Expert Search- The Core

ANALYSIS REVIEW PRESENTATION

PRESENTER: OSCAR WONG

What is Expert Search?

Type: Web-based information retrieval system

Purpose: To help user to find relevant experts based on their expertise from the experts' knowledge base.

User: Student & Researcher

Input: keywords, e.g. artificial intelligence

Output: relevant expert

The Core – Proposed Solution

1. Text Processing

- Tokenization - the stream of text will be chopped into a set of token.
- Posting – create the inverted index to link the token/term to the document.
- Scoring – determine the occurrence rate of the term in the document.

2. Feature selection for auto complete function

- The keywords and the phrases of the document are selected and insert into feature list for auto complete function.
- Then The Clue module can use the feature list to suggest meaningful keywords to user.

The Core – Proposed Solution cont.

3. Text Categorizer

- Determine and categorize all documents into the class or category (e.g. Artificial Intelligence, Multimedia).
- Use defined categories' keywords to match the keywords of the document.

4. Expertise Finder

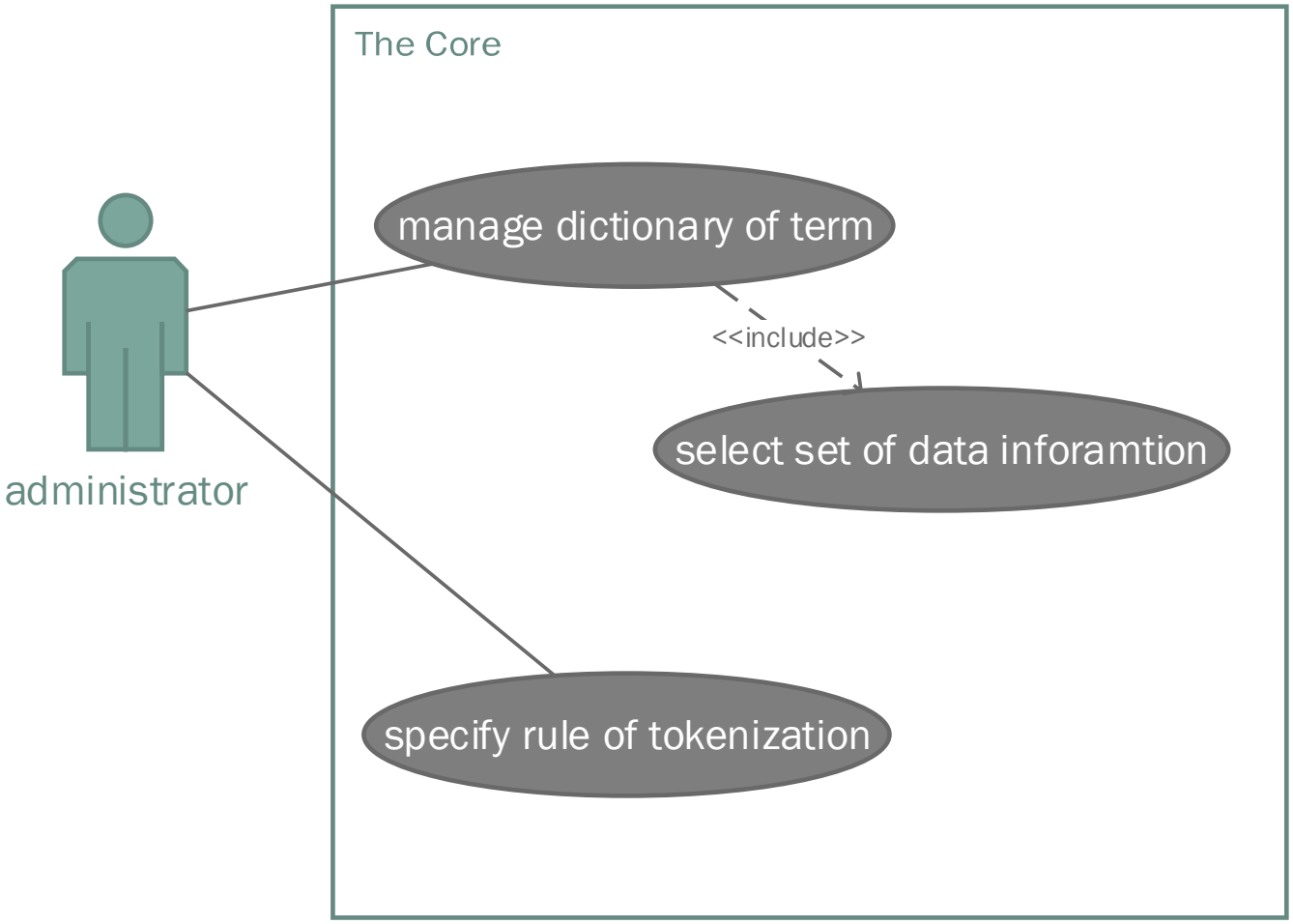
- To find out the important keywords in the document that related to the expert.
- e.g. the term (artificial intelligence) appears many times in the document, that means the relevant expert has some expertise in this term (artificial intelligence).
- This can help The Clue module provide more accuracy result to the user.

Sub-modules of The Core

Sub-module	Functions
Indexer Module	Tokenize stream of word in the document
	Posting and create inverted index
Administrator Module	Term dictionary management
	Setup rule of the tokenization
Feature support for other module	Scoring and term weighting
	Feature selection for auto-complete function
	Text Categorizer
	Expertise Definer

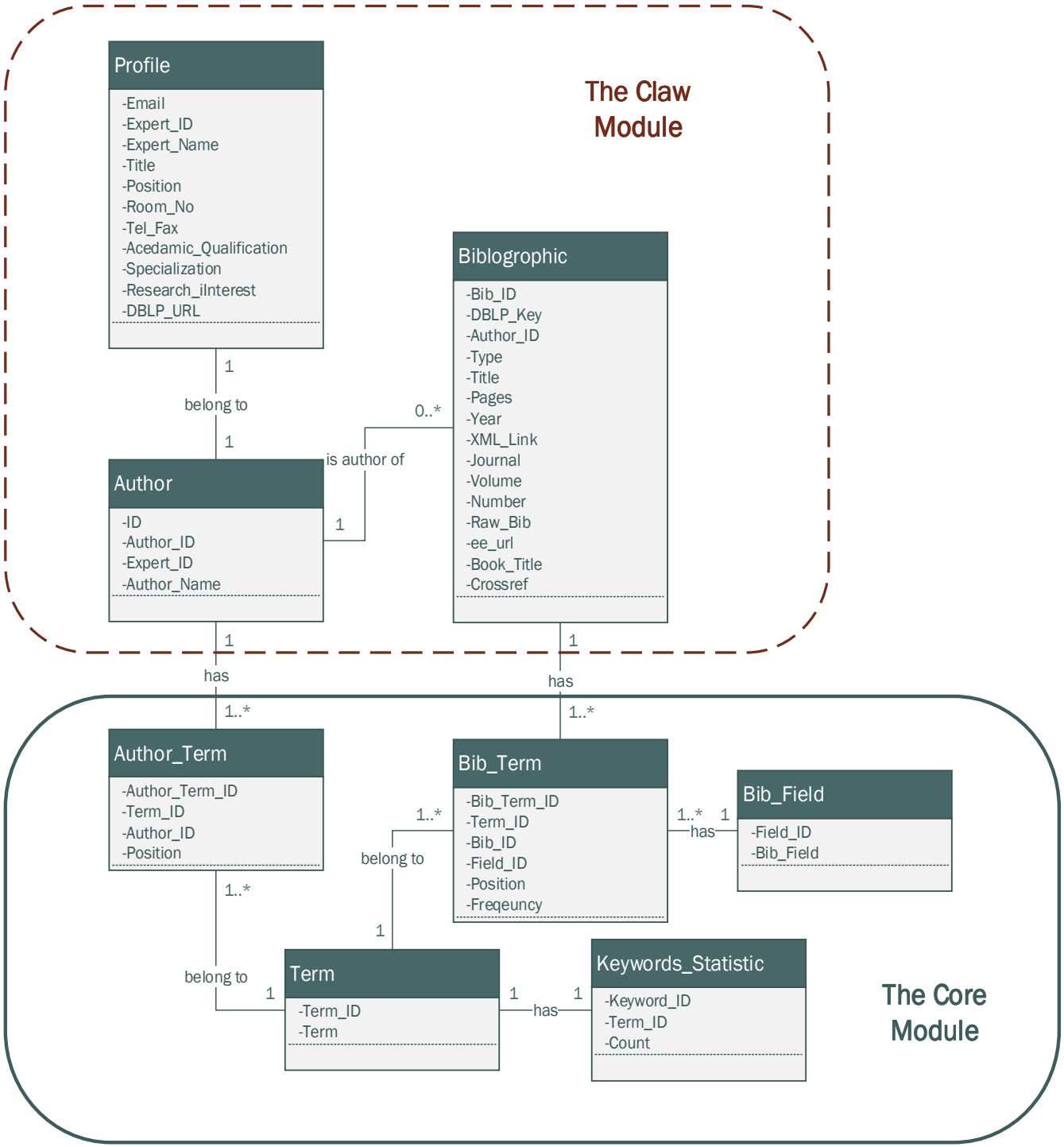
Use Case Diagram

Since The Core module is more focus on the back-end functionality, therefore there is not many use case in The Core module.

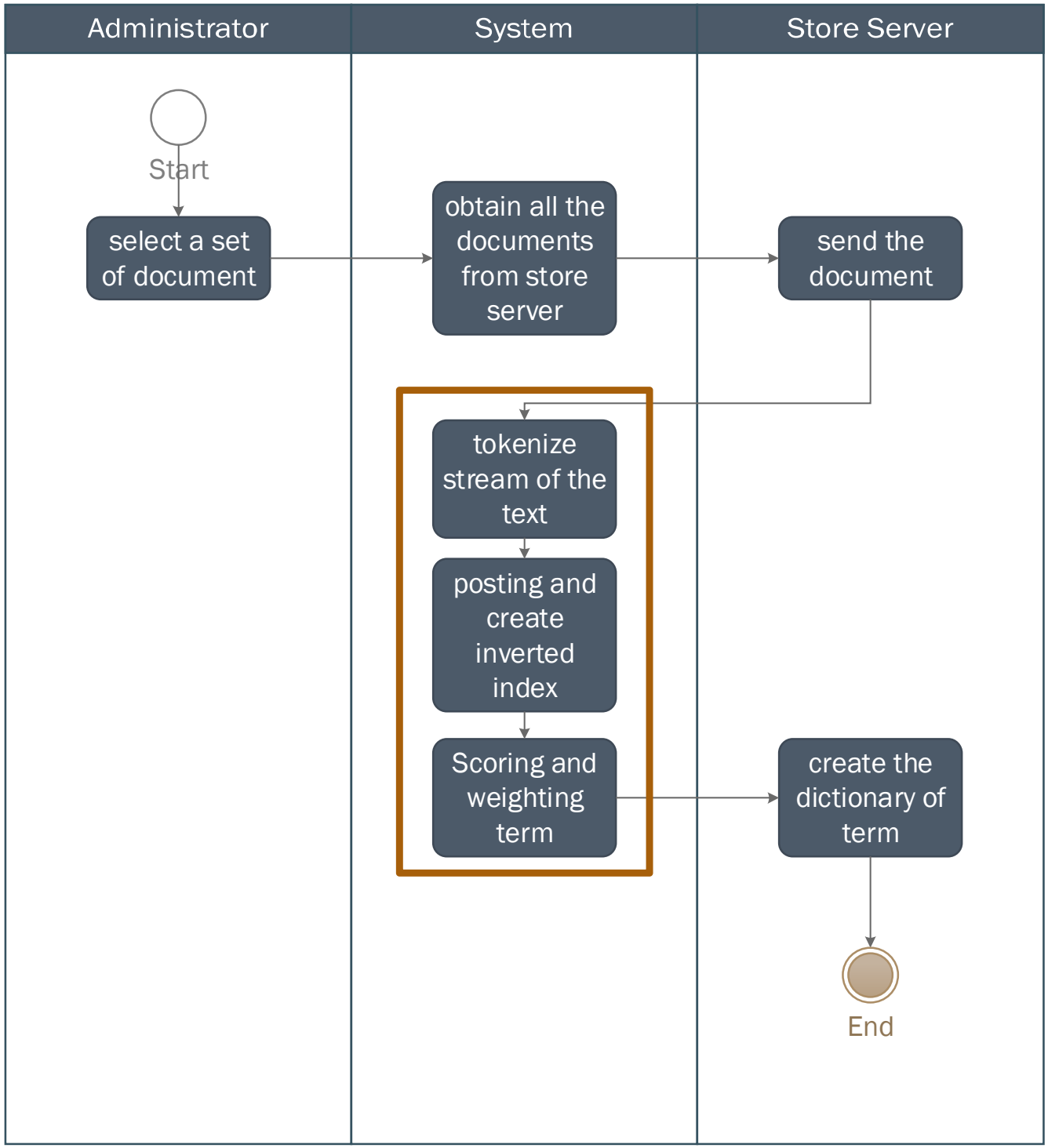


Class Diagram

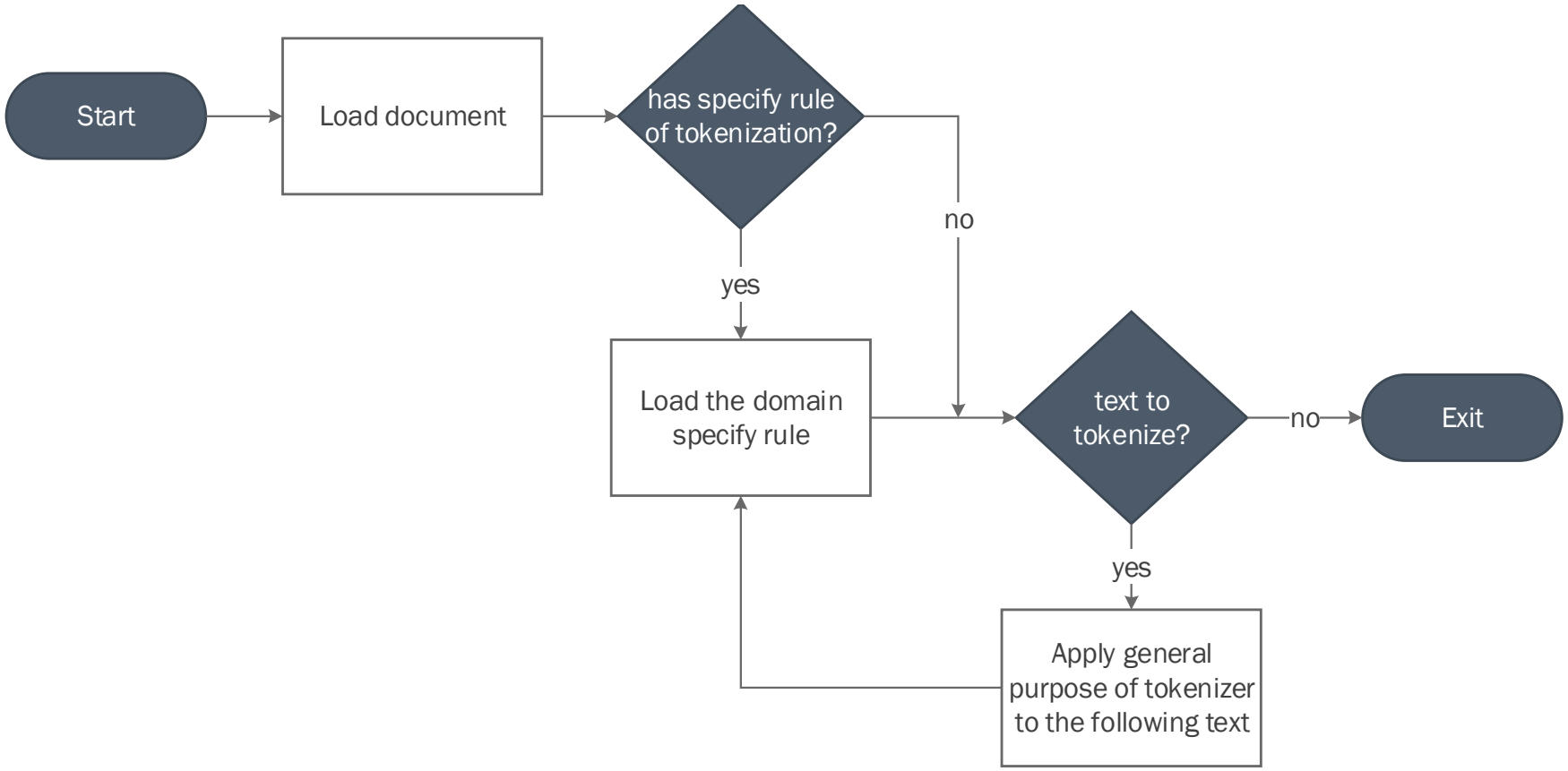
Initially designed class diagram



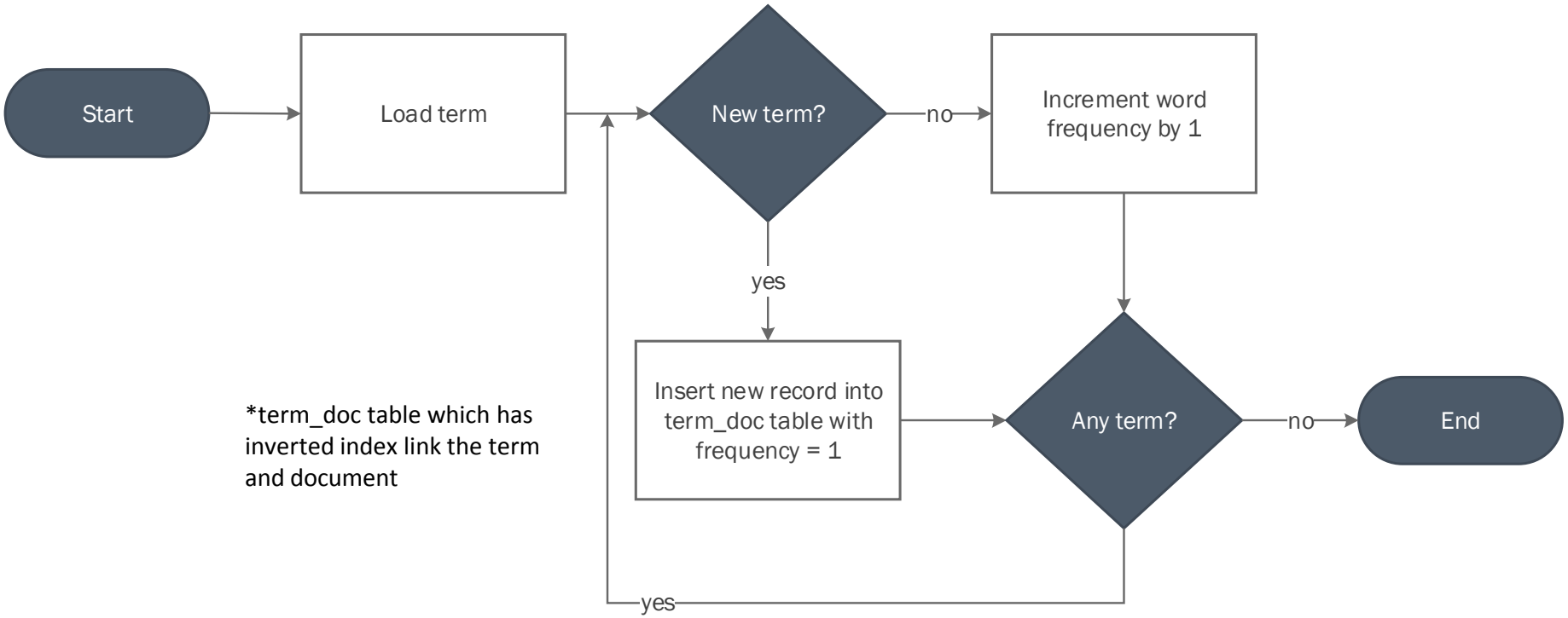
Activity Diagram For Admin to manage term dictionary



Flowchart - Tokenization



Flowchart-Posting & Scoring



Techniques and Algorithms

1. Tokenization
   - **Part-of-speech (POS) Tagging** – assigns parts of speech tag to each word (token), such as noun, verb, adjective, etc.
   - **Named Entity Recognition (NER)** - labels sequences of words in a text which are the names of things such as person's name, company's name, location' name & time.

2. Scoring
   - **Term frequency-inverse document frequency (tf-idf)** - numerical statistic that is intended to reflect how important a word is to a document. The tf-idf value *increases proportionally* to the number of times a word appears in the document.

Techniques and Algorithms cont.

3. Expertise finder

- Use n-gram technique to tokenize and find out the important keywords in the document. To find the expertise of the expert in the document.
- **n-gram technique** – to generates n-gram words from a text.
- E.g.

  text = “Search engine is an artificial intelligence system.“

  2-gram word:

  Array ([0] => Search engine [1] => engine is [2] is an [3] => an artificial [4] => artificial intelligence [5] => intelligence system)

Conclusion

The Core is back-end system which process the data from The Claw and give provide meaningful data to The Clue.

We are still exploring and researching some additional function and better technique and algorithm for improvement and increment of our project to make the system more reliability and feasibility.

Although we are interdependent to each other modules, but we are doing our own work independently and share our responsibility to achieve the common goal.

Thank You