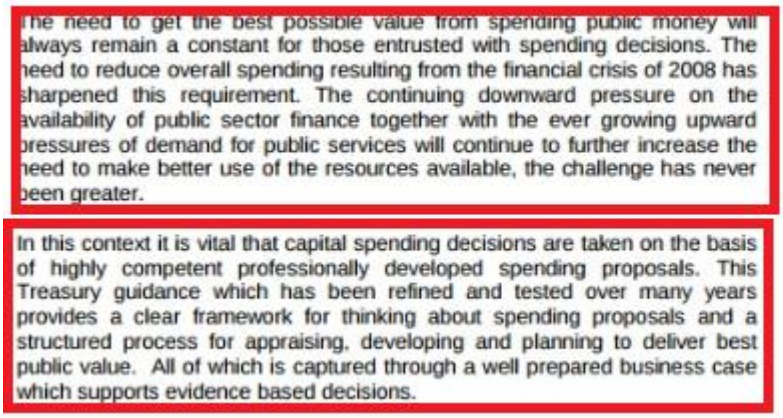
Optimising RAG Model For QA Bot

Sliding Window Approach (Method-1)

Let us first look at the existing approach and the problems associated with it:-

➔ Existing Approach: -
We split the data from the pdf into small chunks of data. Each chunk represents 1 paragraph.

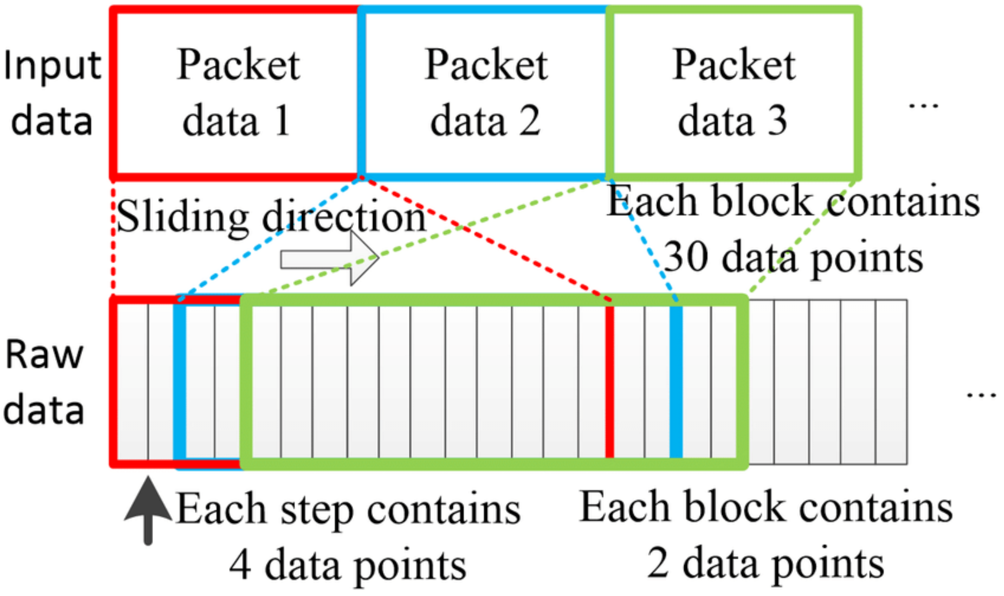


The need to get the best possible value from spending public money will lways remain a constant for those entrusted with spending decisions. The eed to reduce overall spending resulting from the financial crisis of 2008 has harpened this requirement. The continuing downward pressure on the vailability of public sector finance together with the ever growing upward ressures of demand for public services will continue to further increase the eed to make better use of the resources available, the challenge has never een greater.

In this context it is vital that capital spending decisions are taken on the basis of highly competent professionally developed spending proposals. This Treasury guidance which has been refined and tested over many years provides a clear framework for thinking about spending proposals and a structured process for appraising, developing and planning to deliver best public value. All of which is captured through a well prepared business case which supports evidence based decisions.

Problem : The problem with this existing approach is the loss of context between chunks of data.

Solution : Sliding window approach


Input data
Packet data 1
Packet data 2
Packet data 3
...
Sliding direction
Each block contains 30 data points
Raw data
...
Each step contains 4 data points
Each block contains 2 data points

According to the sliding window approach each packet of data starts with the last few data points of the previous packet.

Benefits of using this approach in given assignment :-

- Creates meaningful chunks by finding natural break points
- Maintains context between chunks through overlap
- Stores surrounding context as metadata
- Improves retrieval accuracy by reducing the chance of splitting relevant information
- Helps maintain document coherence and contextual understanding

Dynamic Query Expansion with Semantic Routing (Method-2)

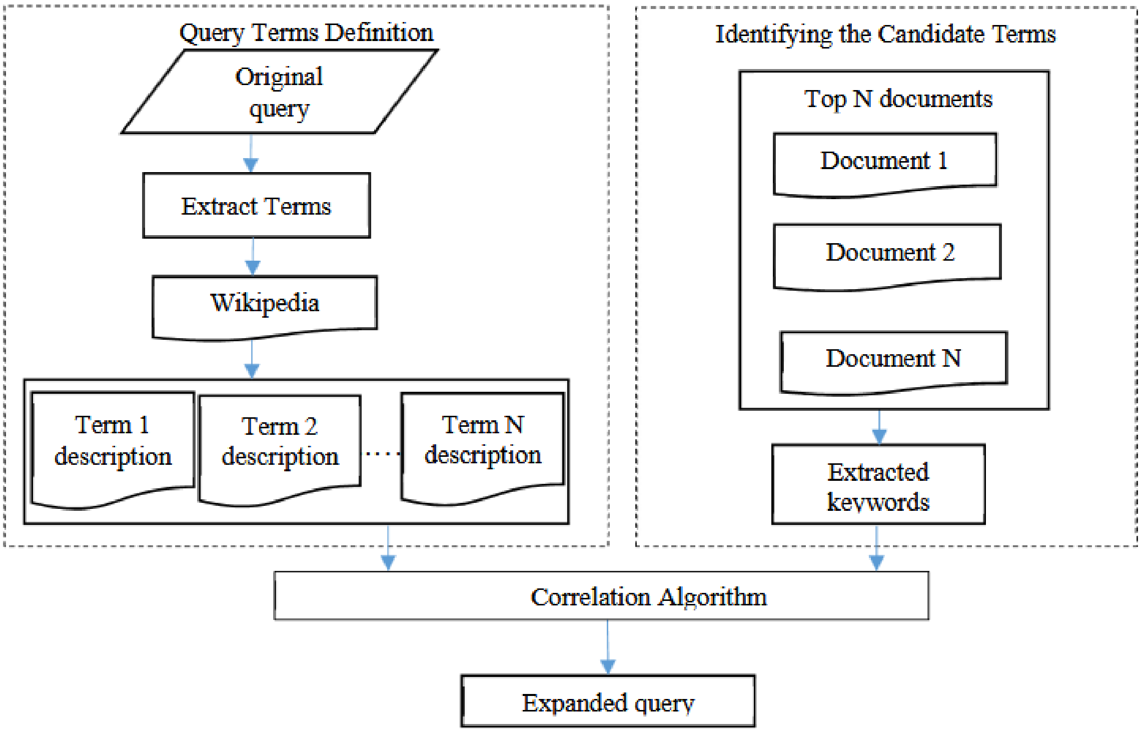


Figure 1: Semantic query expansion method

➔ Existing Approach (From the notebook submitted in form):-

- Takes the users question directly
- Creates a single embedding for the exact question asked
- Searches for matches using just the one embedding
- Returns the top 3 most similar results

➔ For example, if someone asks "Do I need internet?", it only looks for that exact phrase.

➔ Dynamic Query Expansion Approach:-

- First identifies the type of question asked
- Creates multiple variations of the question
- Gives different importance (weights) to each variation
- Combines best results from all variations to get better answers

➔ For example, if someone asks "Do I need internet?", it will also search for:

- "What internet requirements are there?"
- "Is internet connection required?"
- "Internet connection specifications"

➔ Pros of our optimised Dynamic Query Expansion approach:

- Better understanding of questions
- More flexible matching
- Smarter results