunglasses<br>• https://www.curated.com/c/ski_snowboard_bags |

| | | |
|---|---|---|
| | • https://www.onthesnow.co.uk/news/best-snowboarding-boots-how-to-find-the-right-pair/feed/<br>• https://www.curated.com/c/snowboarding/b/arbording/b/jones | https://www.curated.com/c/skiing/b/atomic |
| mountains | • https://www.onthesnow.co.uk/villach-skiing-mountains/projected-openings<br>• https://www.onthesnow.co.uk/villach-skiing-mountains/open-resorts<br>• https://www.britannica.com/browse/Mountains-Volcanoes<br>• https://www.skiinfo.pl/villach-skiing-mountains/warunki-narciarskie<br>• https://www.skiinfo.dk/villach-skiing-mountains/snerapport<br>• https://www.onthesnow.co.uk/villach-skiing-mountains/webcams<br>• https://www.onthesnow.co.uk/villach-skiing-mountains/ski-resorts<br>• https://www.onthesnow.co.uk/rhodope-mountains/skireport<br>• https://www.onthesnow.co.uk/villach-skiing-mountains/projected-closing | • https://www.onthesnow.com/rocky-mountains/lift-tickets<br>• https://www.aspensnowmass.com/four-mountains/aspen-highlands/lift-status<br>• https://www.onthesnow.com/news/snow-science-how-mountains-make-snow/feed/<br>• https://www.onthesnow.co.uk/villach-skiing-mountains/lodging<br>• https://www.onthesnow.com/news/how-mountains-of-california-make-snow/feed/<br>• https://www.onthesnow.com/news/how-the-mountains-of-colorado-make-snow/feed/<br>• https://www.aspensnowmass.com/four-mountains/aspen-highlands<br>• https://www.onthesnow.com/news/how-mountains-of-california-make-snow/ |

| | | |
|---|---|---|
| equipement | • https://thesnowmag.com/category/heli-skiing-articles/heli-alaska/feed/<br>• https://thesnowmag.com/category/heli-skiing-articles/heli-canada/feed/<br>• https://www.onthesnow.co.uk/news/year-round-skiing/<br>• http://www.chamonixskiguide.com/<br>• https://thesnowmag.com/category/heli-skiing-articles/feed/<br>• https://www.onthesnow.co.uk/news/featured-weather/ | • https://www.fis-ski.com/alpine-skiing<br>• https://thesnowmag.com/category/heli-skiing-articles/heli-new-zealand/<br>• https://thesnowmag.com/category/custom-skis/<br>• https://thesnowmag.com/category/park-city/<br>• https://thesnowmag.com/category/heli-skiing-articles/heli-united-states/feed/<br>• https://www.skitalk.com/tags/mogul-skiing/ |
| slopes | • https://www.onthesnow.com/indiana/lodging<br>• https://www.onthesnow.co.uk/news/best-spring-skiing-head-to-north-facing-ski-slopes/<br>• https://www.meribel.net/en/practical-information/slopes/<br>• https://www.onthesnow.com/new-york/brantling-ski-slopes/webcams<br>• https://www.onthesnow.com/new-york/brantling-ski-slopes/trailmap<br>• https://www.onthesnow.com/indiana/perfect-north-slopes/skireport<br>• https://www.onthesnow.com/indiana/perfect-north-slopes/lodging | • https://www.onthesnow.co.uk/news/skiing-in-the-dolomites-alpine-charm-and-jagged-peaks/<br>• https://www.onthesnow.co.uk/salzburg/zell-am-see-schmittenhoehe/ski-report-reviews<br>• https://www.onthesnow.co.uk/bernese-oberland/saanenmoeser/ski-report-reviews<br>• https://www.onthesnow.co.uk/news/guide-to-ski-insurance/<br>• https://www.onthesnow.co.uk/nearby-resorts<br>• https://www.onthesnow.co.uk/news/indoor-skiing-on-revolving-slopes/feed/ |

| | | |
|---|---|---|
| slopes | • https://www.onthesnow.com/indiana/perfect-north-slopes/historical-snowfall<br>• https://www.onthesnow.com/indiana/perfect-north-slopes/lift-tickets<br>• https://www.ski-saintgervais.com/en/notre-offre-de-forfait-saint-gervais-evasion<br>• https://www.onthesnow.com/indiana/perfect-north-slopes/ski-report-reviews<br>• https://www.skiresort.info/snow-reports/sorted/open-slopes/ | • https://www.onthesnow.co.uk/news/uk-dry-ski-slopes/feed/<br>• https://www.onthesnow.com/new-york/brantling-ski-slopes/lodging<br>• https://www.onthesnow.co.uk/news/best-spring-skiing-head-to-north-facing-ski-slopes/feed/<br>• https://www.onthesnow.co.uk/news/europe-snow-report/feed/ |
| terrain | • https://www.aspensnowmass.com/discover/experiences/guides/guide-to-lessons<br>• https://www.onthesnow.com/news/visitors-choice-award-for-best-terrain-park-keystone/<br>• https://www.onthesnow.com/news/jackson-hole-rated-best-all-mountain-terrain-resort/<br>• https://www.onthesnow.com/news/visitors-choice-award-for-best-all-mountain-terrain-telluride/<br>• https://www.onthesnow.com/alaska/eaglecrest-ski-area/reviews<br>• https://www.onthesnow.com/new-jerse | • https://www.onthesnow.com/new-york/bristol-mountain/ski-resort<br>• https://www.onthesnow.com/oregon/mt-bachelor/trailmap<br>• https://www.onthesnow.com/arizona/arizona-snowbowl/trailmap<br>• https://www.onthesnow.com/colorado/durango-mountain-resort/trailmap<br>• https://www.onthesnow.com/new-york/holiday-valley/trailmap |

| | | |
|---|---|---|
| | y/mountain-creek-resort/trailmap<br>● https://www.onthesnow.com/alberta/sunshine-village/trailmap | |
| resorts | ● https://www.onthesnow.co.uk/lillehammer-ski-resorts/open-resorts<br>● https://www.onthesnow.co.uk/lillehammer-ski-resorts/skipass<br>● https://www.onthesnow.co.uk/lillehammer-ski-resorts/projected-closing<br>● https://www.skiresort.info/snow-reports/filter/open-ski-resorts/<br>● https://www.onthesnow.co.uk/andorra/open-resorts<br>● https://www.onthesnow.co.uk/slovenia/open-resorts<br>● https://www.onthesnow.com/australia/open-resorts | ● https://www.onthesnow.co.uk/nidwalden/open-resorts<br>● https://www.onthesnow.co.uk/oberstaufen/open-resorts<br>● https://www.onthesnow.co.uk/romania/open-resorts<br>● https://www.onthesnow.co.uk/sauerland/open-resorts<br>● https://www.onthesnow.co.uk/schwarzwald/open-resorts<br>● https://www.onthesnow.co.uk/ticino/open-resorts<br>● https://www.onthesnow.co.uk/toggenburg/open-resorts |
| gear | ● https://justacoloradogal.com/category/gear/page/2/<br>● https://www.newschoolers.com/forum/thread/931417/My-ski-pant-company<br>● https://www.newschoolers.com/forum/thread/932497/Gear-to-stay-away-from<br>● https://www.newschoolers.com/forum/thread/938051/Park-Skis | ● https://www.mountainheaven.co.uk/blog/cheap-ski-gear/<br>● https://www.newschoolers.com/forum/thread/937016/Misfit-skis-drama<br>● https://www.newschoolers.com/forum/thread/714366/Atomic-2012-2013-Thread<br>● https://www.newschoolers.com/forum/thread/933300/Show-me-your-babies |

| | | |
|---|---|---|
| | • https://www.newschoolers.com/forum/thread/481906/The-answer-to--what-binding- | • https://www.newschoolers.com/forum/thread/935371/Baggy-snowpants<br>• https://www.newschoolers.com/forum/thread/926870/Vs-came-in-today<br>• https://www.newschoolers.com/forum/thread/927598/Post-a-Pic-of-Your-Camera-Gear-Equipment |
| avalanche | • https://www.skitalk.com/tags/avalanche-safety/<br>• http://www.mtavalanche.com/sitemap<br>• http://www.mtavalanche.com/avalanche-incidents<br>• https://avalancheresearch.ca/<br>• https://www.mtavalanche.com/resources<br>• https://www.mtavalanche.com/node/31203<br>• http://www.mtavalanche.com/accidents<br>• https://www.mtavalanche.com/accidents<br>• https://www.mtavalanche.com/glossary | Nan |
| freestyle | • https://newschoolers-on.tumblr.com/tagged/FREESTYLE%20GIF<br>• https://newschoolers-on.tumblr.com/tagged/freestyle%20gif | • https://newschoolers-on.tumblr.com/tagged/nsvideos<br>• https://newschoolers-on.tumblr.com/tagged/photo<br>• https://newschoolers-on.tumblr.com/ta |

| | | |
|---|---|---|
| | • https://newschoolers-on.tumblr.com/tagged/FREESTYLE<br>• https://newschoolers-on.tumblr.com/tagged/VOTD | gged/style%20gif |
| Snow | • https://www.j2ski.com/snow_forecast/France/Les_Menuires_snow.html<br>• https://www.j2ski.com/snow_forecast/France/The_Three_Valleys_snow.html<br>• https://www.j2ski.com/snow_forecast/France/Val_d_Isere_snow.html<br>• https://www.j2ski.com/snow_forecast/France/Courchevel_snow.html<br>• https://www.j2ski.com/snow_forecast/France/Les_Deux_Alpes_snow.html<br>• https://au.j2ski.com/snow_forecast/France/Saint_Martin_de_Belleville_snow.html | • https://www.j2ski.com/snow_forecast/France/Brides_les_Bains_snow_forecast_48.html<br>• https://www.onthesnow.co.uk/bernese-oberland/habkern/weather<br>• https://thesnowmag.com/category/winter-ski-fashion/luxury-ski-wear/feed/<br>• https://www.onthesnow.com/wyoming/snow-king-resort/skireport<br>• https://www.onthesnow.com/vermont/mount-snow/lodging<br>• https://www.j2ski.com/ |
| Skiing techniques | • https://www.onthesnow.co.uk/carinthia/gerlitzen/ski-resort<br>• https://www.onthesnow.co.uk/carinthia/gerlitzen/skireport<br>• http://www.mechanicsofsport.com/skiing/how_to_ski.html<br>• https://www.onthesnow.co.uk/news/ski-instructor-courses-train-qualify-and-work-on-the-slopes/<br>• https://www.lakepla | Nan |

| | | |
|---|---|---|
| | cid.com/story/2023/top-5-reasons-to-take-your-family-skiing-at-whiteface-mountain<br>● https://www.onthesnow.com/news/experience-the-arlberg-austrias-largest-ski-resort/<br>● https://www.onthesnow.com/news/reasons-to-ski-grand-targhee/ | |
| Snowboard tricks | ● https://www.newschoolers.com/forum/thread/874399/Tame-dog-with-stiff-skis<br>● https://www.onthesnow.co.uk/news/indoor-skiing-the-pros-and-cons/ | ● https://www.onthesnow.com/wyoming/snow-king-resort/skireport<br>● https://www.onthesnow.com/vermont/mount-snow/lodging |
| Mountain resorts | ● https://www.onthesnow.com/news/a-truly-all-inclusive-ski-vacation-with-club-med/<br>● https://www.onthesnow.com/new-york/labrador-mt/lift-tickets<br>● https://www.onthesnow.com/new-york/song-mountain/lift-tickets<br>● https://www.morzinesourcemagazine.com/_the_source_files_/<br>● https://www.onthesnow.com/australia/dinner-plain/lift-tickets<br>● https://www.onthesnow.com/british-columbia/powder-king/lift-tickets | ● https://www.onthesnow.com/alaska/eaglecrest-ski-area/lift-tickets<br>● https://www.onthesnow.com/alaska/hilltop-ski-area/lift-tickets<br>● https://www.onthesnow.com/alberta/ski-banff-norquay/lift-tickets<br>● https://www.onthesnow.com/alberta/winsport/lift-tickets<br>● https://www.onthesnow.com/argentina/caviahue/lift-tickets<br>● https://www.onthesnow.com/argentina/chapelco/lift-tickets<br>● https://www.onthesnow.com/argentina |

| | | |
|---|---|---|
| | | /las-lenas/lift-tickets<br>● https://www.onthesnow.com/arizona/arizona-snowbowl/lift-tickets |
| Powder snow | ● https://www.skiguidejapan.com/EN/backcountry/index.html<br>● https://www.skiguidejapan.com/EN/area/geto/geto_resort/index.html<br>● https://www.skiguidejapan.com/EN/area/appi_hachimantai/hachimantai_shimokura/index.html<br>● https://www.skiguidejapan.com/EN/accommodation/index.html<br>● https://www.newschoolers.com/member/iskiatstoke.59236/Videos<br>● https://www.skiguidejapan.com/EN/index.html<br>● https://es.skiinfo.com/colorado/steamboat/webcams | ● https://www.onthesnow.co.uk/tyrol/veneto/skireport<br>● https://www.onthesnow.co.uk/veneto/cortina-dampezzo/ski-report-reviews<br>● https://www.onthesnow.co.uk/vorarlberg/gargellen/ski-resort<br>● https://www.onthesnow.co.uk/vorarlberg/gargellen/skireport<br>● https://www.onthesnow.co.uk/vorarlberg/golm/skireport<br>● https://www.onthesnow.co.uk/styria/galsterbergalm/ski-resort<br>● https://www.onthesnow.co.uk/styria/loser-sandling-altaussee/ski-resort |
| Slope conditions | ● https://www.onthesnow.co.uk/brescia/open-resorts<br>● https://www.onthesnow.co.uk/highland/open-resorts<br>● https://www.onthesnow.co.uk/monterosa/open-resorts<br>● https://www.onthesnow.co.uk/rhoen/open-resorts<br>● https://www.onthesnow.co.uk/abruzzo/ | ● https://www.onthesnow.co.uk/beskid-slaski/open-resorts<br>● https://www.onthesnow.co.uk/czarna-gora/open-resorts<br>● https://www.onthesnow.co.uk/devoluy/open-resorts<br>● https://www.onthesnow.co.uk/eastern-slovakia/open-resorts<br>● https://www.onthes |

| | | |
|---|---|---|
| | open-resorts<br>● https://www.onthesnow.co.uk/apuseni/open-resorts<br>● https://www.onthesnow.co.uk/aragon/open-resorts<br>● https://www.onthesnow.co.uk/belgium/open-resorts | now.co.uk/hautes-pyrenees/open-resorts |
| Terrain parks | ● https://www.ullr.ski/<br>● https://www.ullr.ski/pro/lifts/index.html<br>● https://www.wachusett.com/events-park/terrain-parks/<br>● https://www.onthesnow.com/wisconsin/little-switzerland/reviews | ● https://www.newschoolers.com/forum/thread/925284/TREND-WATCH-21--22<br>● https://www.newschoolers.com/forum/thread/937430/Rails-are-shrinking---<br>● https://www.newschoolers.com/forum/thread/938507/Is-Eileen-Gu-an-industry-plant- |
| Skiing gear | ● https://www.curated.com/c/skiing/b/atomic<br>● https://www.curated.com/c/skiing/b/blizzard<br>● https://www.curated.com/c/skiing/b/fischer | ● https://www.onthesnow.co.uk/devoluy/open-resorts |
| Avalanche safety | ● https://www.skitalk.com/tags/avalanche-safety/<br>● https://www.mountainheaven.co.uk/blog/avalanche-safety-what-you-need-to-know-to-stay-safe/feed/<br>● https://avalancheresearch.ca/<br>● https://mountaineer | ● https://mountaineer.com/activities/climbing/crampons-and-spare-parts/<br>● https://mountaineer.com/activities/climbing/crash-pads/<br>● https://mountaineer.com/activities/climbing/dogbones-sewn-slings-and-lanyards/ |

| | | |
|---|---|---|
| | .com/sports/backcountry-ski/avalanche-safety/<br>● https://caicsignup.avalancheresearch.ca/en/<br>● https://www.mountainheaven.co.uk/blog/ski-schools-in-france-trained-for-avalanche-safety/ | ● https://mountaineer.com/activities/climbing/harnesses/ |
| Skiing holidays | ● https://www.skisolutions.com/<br>● https://www.whitetracks.co.uk/Helicopter_Transfers_Flims_Laax_Switzerland.htm<br>● https://www.onthesnow.co.uk/styria/ramsau-am-dachstein/ski-resort<br>● https://www.onthesnow.co.uk/styria/ramsau-am-dachstein/skireport<br>● https://hu.ski/about/ | ● https://www.tyrol.com/things-to-do/sports/cross-country-skiing<br>● https://www.mountainheaven.co.uk/blog/ski-jackets-the-best-names-to-look-out-for/<br>● https://www.mountainheaven.co.uk/quick-links/ski-jobs |

Testing queries and getting their modified Rocchio query

| **Query** | **Roccio Modified Query** |
|---|---|
| skiing | Skiing resort |
| snowboarding | Snowboarding gear |
| mountains | Mountain vermont skiing resort |
| men/women | Men ski apparel jacket |
| slopes | Slopes skiing trails snow quality |
| terrain | terrain ski mountain snow sports |
| resorts | resorts ski lodges mountain amenities |

| | |
|---|---|
| gear | Skiing gear cost |
| avalanche | avalanche safety precautions skiing |
| freestyle | Freestyle skiing techniques |
| snow | Snow conditions  skiing |
| Skiing techniques | Skiing techniques advanced slope |
| Snowboard tricks | Snowboard tricks olympics |
| Mountain resorts | Mountain resorts budget secluded |
| Powder snow | Powder snow conditions report tips |
| Slope conditions | Slope conditions alps update |
| Terrain parks | Terrain parks safety location |
| Skiing gear | Skiing gear cost sale |
| Avalanche safety | Avalanche safety protocols tips |
| Skiing holidays | Skiing holidays resorts accommodation |

**Observation:**

1. **No Irrelevant Web Pages for Some Queries**
   For certain queries like “avalanche”, “skiing techniques” showed no irrelevant web pages.
2. **Relevance of Irrelevant Web Pages**
   For some queries, like “freestyle” the web pages identified as irrelevant seemed far more or equally relevant as the pages identified as relevant.
3. **We were unable to arrive at a perfect assignment of weights for our queries.**

Considering these issues, we decided not to incorporate the Roccio algorithm into the query expansion.

Instead, we identify relevant documents based on the initial query by dividing the top 50 Solr results into two sets:  the relevant set of 30 documents and the irrelevant set containing 20 documents. Then we calculate the cosine similarity between the terms in the query and the terms in the relevant documents. Terms from the relevant documents that are similar to the terms in the query are selected and added to the query to expand it.

## 2. Pseudo Relevance based Query Expansion:

a. **Association Cluster based Query Expansion:**
It is an expansion technique that is used to enhance the search query by identifying the most frequently co-occurring stems in the relevant documents and works on the underlying principle that these co-occurring terms are associated.

b. **Metric Cluster based Query Expansion:**
The metric cluster based query expansion is based on correlation between the terms or stems in this case. It identifies clusters based on the proximity of the terms with each other.

c. **Scalar Cluster based Query Expansion**

Scalar cluster based Query expansion selects the terms based on the co-occurrence patterns and groups terms together that have similar contextual neighborhoods.

The following is the list of 60 queries to test using pseudo relevance feedback

| **Query** | **Expanded Query** | **Relevant pages fetched** | **Method used** |
|---|---|---|---|
| Skiing | Skiing level winter magazine | 45 | Association |
| Skiing equipment checklist | Skiing equipment checklist level admin commonly | 30 | Metric |
| Skiing slang phrases | Skiing slang phrases part pass year | 25 | Scalar |
| Skiing Accident Statistics | Skiing Accident Statistics earth pass souls | 13 | Association |
| Winter Olympics Skiing | Winter Olympics Skiing tricia earth turn | 22 | Metric |
| Skiing Competition Types | Skiing Competition Types turn run earth | 17 | Association |
| Skiing Gear | Skiing Gear like | 12 | Association |

| | | | |
|---|---|---|---|
| | wood tricia | | |
| Skiing History Timeline | Skiing History Timeline turn organisation seeing | 8 | Metric |
| Skiing Locations | Skiing Locations run reflects directions | 25 | Association |
| Skiing News | Skiing News resort inches humble | 16 | Metric |
| Advanced Skiing Techniques | Advanced Skiing Techniques run levels garbanzo | 32 | Scalar |
| Skiing Trails | Skiing Trails heritage direction pass | 22 | Association |
| Ski Vacation Packages | Ski Vacation Packages run directions unplugged | 11 | Metric |
| Water Skiing | Water Skiing turn run george | 31 | Scalar |
| Skiing Costs | Skiing Costs profit turn bonjour | 24 | Association |
| Freestyle Skiing | Freestyle Skiing potd candles epic | 30 | Metric |
| Skiing Medal Winners | Skiing Medal Winners imperial ranks meribels | 19 | Scalar |
| Skiing Safety Tips | Skiing Safety Tips turn heritage directions | 9 | Association |
| Skiing Legends Stories | Skiing Legends Stories turn interviews local | 21 | Metric |
| Skiing Benefits Overview | Skiing Benefits Overview turn directions policy | 14 | Association |
| Skiing Levels Explained | Skiing Levels Explained | 28 | Scalar |
| Skiing Events Schedule | Skiing Events Schedule run pass local | 32 | Association |
| Ski Resorts | Ski Resorts imperial policy luxury | 16 | Metric |
| Skiing Equipment | Skiing Equipment touring bump stewardship | 18 | Scalar |

| | | | |
|---|---|---|---|
| Skiing Vacation Deals | Skiing Vacation Deals earth run pass | 23 | Association |
| Freestyle Skiing Competitions | Freestyle Skiing Competitions tricia heritage levels | 36 | Metric |
| Skiing Accident Prevention | Skiing Accident Prevention organisations turn islands | 27 | Scalar |
| Olympic Skiers Profiles | Olympic Skiers Profiles zao run handful | 17 | Association |
| Skiing Gear Maintenance | Skiing Gear Maintenance seeing local run | 16 | Metric |
| Skiing Levels Progression | Skiing Levels Progression menu run local | 28 | Scalar |
| Skiing Events Near Me | Skiing Events Near Me loveyouradk local run | 14 | Association |
| Water Skiing Techniques | Water Skiing Techniques menu carp chris | 34 | Metric |
| Skiing Cost Factors | Skiing Cost Factors pass produce profit | 25 | Scalar |
| Freestyle Skiing Jumps | Freestyle Skiing Jumps run earth stable | 12 | Association |
| Skiing Events Today | Skiing Events Today directions local turn | 29 | Metric |
| Skiing Cost Estimate | Skiing Cost Estimate run earth policy | 11 | Association |
| Freestyle Skiing Tricks | Freestyle Skiing Tricks | 31 | Metric |
| Skiing Legends Legacy | Skiing Legends Legacy seeing heritage pioneer | 22 | Scalar |
| Skiing Gear Recommendations | Skiing Gear Recommendations compartment like | 37 | Association |

| | | | |
|---|---|---|---|
| | jokes | | |
| Skiing Levels Assessment | Skiing Levels Assessment run jerry heritage | 15 | Metric |
| Water Skiing Locations | Water Skiing Locations fortified chapman au | 21 | Association |
| Skiing Cost Calculator | Skiing Cost Calculator youre keep app | 18 | Metric |
| Freestyle Skiing Styles | Freestyle Skiing Styles keep thread epic | 9 | Scalar |
| Skiing Gear Upgrades | Skiing Gear Upgrades america new prints | 33 | Association |
| Skiing Levels Evaluation | Skiing Levels Evaluation Evaluation zermatt procedures program | 21 | Metric |
| Skiing Events This Weekend | Skiing Events This Weekend march fall prints | 8 | Association |
| Freestyle Skiing Techniques | Freestyle Skiing Techniques | 13 | Association |
| Skiing Achievements | Skiing Achievements | 17 | Metric |
| Skiing Gear Comparison | Skiing Gear Comparison | 27 | Scalar |
| Skiing Levels Progress | Skiing Levels Progress | 26 | Association |
| Skiing Events Nearby | Skiing Events Nearby | 14 | Metric |
| Water Skiing Basics Guide | Water Skiing Basics Guide | 7 | Scalar |
| Skiing Cost Analysis | Skiing Cost Analysis | 19 | Association |
| Freestyle Skiing Competitors | Freestyle Skiing Competitors | 24 | Metric |

**Observation:**

1.In most of the queries, Association method seems to give better and more relevant results than scalar and metric clustering.

2. Scalar and metric clustering based query expansion is taking a long time to give the results.

Only few of the the local document ids, local stem set and local vocabulary set are provided for each query

**Query 1: skiing gear**

**Local document ids:**

b6953acee57e1479bfcc87103dba438
8c7f362735399f5f5e57ed97cf9ed94a
032359deb3293270633f4ede65a65952
67eb00cf5feca9e91e40b6dab2be1f14
8ec257cb0d155724458c59fa38e53ed3
850c0cd1118eb78e16860607466cf4fd
701bae5427754f82f881b5e7f4617274
0f01b02e2aa3ed9c351433a9a24aa53e
c408b6c5b403141a87689feb54c3539e
9c16461a9988d1a8a2118b9522c3e5b9
cab25aa70dde248e9b6589ab6a498212
9f6bb84c769d2ae8ce168a45b9a866c1
df162c4c22c404c90cdf4ddc81b6a6ee
dcfecdc5629798f599fe8079b205c719
81de2f3776c6238527366c44b9ddd51f
eaf970f1c23845c59d07280cd07ef248
fb39d011e21f1278ef377a365c47578a
64cf8b7deac94c31ec06fabdaf551d37
fc3ecf9b8a38638a98978803beb55258
6bc320a35044c47967513647dbde91df
6b64de68adeeb8ac6262699fda7c62a1
524f154895bd4adac911d6d3df77720a

**Local stem set:**

'hinterthierse' ['hinterthiersee'], 'thierse' ['thiersee'], 'skisport' ['skisport'], 'hartl' ['hartl'], ‘rie' ['ried'], 'skiduthyrn' ['skiduthyrning'], 'hyra' ['hyra'], 'skidor' ['skidor'], 'och' ['och'], 'tyrol' ['tyrol'], 'se' ['se'], 'wilder' ['wilder'], 'kaiser' ['kaiser'], 'webbkameror' ['webbkameror'], 'direktrapport' ['direktrapporter'], 'boend' ['boende'], 'planera' ['planera'], 'din' ['din'], 'resa' ['resa'], 'skidkurs' ['skidkurser'], 'vuxna' ['vuxna'], 'barn'

['barn'], 'kontrollera' ['kontrollera'], 'priser' ['priser'], 'med' ['med'], 'billigar' ['billigare'], 'erbjudand' ['erbjudande'], 'fuch' ['fuchs'], 'brixen' ['brixen'], 'thale' ['thale'], 'liftweg' ['liftweg'], 'mobil' ['mobiler', 'mobile'], 'michel' ['michel'], 'kirchberg' ['kirchberg'], 'deliveri' ['delivery'], 'materi' ['material'], 'accommod' ['accommodation'], 'rainer' ['rainer'], 'scheffau' ['scheffau'], 'hervi' ['hervis'], 'johann' ['johann'], 'rudi' ['rudi'], 'gerri' ['gerry'], 'aschauer' ['aschauer'], 'hochzillert' ['hochzillertal']
'hollau' ['hollaus'], 'stumm' ['stumm'], 'bike' ['bike'], 'vital' ['vital'], 'kaltenbach' ['kaltenbach'], 'schmiedau' ['schmiedau'], 'stock' ['stock'], 'ww' ['ww'], 'walter' ['walter'], 'maurach' ['maurach'], 'achense' ['achensee'], 'nordic' ['nordic'], 'rofan' ['rofan'], 'achenseestr' ['achenseestr'], 'snow' ['snow']

**Local vocabulary set: Displaying few vocabulary**

'Photography', 'personalise', ‘accompanying', 'material', 'outerwear', 'contain', 'delighted', 'edit', 'advertise', 'form', 'ambassadors', 'probe', 'claude', 'slope', 'es', 'envy', 'support', 'tweet', 'include', 'sites', 'eyeballs', 'history', 'waterville', 'alto', 'ask', 'especially', 'struggle', 'sportsways', 'river', 'take', 'born', 'pro', 'completely', 'live', 'skimming', 'women', 'placid', 'cat', 'preferences', 'okay', 'wordpress', 'runtothefinish', 'incredibly', 'oct', 'hut', 'stuck', 'means', 'leases', 'orbital', 'slang', 'reports', 'impact', 'online', 'speedy', 'display', 'decide', 'verbal', 'machine', 'podcasts', 'value', 'visit',

'ballclapper', 'priceless', 'goggles', 'search', 'encouraging', 'deal', 'renter', 'patrick', 'service', 'blister', 'instead', 'markets', 'black', 'essentials', 'mogul', 'stickers', 'telluride', 'refine', 'youd', 'know', 'developed', 'charged', 'pigon_racer', 'open', 'sitemap', 'offers', 'right', 'update', 'sure', 'edited', 'people', 'self', 'company', 'fresh', 'change', 'painful', 'accounting', 'ahead', 'channel', 'wasnt', 'reply', 'shops', 'crested', 'towards', 'poster', 'theyd', 'tips', 'independent', 'prix', 'pay', 'phone', 'act', 'knows', 'audience', 'year', 'avoid', 'cost', 'talking', 'merino', 'kid', 'language', 'commission', 'fashion', 'barnes', 'town', 'javascript', 'burton', 'mike'

**Local stem set:**

'zobraz' ['zobraz'], 'easter' ['easter'], 'holiday' ['holiday', 'holidays'], 'guildfordskislopecouk' ['guildfordskislopecouk'], 'skip' ['skip'], 'content' ['content'], 'slope' ['slopes', 'slope'], 'donut' ['donutting'], 'parti' ['parties'], 'info' ['info'], 'find' ['find'], 'due' ['due'], 'staff' ['staff'], 'close' ['closed', 'close'], 'monday' ['monday'], 'april' ['april'], 'th' ['th'], 'open' ['opening', 'open'], 'tuesday' ['tuesday'], 'current' ['currently', 'current'], 'remaind' ['remainder'], 'period' ['period'], 'ndrd' ['ndrd'], 'fulli' ['fully'], 'apart' ['apartments', 'apart'], 'public' ['public'], 'session' ['sessions'], 'daytim' ['daytime'], 'even' ['even', 'evenings', 'evening'], 'morn' ['morning'], 'pm' ['pm'], 'bore' ['boring'], 'stuff' ['stuff'], 'must' ['must'], 'advanc' ['advance'], 'enquir' ['enquire'], 'pleas' ['please'], 'enquiri' ['enquiry'], 'form' ['form'], 'karin' ['karin'], 'bradburi' ['bradbury'], 'share' ['share'], 'choos' ['choose'], 'platform' ['platform'], 'facebook' ['facebook'], 'x' ['x'], 'reddit' ['reddit'], 'linkedin' ['linkedin'], 'tumblr' ['tumblr'], 'pinterest' ['pinterest'], 'vk' ['vk'], 'email' ['email'], 'copyright' ['copyright'], 'christ' ['christs'], 'colleg' ['college'], 'club' ['club'], 'sitemap' ['sitemap'], 'legal' ['legal'], 'page' ['pages', 'page'], 'load' ['load'], 'link'

['link'], 'go' ['go', 'going'], 'octob' ['october'], 'half' ['half'], 'look' ['looking'], 'someth' ['something'], 'bring' ['bring'], 'privat' ['private'], 'week' ['week'], 'upward' ['upwards'], 'practic' ['practice'], 'instructor' ['instructors', 'instructor'], 'given' ['given'], 'one' ['one'], 'passport' ['passports', 'passport'], 'give' ['give'], 'call' ['call'], 'minimum' ['minimum'], 'age' ['age'], 'accompani' ['accompanied'], 'per' ['per'], 'person' ['person'], 'everi' ['every'], 'time' ['times', 'time']

**Query 2: skiing resort**

**Local Document Set:**
81de2f3776c6238527366c44b9ddd51f
eaf970f1c23845c59d07280cd07ef248
fb39d011e21f1278ef377a365c47578a
64cf8b7deac94c31ec06fabdaf551d37
fc3ecf9b8a38638a98978803beb55258
6bc320a35044c47967513647dbde91df
6b64de68adeeb8ac6262699fda7c62a1
524f154895bd4adac911d6d3df77720a
60f4df1a17154e65dd028570f02617a0
2be3849678b7c891641004fcf84ed051
ebf323c219ea9edb62951b0d217ed507
9fde3613023f0ec5019827226d6c78cf
4d04ae853da46b1f0c26d396b0d1eff8
67b6e0d88c453ef0919a4182b2ad5074
648a0c52e4b47c2e4547f7f8dc4ec405
d6174b4c0433ab21e1bf48263a3b18ae
db6953acee57e1479bfcc87103dba438
8c7f362735399f5f5e57ed97cf9ed94a
032359deb3293270633f4ede65a65952
67eb00cf5feca9e91e40b6dab2be1f14
8ec257cb0d155724458c59fa38e53ed3
850c0cd1118eb78e16860607466cf4fd
701bae5427754f82f881b5e7f4617274
0f01b02e2aa3ed9c351433a9a24aa53e
c408b6c5b403141a87689feb54c3539e
9c16461a9988d1a8a2118b9522c3e5b9
cab25aa70dde248e9b6589ab6a498212
9f6bb84c769d2ae8ce168a45b9a866c1

**Local vocabulary set: Displaying few vocabulary**

'mobile', 'page', 'era', 'collecting', 'ballet', 'rebell', 'becoming', 'lightweight', 'posted', 'promotion', 'oldmanski', 'drop', 'skip', 'team', 'wherewhat', 'país', 'resort', 'passion', 'muscle', 'joy', 'wild', 'majics', 'innovative', 'freakschoolers', 'email', 'mesa', 'plastic', 'guests', 'nah', 'literally', 'guides', 'help', 'shot', 'spect', 'bumps', 'edges', 'anyone', 'magnetise', 'narrow', 'joal', 'inspiredmedia', 'rainbow', 'saw', 'government', 'guest', 'holiday', 'waterproof', 'enough', 'notorious', 'age', 'sometimes', 'knowledge', 'resident', 'skissnowboards', 'attend', 'sun', 'heuga', 'chart', 'wanted', 'commissions', 'dsa_departmentwinter', 'message', 'safety', 'grits', 'news', 'journal', 'routes', 'helitrax', 'atomic', 'say', 'discounts', 'climbing', 'insider', 'dash', 'muffman', 'iceland', 'jan', 'nonetheless', 'already', 'foremost', 'recent', 'side', 'purchased', 'enable', 'pond', 'norway', 'charger', 'lift', 'save', 'tent', 'chance', 'weekend', 'seems', 'nervous', 'great', 'cookies', 'heres', 'adrenaline', 'greater', 'someone', 'boot', 'theres', 'thought', 'covet', 'bookstores', 'beanie', 'skitalk', 'plake', 'rentals', 'starglass', 'shutterstock', 'uk', 'bigpurpleskisuit', 'versus', 'rider', 'category', 'founder', 'therefore', 'exchanges', 'coloradodogfart', 'ease', 'memory', 'using', 'archive', 'containing', 'chunderface', 'httpswwwpugskicomforumsheritage', 'cooper', 'hassle', 'jacket', 'ive', 'dances', 'czechoslovakia', 'original', 'feet', 'bikes', 'posters', 'cdubdashrdr', 'furniture', 'norsquip', 'ultimate', 'a_platic_bag', 'hiking', 'web', 'karen_skier', 'independent', 'prix', 'pay', 'phone', 'act', 'knows', 'audience', 'year', 'avoid', 'cost', 'talking', 'merino'

**Local stem set:**

‘slot’ [‘slots'], ‘Also’ [‘also'], ‘Offic’ [‘office'], ‘hour’ ['hours'], 'friday' ['friday'], 'want' ['want'], 'obvious' ['obviously'], 'anytim' ['anytime'], 'like' ['like'], 'swisspassescom' ['swisspassescoms', 'swisspassescom'], 'swiss' ['swiss'], 'rentalhir' ['rentalhire'], 'resort' ['resorts', 'resort']

**Query 3:ski board**

**Local Document set:**

81de2f3776c6238527366c44b9ddd51f
eaf970f1c23845c59d07280cd07ef248
fb39d011e21f1278ef377a365c47578a
64cf8b7deac94c31ec06fabdaf551d37
fc3ecf9b8a38638a98978803beb55258
6bc320a35044c47967513647dbde91df
6b64de68adeeb8ac6262699fda7c62a1
524f154895bd4adac911d6d3df77720a
60f4df1a17154e65dd028570f02617a0

2be3849678b7c891641004fcf84ed051
ebf323c219ea9edb62951b0d217ed507
9fde3613023f0ec5019827226d6c78cf
4d04ae853da46b1f0c26d396b0d1eff8
67b6e0d88c453ef0919a4182b2ad5074
648a0c52e4b47c2e4547f7f8dc4ec405
d6174b4c0433ab21e1bf48263a3b18ae
db6953acee57e1479bfcc87103dba438
8c7f362735399f5f5e57ed97cf9ed94a
032359deb3293270633f4ede65a65952
67eb00cf5feca9e91e40b6dab2be1f14
8ec257cb0d155724458c59fa38e53ed3
850c0cd1118eb78e16860607466cf4fd
701bae5427754f82f881b5e7f4617274
0f01b02e2aa3ed9c351433a9a24aa53e
c408b6c5b403141a87689feb54c3539e
9c16461a9988d1a8a2118b9522c3e5b9
cab25aa70dde248e9b6589ab6a498212
9f6bb84c769d2ae8ce168a45b9a866c1

**Local vocabulary set: Displaying few vocabulary**

'permit', 'lyžeřské', 'nabídku', 'checked', 'grangettes', 'liftweg', 'family', 'pm', 'planuj', 'kelchsau', 'se', 'really', 'shop', 'latest', 'neben', 'juwel', 'join', 'opening', 'taniacourchevel', 'da', 'och', 'dospělí', 'mostly', 'weekdays', 'mit', 'unneccesary', 'excursions', 'skischulen', 'whilst', 'děti', 'certain', 'planner', 'magazin', 'überblick', 'dorf',

Query Expansion - Vidya Sreekumar (VXS220066)

'skis', 'skipasů', 'money', 'til', 'sk', 'story', 'taken', 'count', 'pass', 'working', 'driving', 'week', 'urban', 'österrike', 'parties', 'remainder', 'tyrolsko', 'sturtevants', 'view', 'load', 'newslettersblogs', 'lessons', 'gallery', 'number', 'poll', 'sport', 'skileje', 'getting', 'anytime', 'medical', 'thiersee', 'ziller', 'sv', 'time', 'swiss', 'last', 'kinder', 'flight', 'combine', 'valley', 'viaje', 'financially', 'hours', 'karnetów', 'og', 'march', 'ie', 'aid', 'guildford', 'oferta', 'tras', 'førstehåndsrapporter', 'packages', 'slots', 'karin', 'menu', 'priser', 'hervis', 'snörapport', 'things', 'connys', 'søk', 'fully', 'month', 'order', 'der', 'fly', 'rights', 'markbachjochbahn', 'wont', 'fr', 'apartments', 'taniej', 'events', 'forest', 'vær', 'like', 'details', 'overnatting', 'closed', 'tirol', 'term', 'locaties', 'vital', 'early', 'das', 'book', 'château', 'whats', 'bundesstrasse', 'hotel', 'nearby', 'adults', 'well', 'cz', 'schmiedau', 'seasonaires', 'maps', 'informes', 'soleil', 'exposure', 'season', 'hollaus', 'berglift', 'av', 'contact', 'half', 'schneebericht', 'otp', 'possibility', 'coach', 'thyon', 'raporty', 'powder', 'sprawdź', 'schneeberichte', 'due', 'alpinschiverleih', 'need', 'though', 'pertisau', 'roku', 'plan', 'tratí', 'kleinriedstraße', 'every', 'pricing', 'mosses', 'main', 'allow', 'wypożyczalnie', 'eases', 'go', 'precios', 'zprávy', 'leave'

**Local stem set:**

'report' ['reports', 'report'], 'plan' ['plan', 'planning'], 'relat' ['related'], 'nearbi' ['nearby'], 'weather' ['weather'], 'first' ['first'], 'hand' ['hand'], 'overview' ['overview'], 'review' ['reviews'], 'pist' ['piste'], 'map' ['map', 'maps'], 'planner' ['planner'], 'lift' ['lifts', 'lift'], 'ticket' ['tickets'], 'children' ['children'], 'year' ['year', 'years'], 'search' ['search'], 'check' ['check', 'checked'], 'price' ['prices', 'pricing'], 'avail' ['availability', 'available'], 'view' ['view'], 'deal' ['deal'], 'lesson' ['lessons'], 'skischulen' ['skischulen'], 'materialzustellung' ['materialzustellung'], 'waar' ['waar'], 'kun' ['kun'], 'je' ['je'], 'skimateria' ['skimateriaal'], 'huren' ['huren'], 'steinberg' ['steinberg'], 'een' ['een'], 'overzicht' ['overzicht'], 'van' ['van'], 'skiverhuur' ['skiverhuur'], 'locati' ['locaties'], 'sneeuwrapport' ['sneeuwrapport'], 'gelinkt' ['gelinkt'], 'aan' ['aan'], 'achenseetirol' ['achenseetirol'], 'oostenrijk' ['oostenrijk'], 'dichtbij' ['dichtbij'], 'christlum' ['christlum'], 'pertisau' ['pertisau'], 'weer' ['weer'], 'gebruikersrapport' ['gebruikersrapport'], 'verblijf' ['verblijf'], 'jouw' ['jouw'], 'accommodato' ['accommodatoe'], 'volwassen' ['volwassen'], 'kinderen' ['kinderen'], 'onder' ['onder'], 'jaar' ['jaar'], 'zoek' ['zoek'], 'control' ['controleer'], 'prijzen' ['prijzen'], 'en' ['en'], 'beschikbaarheid' 'beschikbaarheid'], 'bij' ['bij'], 'onz' ['onze'], 'goedkop' ['goedkoper'], 'bekijk' ['bekijk'], 'aanbiedingen' ['aanbiedingen'], 'skiutlei' ['skiutleie'], 'leie' ['leie'], 'av' ['av'], 'langrenn' ['langrenn'], 'og' ['og'], 'skiinfono' ['skiinfono'], 'tilknyttet' ['tilknyttet'], 'webkamera' ['webkamera'], 'overnat' ['overnatting'], 'planlegg' ['planlegg'], 'tur' ['tur'], 'innkvart' ['innkvartering'], 'voksn' ['voksne'], 'sjekk' ['sjekk'], 'billiger' ['billigere'], 'tilbud' ['tilbud'], 'skitim' ['skitimer'], 'nart' ['nart'], 'pogoda' ['pogoda'], 'planuj' ['planuj'], 'wyjazd' ['wyjazd'], 'w' ['w'], 'kameri' ['kamery'], 'raporti' ['raporty'], 'ze' ['ze'], 'stoku' ['stoku'], 'opini' ['opinie'], 'mapa' ['mapa'], 'tra' ['tras'], 'opadi' ['opady'], 'noclegi' ['noclegi'], 'ceni' ['ceny'], 'zaplanuj' ['zaplanuj'], 'zakwaterowani' ['zakwaterowanie'], 'osobi' ['osoby'], 'dzieci' ['dzieci'], 'roku' ['roku'], 'szukaj' ['szukaj'], 'u' ['u'], 'naszego' ['naszego'], 'partnera' ['partnera'], 'taniej' ['taniej'], 'zobacz' ['zobacz'], 'narciarski' ['narciarskie'], 'dostawa' ['dostawa'], 'miejsca' ['miejsca'], 'zakwaterowania' ['zakwaterowania'], 'alquil' ['alquiler'], 'equipo' ['equipos'], 'niev' ['nieve'], 'relacionada' ['relacionadas'], 'cercano' ['cercano'], 'tiempo' ['tiempo'], 'inform' ['information', 'informes'], 'alojamiento' ['alojamiento'], 'planifica' ['planifica'], 'tu' ['tu'], 'viaj' ['viaje'], 'adulto' ['adultos'], 'menor' ['menores'], 'buscar' ['buscar'], 'consult' ['consulte'], 'precio' ['precios'], 'disponibilidad' ['disponibilidad'], 'con' ['con']

**Collaboration with the UI:**

1.The query is received from the UI through an API call.

2.We receive the query and the query expansion method chosen by the user as the parameters to the query expansion part of the code. Then the documents(Solr results) are retrieved from the Solr collection. These documents are considered as local documents, and local vocabulary sets and local stem sets are created.

3. Based on the query expansion method(Association, Scalar or Metric)  specified by the user an expanded query is formed and this query is fed to the relevance models
4. The relevance model gets the result for the expanded query from the Solr indexed collection which is then passed to the UI to be visible to the user.

Aishwarya Vinod Menon (AXV220062), Sowmya Sivaramakrishnan (SXS230043), Anusha Gupta (AXG230026), Hanisha Anil Mohinani (HXM220089), Vidya Sreekumar (VXS22066)

## ***7. Discussion***

Upon completing our project, we've acquired a comprehensive understanding of the multifaceted elements necessary for a functional search engine. Our efforts resulted in a basic search tool capable of providing somewhat relevant results by applying the techniques taught in class. This endeavor not only provided us with practical insights into search engine development but also underscored the vastness of the field of Information Retrieval. A comparison of our outcomes with the prowess of major search engines like Google and Bing revealed the considerable gap between our efforts and the complexities involved in maintaining such widely-used platforms. It's evident that our project only scratched the surface of what's needed for a successful and current search engine.

## ***8. Conclusion***

In conclusion, for this project, we created a search engine — using the knowledge we gained throughout this class and some open-source programs — that focuses on information related to Religions. The five of us split the 5 different tasks required to build the search engine and collaborated to combine our individual parts into a final product that can provide relatively accurate information about Skiing. While there were many challenges along the way, we were able to overcome them and create a good search engine. This project provided us with an opportunity to apply the knowledge we gained in this class in a practical way and boosted our understanding of the theoretical concepts we learned.