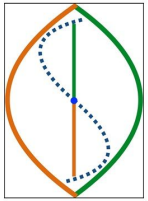


Data Sciences
and
Analytics Centre

INTERNATIONAL INSTITUTE OF
INFORMATION TECHNOLOGY
HYDERABAD

Extracting Evidence from Detective Novels

Aditya Motwani, Aayush Naik, Kamalakar Karlapalem

IE in Detective Novels

A mine of information stored in a narrative.

Information retrieval (IR) - ***only as good as*** the query terms used,
limited to the background knowledge of the case + extended search terms from the investigator's personal experience.

Explore the problem of extracting evidence summaries from detective novels.

Aim to reduce the ***search space*** for easier analysis and potential applications of visualisation in such narratives.

IE in Detective Novels

Standard Approach
Train state-of-the-art deep networks

Limitations
Require a very large labelled training corpus of annotated data.
Difficult to extract data in a supervised way

Our Approach
Extract evidence summaries unsupervised method
Doesn’t require large corpus of annotated data
Exploits the underlying semantic structure in long text

Our Approach

Approaching the problem ***backwards.***

Doesn't identify all clues, and then find the culprit.

Finds the culprit declaration paragraph then works towards finding all the evidence summaries for all central characters.

Identify all potential evidence sentences for each character, then evaluate how useful these summaries are for culprit identification v/s a baseline.

Pre-processing task

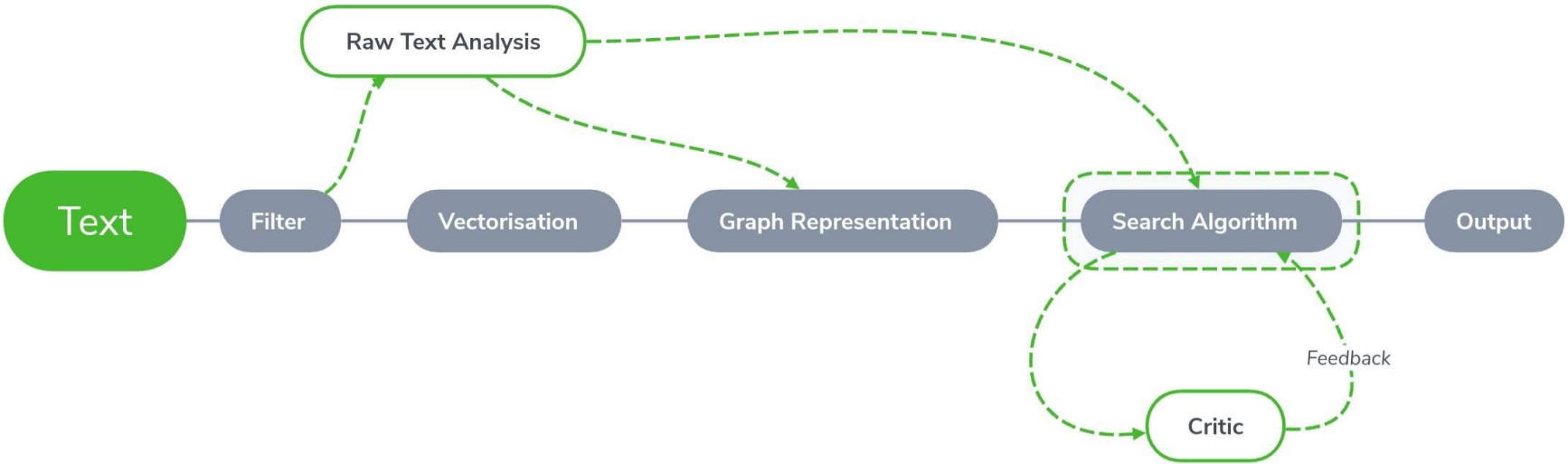
Collected all novel corpus data, and removed all breaks in forms of chapters, paragraphs.

Unit of processing - **sentences.**

Resolved common character names and pronouns, removed stop words, removed footnotes, chapter names, appendices and other supplementary information not directly related to the story.

Converted all sentences to **SpaCy vectors** for further processing.

Pipeline


Raw Text Analysis
Text
Filter
Vectorisation
Graph Representation
Search Algorithm
Output
Feedback
Critic

Two components

1) **Culprit Identification Paragraph**

   We identify the paragraph where the culprit was revealed. This is done by ***culprit binning algorithm***.

2) **Character specific evidence summaries**

   After identification of culprit paragraph, we extract the evidence summaries for each character, working on our graph based search. We employ the use of an ***Augmented A* search,*** a graph based search algorithm.

Culprit Identification Paragraph

Motivation

Frequency of mention of the culprit name increases in the paragraph after they are revealed. Our algorithm aims to exploit this intuition.

Challenge

Find the optimal size of “paragraph”. This paragraph, may not correspond to the traditional demarcated paragraph, hence we find the optimal size of ***bucket.***

Approach

Iteratively bin sentences into different sizes of bins, then find the optimal size of bin where the frequency of occurrence of any character is highest.

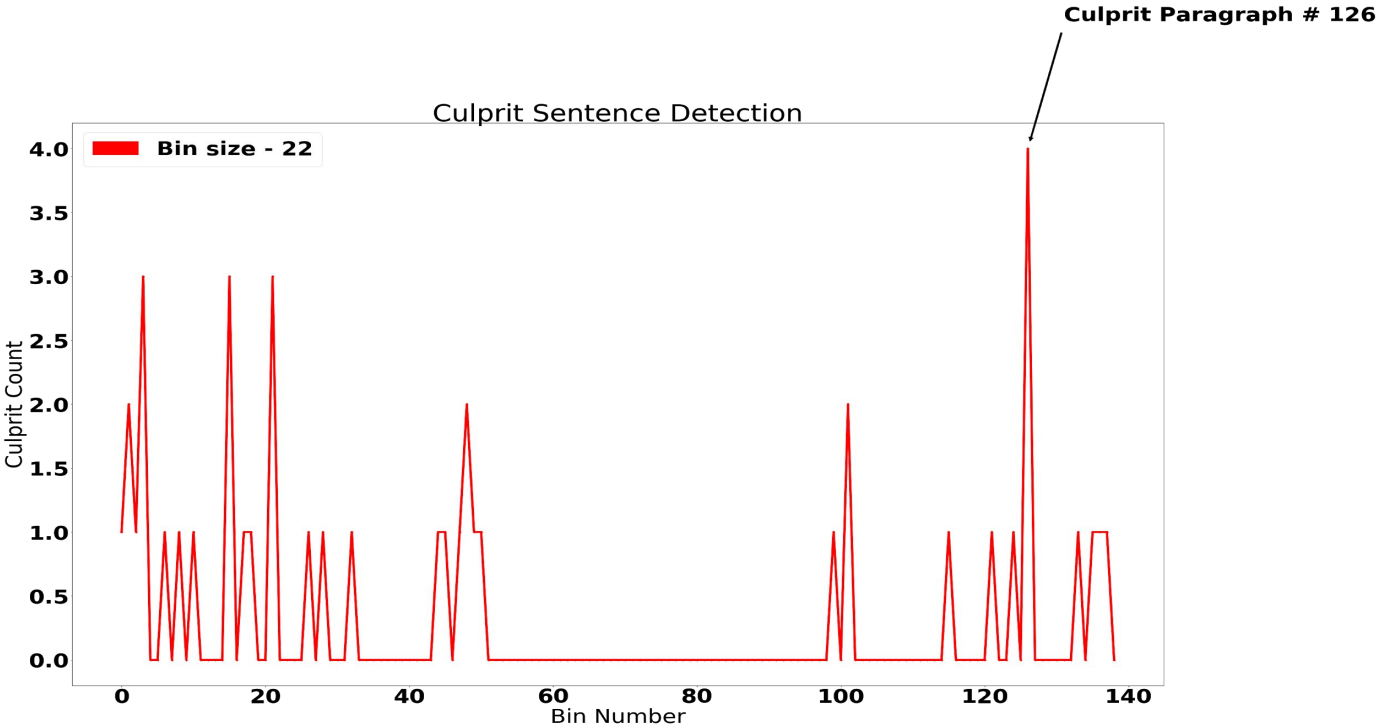
-

Culprit binning algorithm

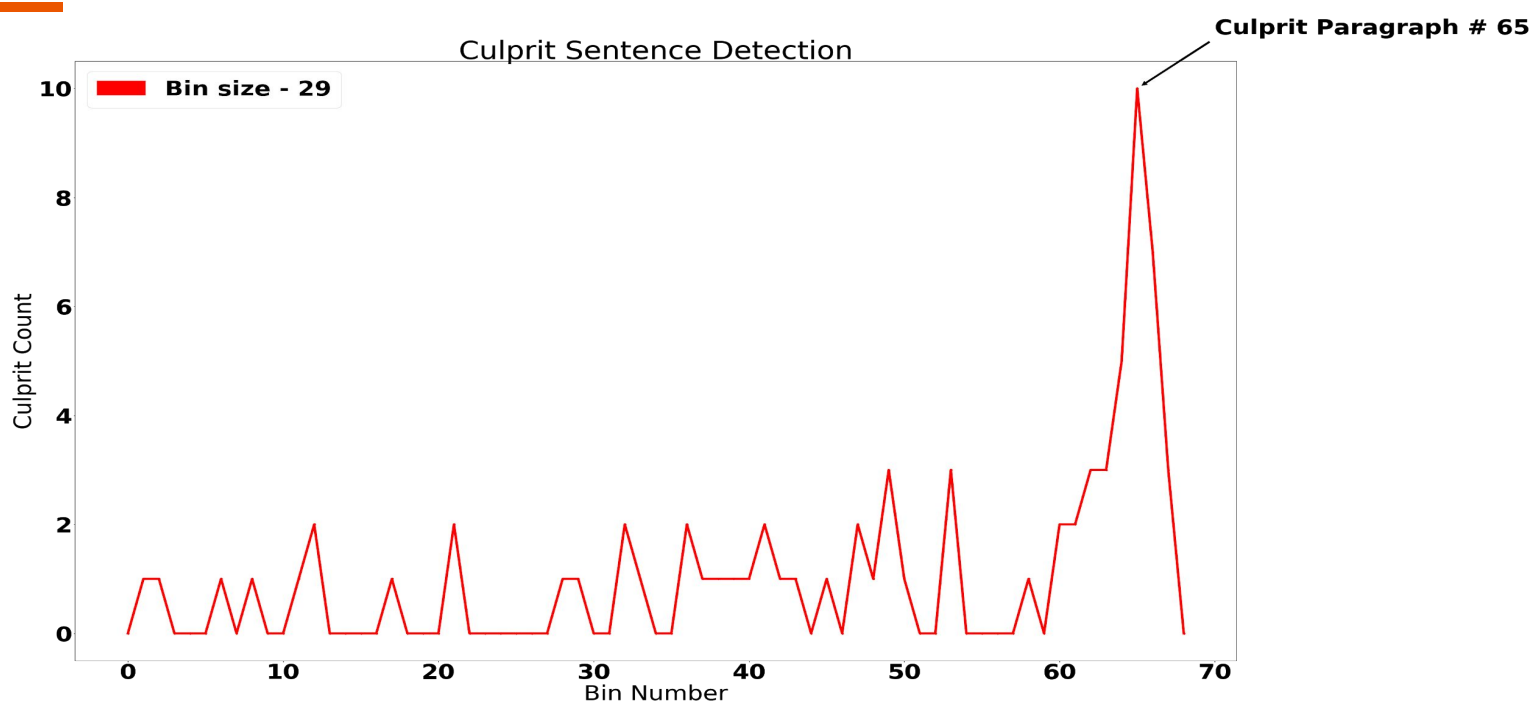
```
Initialise: bucket_size = min_bucket_size, max_mentions = 0, optimal_bucket_size
= min_bucket_size, culprit_bucket = None

While: bucket_size < max_bucket_size
     Buckets = empty_list()
     Divide list of sentences into buckets of size bucket_size
     Buckets = [bucket size*i to bucket size*(i+1)] : for i in range(number of
buckets)]
     mentions = Count of times culprit is mentioned in each bucket
     If: (mentions > max_mentions) then
          max_mentions = mentions
          optimal_bucket_size = bucket_size
          culpit_bucket = bucket
     end
     bucket_size ++
  end
  return optimal_bucket_size, culprit_bucket
```

Mr Franklin Clarke in the ABC Murders



Mr Stapleton in the Hound of Baskerville



Augmented A* Search

Our Augmented A* algorithm in two parts - a general graph search algorithm [Algorithm (2)] that takes in a strategy as input and the augmented A* strategy [Algorithm (3)].

```
Algorithm 2: Graph Search (Sentences doc, Sentence start, Sentence goal, Strategy strategy)
Initialise:
        Sentence current = start
        Set(Sentence) frontier = NULL
        Set(Sentence) visited = NULL
while: current != goal do
    yield current
    visited = visited ∪ {current}
    Set(sentence) new_nodes = Set{neighbours(current)} - visited - frontier
    frontier = frontier ∪ new_nodes
    next_node, dist_from_start = strategy(doc,current,frontier,goal,dist_from_start)
    current = next_node
    frontier = frontier - {current}
    end
end
```

Augmented A* Search

```
Algorithm 3: A* Strategy (Sentences doc, Sentence current, Set(Sentence) frontier,
Sentence goal, Map distance_from_start, Float α, Float β, Function goal_heuristic,
Function distance)
    if distance_from_start is empty: then
            distance_from_start = ∞ : sent for all sent in doc
            distance_from_start[current] = 0
    end
    forall sentence neighbour in neighbours(current) do
        dist_from_start[neighbour] = min(distance_from_start[neighbour],
                distance_from_start[current]+distance(current,neighbour))
    end
    sentence next_node = argmin(α*distance_from_start[x] +
    β*goal_heuristic(goal,x)) for all x in sentences
    ## get the next sentence in the summary
    return next_node,dist_from_start
end
```

Methods of evaluation

Baseline- (lexrank)
Graph based summarisation method to create evidence based summaries.

The baseline output is a generic character summary of the character, not explicitly focusing on evidence.

Evaluators
10 evaluators were tasked with judging how useful were the evidence summaries in identifying the culprit of the case.

Our aim was to compare our results in terms of information provided by our method v/s the baseline in identifying the culprit.

Methods of evaluation

- Employed **10 evaluators** who have read all the books or the long summaries.
- Long summaries were **extracted from wikipedia or sparknotes**.
- Summaries presented were for 3 main characters ***apart*** from the detective.
- Evaluators reported if the summaries for different characters ***assisted in identifying*** the culprit.
- Responses grouped into ***three categories*** depending on the degree of information provided through the evidence summaries.

Evaluation Metric

The responses of the participants were broadly categorized into three classes

CI (*Culprit identification)* The evidence summaries were useful in finding the right culprit

NCI (Non *Culprit identification)* The evidence summaries are misleading and direct the crime to a non-culprit character

NA (Not Available) The evidence summaries are not informative in labelling any character as culprit

Results

We compare the results from our method v/s the lexrank generated summaries from our evaluators.

- As compared to LexRank, our method has a **higher CI percentage** across all categories.
  - For the CI category, highest percentage for LexRank is 30% (Eye of Needle), whereas it is 80% for our method (Red Dragon).

- **LexRank** results were dominated by **NA category,** and not so for our Augmented A* method.

- Our **Augmented A*** method also had a **greater degree** of readers for **NCI** v/s LexRank.

Results

For results from both the methods, we also draw these analysis

1. Novels with a high percentage of readers who fall in **N A**, can be said to provide negligible information about the culprit before revelation - **HARD**.

2. Novels with high percentage of readers who fall in **CI** are the ones which make culprit guessing easy and do not have sufficient obfuscation **- TRIVIAL**.

3. Novels with high percentage of readers who fall in **NCI** seem to provide only subtle obfuscated information about the culprit - **ELUSIVE**.

Conclusion & Future Work

- We explore the problem of extracting evidence summaries for characters in a novel.
- As there is no existing annotated data, we explore an unsupervised method of learning.
- We use an augmented A* algorithm which takes in an implicit graph representation and finds an optimal path to the culprit reveal sentence.

Future work, we would like to explore abstractive summarization as a technique to generate coherent evidence summaries.

Domain of voice-activated agents, which would provide culprit narratives for humans.