Segmentation

1.1 Pragmatic Segmenter

Language :- Ruby on Rails
Description :- pragmatic segmenter does not use any machine learning techniques
Pragmatic segmenter is specifically used for the purpose of segmenting texts for use in translation related applications. Some example of application are :-

- Removes “table of contents” style long string of periods
- keeps parentheticals, quotations, and parentheticals or quotations within a sentence as one segment for clarity even though technically there may be multiple grammatical sentences within the segment
- segmenter does not have a rule

1.2 Punkt Sentence Segmenter

Language :- Python
Description :- Punkt tokenizer uses unsupervised machine learning algorithm
This tokenizer divides a text into a list of sentences by using an unsupervised algorithm to build a model for abbreviation words, collocations, and words that start sentences. It must be trained on a large collection of plaintext in the target language before it can be used.
Punkt is designed to learn parameters (a list of abbreviations, etc.) unsupervised from a corpus similar to the target domain.

2.0 Quantitative Evaluation

```
1 Sardines" is the first episode of the first series of the British black comedy anthology series Inside No. 9. Written by Steve Pemberton and Reece Shearsmith, it
  premiered on BBC Two and BBC Two HD on 5 February 2014. In the episode, a group of adults play sardines at an engagement party. Rebecca, the bride-to-be, finds a
  boring man named Ian in a wardrobe; he introduces himself as a colleague of Jeremy, Rebecca's fiancé. The pair are subsequently joined by family, friends and
  colleagues of Rebecca and Jeremy. As more people enter the room and step into the wardrobe, secrets shared by some of the characters are revealed, with various
  allusions to incestuous relationships, child sexual abuse, and adultery. The humour is both dark and British, with references to past unhappiness and polite but
  awkward interactions.
2
3 The story takes place entirely in the bedroom of a country house, with much of the filming taking place inside the wardrobe. Pemberton and Shearsmith wrote the
  episode with the intention of evoking a feeling of claustrophobia in viewers. In addition to the writers, the episode starred Katherine Parkinson, Tim Key, Luke
  Pasqualino, Ophelia Lovibond, Anne Reid, Julian Rhind-Tutt, Anna Chancellor, Marc Wootton, Ben Willbond and Timothy West. The cast and writing were praised by
  television critics, and the episode was chosen as pick of the day in a number of publications. On its first showing, "Sardines" was watched by 1.1 million viewers,
  which was 5.6% of the audience.
4
5 The comedy writers and actors Steve Pemberton and Reece Shearsmith, who had previously worked together on The League of Gentlemen and Psychoville, took inspiration
  for Inside No. 9 from "David and Maureen", the fourth episode of the first series of Psychoville, which was in turn inspired by Alfred Hitchcock's Rope. "David and
  Maureen" took place entirely in a single room, and it was filmed in only two shots.[2] The writers were keen to explore other stories in this bottle episode or TV
  play format, and Inside No. 9 allowed them to do this.[3] At the same time, the concept of Inside No. 9 was a "reaction" to Psychoville, with Shearsmith saying that
  "We'd been so involved with labyrinthine over-arcing, we thought it would be nice to do six different stories with a complete new house of people each week. That's
  appealing, because as a viewer you might not like this story, but you've got a different one next week."[4]
```

Fig 1 :- Input

There are total 18 sentences in the input text

2.1 Pragmatic Segmenter

2.1.1 Output

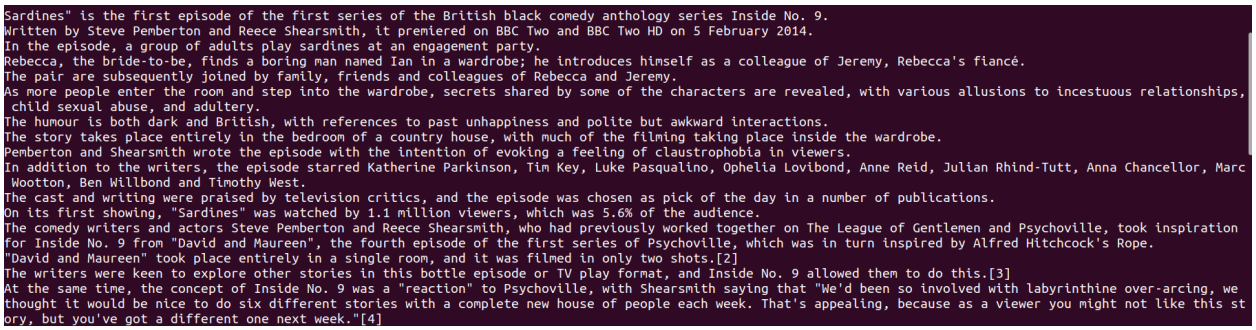

Fig 2 :- Pragmatic Tokenizer output

All the sentence were perfectly segmented, there was no errors. All 18 sentences was perfectly segmented.

2.2 Punkt Segmenter

2.2.1 Output

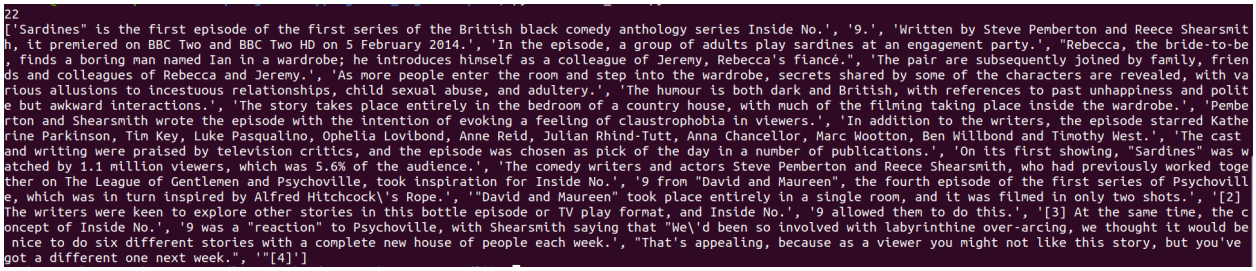

Fig 3 :- Punkt Tokenizer output

Punkt tokenizer did not segmented perfectly compared to pragmatic. Punkt gave 22 sentences which is more than the original given text.

3.0 Error Evaluation

3.1 Punkt Segmenter

- It failed to detect short forms like “No.”
- It detected “[4]” as new line.