Summarizing Literature with Natural Language Processing Models

Anna Rutledge
Data Scientist

A Doll's House

Henrik Ibsen

Study Guide | Full Text

A Doll's House is a play by Henrik Ibsen that was first performed in 1879. Explor
analysis of Nora Helmer, plot summary, and important quotes.

Summary

Read our full plot summary and analysis of *A Doll's House*, chapter-b
breakdowns, and more.

Summary & Analysis

Use a neural network model for natural language processing to create summaries from literary texts

Roadmap

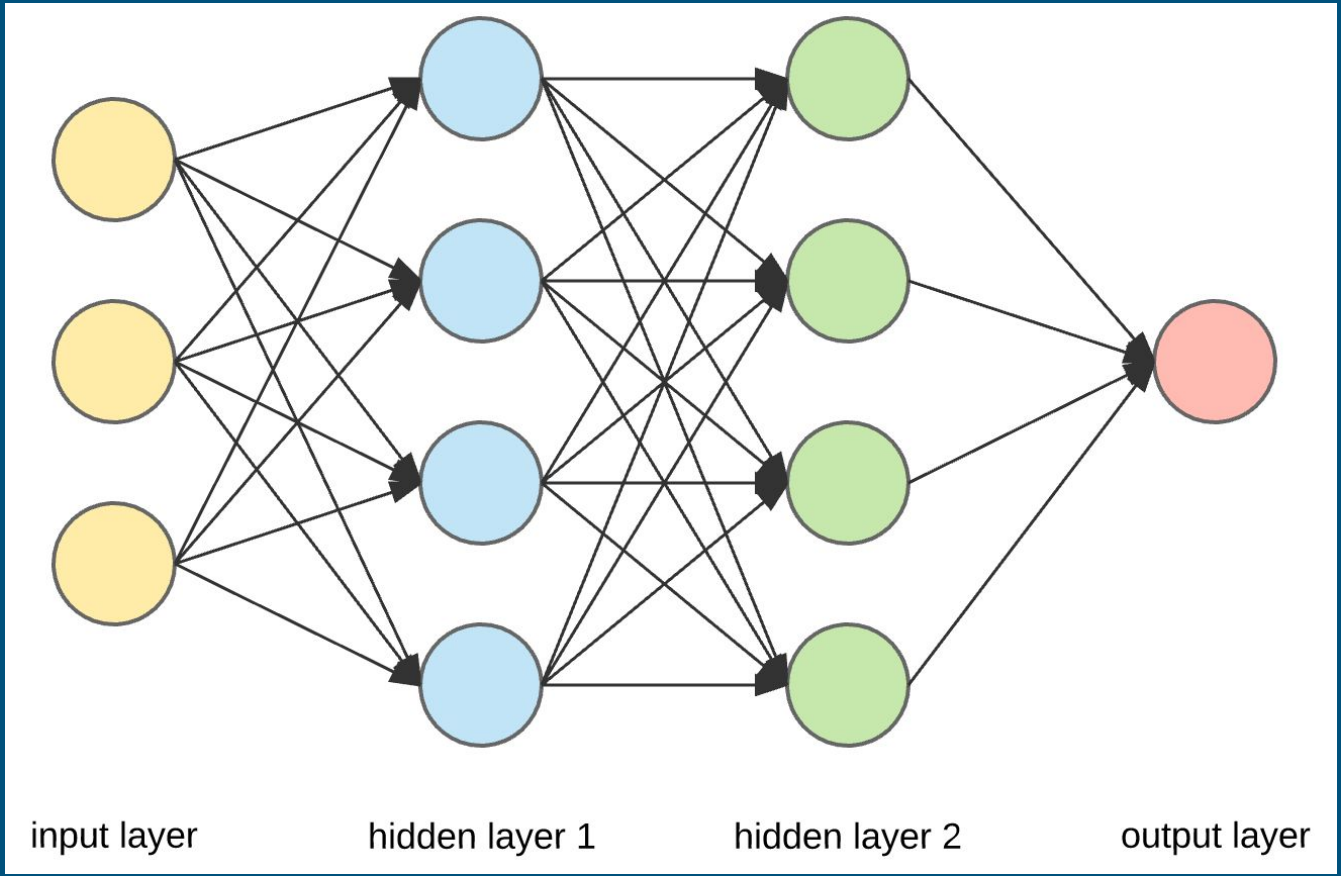
- Problem Statement
- Background & Definitions
- Data Science Process
- Conclusions/Results
- Future Directions for Improvement

Use a neural network model for natural language processing to create summaries from literary texts

‘Natural language processing (NLP) refers to the branch of computer science—and more specifically, the branch of artificial intelligence or AI—concerned with giving computers the ability to understand text and spoken words in much the same way human beings can.’

Source: https://www.ibm.com/cloud/learn/natural-language-processing

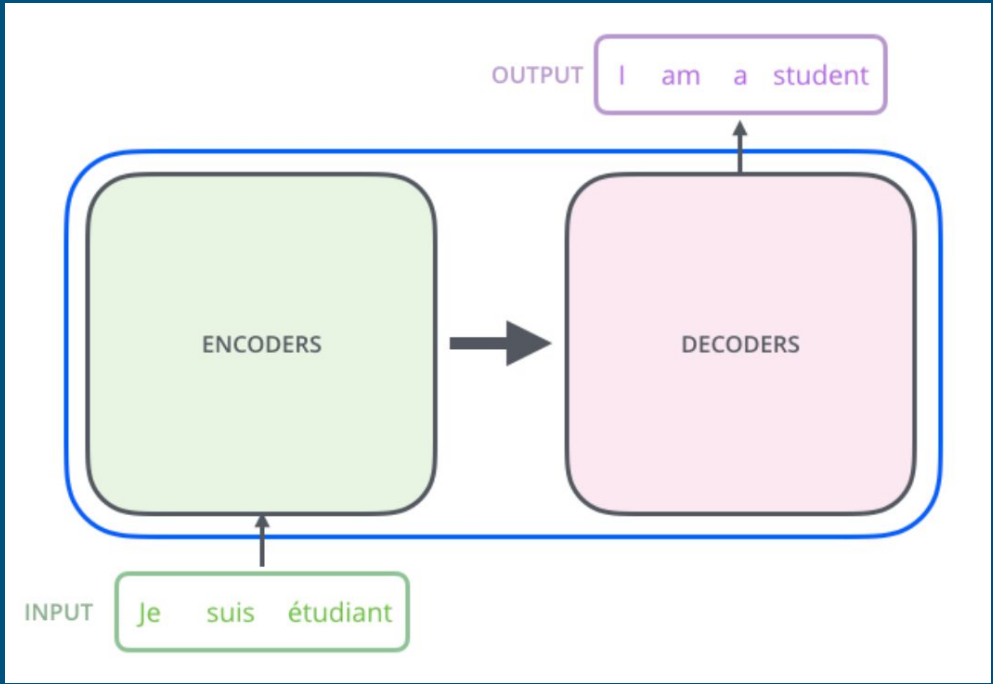
Neural Network



Hugging Face

- Publicly available datasets
- Transformers library

Transformers



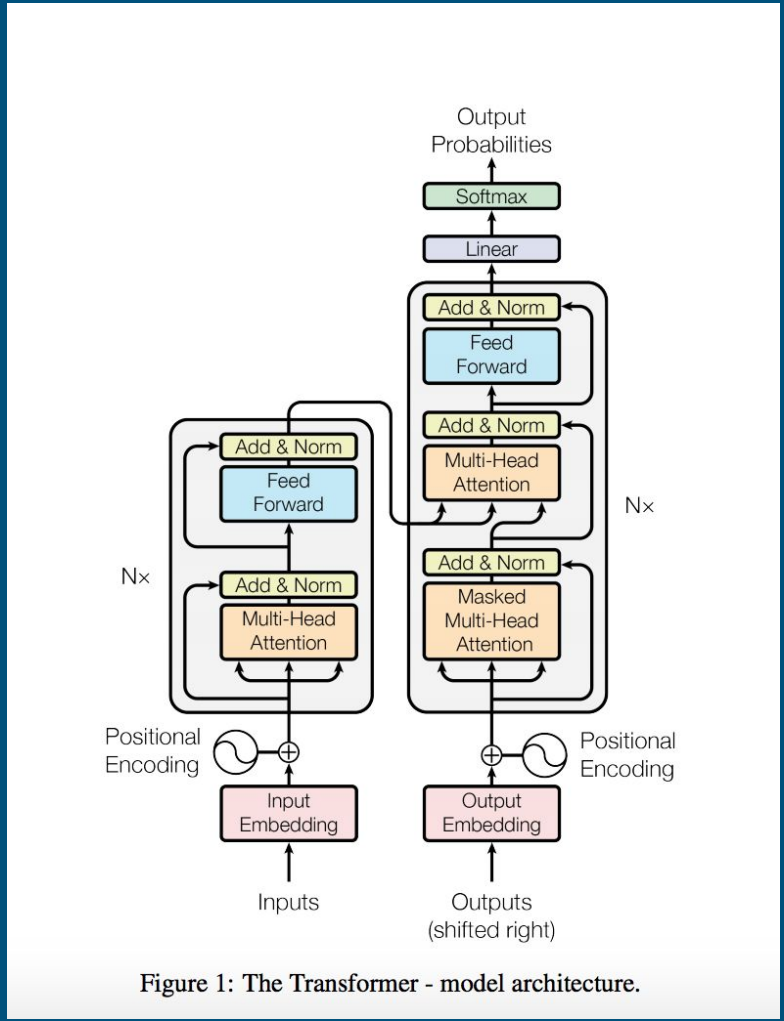


Figure 1: The Transformer - model architecture.

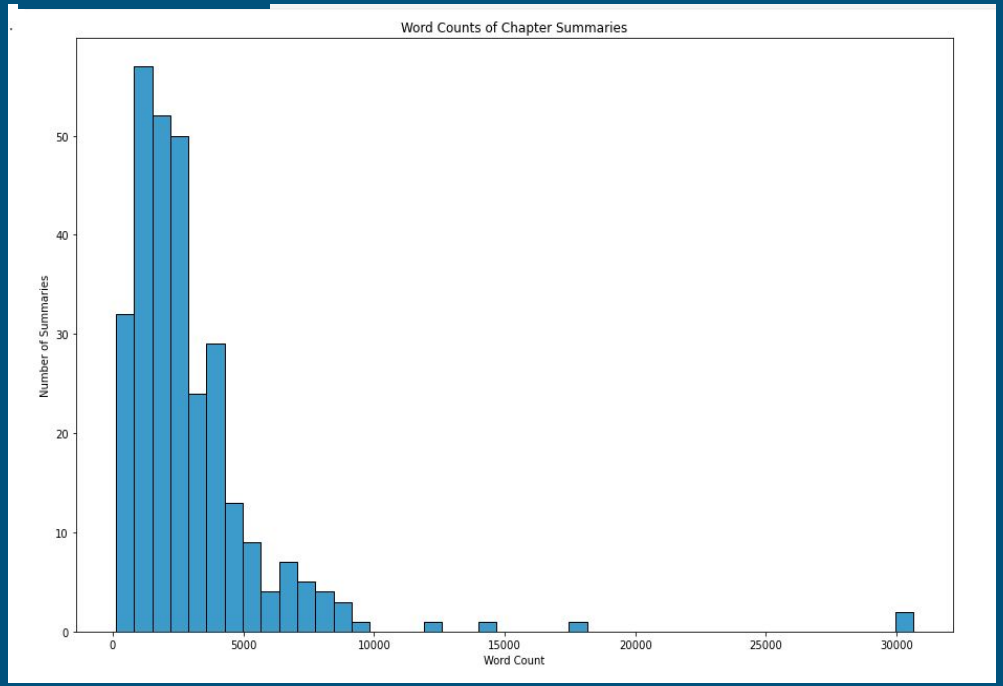
- Data Collection, Cleaning, & Analysis
- Baseline Model
- Data Encoding
- Transformers Model Training & Tuning
- Model evaluation

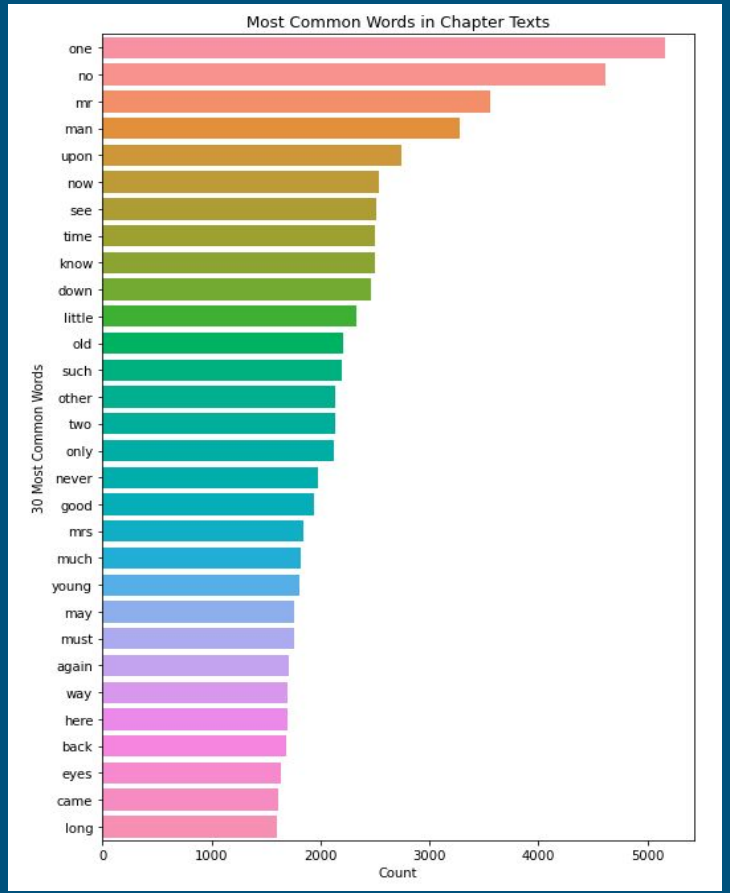
Data Collection & Cleaning

```
    try:
        full_text_url = full_texts_to_scrape[title]
        ft_response = requests.get(full_text_url)
        time.sleep(1)
        if ft_response.status_code==200:
            ft_html = ft_response.text
            ft_soup = BeautifulSoup(ft_html, 'lxml')
            text = ''
            for li in ft_soup.find('ul').find_all('li'):
                full_text_section_url = 'http://www.fullbooks.com/' + li.find('a')['href']
                fts_response = requests.get(full_text_section_url)
                time.sleep(1)
                if fts_response.status_code==200:
                    fts_html = fts_response.text
                    fts_soup = BeautifulSoup(fts_html, 'lxml')
                    for section in fts_soup.find_all('font', {'face':'Arial'}):
                        text = text + section.text
            more_titles.append(title)
            more_texts.append(text)
            print(f'text {i} added')

    except:
        print(f'error with text {i} - {title}')
```

```
# In chapter summaries and texts - replace '\n' , '\t' with a sp

to_replace = ['\n', '\t', '  ', '    ', '     ']

for i in df.index:
    for char in to_replace:
        df.iloc[i]['chapter_summary'] = df.iloc[i]['chapter_summ
        df.iloc[i]['chapter_text'] = df.iloc[i]['chapter_text'].
```

Exploratory Data Analysis





Baseline Model

It is a truth universally acknowledged, that a single man in possession of a good fortune, must be in want of a wife.

However little known the feelings or views of such a man may be on his first entering a neighbourhood, this truth is so well fixed in the minds of the surrounding families, that he is considered the rightful property of some one or other of their daughters.

"My dear Mr. Bennet," said his lady to him one day, "have you heard that Netherfield Park is let at last?"

Mr. Bennet replied that he had not.

"But it is," returned she; "for Mrs. Long has just been here, and she told me all about it."

Mr. Bennet made no answer.

"Do you not want to know who has taken it?" cried his wife impatiently.

"YOU want to tell me, and I have no objection to hearing it."

This was invitation enough.

"Why, my dear, you must know, Mrs. Long says that Netherfield is taken by a young man of large fortune from the north of England; that he came down on Monday in a chaise and four to see the place, and was so much delighted with it, that he agreed with Mr. Morris immediately; that he is to take possession before Michaelmas, and some of his servants are to be in the house by the end of next week."

1 It is a truth universally acknowledged, that a single man in possession of a good fortune, must be in want of a wife.

2 However little known the feelings or views of such a man may be on his first entering a neighbourhood, this truth is so well fixed in the minds of the surrounding families, that he is considered the rightful property of some one or other of their daughters.

3 "My dear Mr. Bennet," said his lady to him one day, "have you heard that Netherfield Park is let at last?"

Model Evaluation: Rouge Scores

Summary vs Generated Summary:

$$\text{Rouge-n Recall} = \frac{\text{number of matching groups of } n \text{ words}}{\text{number of words in reference summary}}$$

$$\text{Rouge-n Precision} = \frac{\text{number of matching groups of } n \text{ words}}{\text{number of words in generated summary}}$$

RougeL = longest matching sequence of words

Baseline Rouge Scores

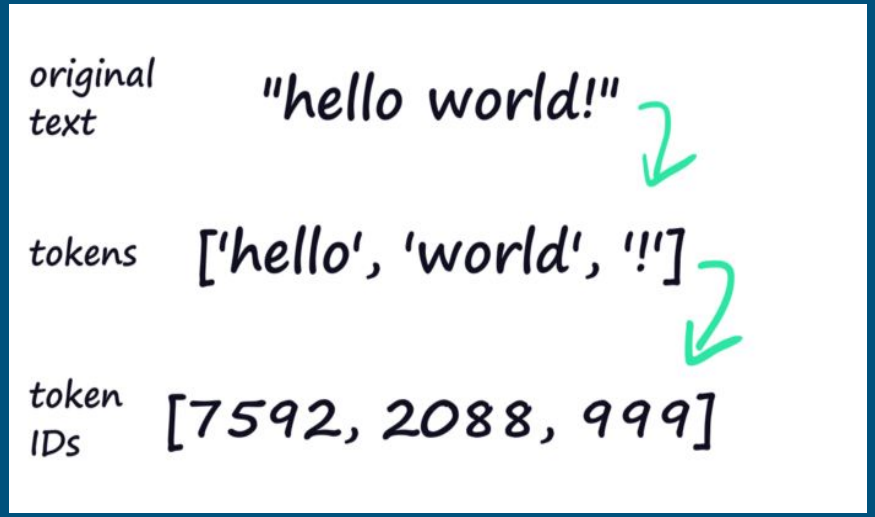
Rouge1 F1-Score: 12.38

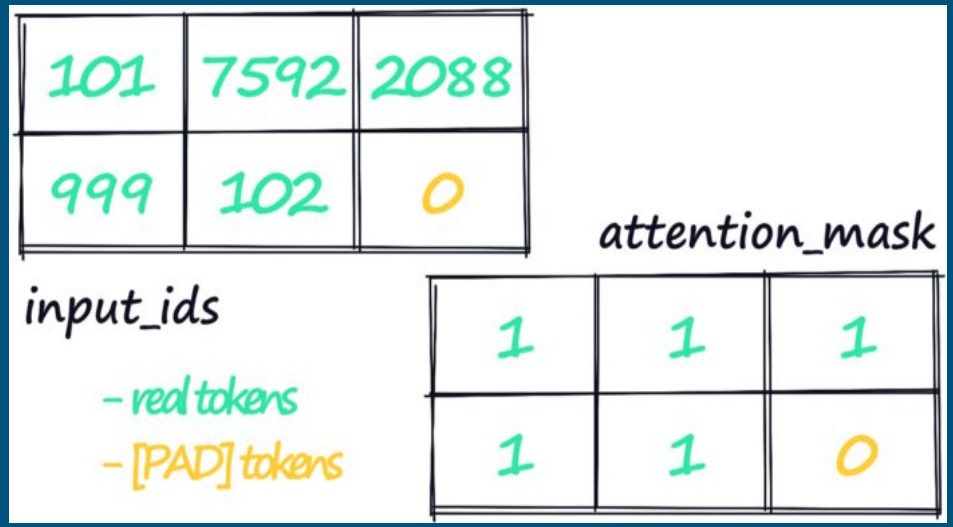
Rouge2 F1-Score: 1.33

RougeL F1-Score: 7.33

Data Encoding with Tokenizers

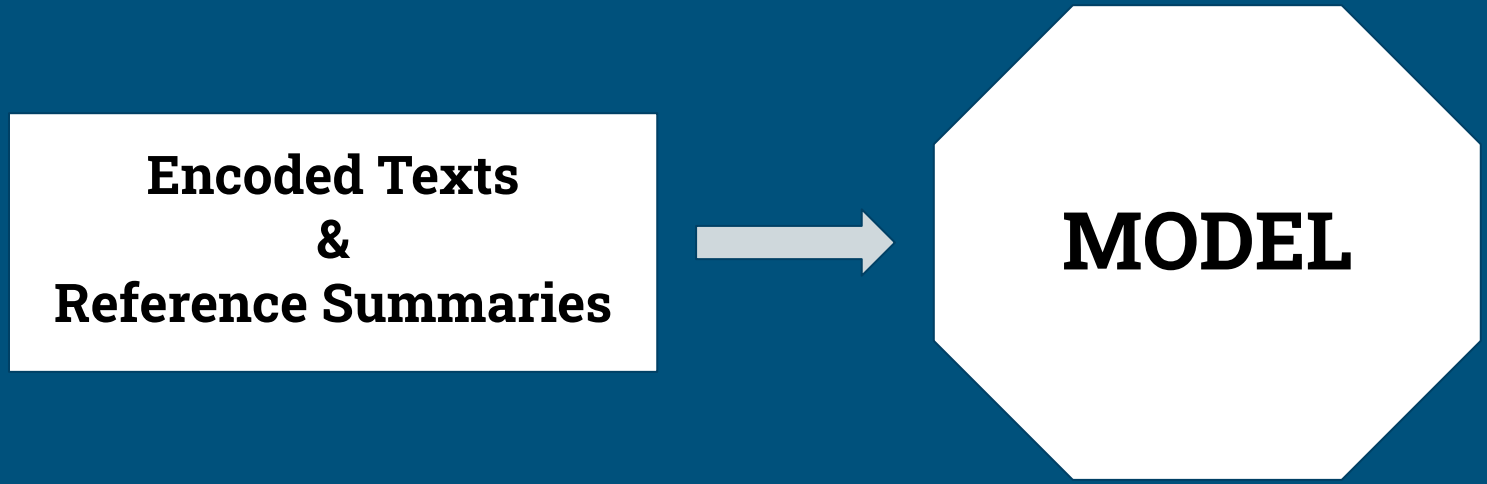
```
model_checkpoint = "google/mt5-small"
tokenizer = AutoTokenizer.from_pretrained(model_checkpoint)
```





Fine Tune & Train Model

- model = TFAutoModelForSeq2SeqLM.from_pretrained(model_checkpoint)
- Tune model hyperparameters
- model.fit(tf_train, validation_data=tf_validation, epochs=8)



Test & Evaluate Model

Reference

A sheer mountainside leads down to the Seventh Circle; they can only get down it where a landslide has sha

Model-generated

the mountains of the mountains of the mountains of the mountains of the

Reference

Next, the UM uses the analogy of the road to present his conclusion that man is afraid of completing his goal. Man

Model-generated

Gentlemen, I am joking, and I am joking

Rouge1: 8.04
Rouge2: 0.0
RougeL: 8.04

Repeat Process!

Fine Tuning, Fitting, & Testing

- Maximum input lengths
- Training time
- Learning rate
- Weight decay rate
- Model checkpoint
- etc..

Refined Model

Fagin decides that he wants to go back to his house, and he runs into a field in south London. He also tells him that he would be happy with him. The men they want to know what is going to be the next day, and Fagin decided that it is

In the first day of the year, the Carpanto shopkeeper in Winesburg, has become a tenant in the town, where he is married to a young woman. When she returns to his house, he runs a bar bar. After she was married, she

The Sherlock Holmes and Holmes are in hospital after being found guilty of the murder of Sherlock Holmes, who has been given a two-and-a-half years ago. Holmes returns to his room for a few minutes later, the Sherlock Holmes tries to get into the room. He says that

Rouge1: 24.11

Rouge2: 5.0

RougeL: 14.18

Directions for Improvement

- Continued fine tuning & training
- Research & address specific issues in model
- Further data cleaning
- More data!

Thank you!

Any Questions?