The University of Texas at Dallas
CS 6322
Information Retrieval
Spring 2025

Class Project Report

Project Title: Search Engine for Cats

Group number: 2

Students:

Avaneesh Ramaseshan Baskaraswaminathan, axb220230@utdallas.edu
Venkata Subbaiah Pavan Karthik Navuluru, vxn220052@utdallas.edu
Dhanyan Muralidharan, dtm220000@utdallas.edu
Varsha Viswanathan, vxv230013@utdallas.edu
Sharon T Alexander sxa230048@utdallas.edu

1. The Problem

The main goal of our project was to develop a **Search Engine for Cats** that efficiently crawls, indexes, and ranks web pages related to cats. We focused on gathering web pages that provide useful, credible, and varied content related to cats, such as educational articles, adoption sites, grooming tips, and cat health resources. Our search engine leverages advanced relevance models like **vector space models**, **PageRank**, **HITS**, clustering, and **query expansion techniques** to improve search results.

Architecture of the Search Engine

The search engine was developed using a modular approach with each team member responsible for specific components. Below is an overview of the architecture:

- **Crawler (Varsha Viswanathan, VXV230013)**: Responsible for crawling **132,675 web pages** related to cats. The crawler utilizes the **Apache Nutch** framework for incremental crawling and handles URL extraction, crawling, and parsing. The web graph is also generated at this stage.

- **Indexer and Relevance Models (Venkata Subbaiah Pavan Karthik Navuluru, VXN220052)**: This part involves indexing the crawled data and implementing relevance models. The relevance models include:

    - **Vector Space Model**: A traditional information retrieval model based on term frequency and inverse document frequency (TF-IDF).

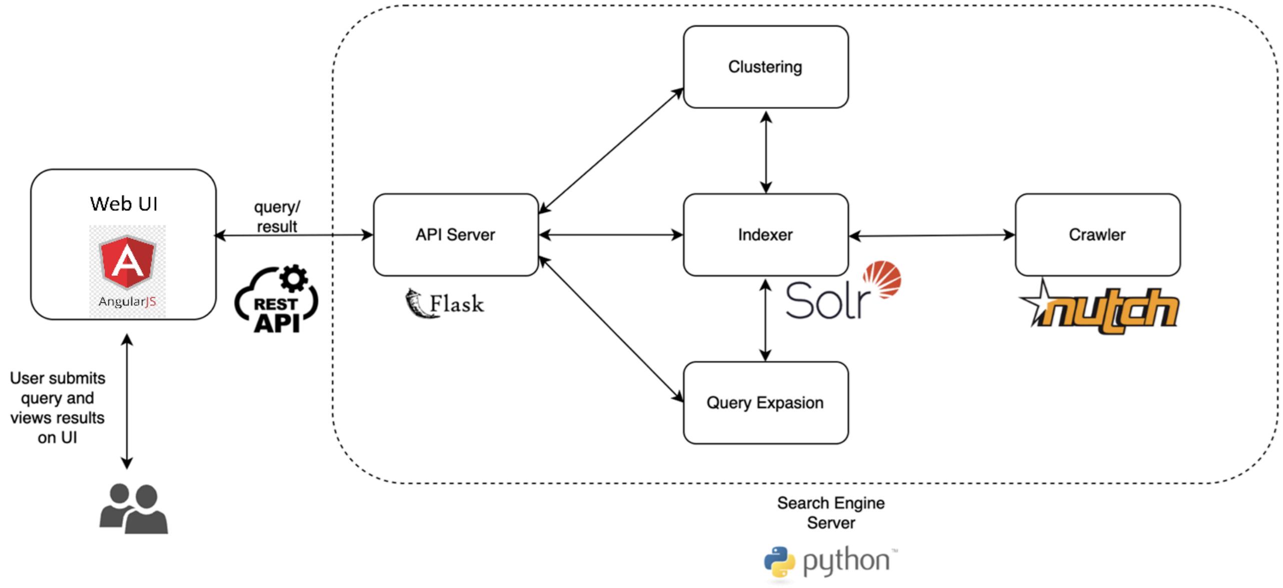
    - **PageRank and HITS**: Link-based models that help rank web pages based on their link structures.

    - **Combination of models**: Combining vector space with link-based models for improved accuracy.

- **User Interface (Avaneesh Ramaseshan Baskaraswaminathan, AXB220230)**: This component provides the user interface (UI) for the search engine. It displays the search results, relevance results, clusters, and expanded queries. The UI also integrates external search engines like Google and Bing for comparison. Built using Angular frontend framework, the APIs are hosted in a Python Flask backend to maintain accessibility.

- **Clustering (Dhanyan Muralidharan, DTM220000)**: The clustering component uses **flat clustering** and **agglomerative clustering** techniques. These clustering results are integrated with the relevance models to improve search results.

- **Query Expansion (Sharon T Alexander, SXA230048)**: Query expansion techniques are used to enhance the user search experience. The **Rocchio algorithm** was used for expanding the query based on the top documents retrieved. Additionally, clustering methods like **association**, **metric**, and **scalar clustering** were applied for further expansion.



What We Learned and Our Experience

- **Tech Stack Choices**: Initially, we considered using a **.NET framework** for the crawler and indexer. However, we faced limitations, particularly in terms of the indexing speed and performance. This led us to switch to **Apache Nutch** for crawling and **Solr** for indexing, both of which provided a much more scalable solution. The transition to these open-source tools significantly improved the performance of the system.

- **Challenges**: A major challenge we faced was related to **anchor text extraction** while using the **BeautifulSoup** library for parsing. We realized that BeautifulSoup's handling of anchor points wasn't accurate enough for our needs, leading to improper indexing and incomplete results.

- **Resolution**: We ultimately decided to focus on using **Nutch** with **Solr** to overcome the indexing and crawling limitations. By setting up Apache Nutch for crawling and integrating it with Solr for indexing and searching, we achieved a much more efficient pipeline for large-scale web crawling and indexing.

---

2. Crawling - Varsha Viswanathan (VXV230013)

Web Crawling Process

In this project, the main task was to crawl a set of web pages related to cats, including resources on cat care, adoption, health, and general information. I used **Apache Nutch** for web crawling, an open-source, scalable, and flexible web crawler. This was the first step towards gathering the necessary data for our search engine.

Gathering the Web Pages

To kickstart the crawling process, we initially selected **200 seed URLs** that provided a broad range of cat-related content. These seed URLs were carefully chosen to cover a variety of categories, including educational content, adoption sites, and health blogs. The URLs were divided into four groups of 50 URLs each, and the crawling was done iteratively in four cycles.

During each cycle, the crawler would visit the pages from the seed list and follow hyperlinks from these pages to discover additional pages. This iterative crawling approach helped gather a diverse set of data from a wide array of sources.

The seed URLs used for the crawling included reputable websites such as:

- **Pet Adoption Websites**: Example – Adopt-a-Pet, which provides adoption listings and resources for potential cat owners.
- **Cat Care Blogs**: Example – Kitten Lady, which provides advice on fostering and caring for kittens.
- **Cat Health Resources**: Example – ASPCA, which provides pet care information including health, safety, and grooming tips.
- **Cat Educational Websites**: Example – Cattime, Catster, and The Cat Network, which offer general cat care advice, adoption tips, and health resources.

After gathering these 200 seed URLs mentioned below, the crawler would explore links on these pages, recursively crawling through related documents. This expanded the web graph and included relevant content for indexing.

-Varsha vishwanathan (VXV230013)

Seed links used:

https://www.wikipedia.org/wiki/Cat

https://www.cats.org.uk/

https://www.purina.com/cats

https://www.petfinder.com/cats/

https://www.catsProtection.org/

https://www.thecatnetwork.org/

https://www.aspca.org/pet-care/cat-care

https://www.adoptapet.com/cats

https://www.petco.com/shop/en/petcostore/category/cats https://www.meowfoundation.com/

https://www.britannica.com/animal/cat-mammal

https://www.catster.com/

https://www.cats.about.com/

https://www.kittens-rescue.com/

https://www.cattime.com/

https://www.peta.org/issues/animals-in-entertainment/cats-in-entertainment/

https://www.akc.org/dog-breeds/cat-breeds/

https://www.catworld.co.uk/

https://www.washingtonpost.com/news/inspired-life/wp/2018/12/19/the-most-popular-cats-breeds-in-the-u-s-according-to-the-cat-fanciers-association/

https://www.theholidayspot.com/cats/

https://www.cat-lovers.co.uk/

https://www.lovemypet.com/cats/

https://www.catsandmeows.com/

https://www.petmd.com/cat

https://www.goodhousekeeping.com/pets/cats/

https://www.mypet.com/cats

https://www.vetstreet.com/cats

https://www.americanhumane.org/fact-sheet/cats/

https://www.meowbox.com/

https://www.nationalgeographic.com/animals/mammals/facts/domestic-cat

https://www.petplan.com/pet-care/cat-care/cat-health-101/

https://www.dailypaws.com/cats

https://www.catsrule.com/

https://www.catswhiskers.com/

https://www.smithsonianmag.com/science-nature/cats-are-not-domestic-animals-130982452/

https://www.southernliving.com/pets/cat-breeds

https://www.ehow.com/animals/pets/cats/

https://www.reddit.com/r/cats/

https://www.meow.com/

https://www.kittysnuggles.com/

https://www.worldwidecat.com/

https://www.purrfectpost.com/

https://www.catify.com/

https://www.catacult.com/

https://www.groomers.co.uk/cat-grooming/

https://www.petfinder.com/cat-adoption/

https://www.litter-robot.com/blog/

https://www.shopee.com.my/cats

https://www.catsafari.com/

https://www.cat-care-tips.com/

https://www.nature.com/articles/s41598-019-39655-9

https://www.cats.org.nz/

https://www.catbreedslist.com/

https://www.catalystpets.com/

https://www.humanesociety.org/resources/cats

https://www.cattitude.com/

https://www.felinefury.com/

https://www.kittygenius.com/

https://www.catworld.com.au/

https://www.bestfriends.org/cats

https://www.ehow.com/animals/pets/cats/how-to-train-cats

https://www.catsanddogs.org/

https://www.petcha.com/cats/

https://www.purrfectfur.com/

https://www.mysweetcat.com/

https://www.catrevolution.com/

https://www.dailypetcare.com/cat

https://www.purrrfect.com/

https://www.huffpost.com/topic/cats

https://www.vetstreet.com/cats/cat-breeds

https://www.wildcatconservancy.org/

https://www.pawprintsandpurrs.com/

https://www.cattherapy.org/

https://www.catsmeow.net/

https://www.houzz.com/photos/query/cats

https://www.pawpals.com/cats/

https://www.coolcats.com/

https://www.dogandcatlovers.com/

https://www.cats2love.com/

https://www.catinthebox.com/

https://www.acatslife.com/

https://www.animalplanet.com/pets/cats

https://www.petexpress.com/cats/

https://www.catsattack.com/

https://www.pethub.com/cats

https://www.catswhisperer.com/

https://www.furvana.com/

https://www.cattocanine.org/

https://www.adopt-a-pet.com/cats

https://www.nursethecat.com/

https://www.felinecareclinic.com/

https://www.catfoster.com/

https://www.catgiving.com/

https://www.purrfectrescue.com/

https://www.catsunlimited.org/

https://www.catwatch.com/

https://www.felinepets.org/

https://www.kittentown.com/

https://www.catscare.org/

https://www.catsthatcare.org/

https://www.meowmax.com/

https://www.catcarecenter.com/

https://www.handsforcats.org/

https://www.petplace.com/cats/

https://www.purrfectcause.com/

https://www.furkitty.com/

https://www.thedailymeow.com/

https://www.purrfectpaw.com/

https://www.catcareassociation.com/

https://www.dailytail.com/cats/

https://www.tailsofmeow.com/

https://www.catalystfeline.com/

https://www.catacademy.com/

https://www.catpaws.com/

https://www.thekittenrescue.com/

https://www.catcharity.org/

https://www.felinefoster.org/

https://www.catexpansion.com/

https://www.meowpalace.com/

https://www.catpromises.com/

https://www.felinehouse.com/

https://www.thecatcastle.com/

https://www.felinefosterhome.com/

https://www.catfriends.com/

https://www.purrrfectcat.com/

https://www.kittycorner.org/

https://www.catshelter.org/

https://www.cataid.com/

https://www.catsinneed.com/

https://www.pawpalsrescue.com/

https://www.purrfectanimals.com/

https://www.kittykittens.com/

https://www.catwalkers.com/

https://www.catpack.com/

https://www.catclubhouse.com/

https://www.kittensinc.org/

https://www.thekittycorner.com/

https://www.catcoach.com/

https://www.felinefamily.org/

https://www.catscareclinic.com/

https://www.kittypedia.com/

https://www.purrrfecthelp.com/

https://www.felineworld.com/

https://www.kittycornerstore.com/

https://www.catmojo.com/

https://www.catwise.com/

https://www.catstream.com/

https://www.kittywatch.com/

https://www.petmd.com/cat

https://www.thepurrfectmatch.com/

https://www.catwatch.com/

https://www.felinefrenzy.com/

https://www.purrfectpaw.com/

https://www.catclinic.com/

https://www.felinecareclinic.com/

https://www.catcorner.com/

https://www.catacademy.com/

https://www.felineworld.com/

https://www.purrrfectcat.com/

https://www.catpaws.com/

https://www.kittycornerstore.com/

https://www.purrfectcause.com/

https://www.catfoster.com/

https://www.felinepets.org/

https://www.tailsofmeow.com/

https://www.catcareassociation.com/

https://www.purrrfectrescue.com/

https://www.felinefriends.org/

https://www.kittentown.com/

https://www.catfriends.com/

https://www.kittyadopt.com/

https://www.catsthatcare.org/

https://www.pawpalsrescue.com/

https://www.catpromises.com/

https://www.catwalkers.com/

https://www.felinefamily.org/

https://www.catcoaching.com/

https://www.catcoach.com/

https://www.catpack.com/

https://www.catcornerstore.com/

https://www.purrfectanimals.com/

https://www.thekittycorner.com/

https://www.felinefury.com/

https://www.catwise.com/

https://www.catalystpets.com/

https://www.felineharmony.com/

https://www.catwalkers.com/

https://www.purrfectplay.com/

https://www.catsalliance.org/

https://www.catcentral.com/

https://www.catcafe.org/

https://www.furkitty.com/

https://www.catwise.com/

https://www.catstream.com/

https://www.kittywatch.com/

https://www.catlibrary.com/

https://www.cats.org/

https://www.catbreedslist.com/

https://www.catloverscorner.com/

https://www.thecatcafe.com/

https://www.felinezone.com/

https://www.catfosterhome.org/

https://www.kittensforadoption.com/

https://www.felinerescue.org/

https://www.catsafe.org/

https://www.thekittenkitchen.com/

https://www.catsofinstagram.com/

https://www.catsndogs.org/

https://www.care2.com/cats

https://www.adoptacathouse.org/

https://www.meowbox.com/

https://www.felinefoundation.org/

https://www.catsoftheworld.com/

https://www.petsmart.com/cats/

https://www.catadvisor.com/

https://www.catcareadvice.com/

https://www.catrevolution.com/

https://www.purrfectmatch.com/

Web Pages Crawled

The total number of web pages crawled and used for the project was **approximately 135,000**. The specific pages crawled were related to various topics such as:

- Adoption sites offering listings of cats available for adoption
- Health websites containing information on common cat diseases and treatments
- Blogs focused on kitten care, general pet health, and advice for cat owners
- Informational resources from well-known animal welfare organizations like the ASPCA and PETA

Some of the key sources we utilized for our crawl include:

1. **Adopt-a-Pet** (Pet adoption platform)
2. **Kitten Lady** (Kitten care blog)
3. **ASPCA** (Pet care resources and guidelines)
4. **Cattime** (General cat information site)
5. **PETA** (Animal rights and health organization)

De-duplication and Crawl Management

One of the major challenges in web crawling is ensuring that duplicate pages are not crawled repeatedly. To solve this problem, we implemented a **URL filtering mechanism** within the Apache Nutch framework. This mechanism checks for duplicate URLs and prevents the crawler from revisiting the same page. By storing previously visited URLs and their metadata, the crawler ensures that only unique URLs are processed.

Additionally, we incorporated an incremental crawling strategy. This means that the system does not start from scratch each time; instead, it continues to crawl new pages while avoiding revisiting old ones. This also helped in reducing unnecessary load on the web servers.

Hyperlink Information for Indexing

The collected data from the crawler, including **hyperlink information**, was passed to the student responsible for **index creation and relevance modeling**. The hyperlinks between the pages were crucial for building the **web graph**, which was necessary for implementing link-based relevance models such as **PageRank** and **HITS**.The hyperlinks extracted during the crawling process were formatted in a structured manner and included as metadata in the Solr index. This was done by ensuring that the **inlinks** and **outlinks** (i.e., links pointing to and from each web page) were properly stored in the index for use in the relevance models.

I also made sure that this data was formatted as **JSON** to align with the Solr requirements, ensuring that it could be processed and indexed effectively. The JSON format allowed us to store each web page's metadata, including the page's **URL**, **title**, **content**, and **link structure**.

-Varsha vishwanathan (VXV230013)

Configuring Solr to Count for Scores

The Solr server used in the project was configured to accept and index the **boost** values and other metadata extracted from the crawled pages. I modified the Solr schema to include fields that could store the **inlink** and **outlink** counts, **boost scores**, and other necessary attributes.

The boost score, which indicates the relevance of a web page in relation to the query, was particularly important for the relevance models. To store this score properly, I made sure to adjust the configuration of Solr to use numeric fields for the **boost** values. This allowed Solr to correctly interpret the values during indexing and provide accurate search results based on the relevance of the pages.

I also configured Solr to index **links** and **scores** for future use in the relevance modeling and ranking process. The data collected from the crawling process, including these values, was passed directly to Solr for indexing.

Summary

In summary, the web crawling module was responsible for gathering approximately **135,000 web pages** related to cats. The URLs for the crawl were sourced from a set of **200 seed URLs**, which were crawled iteratively in four cycles of 50 URLs each. A de-duplication mechanism was implemented to avoid repeated crawling of the same URLs, and hyperlink information was passed to the indexing module for further processing. The Solr server was properly configured to handle the boost values and link-based metadata, which played a crucial role in the relevance models.

By implementing these processes, we ensured that the crawler efficiently gathered high-quality, relevant data for the search engine, contributing to the overall performance and accuracy of our project.

-Varsha vishwanathan (VXV230013)

3. Indexing and Relevance – Venkata Subbaiah Pavan Karthik Navuluru (VXN220052)

Indexing Process:

- Used Apache Solr with input from crawldb, linkdb, segments

Commands used :

1) /bin/crawl -i -D solr.server.url=http://localhost:8983/solr/nutch -s ${NUTCH_RUNTIME_HOME}/urls ${NUTCH_RUNTIME_HOME}/crawl 10

   Where -i indicates indexing data after crawling.

2) bin/nutch index crawl/crawldb -linkdb crawl/linkdb -dir crawl/segments -filter -normalize -deleteGone
3) bin/nutch webgraph -filter -normalize -segmentDir crawl/segments/ -webgraphdb crawl/ → construction of web graph

Next Steps taken:

Constructed a directed web graph by combining URL data from Solr's indexed documents and the Apache Nutch-generated inlink database (part-r-00000). Edges were added between source and target URLs only when both were confirmed to be present in the Solr dataset, ensuring consistency between the indexed content and the graph structure. Using this graph, we computed two key link analysis metrics: HITS (Hyperlink-Induced Topic Search) authority scores and PageRank values, both of which help identify the most influential or authoritative pages in the network. The results were filtered to retain only Solr-matched URLs, and the highest scoring URLs by authority and PageRank were identified and recorded for further evaluation of link-based relevance.

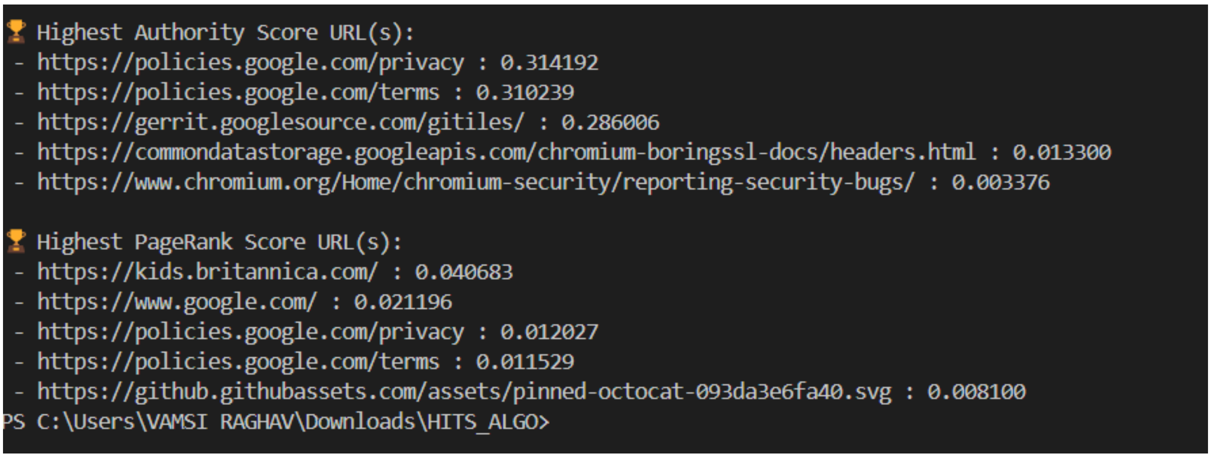
Web statistics :

- Nodes : 75991
- Links : 233242
- Max inlinks : 10000
- Max out Links : 81

Relevance Models:

.Page Rank and HITS via Python networkx library and Solr indexed data.

-Venkata Subbaiah Pavan Karthik Navuluru (vxn220052)



```
🏆 Highest Authority Score URL(s):
- https://policies.google.com/privacy : 0.314192
- https://policies.google.com/terms : 0.310239
- https://gerrit.googlesource.com/gitiles/ : 0.286006
- https://commondatastorage.googleapis.com/chromium-boringssl-docs/headers.html : 0.013300
- https://www.chromium.org/Home/chromium-security/reporting-security-bugs/ : 0.003376

🏆 Highest PageRank Score URL(s):
- https://kids.britannica.com/ : 0.040683
- https://www.google.com/ : 0.021196
- https://policies.google.com/privacy : 0.012027
- https://policies.google.com/terms : 0.011529
- https://github.githubassets.com/assets/pinned-octocat-093da3e6fa40.svg : 0.008100
PS C:\Users\VAMSI RAGHAV\Downloads\HITS_ALGO>
```

Highest hit score : 0.314192 URL : https://policies.google.com/privacy

Highest page rank score : 0.040683 URL : https://kids.britannica.com/

Collaboration with UI : Collaborated with the UI student to ensure that the relevance models I developed—HITS (authority and hub scores) and PageRank—were accurately integrated into the search engine interface. I generated these scores by constructing a web graph from the Nutch `inlinkdb`, filtering nodes to include only URLs present in the Solr index. Once the authority and PageRank scores were computed using NetworkX, I provided the output in structured `.txt` files formatted as JSON. These files were then used by the UI student to reorder Solr query results dynamically based on relevance scores, allowing the interface to display the most authoritative or important pages at the top. Our collaboration ensured that users received more meaningful search results backed by graph-based ranking.

Collaboration with Clustering and QE

→ Tested 20 queries collaboratively with QE and Clustering teammates;

→ Observed better semantic performance with HITS

-Venkata Subbaiah Pavan Karthik Navuluru (vxn220052)

4. User Interface – Avaneesh Ramaseshan Baskaraswaminathan (AXB220230)

Describe how you have designed the interface

For our search engine focusing on cats, we designed an intuitive user interface using the Angular framework for the frontend and Flask for the backend. The interface features a simple search bar where users can input their queries. Below the search bar, we have 8 options in the form of radio buttons to choose the ranking method, and a submit button. Results are displayed in 3 separate sections:

1. 1st showing the results of our custom search engine on cats
2. 2nd showing results from google.com
3. 3rd showing results from bing.com

The Google and Bing iframes start with their respective home pages, while our custom iframe starts out empty. To get the results, we make a GET request with the query and type parameters to an API that connects to the relevance models (PageRank, HITS) and to clustering and query expansion modules. The API uses these parameters to return a JSON file with ranked results.

The 8 radio button options provided on the interface are:

1. Page Rank – API returns JSON with PageRank-based relevance
2. HITS – API returns JSON with HITS-based relevance
3. Flat Clustering – JSON with Flat Clustering results
4. Hierarchical Clustering (Single Linkage)
5. Hierarchical Clustering (Complete Linkage)
6. Metric Query Expansion
7. Association Query Expansion
8. Scalar Query Expansion

We ran extensive tests using a mix of queries provided by the clustering module and ones created internally. We used 50 queries for clustering relevance evaluation and 20 queries for testing the Rocchio algorithm and its expanded results.

How you have worked with the student that has generated the index

I collaborated with Pavan Karthik to develop the API and logic required to fetch search results. We exposed the indexing and ranking methods as backend functions with query and type as parameters. These were called from the frontend using Angular over HTTP to generate live search results. We tested 20 queries together.

Number of queries used for testing

We used 20 total queries to test the search engine post-indexing and building relevance models. Examples of these queries:

1. Queries tested in collaboration with the relevance model developer:
2. Cat breeds
3. Feline diabetes symptoms
4. Best cat litter types
5. Vaccination schedule for cats
6. Grooming long-haired cats
7. Common cat illnesses

Queries I created independently:

1. DIY cat toys
2. Funny cat memes
3. Signs of a happy cat
4. Cat behavior explained
5. Why do cats purr
6. Are cats nocturnal
7. Adopting senior cats
8. Cat food for kittens
9. How to litter train a cat

These queries helped cover the full range of features – including base ranking, clustering, and expansion models. We selected them to mimic what a real cat owner or researcher might search for.

Collaboration with the student that produced clusters

I worked with Dhanyan to integrate clustering output into the frontend. We exported cluster data via flat files and indexed formats. I created frontend logic that uses this clustering metadata to group search results by cluster ID. The results were labeled and styled distinctly to improve readability.

Comparison with Google and Bing

Search result relevance:
Google and Bing perform better overall due to larger indexes, but our engine retrieves more focused and topical results for cat-specific queries.
Search result speed:
Our engine is faster locally due to reduced latency and direct file-based data access.
User experience:
Google and Bing offer more comprehensive UI features, but our interface is simplified and highly configurable by students.

Privacy and Security:
Our system lacks encryption or secure login, unlike Google and Bing. This could be improved in future versions.

How clustering results were used in the UI

We allow users to select clustering modes (Flat, Single-Link, Complete-Link). Once selected, the backend preloads cluster labels and attaches them to results. These are displayed in collapsible or grouped sections in the frontend for better semantic understanding.

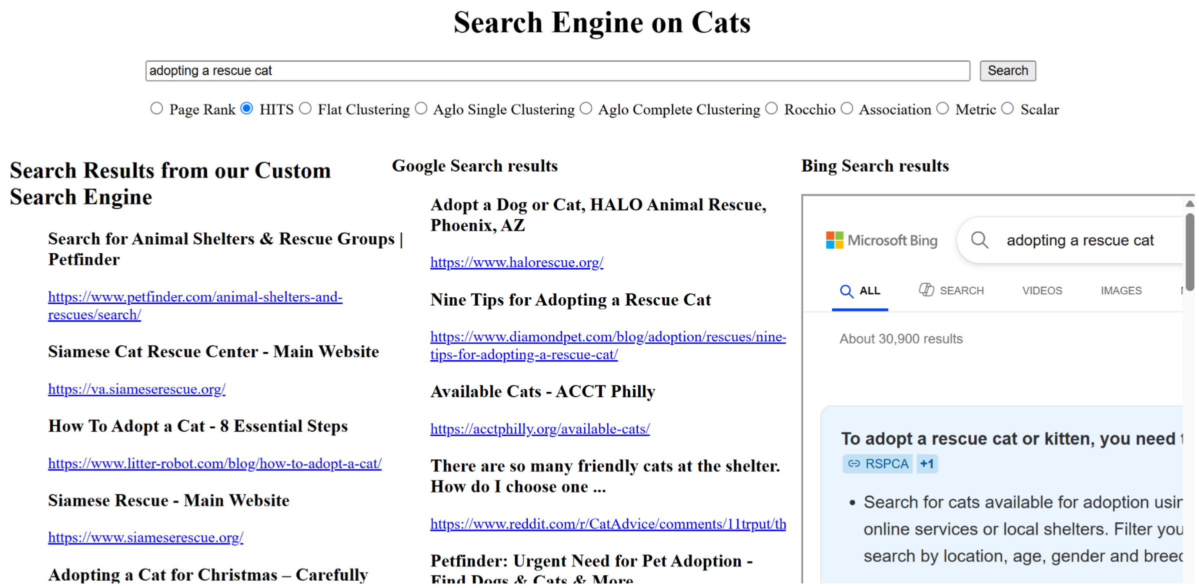
Query selection for demonstration

For demonstration, we selected queries that:

1. Yielded high-quality, verifiable results
2. Represented different search engine modules (ranking, clustering, expansion)
3. Would be relatable to general users or cat owners

Demo Queries

1. Adopting a rescue cat - using HITS



Search Engine on Cats

adopting a rescue cat | Search

○ Page Rank ● HITS ○ Flat Clustering ○ Aglo Single Clustering ○ Aglo Complete Clustering ○ Rocchio ○ Association ○ Metric ○ Scalar

Search Results from our Custom Search Engine

Search for Animal Shelters & Rescue Groups | Petfinder

https://www.petfinder.com/animal-shelters-and-rescues/search/

Siamese Cat Rescue Center - Main Website

https://va.siameserescue.org/

How To Adopt a Cat - 8 Essential Steps

https://www.litter-robot.com/blog/how-to-adopt-a-cat/

Siamese Rescue - Main Website

https://www.siameserescue.org/

Adopting a Cat for Christmas – Carefully

Google Search results

Adopt a Dog or Cat, HALO Animal Rescue, Phoenix, AZ

https://www.halorescue.org/

Nine Tips for Adopting a Rescue Cat

https://www.diamondpet.com/blog/adoption/rescues/nine-tips-for-adopting-a-rescue-cat/

Available Cats - ACCT Philly

https://acctphilly.org/available-cats/

There are so many friendly cats at the shelter. How do I choose one ...

https://www.reddit.com/r/CatAdvice/comments/11trput/th

Petfinder: Urgent Need for Pet Adoption - Find Dogs & Cats & More

Bing Search results

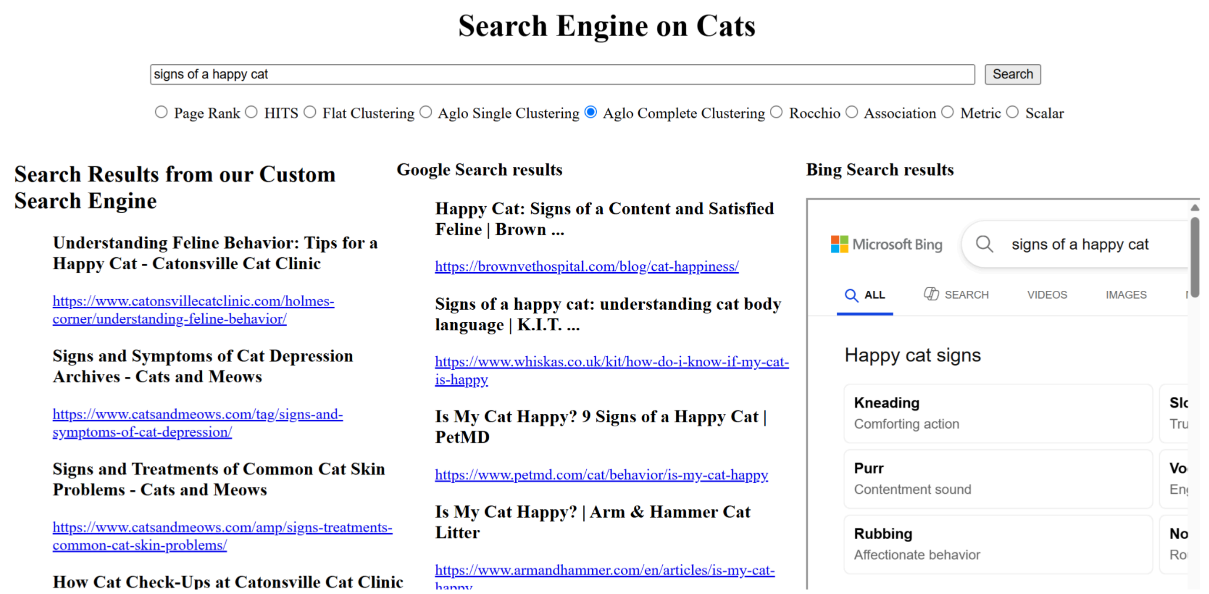
About 30,900 results

To adopt a rescue cat or kitten, you need t

RSPCA +1

- Search for cats available for adoption usir online services or local shelters. Filter you search by location, age, gender and breed

2. Signs of a happy cat - using Aglommerative Complete Clustering



Search Engine on Cats

signs of a happy cat | Search

○ Page Rank ○ HITS ○ Flat Clustering ○ Aglo Single Clustering ● Aglo Complete Clustering ○ Rocchio ○ Association ○ Metric ○ Scalar

Search Results from our Custom Search Engine

Understanding Feline Behavior: Tips for a Happy Cat - Catonsville Cat Clinic

https://www.catonsvillecatclinic.com/holmes-corner/understanding-feline-behavior/

Signs and Symptoms of Cat Depression Archives - Cats and Meows

https://www.catsandmeows.com/tag/signs-and-symptoms-of-cat-depression/

Signs and Treatments of Common Cat Skin Problems - Cats and Meows

https://www.catsandmeows.com/amp/signs-treatments-common-cat-skin-problems/

How Cat Check-Ups at Catonsville Cat Clinic

Google Search results

Happy Cat: Signs of a Content and Satisfied Feline | Brown ...

https://brownvethospital.com/blog/cat-happiness/

Signs of a happy cat: understanding cat body language | K.I.T. ...

https://www.whiskas.co.uk/kit/how-do-i-know-if-my-cat-is-happy

Is My Cat Happy? 9 Signs of a Happy Cat | PetMD

https://www.petmd.com/cat/behavior/is-my-cat-happy

Is My Cat Happy? | Arm & Hammer Cat Litter

https://www.armandhammer.com/en/articles/is-my-cat-happy

Bing Search results

Microsoft Bing | signs of a happy cat

ALL SEARCH VIDEOS IMAGES

Happy cat signs

Kneading
Comforting action

Purr
Contentment sound

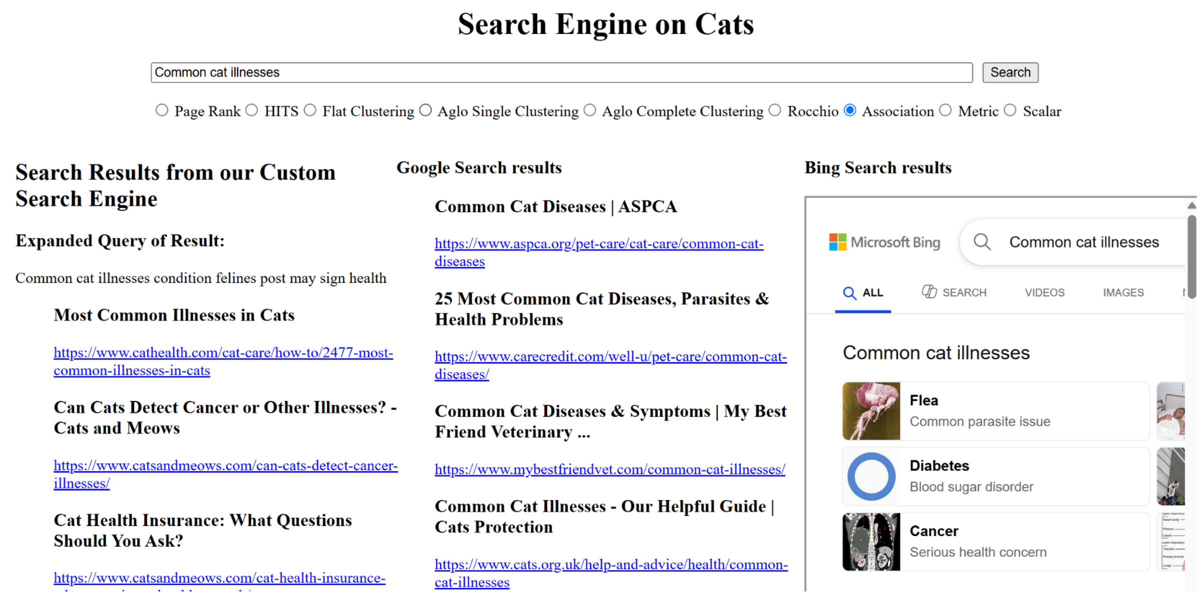
Rubbing
Affectionate behavior

Slo
Tru

Vo
En

No
Ro

3. Common cat illnesses - using Association Query Expansion



Search Engine on Cats

Common cat illnesses | Search

○ Page Rank ○ HITS ○ Flat Clustering ○ Aglo Single Clustering ○ Aglo Complete Clustering ○ Rocchio ● Association ○ Metric ○ Scalar

Search Results from our Custom Search Engine

Expanded Query of Result:

Common cat illnesses condition felines post may sign health

Most Common Illnesses in Cats

https://www.cathealth.com/cat-care/how-to/2477-most-common-illnesses-in-cats

Can Cats Detect Cancer or Other Illnesses? - Cats and Meows

https://www.catsandmeows.com/can-cats-detect-cancer-illnesses/

Cat Health Insurance: What Questions Should You Ask?

https://www.catsandmeows.com/cat-health-insurance-

Google Search results

Common Cat Diseases | ASPCA

https://www.aspca.org/pet-care/cat-care/common-cat-diseases

25 Most Common Cat Diseases, Parasites & Health Problems

https://www.carecredit.com/well-u/pet-care/common-cat-diseases/

Common Cat Diseases & Symptoms | My Best Friend Veterinary ...

https://www.mybestfriendvet.com/common-cat-illnesses/

Common Cat Illnesses - Our Helpful Guide | Cats Protection

https://www.cats.org.uk/help-and-advice/health/common-cat-illnesses

Bing Search results

Microsoft Bing | Common cat illnesses

ALL SEARCH VIDEOS IMAGES

Common cat illnesses

Flea
Common parasite issue

Diabetes
Blood sugar disorder

Cancer
Serious health concern

5. Clustering – Dhanyan Muralidharan (DTM220000)

Describe how you have designed the flat clustering – how many predefined clusters did you select, and why (10 points)?

The design of the flat clustering starts with pre-processing the clustered results, which are presented in columns—URL, title, and content.

The following steps make up the **pre-processing phase:**

For each document, we:

- Get the URL and `content`.
- Tokenize the content.
- Clean tokens (remove stopwords, punctuation, lowercase).
- Rejoin cleaned tokens into a cleaned text string.

Following the data pre-processing phase, we vectorize the page content using **TF-IDF Vectorization.** We ignore terms that appear in less than 10% or more than 50% of documents. We also ignore stopwords.

Now, the TF-IDF matrix's dimensionality is reduced using Truncated SVD, and the vectors are normalized to unit length. Thus, we end up with a dense, low-dimensional representation of the documents, ready for clustering.

Identifying the Optimal Number of Clusters:

I employed a combination of two methods to determine the optimal K value, namely:

1) **Silhouette Score:** A clustering quality evaluation metric that quantifies how well each data point fits into its assigned cluster.

   Mathematically,

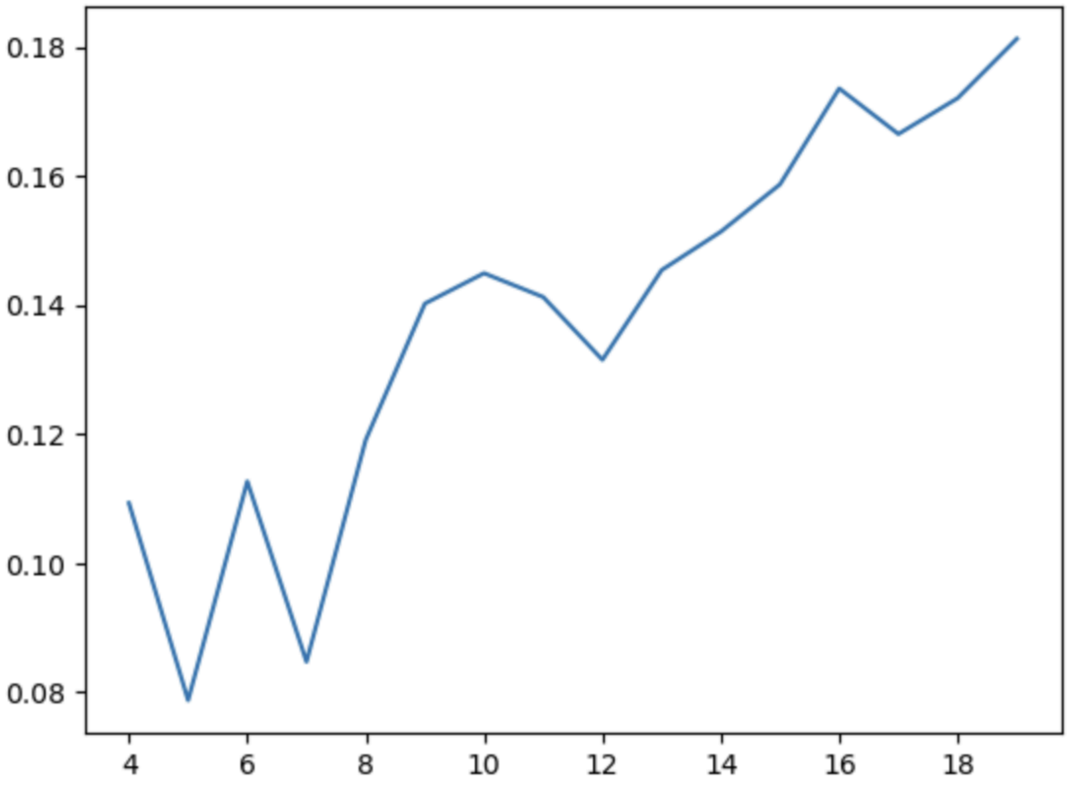
   $$\text{Silhouette}(i) = b(i) - a(i) / \max\{a(i), b(i)\}$$

   Where:

   - a(i) denotes the average distance from point i to all other points in the same cluster (intra-cluster distance).

   - b(i) denotes the lowest average distance from point i to all points in any other cluster (nearest-cluster distance).

Below is the silhouette score graph for my case:



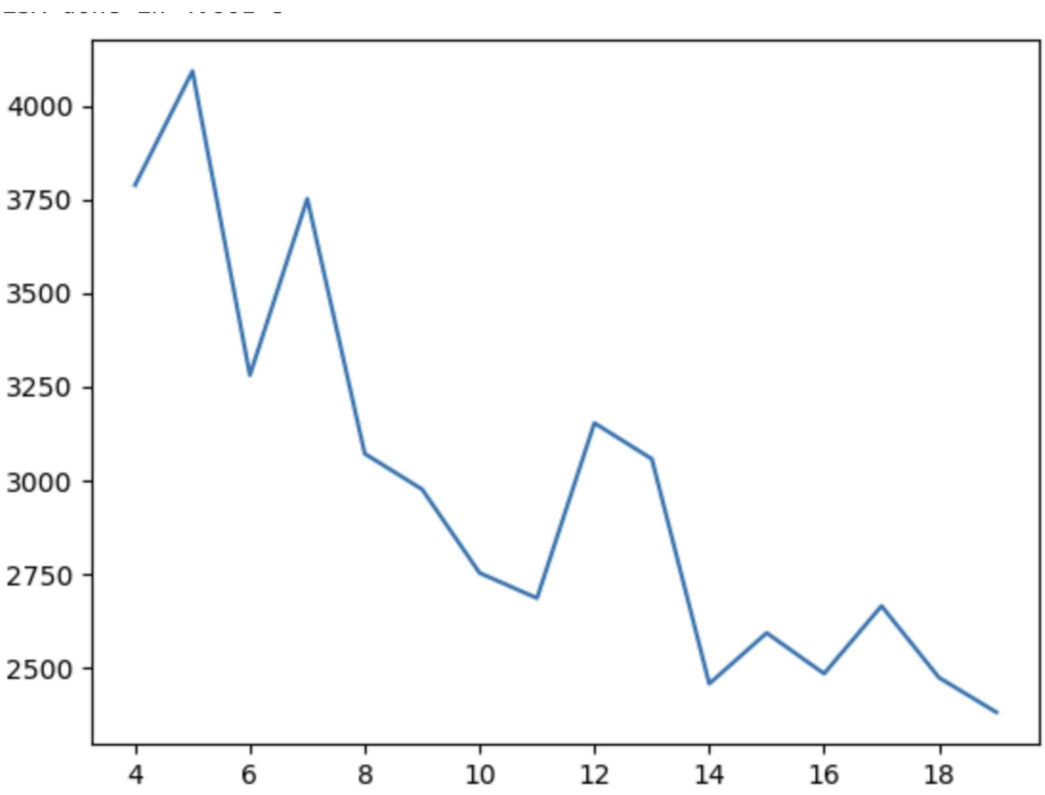
Silhouette score vs K value

(Graph 1)

2) **Sum of Squared Errors:** A measure that quantifies the compactness of the clusters. Also called Inertia, this method measures the total squared distance between each point and its assigned cluster center.

   The intended approach is to plot the K value vs. SSE graph and identify the K value that corresponds to a sharp “elbow.” After this point, increasing the K value gives diminishing returns.

   For my case, here is the Elbow curve.



SSE vs K value

(Graph 2)

Key inferences from the graphs:

1) A steadily increasing silhouette score in Graph-1 would often indicate that the data doesn’t have strong natural clusters. The increase in the score may be attributed to overfitting. Thus, the best course of action would be to pick a “knee point” from where the score starts to increase more slowly. In Graph 1, this point is somewhere between 10 and 12.
2) In Graph 2, we are looking for an elbow point where there is a sharp fall in the rate of decrease (around 10 - 11 fits the description).

Thus, by combining the inferences from SSE and silhouette score graphs, I decided on K = 11.

What did you do with the results of clustering – did you incorporate them in the relevance models – and did you provided to the user interface results that were obtained when clustering is used? (8 points)

I utilized the results of clustering in the following manner:

1) **Cluster-level Vectorization of Document Content**
   To represent each cluster in a form that could be meaningfully compared to user queries, I performed a cluster-level vectorization operation. Specifically, I used the textual content within each of the 10 clusters formed during the clustering phase and applied TF-IDF-based vectorization. This involved aggregating the textual data within a cluster, then transforming this aggregated content into a dense vector representation.

2) **Query Vectorization using the Same Embedding Scheme**
   When a user submits a search query through the interface, it undergoes the exact same preprocessing and vectorization pipeline as the document clusters. This includes tokenization, stopword removal, and transformation using the pre-fitted TF-IDF vectorizer. By embedding the query in the same feature space as the clusters, the system ensures compatibility and consistency in semantic representation between the user query and the precomputed cluster vectors.

3) **Cosine Similarity Computation for Cluster Matching**
   After obtaining a vector representation for the user query, a pairwise cosine similarity operation is performed between this query vector and each of the 10 cluster vectors. Cosine similarity measures the angular distance between vectors in high-dimensional space, offering a normalized similarity metric regardless of document length. The cluster whose vector has the highest cosine similarity score with the query vector is deemed the most relevant to the user's intent, and is therefore selected for result retrieval.

4) **Retrieval and Randomization of Top Results from the Most Relevant Cluster**
   Once the most relevant cluster is identified based on cosine similarity, the system proceeds to select a subset of results from it. From the pool of documents belonging to this cluster, the top 15 URLs are randomly sampled to ensure a diverse set of documents is shown to the user. These selected URLs are then sent to the front-end application and displayed to the user upon the click of the "Flat Clustering" button. This mechanism provides a quick and intuitive way for users to explore topic-specific documents without having to navigate through irrelevant information.

Following this methodology enabled me to incorporate the clustering results in the relevance models. These results were passed on to the User Interface to be availed through radio buttons.

How did you use the results of agglomerative clustering (7 points) – how many clusters did you obtain (5 points)? How were they presented on the user interface (2 points)?

Agglomerative clustering uncovers potential hierarchical relationships between documents that flat clustering might miss. This approach is beneficial when the data might have nested or subtle topic groupings that don't appear clearly with methods like KMeans.

I used Agglomerative Clustering with Single and Complete linkages and Euclidean distance on the same LSA-reduced TF-IDF vectors that were prepared for flat clustering. This method builds a tree (dendrogram) of clusters by iteratively merging the two closest groups.

To use the results effectively, I set a distance threshold that controls how similar two clusters must be to be merged. This allows for automatic determination of the number of clusters based on the data's natural hierarchy rather than a fixed K.

Each document was assigned to a cluster based on this hierarchy. The content of each cluster was vectorized, and cosine similarity was again used to match the user query to the most semantically similar cluster, similar to the flat clustering approach

Number of Clusters:

Using visual inspection of the dendrogram and the inconsistency method, I determined that the optimal number of clusters for the agglomerative method was 10. This value best balanced granularity with semantic cohesion. The choice was validated by observing cluster compactness, distinctiveness, and interpretability of grouped topics. Fewer clusters resulted in overly broad groups, while more than 6 led to fragmentation of coherent subtopics.

User Interface:

The agglomerative clustering results were integrated into the user interface using two dedicated radio buttons labeled Agglomerative single and Agglomerative Complete. When these options are selected, the query goes through the same preprocessing and vectorization pipeline and the results from the most similar agglomerative cluster are retrieved and shown.

This setup allows seamless switching between flat and hierarchical clustering results, giving users more control over the specificity of search outcomes.

How many queries did you experiment with – such that clustering could be used to improve the results of your search engine (3 points)?

I experimented with over 40 different search queries to evaluate and fine-tune the impact of clustering (both flat and agglomerative) on search relevance. These queries varied in complexity and specificity, including:

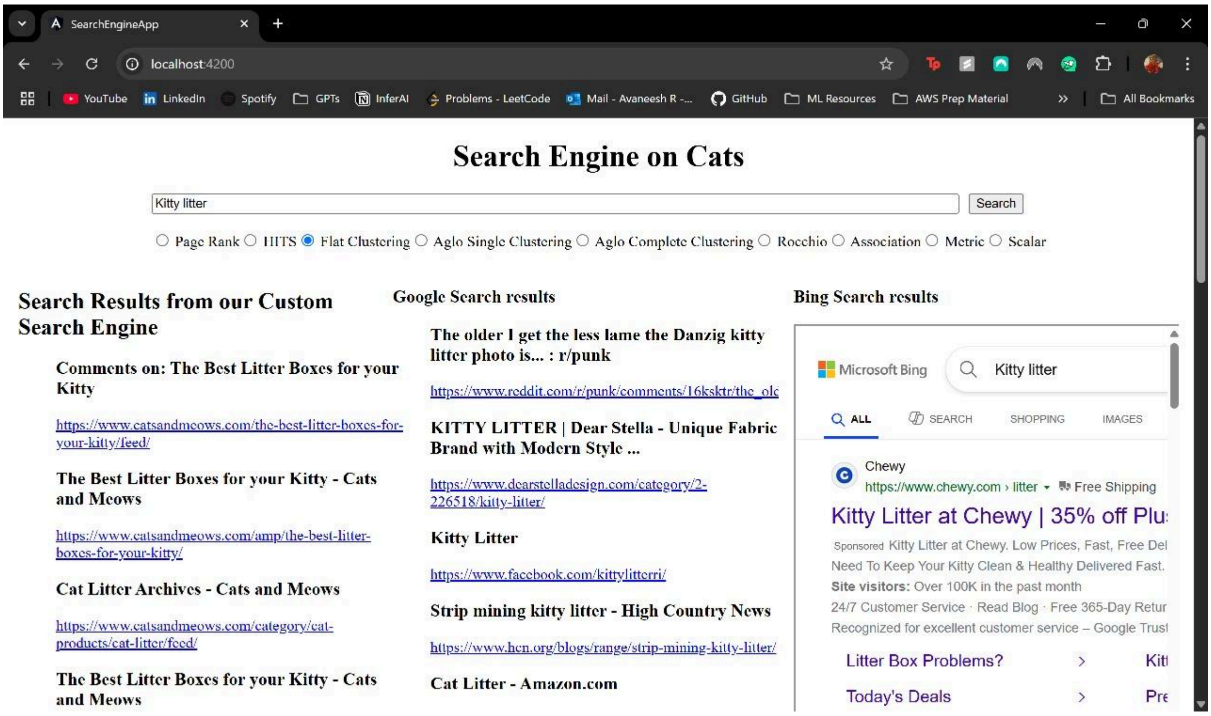
- Kitty Litter Cleaning process

- Lifespan of a Persian Cat
- Cat shedding seasons
- Why does a sphynx cat not grow hair?
- Cat common eye colors
- Breakfast ideas for cats
- Number of cat breeds, to name a few!

These tests helped assess which clustering approach yielded better relevance, topic diversity, and user experience for various kinds of search intents.

Provide three examples of the queries and the results produced by your search engine and the clusters that you have created (5 points).

Example 1: Kitty Litter



Search Engine on Cats

Kitty litter [Search]

○ Page Rank ○ HITS ● Flat Clustering ○ Aglo Single Clustering ○ Aglo Complete Clustering ○ Rocchio ○ Association ○ Metric ○ Scalar

Search Results from our Custom Search Engine

Comments on: The Best Litter Boxes for your Kitty

https://www.catsandmeows.com/the-best-litter-boxes-for-your-kitty/feed/

The Best Litter Boxes for your Kitty - Cats and Meows

https://www.catsandmeows.com/amp/the-best-litter-boxes-for-your-kitty/

Cat Litter Archives - Cats and Meows

https://www.catsandmeows.com/category/cat-products/cat-litter/feed/

The Best Litter Boxes for your Kitty - Cats and Meows

Google Search results

The older I get the less lame the Danzig kitty litter photo is... : r/punk

https://www.reddit.com/r/punk/comments/16ksktr/the_olc

KITTY LITTER | Dear Stella - Unique Fabric Brand with Modern Style ...

https://www.dearstelladesign.com/category/2-226518/kitty-litter/

Kitty Litter

https://www.facebook.com/kittylitterri/

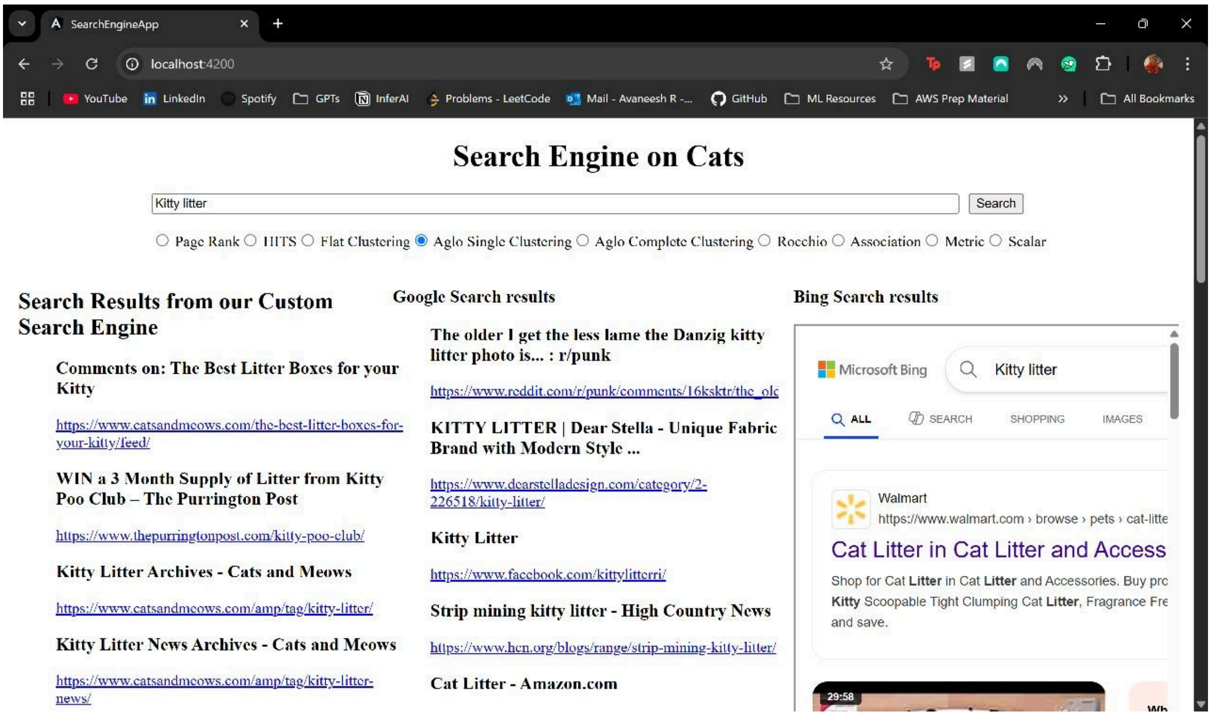
Strip mining kitty litter - High Country News

https://www.hcn.org/blogs/range/strip-mining-kitty-litter/

Cat Litter - Amazon.com

Bing Search results

Chewy
https://www.chewy.com › litter ▾ Free Shipping
Kitty Litter at Chewy | 35% off Plu
Sponsored Kitty Litter at Chewy. Low Prices, Fast, Free Del
Need To Keep Your Kitty Clean & Healthy Delivered Fast.
Site visitors: Over 100K in the past month
24/7 Customer Service · Read Blog · Free 365-Day Retur
Recognized for excellent customer service – Google Trust
Litter Box Problems? › Kitt
Today's Deals › Pre



Search Engine on Cats

Kitty litter | Search

○ Page Rank ○ HITS ○ Flat Clustering ● Aglo Single Clustering ○ Aglo Complete Clustering ○ Rocchio ○ Association ○ Metric ○ Scalar

Search Results from our Custom Search Engine

Comments on: The Best Litter Boxes for your Kitty

https://www.catsandmeows.com/the-best-litter-boxes-for-your-kitty/feed/

WIN a 3 Month Supply of Litter from Kitty Poo Club – The Purrington Post

https://www.thepurringtonpost.com/kitty-poo-club/

Kitty Litter Archives - Cats and Meows

https://www.catsandmeows.com/amp/tag/kitty-litter/

Kitty Litter News Archives - Cats and Meows

https://www.catsandmeows.com/amp/tag/kitty-litter-news/

Google Search results

The older I get the less lame the Danzig kitty litter photo is... : r/punk

https://www.reddit.com/r/punk/comments/16ksktr/the_olc

KITTY LITTER | Dear Stella - Unique Fabric Brand with Modern Style ...

https://www.dearstelladesign.com/category/2-226518/kitty-litter/

Kitty Litter

https://www.facebook.com/kittylitterri/

Strip mining kitty litter - High Country News

https://www.hcn.org/blogs/range/strip-mining-kitty-litter/

Cat Litter - Amazon.com

Bing Search results

Microsoft Bing | Kitty litter

ALL | SEARCH | SHOPPING | IMAGES

Walmart
https://www.walmart.com › browse › pets › cat-litte

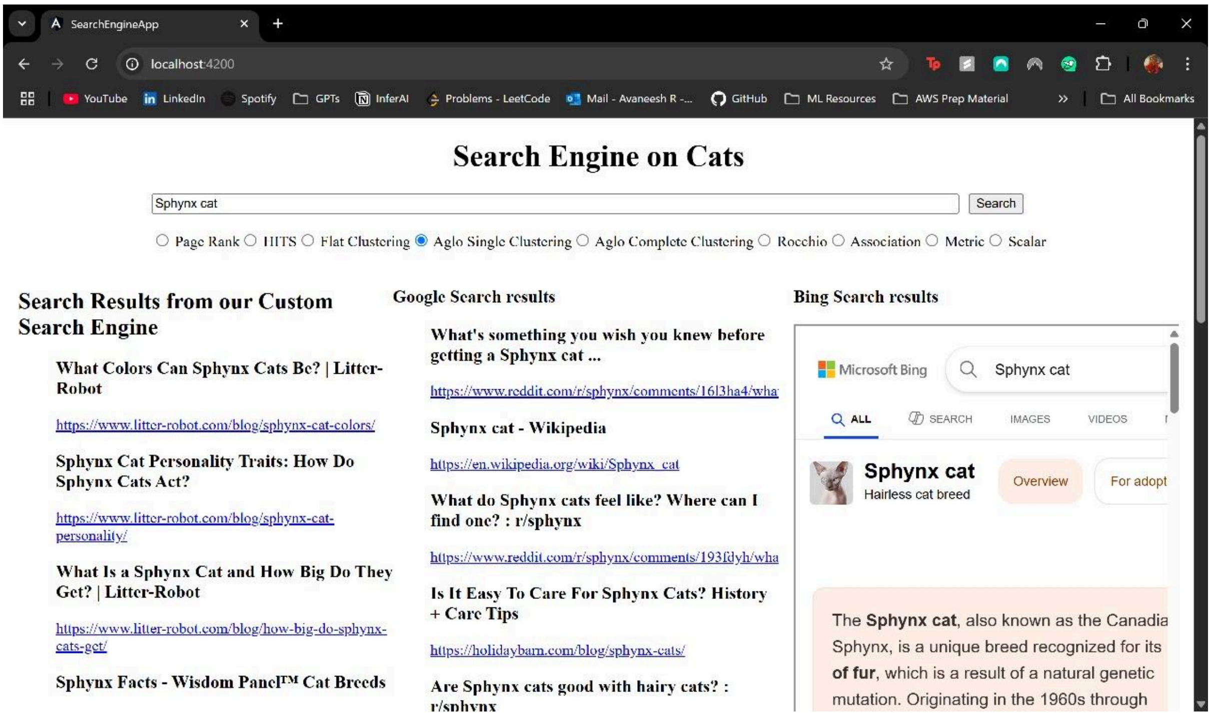
Cat Litter in Cat Litter and Access

Shop for Cat **Litter** in Cat **Litter** and Accessories. Buy pro **Kitty** Scoopable Tight Clumping Cat **Litter**, Fragrance Fre and save.

29:58

Wh

Example 2: Sphynx Cat



Search Engine on Cats

Sphynx cat | Search

○ Page Rank ○ HITS ○ Flat Clustering ● Aglo Single Clustering ○ Aglo Complete Clustering ○ Rocchio ○ Association ○ Metric ○ Scalar

Search Results from our Custom Search Engine

What Colors Can Sphynx Cats Be? | Litter-Robot

https://www.litter-robot.com/blog/sphynx-cat-colors/

Sphynx Cat Personality Traits: How Do Sphynx Cats Act?

https://www.litter-robot.com/blog/sphynx-cat-personality/

What Is a Sphynx Cat and How Big Do They Get? | Litter-Robot

https://www.litter-robot.com/blog/how-big-do-sphynx-cats-get/

Sphynx Facts - Wisdom Panel™ Cat Breeds

Google Search results

What's something you wish you knew before getting a Sphynx cat ...

https://www.reddit.com/r/sphynx/comments/16l3ha4/wha

Sphynx cat - Wikipedia

https://en.wikipedia.org/wiki/Sphynx_cat

What do Sphynx cats feel like? Where can I find one? : r/sphynx

https://www.reddit.com/r/sphynx/comments/193fdyh/wha

Is It Easy To Care For Sphynx Cats? History + Care Tips

https://holidaybarn.com/blog/sphynx-cats/

Are Sphynx cats good with hairy cats? : r/sphynx

Bing Search results

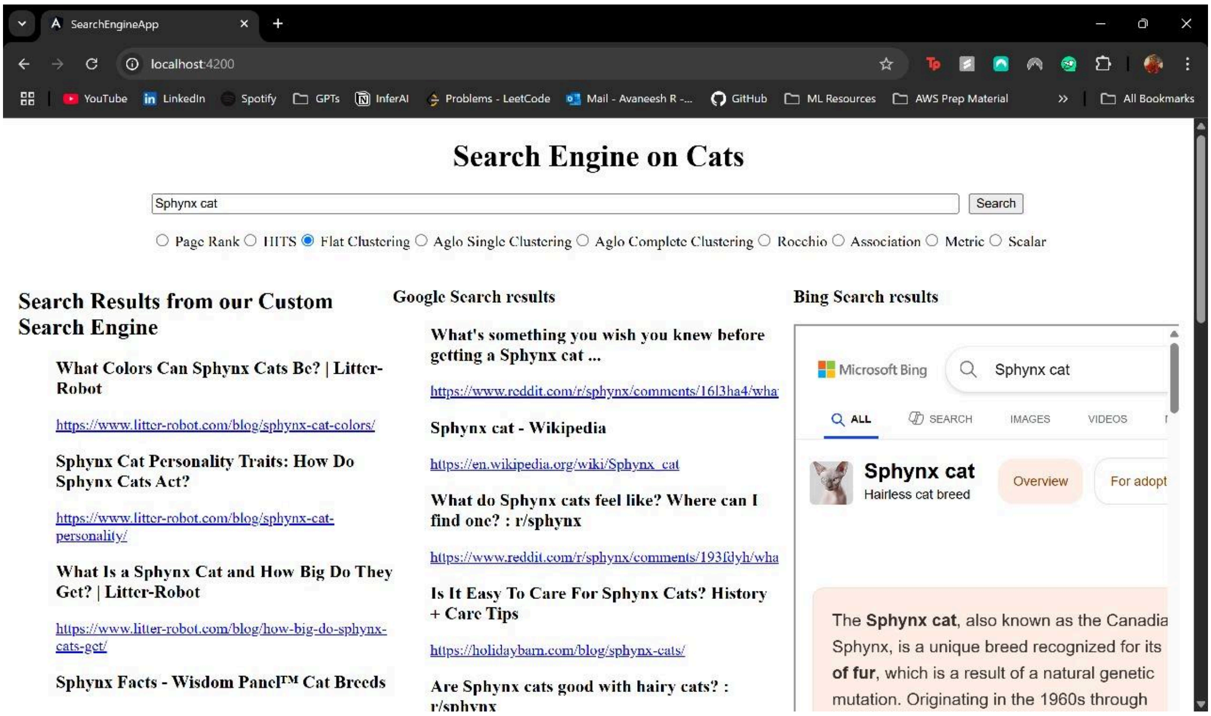
Microsoft Bing | Sphynx cat

ALL | SEARCH | IMAGES | VIDEOS

Sphynx cat
Hairless cat breed

Overview | For adopt

The **Sphynx cat**, also known as the Canadia Sphynx, is a unique breed recognized for its **of fur**, which is a result of a natural genetic mutation. Originating in the 1960s through



Search Engine on Cats

Sphynx cat Search

○ Page Rank ○ HITS ● Flat Clustering ○ Aglo Single Clustering ○ Aglo Complete Clustering ○ Rocchio ○ Association ○ Metric ○ Scalar

Search Results from our Custom Search Engine

What Colors Can Sphynx Cats Be? | Litter-Robot

https://www.litter-robot.com/blog/sphynx-cat-colors/

Sphynx Cat Personality Traits: How Do Sphynx Cats Act?

https://www.litter-robot.com/blog/sphynx-cat-personality/

What Is a Sphynx Cat and How Big Do They Get? | Litter-Robot

https://www.litter-robot.com/blog/how-big-do-sphynx-cats-get/

Sphynx Facts - Wisdom Panel™ Cat Breeds

Google Search results

What's something you wish you knew before getting a Sphynx cat ...

https://www.reddit.com/r/sphynx/comments/16l3ha4/wha

Sphynx cat - Wikipedia

https://en.wikipedia.org/wiki/Sphynx_cat

What do Sphynx cats feel like? Where can I find one? : r/sphynx

https://www.reddit.com/r/sphynx/comments/193fdyh/wha

Is It Easy To Care For Sphynx Cats? History + Care Tips

https://holidaybarn.com/blog/sphynx-cats/

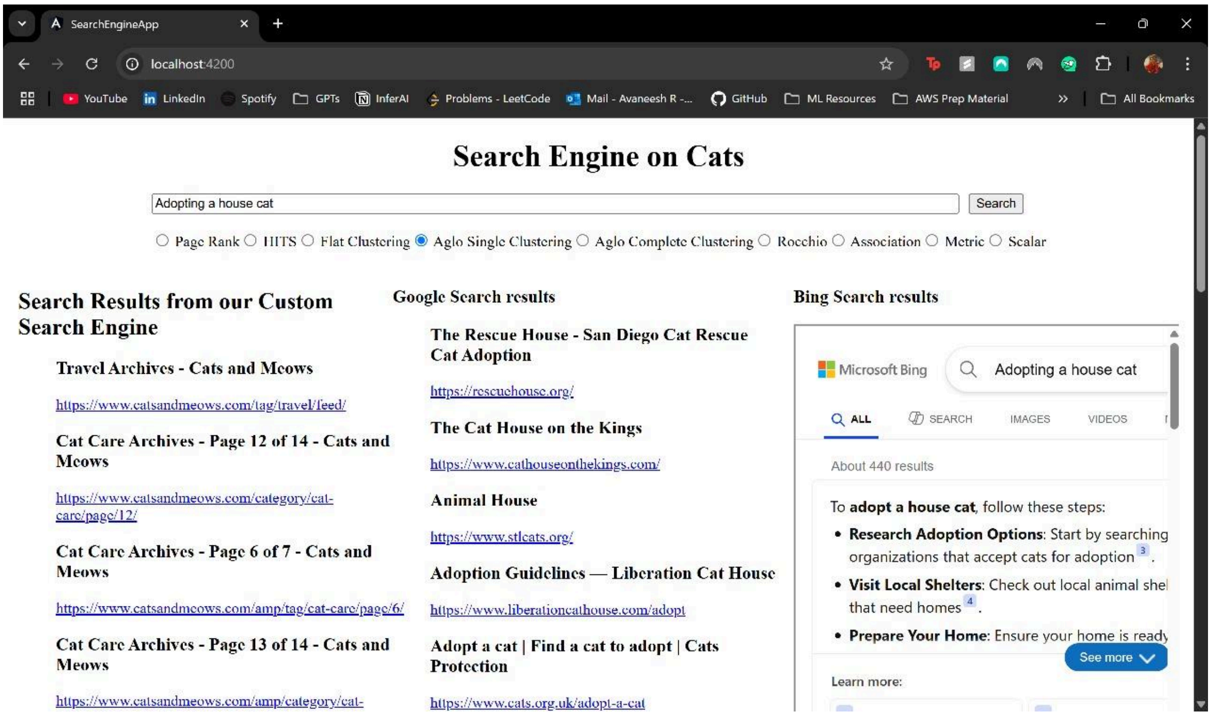
Are Sphynx cats good with hairy cats? : r/sphynx

Bing Search results

The **Sphynx cat**, also known as the Canadia Sphynx, is a unique breed recognized for its **of fur**, which is a result of a natural genetic mutation. Originating in the 1960s through

---

Example 3: Adopting a house cat



Search Engine on Cats

Adopting a house cat Search

○ Page Rank ○ HITS ○ Flat Clustering ● Aglo Single Clustering ○ Aglo Complete Clustering ○ Rocchio ○ Association ○ Metric ○ Scalar

Search Results from our Custom Search Engine

Travel Archives - Cats and Meows

https://www.catsandmeows.com/tag/travel/feed/

Cat Care Archives - Page 12 of 14 - Cats and Meows

https://www.catsandmeows.com/category/cat-care/page/12/

Cat Care Archives - Page 6 of 7 - Cats and Meows

https://www.catsandmeows.com/amp/tag/cat-care/page/6/

Cat Care Archives - Page 13 of 14 - Cats and Meows

https://www.catsandmeows.com/amp/category/cat-

Google Search results

The Rescue House - San Diego Cat Rescue Cat Adoption

https://rescuehouse.org/

The Cat House on the Kings

https://www.cathouseonthekings.com/

Animal House

https://www.stlcats.org/

Adoption Guidelines — Liberation Cat House

https://www.liberationcathouse.com/adopt

Adopt a cat | Find a cat to adopt | Cats Protection

https://www.cats.org.uk/adopt-a-cat

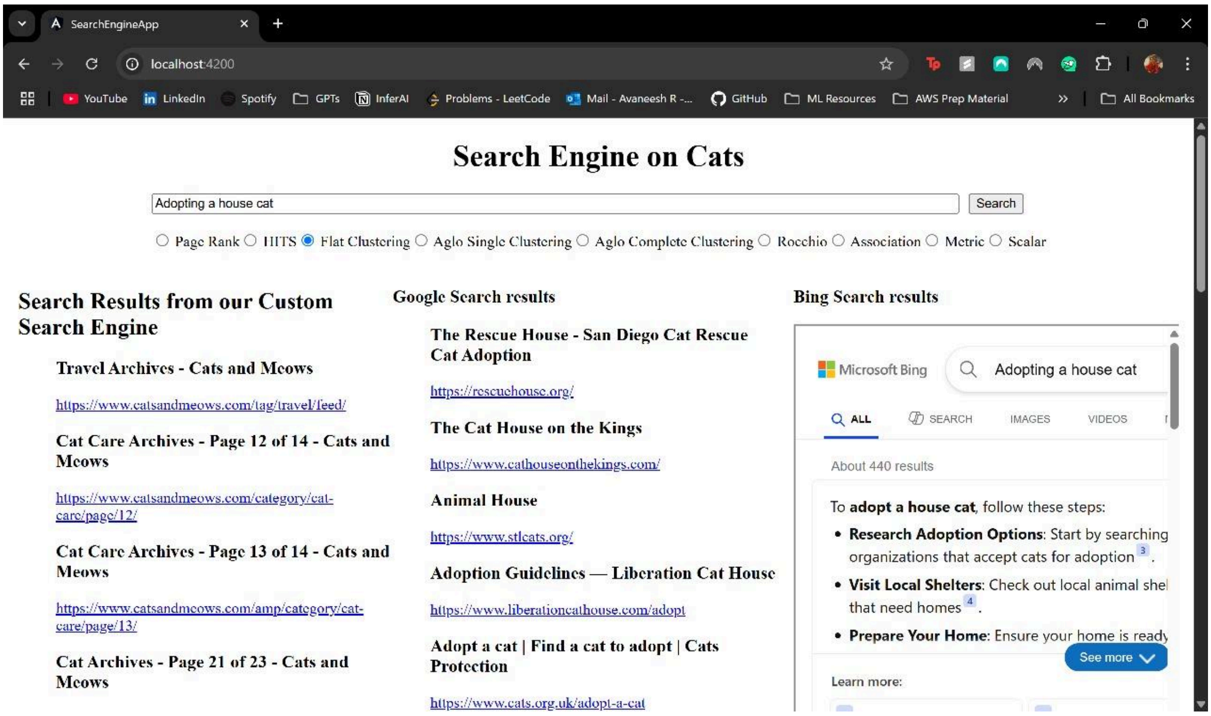
Bing Search results

About 440 results

To **adopt a house cat**, follow these steps:

- **Research Adoption Options**: Start by searching organizations that accept cats for adoption [3].
- **Visit Local Shelters**: Check out local animal shel that need homes [4].
- **Prepare Your Home**: Ensure your home is ready

Learn more:



Search Engine on Cats

Adopting a house cat [Search]

○ Page Rank ○ HITS ● Flat Clustering ○ Aglo Single Clustering ○ Aglo Complete Clustering ○ Rocchio ○ Association ○ Metric ○ Scalar

Search Results from our Custom Search Engine

Travel Archives - Cats and Meows

https://www.catsandmeows.com/tag/travel/feed/

Cat Care Archives - Page 12 of 14 - Cats and Meows

https://www.catsandmeows.com/category/cat-care/page/12/

Cat Care Archives - Page 13 of 14 - Cats and Meows

https://www.catsandmeows.com/amp/category/cat-care/page/13/

Cat Archives - Page 21 of 23 - Cats and Meows

Google Search results

The Rescue House - San Diego Cat Rescue Cat Adoption

https://rescuehouse.org/

The Cat House on the Kings

https://www.cathouseonthekings.com/

Animal House

https://www.stlcats.org/

Adoption Guidelines — Liberation Cat House

https://www.liberationcathouse.com/adopt

Adopt a cat | Find a cat to adopt | Cats Protection

https://www.cats.org.uk/adopt-a-cat

Bing Search results

Adopting a house cat

About 440 results

To **adopt a house cat**, follow these steps:

- **Research Adoption Options**: Start by searching organizations that accept cats for adoption [3] .
- **Visit Local Shelters**: Check out local animal shel that need homes [4] .
- **Prepare Your Home**: Ensure your home is ready

Learn more:

6. Query Expansion and Relevance Feedback – Sharon T Alexander (SXA230048)

Rocchio Algorithm:
The 20 queries were selected based on the number of query parameters, spelling mistakes, breed of cats and queries containing words that are animals (outside of cat domain).

1) 4 queries with one keyword for the data that was crawled

    a) cat

    b) kitty

    c) bobcat

    d) tiger

2) 4 queries with two keywords for the data that was crawled

    a) black cat

    b) cat breed

    c) persian cat[1]

    d) cat litter

3) 4 queries with spelling mistakes in the query words for the data that was crawled

    a) sphinx

    b) Manaki Neko

    c) tuxedo kat

    d) germny cat

4) 4 queries with other animal names

    a) dog

    b) rat

    c) horse

    d) bird

[1] Sharon T Alexander (SXA230048)

5) 4 queries with the words that are different species of cat

a) manx cat

b) munchkin cat

c) kinkalow cat

d) singapura cat

Examples of relevant web pages:

Query	Relevant Web Pages
cat	https://www.litter-robot.com/eu/de/blog/cat-tips/ https://www.catsandmeows.com/cat-care-tips-for-bathing-your-cat/ https://www.cathealth.com/cat-care/how-to/2197-cat-costumes-purr-ific-or-cat-astrophe
kitty	https://www.kittyofangels.org/camp-kitty.html https://www.krazyforkats.org/kitty-safety https://www.catsandmeows.com/amp/tag/kitty-litter/
black cat	https://www.litter-robot.com/blog/black-cat-breeds/ https://americanfolklore.net/folklore/2014/03/why_is_a_black_cat_bad_luck.html https://www.catsandmeows.com/tag/black-cats/
tuxedo kat	https://www.litter-robot.com/blog/7-tuxedo-cat-facts/

	https://www.catster.com/lifestyle/tuxedo-cat-facts/[2]
rat	https://www.animallama.com/rats/page/2 https://www.britannica.com/animal/sand-rat https://www.animallama.com/rat-care-guide/
munchkin cat	https://www.litter-robot.com/blog/munchkin-cat-personality/ https://munchkinkittenstore.com/category/training/

[3]

Examples of irrelevant web pages:

Query	Irrelevant Web Pages
Kitty	https://www.kittygenius.com/category/fashion/feed/ https://www.kittygenius.com/category/events/feed/
black cat	https://kristenlevine.com/black-dog-day/
tuxedo kat	https://bestfriends.org/stories/features/mayim-bialik-fox-best-friends-team-save-cat https://www.krazyforkats.org/the-kat-house https://www.missinganimalresponse.com/lost-dog-behavior

[2]

[3] Sharon T Alexander (SXA230048)

Query	Expanded Query
cat	cat domestic feline
kitty	kitty kitten cat [4]
bobcat	bobcat lynx rufus[5]
tiger	tiger big cat
black cat	black cat fur coat superstition halloween
cat breed	cat breed pedigree cfa persian siamese
persian cat	persian cat longhair breed round face
cat litter	cat litter clumping odor control clay
sphinx	sphinx cat hairless
manaki neko	manaki neko beckoning cat japanese
tuxedo kat	tuxedo kat black white coat pattern

[4] Sharon T Alexander (SXA230048)

[5] Sharon T Alexander (SXA230048)

germny cat	germny cat german domestic feline breeds
dog	dog canis familiaris
rat	rat rattus rodent
horse	horse equus ferus [6]
bird	bird avian feathers[7]
manx cat	manx cat tailless isle man breed
munchkin cat	munchkin cat short legs dwarf mutation
kinkalow cat	kinkalow cat dwarf breed munchkin american
singapura cat	singapura cat small breed ticked coat

The approach was used for the 20 cat-related queries shown in the table above. The results obtained showed varying degrees of success. Following were some of the findings:

1. Some queries expanded with accurate taxonomic terms – like "bobcat lynx rufus" and "tiger big cat" – which provided precise scientific context but sometimes missed popular content that casual searchers might seek.
2. Some expanded queries introduced potential ambiguity – "sphinx cat hairless" could retrieve results about the Egyptian Sphinx alongside the hairless cat

[6]

[7]

breed. Similarly, "manaki neko beckoning cat japanese" corrected the misspelling but introduced cultural terminology.

3. Some expanded queries successfully captured essential characteristics – like "persian cat longhair breed round face" which highlighted key physical traits. These expansions produced relevant and focused results.
4. Breed-specific queries like "munchkin cat short legs dwarf mutation" added medical or genetic terminology that might retrieve technical content rather than general pet owner information.
5. Non-cat animal queries ("dog canis familiaris", "rat rattus rodent") expanded with scientific classifications, creating a different emphasis than the cat-focused queries which tended toward physical traits and breed characteristics.

These observations indicate that query expansion effectiveness varies based on the original term's specificity, potential ambiguities, and the balance between technical and common terminology.

Pseudo Relevance Feedback:

The 50 cat-related queries were structured across seven thematic categories (breeds, traits, health, behavior, etc.) and expanded using pseudo-relevance feedback (PRF) principles. Observations from the results include:

Query Structure

Breed-Specific Focus: Queries like "Persian cat lifespan" retrieved breed specific information more than the lifespan ones (e.g., Persian Cats: Facts, Personality, and Breed Guide, Persian Cat Lifespan: How Long Do Persian Cats Live?).

Trait Expansion: Terms like "hypoallergenic" drew from breed directories and genetic markers (e.g., Are Bengal Cats Hypoallergenic? | Litter-Robot).

Wild/Hybrid Terms: Queries such as "Savannah cat generations" used taxonomy from classification datasets and legal ownership debates in retrieval results (e.g., F1 Savannah Cats: Are They Legal & Savannah Cat Price).

Key Observations

Precision vs. Ambiguity: Unambiguous terms (e.g., "Maine Coon size") achieved higher document relevance. Ambiguous expansions (e.g., "sphinx cat") suffered from query drift, mixing Egyptian monuments, Celebrities e.t.c. with cat breeds.

Genetic Terminology Impact: Queries expanded with medical terms (e.g., "Munchkin spinal health") skewed toward blogs and articles from shops and insurance providers, reducing usability from an academic standpoint.

Challenges

Spelling Errors: Misspelled terms (e.g., "Manaki Neko") required manual correction to align with cultural references, as automated PRF amplified noise.

Hybrid Queries: Wildcat terms (e.g., "Chausie jungle ancestry") triggered inconsistent results due to sparse corpus data.

These findings indicate that the success hinges on query clarity and corpus diversity

[8]

Examples of queries for associative clustering:

1. Query: black cat
    a. Local document set: https://github.com/sharona1ex/IR-Project/blob/main/black%20cat_local_document_association.json
    b. Local Vocabulary, Local Stem and their vocabulary is in the following JSON file. The following is the structure:

```
{"local_vocab": <local vocabulary>,
      "local_stem": <local stems>,
      'stem_vocabulary':<vocabulary for stem>,
      "correlation values": correlation
values,
      "expanded_query": <expanded query>
}
```

    https://github.com/sharona1ex/IR-Project/blob/main/black%20cat_association.json
    c. Expanded query: black cat halloween name home stories
2. Query:Persian cat lifespan
    a. Local document set: https://github.com/sharona1ex/IR-Project/blob/main/Persian%20cat%20lifespan_local_document_association.json
    b. Local Vocabulary, Local Stem and their vocabulary is in the following JSON file. https://github.com/sharona1ex/IR-Project/blob/main/Persian%20cat%20lifespan_association.json
    c. Expanded query: Persian cat lifespan eye need long live year pet
3. Query: Savannah cat generations
    a. Local document set: https://github.com/sharona1ex/IR-Project/blob/main/Savannah%20cat%20generations_local_document_association.json

[8] Sharon T Alexander (SXA230048)

   b. Local Vocabulary, Local Stem and their vocabulary is in the following JSON file. https://github.com/sharona1ex/IR-Project/blob/main/Savannah%20cat%20generations_association.json
   c. Expanded query:Savannah cat generations f breed allow domesticated breeder allows

Examples of queries for metric clustering:

1. Query:black cat
   a. Local document set: https://github.com/sharona1ex/IR-Project/blob/main/black%20cat_local_document_metric.json

   b. Local Vocabulary, Local Stem and their vocabulary is in the following JSON file. https://github.com/sharona1ex/IR-Project/blob/main/black%20cat_metric.json

   c. Expanded query:black cat national amazing theft
2. Query:Persian cat lifespan
   a. Local document set: https://github.com/sharona1ex/IR-Project/blob/main/Persian%20cat%20lifespan_local_document_metric.json

   b. Local Vocabulary, Local Stem and their vocabulary is in the following JSON file. https://github.com/sharona1ex/IR-Project/blob/main/Persian%20cat%20lifespan_metric.json

   c. Expanded query:persian cat lifespan anna partially baker
3. Query: Savannah cat generations
   a. Local document set: https://github.com/sharona1ex/IR-Project/blob/main/Savannah%20cat%20generations_local_document_metric.json

   b. Local Vocabulary, Local Stem and their vocabulary is in the following JSON file. https://github.com/sharona1ex/IR-Project/blob/main/Savannah%20cat%20generations_metric.json

   c. Expanded query:savannah cat generations next adulthood captivity

Examples of queries for scalar clustering:

4. Query:black cat
   a. Local document set:

        https://github.com/sharona1ex/IR-Project/blob/main/black%20cat_local_document_scalar.json

    b. Local Vocabulary, Local Stem and their vocabulary is in the following JSON file.[9]
       [10]https://github.com/sharona1ex/IR-Project/blob/main/black%20cat_scalar.json

    c. Expanded query:black cat halloween anime home adopt

5. Query:Persian cat lifespan
    a. Local document set:
       https://github.com/sharona1ex/IR-Project/blob/main/Persian%20cat%20lifespan_local_document_scalar.json

    b. Local Vocabulary, Local Stem and their vocabulary is in the following JSON file. (The file is larger so we have zipped it)
       https://github.com/sharona1ex/IR-Project/blob/main/Persian%20cat%20lifespan_scalar.zip

    c. Expanded query:Persian cat lifespan medium size breed eye ragdoll home

6. Query: Savannah cat generations
    a. Local document set:
       https://github.com/sharona1ex/IR-Project/blob/main/Savannah%20cat%20generations_local_document_scalar.json

    b. Local Vocabulary, Local Stem and their vocabulary is in the following JSON file.
       https://github.com/sharona1ex/IR-Project/blob/main/Savannah%20cat%20generations_scalar.zip

    c. Expanded query:Savannah cat generations snare huge domestic serval removed consider

Search engine results

Results from the search engine are available in the local document set JSON file provided in the above file links.

Collaboration with UI and relevance model:

1. From the API (URL) call we receive the original query and the cluster method type to be used.

[9]

[10]

2. Based on these parameters, the relevant model provides top 20 documents to the query expansion python code. These results are then used as local documents to form local vocabulary , stem sets.
3. Based on the query parameter the cluster method out of association, metric and scalar is called to return expanded query to relevance model.
4. The relevance model then hits solar in order to receive results for expanded query.
5. These results are then passed to the front end and displayed.

Query selection for demo:

For the demo, the query "black cat" was used since it has two words and can be enhanced with other words. It is tested for association, metric and scalar cluster method.[11]

---

[11] Sharon T Alexander (SXA230048)

7. Discussion - Avaneesh Ramaseshan Baskaraswaminathan, Venkata Subbaiah Pavan Karthik Navuluru, Dhanyan Muralidharan, Varsha Viswanathan and Sharon T Alexander

(axb220230@utdallas.edu, vxn220052@utdallas.edu, dtm220000@utdallas.edu, vxv230013@utdallas.edu, sxa230048@utdallas.edu)

Throughout the development of the "Search Engine for Cats," our team faced multiple challenges and learning opportunities, especially in adapting to new technologies and optimizing processes. The project was initially set up using .NET, but we encountered limitations with indexing performance, which led us to shift to a more robust solution using Apache Nutch and Solr. These technologies provided better support for crawling large-scale datasets and faster indexing, allowing us to effectively handle our growing collection of cat-related web pages.

The key decisions and challenges during the process included:

1. **Crawling and Data Collection**: Varsha Viswanathan took responsibility for crawling 100,000 web pages related to cats. Initially, we used Python's BeautifulSoup library; however, we faced issues with relevance due to inaccurate anchor points in many pages. This led us to switch to Apache Nutch, which provided better support for large-scale crawling and link analysis. The initial crawling setup was slow, but optimizations were made to speed it up by tweaking Nutch's fetcher and thread configurations.

2. **Indexing and Relevance Models**: Pavan Karthik Navuluru was responsible for indexing the crawled data and developing relevance models. After testing various models, we found that using both vector space models and combining them with link analysis models (PageRank and HITS) significantly improved the relevance of the search results. Apache Solr was used as the indexing backend, and its integration with Nutch allowed seamless crawling and indexing, boosting overall system performance.

3. **Query Expansion and Clustering**: Sharon T Alexander developed a query expansion module based on the Rocchio algorithm. Additionally, Dhanyan Muralidharan worked on clustering the crawled data using flat and agglomerative clustering techniques. Clustering helped to group similar documents and enhanced the accuracy of query results. The collaboration between clustering and relevance models allowed for a more personalized search experience, especially for ambiguous queries.

4. **Web Interface**: Avaneesh Ramaseshan Baskaraswaminathan developed the graphical user interface (GUI) of the search engine, which displayed search results, relevance models, clustering results, and expanded queries. This module made the search engine

interactive, allowing users to compare results from various models and perform real-time querying.

Each team member collaborated closely, ensuring the integration of various components into a cohesive search engine. Despite challenges related to optimizing Nutch's crawl settings and handling large-scale data, the project resulted in a fully functional search engine that effectively indexes and presents relevant content related to cats.

---

8. Conclusion - Avaneesh Ramaseshan Baskaraswaminathan, Venkata Subbaiah Pavan Karthik Navuluru, Dhanyan Muralidharan, Varsha Viswanathan and Sharon T Alexander

(axb220230@utdallas.edu, vxn220052@utdallas.edu, dtm220000@utdallas.edu, vxv230013@utdallas.edu, sxa230048@utdallas.edu)

The "Search Engine for Cats" project successfully implemented a search engine capable of crawling, indexing, and providing relevant search results based on various relevance models. We utilized Apache Nutch for crawling, Solr for indexing, and implemented clustering techniques to enhance search result accuracy. Despite the challenges encountered during the project, such as the limitations of BeautifulSoup and issues with slow indexing, the switch to Apache Nutch and Solr improved performance significantly. Our final system allows users to query cat-related content effectively, with results displayed based on multiple relevance models, clustering, and query expansion.

Our approach to combining multiple models, including vector space, PageRank, HITS, and clustering, is what differentiates our search engine from simpler models. Through iterative development and collaboration, our team was able to refine the system, optimize performance, and produce a scalable solution for cat-related web searches. The project has not only contributed to learning how to build a real-world search engine but also provided a deeper understanding of how web crawling, indexing, and relevance modeling work together to serve accurate results.

Future Work:

- Implement additional query expansion techniques.
- Improve the clustering algorithm by experimenting with different methods.
- Scale the system to handle even larger datasets.