Exploring monitoring methods using text clustering and LLM

Enhancing Data Organization and Exploration Across Domains

Minh Doan[1], Sam Cimino[2], Tomas Bird[1], Jenn Bayer[2]

1) Department of Fisheries and Oceans, Canada

2) US Geological Survey

1. Introducing Bob the Biologist

Biologist at Department of Fisheries and Oceans of Canada

Responsible for monitoring health of fish habitat in streams

What monitoring methods should Bob use given:

His organization's objectives

His constraints?

2. Questions he asks himself

How do I read through all these information?

Bob would be able to get through 5-10 of the long paper per day and he would be burnt out by the weekend

How does any of these related to me or my goal?

Bob is confused at how the method proposed in the paper can help him given his constraints

I do not understand what this paper is talking about?

Bob is lost reading the paper

3. DB example - MonitoringResources.org

Repository of 1000’s of methods

Goldmine of knowledge

But…

Volume of information is too big to read

There’s too much **Diversity** of information

A lot of it is very **Complex** and outside of his experience

4. AI can Help!

AI:

- Faster at reading
- Unbiased by the content and able to view the data objectively
- Able to repeat what they have done on any scale at any time
- Can be cheap! [less than Grande Capuchino with two pumps of hazelnut syrup and a Boston cream donut per day]

Bob:

- He can swim and dive into the river without shorting himself

5. Sneak peak into how AI handle text data

Consider how AI can turn text data into numbers with tokenization

AI turns words into numerical tokens

- Tokens are vectors (with many rows)

Based on tokens, AI can do many things including:

- Summarization:
  - Find the average between all vectors, and then generate words that would encompass all those vectors
- Clustering:
  - Find distance between all vectors, then perform mathematical graph algorithms to find clusters of vectors

6. What AI will be doing in for Bob

Summarization:

- AI will read through all the methods for Bob from a data source (Monitoring Resources)
- AI will summarize them all so Bob doesn't have to read through too much

Clustering:

- AI can tokenize the data and use math algorithm on them to find clusters of topics
- AI can do this repeatedly until Bob can ensure all topics are independent from each other
- Bob can then find the method within the topic of interest based on their similarity score to his need

7.Interactive Example – MASS Dashboard

Questions and Answers

Exploring data through clustering and LLM

Enhancing Data Organization and Exploration Across Domain

Tomas Bird and Minh Doan