UBC Capstone Project 2023

Masters of Data Science

Computational Linguistics

About the UBC Capstone Project

Objective: students apply their knowledge to help partner organization solve a business problem.

Duration: 2 months (May - June)

Application process:

- November – online proposal.
- January – meeting with students.
- February – decisions announced.

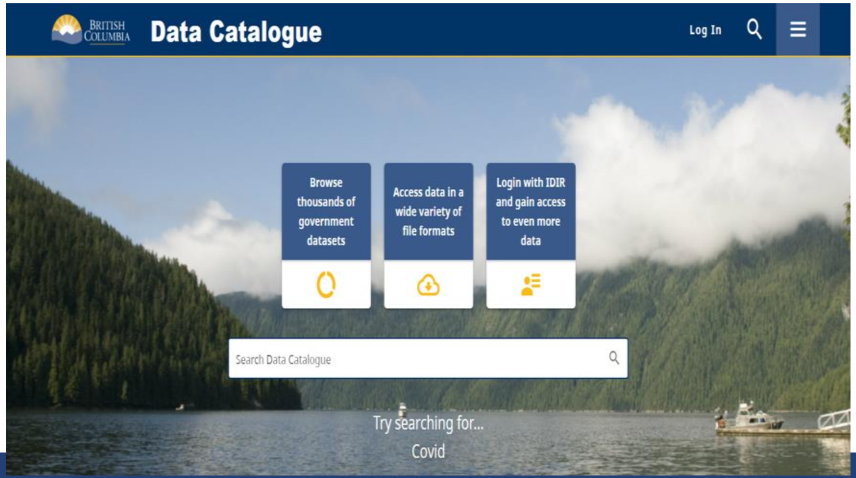
Business Problem

Problem Statements

1. The **search function is limited** in its ability to display all the relevant datasets.
2. The **system is limited in using analytics to portray a big picture** understanding of the datasets in the catalogue.

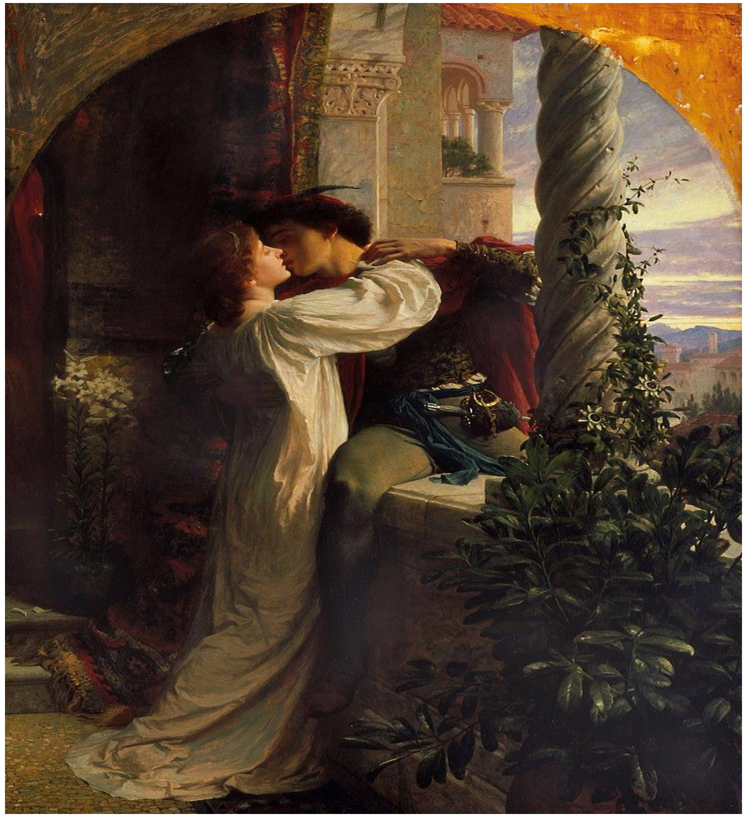
Opportunities

1. Explore options to **enhance the search function** in the Data Catalogue
2. Seek an independent perspective on **how to visualize the data** and **improve the experience for users**


British Columbia
Data Catalogue
Log In
Browse thousands of government datasets
Access data in a wide variety of file formats
Login with IDIR and gain access to even more data
Search Data Catalogue
Try searching for...
Covid

Causes of the Problem

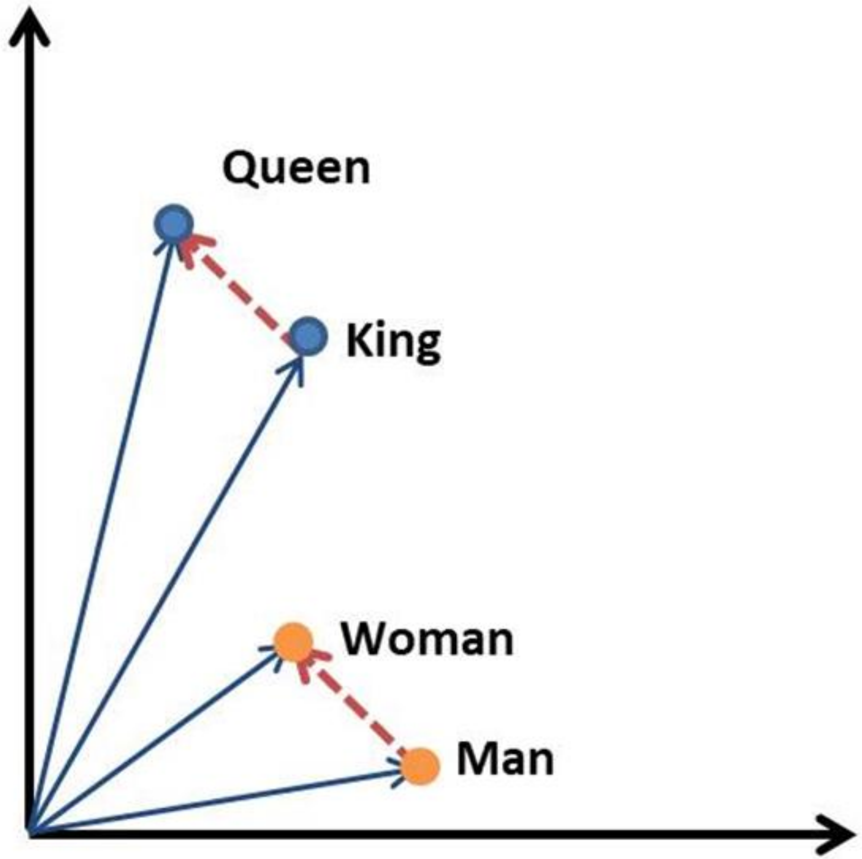
- Search engines (ad hoc IR systems)
- Vocabulary: Mismatch Problem
  - Example: A tragic love story
- Search engine as a binary test:
  - False positives
  - False negatives

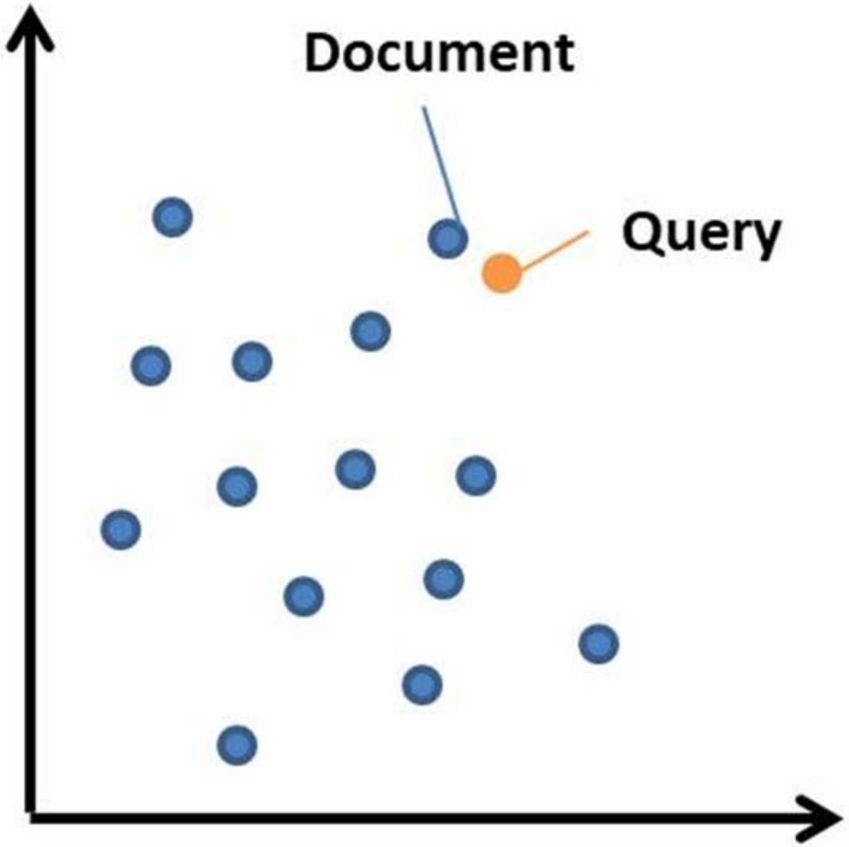


Solution

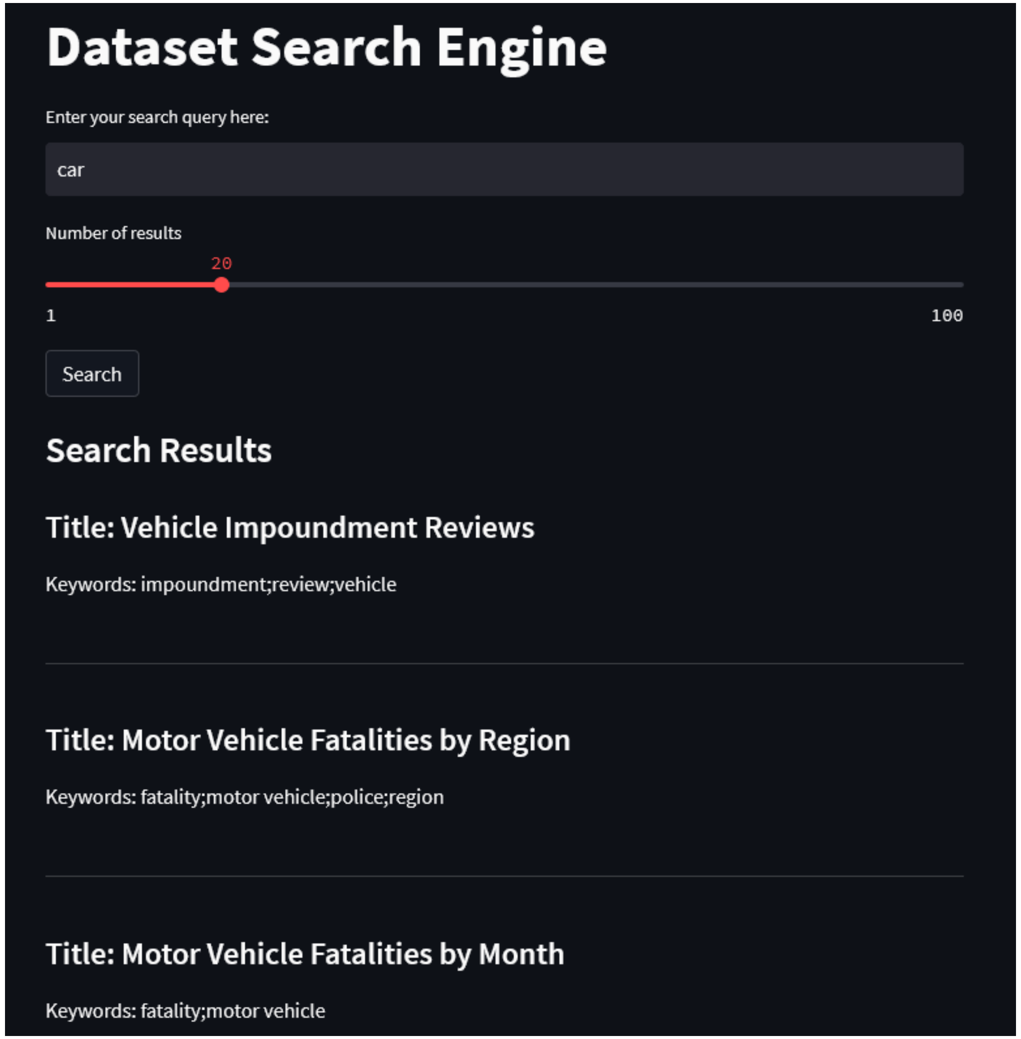
- Semantic Search (embedding learning)
- Example
  - A bottle of tezguino is present at the table.
  - Tezguino is highly popular among individuals.
  - Consumption of tezguino induces intoxication.
  - Corn is used in the production of tezguino.

Solution





Demo – Increases Accuracy



Dataset Search Engine

Enter your search query here:

car

Number of results

20

1 100

Search

Search Results

Title: Vehicle Impoundment Reviews

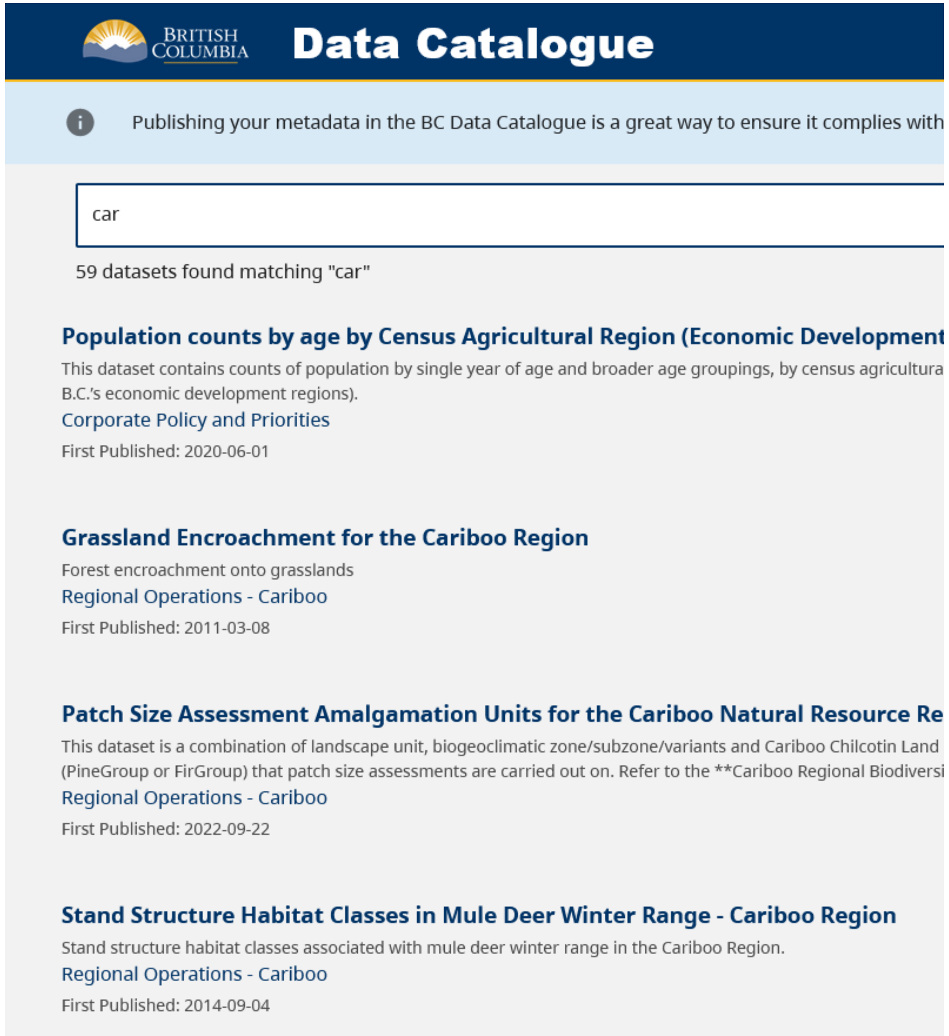
Keywords: impoundment;review;vehicle

Title: Motor Vehicle Fatalities by Region

Keywords: fatality;motor vehicle;police;region

Title: Motor Vehicle Fatalities by Month

Keywords: fatality;motor vehicle



Data Catalogue

Publishing your metadata in the BC Data Catalogue is a great way to ensure it complies with

car

59 datasets found matching "car"

Population counts by age by Census Agricultural Region (Economic Development

This dataset contains counts of population by single year of age and broader age groupings, by census agricultura
B.C.'s economic development regions).

Corporate Policy and Priorities

First Published: 2020-06-01

Grassland Encroachment for the Cariboo Region

Forest encroachment onto grasslands

Regional Operations - Cariboo

First Published: 2011-03-08

Patch Size Assessment Amalgamation Units for the Cariboo Natural Resource Re

This dataset is a combination of landscape unit, biogeoclimatic zone/subzone/variants and Cariboo Chilcotin Land
(PineGroup or FirGroup) that patch size assessments are carried out on. Refer to the **Cariboo Regional Biodiversi

Regional Operations - Cariboo

First Published: 2022-09-22

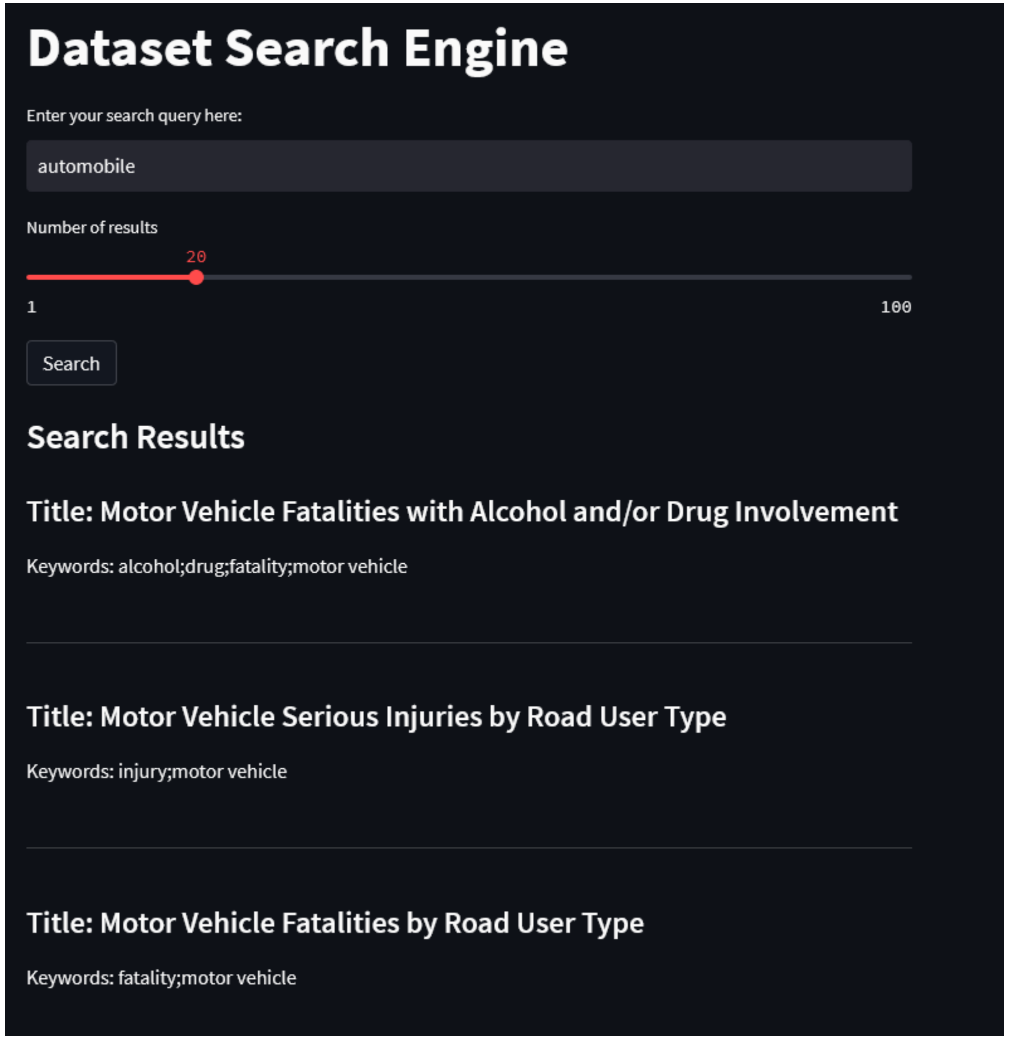
Stand Structure Habitat Classes in Mule Deer Winter Range - Cariboo Region

Stand structure habitat classes associated with mule deer winter range in the Cariboo Region.

Regional Operations - Cariboo

First Published: 2014-09-04

Demo – Understands Synonyms



Dataset Search Engine

Enter your search query here:

automobile

Number of results

Search Results

Title: Motor Vehicle Fatalities with Alcohol and/or Drug Involvement

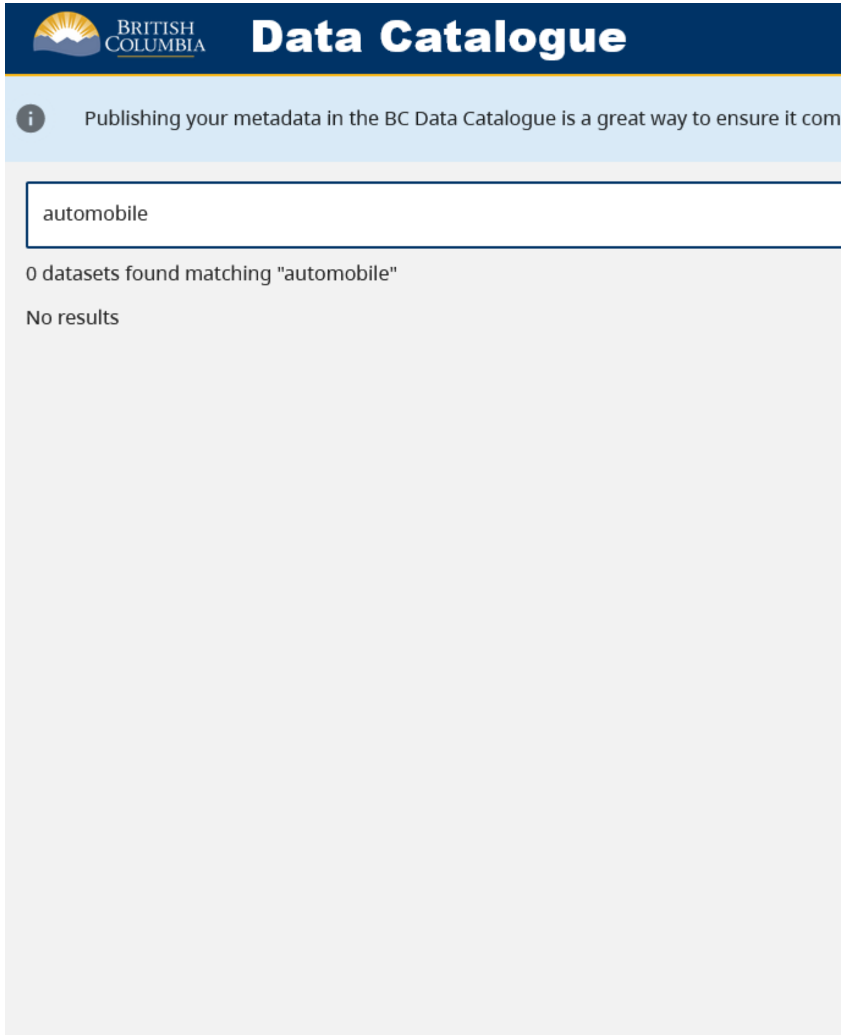
Keywords: alcohol;drug;fatality;motor vehicle

Title: Motor Vehicle Serious Injuries by Road User Type

Keywords: injury;motor vehicle

Title: Motor Vehicle Fatalities by Road User Type

Keywords: fatality;motor vehicle



Data Catalogue

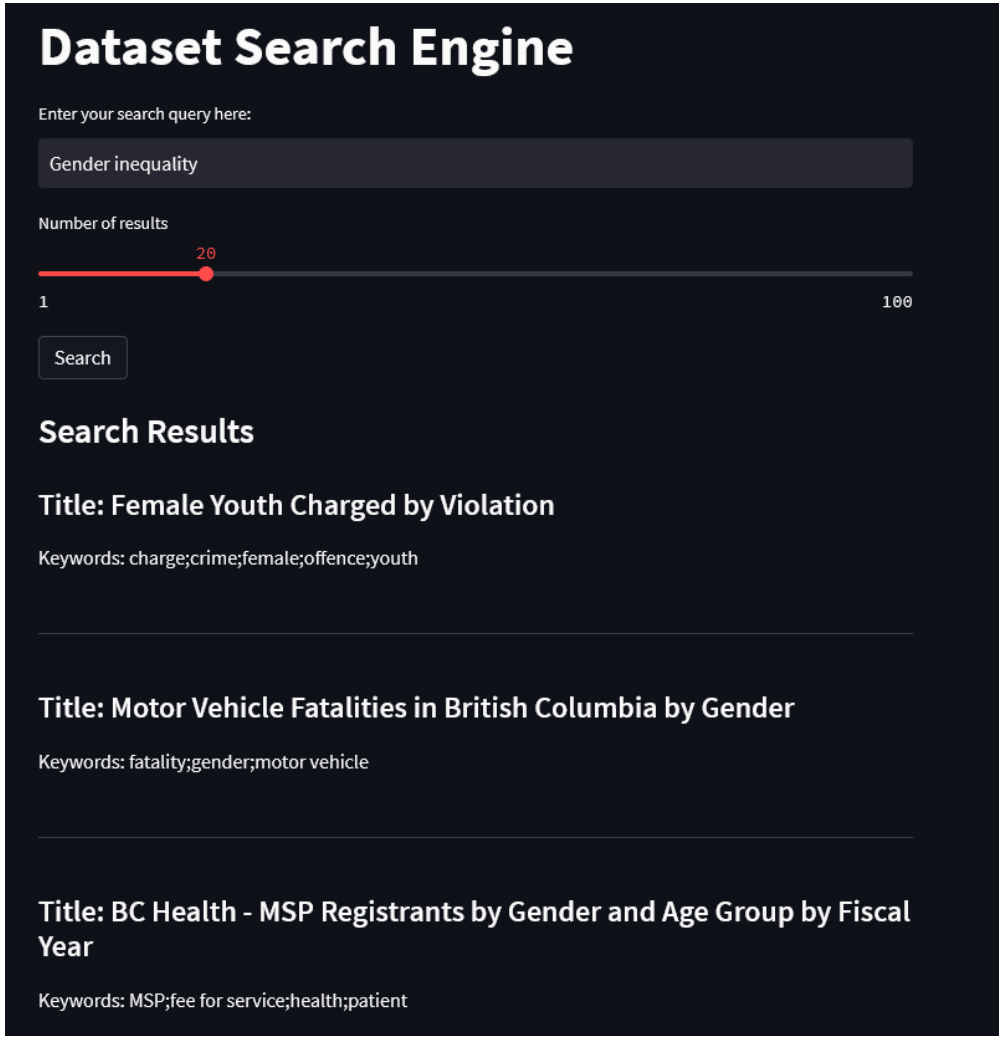
Publishing your metadata in the BC Data Catalogue is a great way to ensure it com

automobile

0 datasets found matching "automobile"

No results

Demo – Understands Phrases



Dataset Search Engine

Enter your search query here:

Gender inequality

Number of results

Search Results

Title: Female Youth Charged by Violation

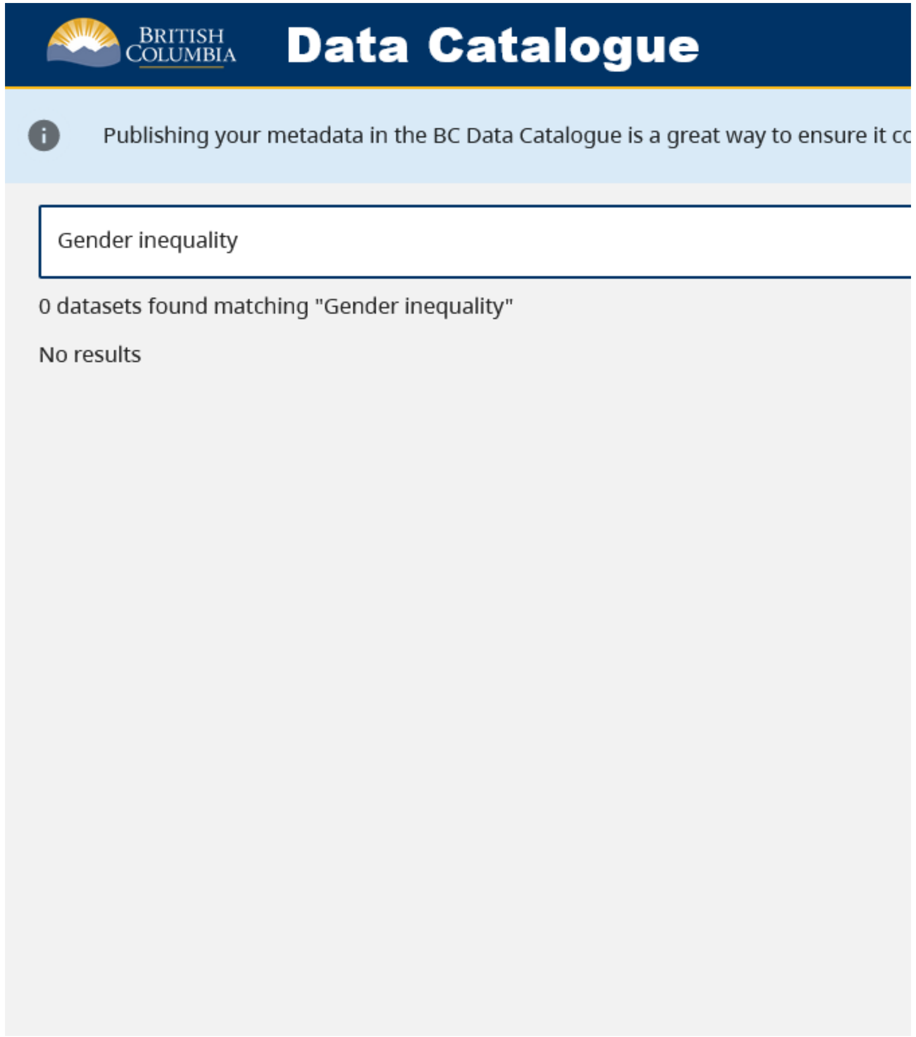
Keywords: charge;crime;female;offence;youth

Title: Motor Vehicle Fatalities in British Columbia by Gender

Keywords: fatality;gender;motor vehicle

Title: BC Health - MSP Registrants by Gender and Age Group by Fiscal Year

Keywords: MSP;fee for service;health;patient



Data Catalogue

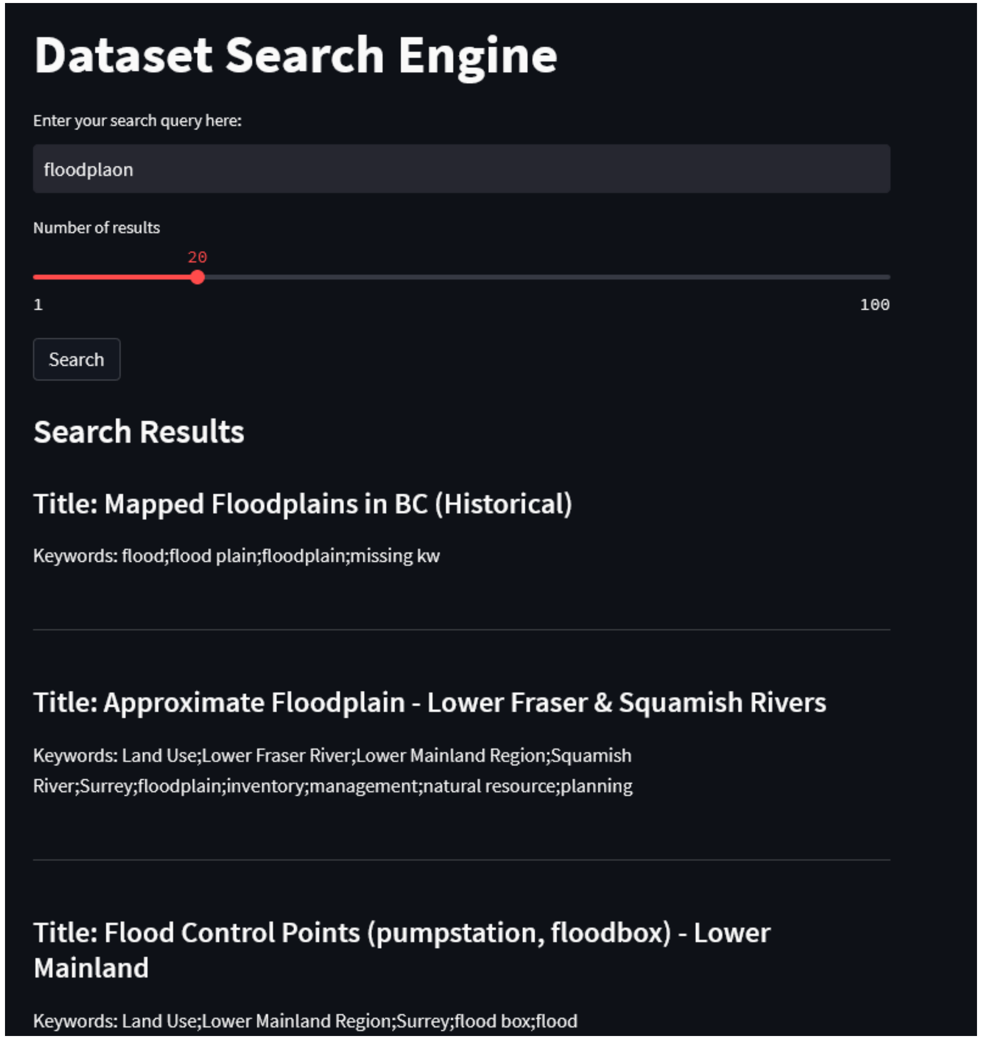
Publishing your metadata in the BC Data Catalogue is a great way to ensure it cc

Gender inequality

0 datasets found matching "Gender inequality"

No results

Demo – Handles Spelling Errors



Dataset Search Engine

Enter your search query here:

floodplaon

Number of results

20

Search Results

Title: Mapped Floodplains in BC (Historical)

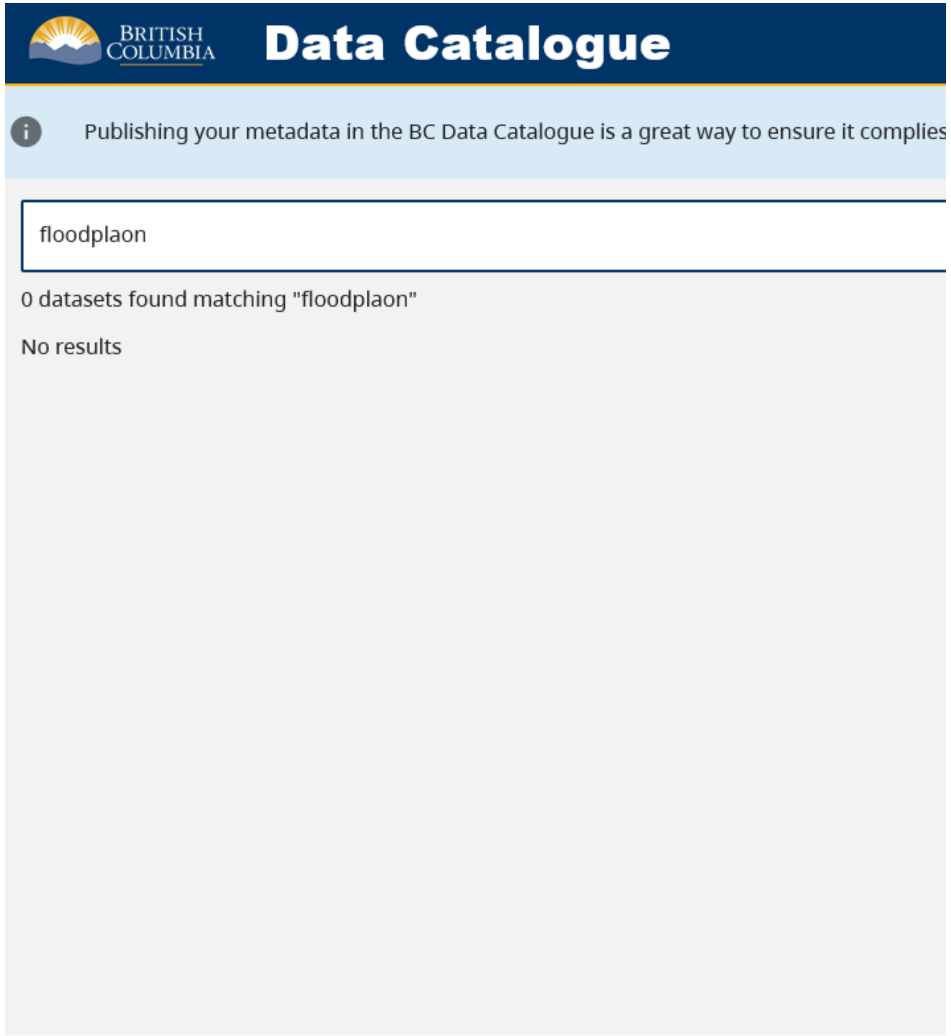
Keywords: flood;flood plain;floodplain;missing kw

Title: Approximate Floodplain - Lower Fraser & Squamish Rivers

Keywords: Land Use;Lower Fraser River;Lower Mainland Region;Squamish River;Surrey;floodplain;inventory;management;natural resource;planning

Title: Flood Control Points (pumpstation, floodbox) - Lower Mainland

Keywords: Land Use;Lower Mainland Region;Surrey;flood box;flood



Data Catalogue

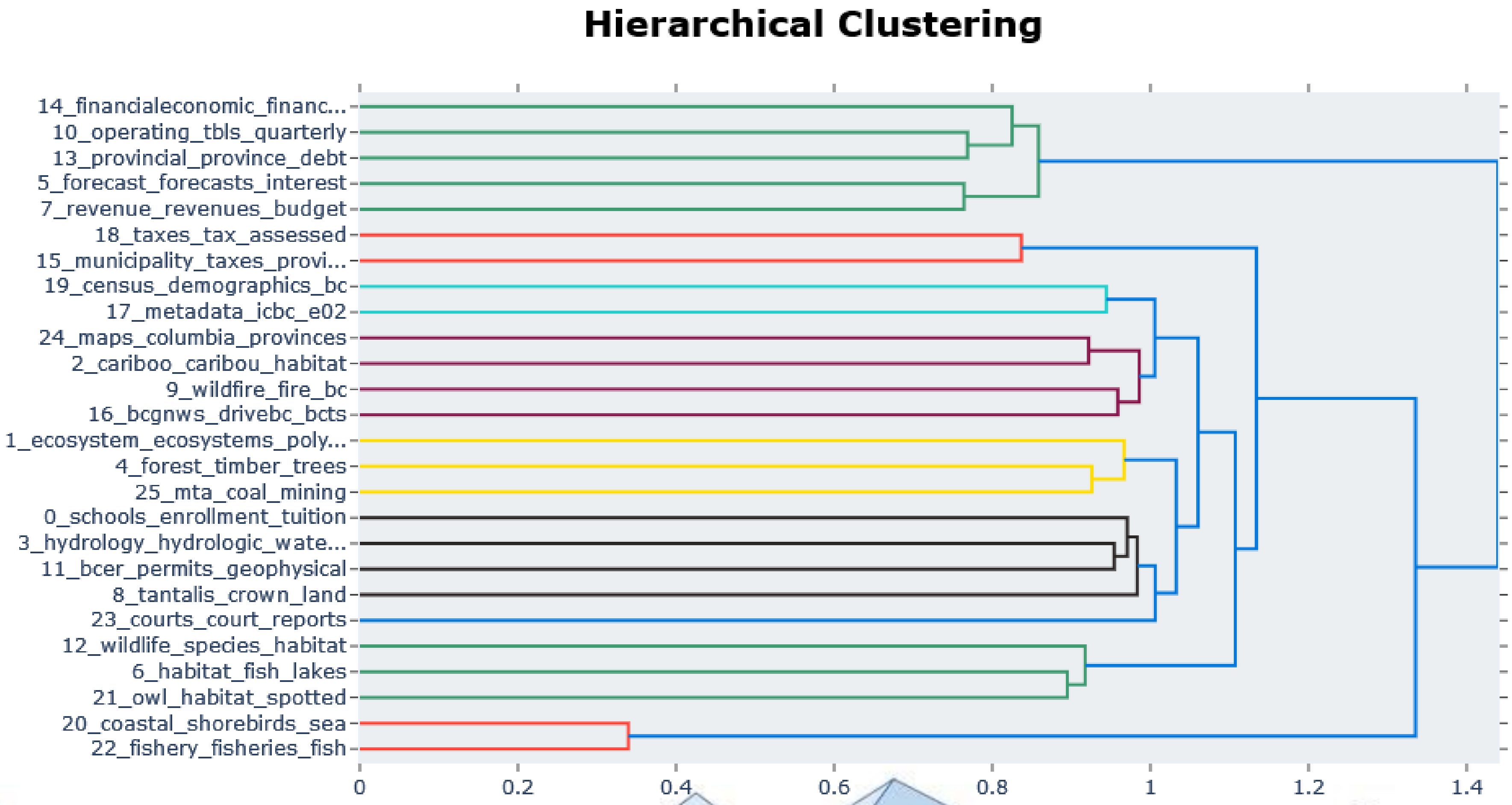
Publishing your metadata in the BC Data Catalogue is a great way to ensure it complies

floodplaon

0 datasets found matching "floodplaon"

No results

Topic Modeling



Implementation Plan

- **Finish upgrade** to ckan and solr (search platform)
- **Integrate the enhanced search solution** with new version of solr
- **Explore topic modeling as an enhancement** to the BC Data Catalogue

Lessons Learned Along the Way

BCDS

- Personal information cannot be collected from the web.
- Need to prepare a “gold standard” dataset and evaluation metrics in advance.

UBC Students

- How things work in the industry
- How we can apply our knowledge
- Learned BC Data Service work ethics and tools

Benefits

BCDS

- Collaboration opportunity for BC Stats and DSS.
- Established project documentation to facilitate future capstone projects
- Got a fresh perspective on enhancing the product
- Learned about practical applications of computational linguistics

UBC Students

- Learned new skills
- Worked with real-world data
- Teamwork experience

How Labour Intensive Was It?

BCDS

- 35 hours for proposal and project set-up.
- 6 hours for weekly meetings
- 15 hours for ad hoc student support

UBC Students

- 6 weeks full time work

BRITISH COLUMBIA

Demo

Questions

References

- Furnas, G. W., Landauer, T. K., Gomez, L. M., and Dumais, S. T. (1987). The vocabulary problem in human-system communication. Communications of the ACM 30(11), 964971.
- Jurafsky, D., & Martin, J. H. (2020). Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Pearson.
- Lin, J., Nogueira, R., and Yates, A. (2021). Pretrained transformers for text ranking: BERT and beyond. arXiv:2010.06467v3