Summarization of GitHub Issues

Supervisor: Dr. Phuong T. Nguyen
Student: Moldir Koishybayeva

Problem

To employ different Natural Language Processing Techniques to generate a GitHub issue title

Table of contents

01 Dataset

iTape Project’s Dataset

Dataset: 922 730 issue samples, 1.56 GB

	number	title	body	repo
0	35557	Add test to disallow extra characters in Appli...	In the challenge [Applied Accessibility: Stand...	freeCodeCamp/freeCodeCamp
1	14968	Running pytorch 1.0.0 on aws lambda	## ❓ Questions and Help\r\n\r\nHello,\r\n\r\nI...	pytorch/pytorch
2	18448	Uncaught TypeError: Cannot read property 'endI...	[Enter steps to reproduce:]\r\n\r\n1. ...\r\n2...	atom/atom
3	37086	Ability to enable/disable replication controller	At present there isn't any way to temporarily ...	kubernetes/kubernetes
4	5531	[gatsby-plugin-sharp] Support Default Configur...	## Summary\r\n\r\nAt present there is no way t...	gatsbyjs/gatsby

02

Data Preprocessing

RegEx 

Issue Content column

- code snippets
- image link
- hyperlink
- url
- new line

Issue Title Column

- tags
- emphasis

Sample Issue

In the challenge **[Applied Accessibility: Standardize Times with the HTML5 datetime Attribute]**

(https://www.freecodecamp.rocks/learn/responsive-web-design/applied-accessibility/standardize-times-with-the-html5-datetime-attribute), if an extra character/characters is/are added in the code between the `</sup>` and `</time>` tags, the challenge still passes. The buggy code looks like this:

```html
<p>Thank you to everyone for responding to Master Camper Cat's survey. The best day to host the vaunted Mortal Kombat tournament is <time datetime="2016-09-15">Thursday, September 15<sup>th</sup>randomweirdcharacters</time>. May the best ninja win!</p>
```

Note the `randomweirdcharacters` part, we want a test to disallow it.

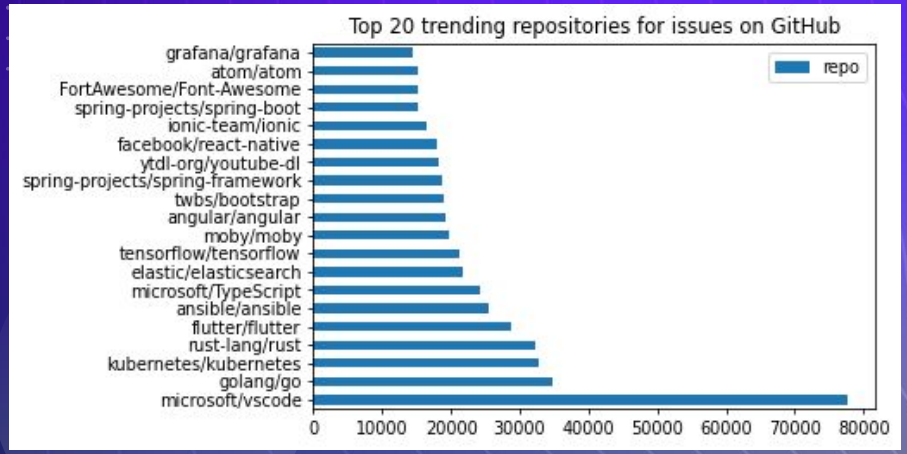
Refer to the discussion on PR #35553

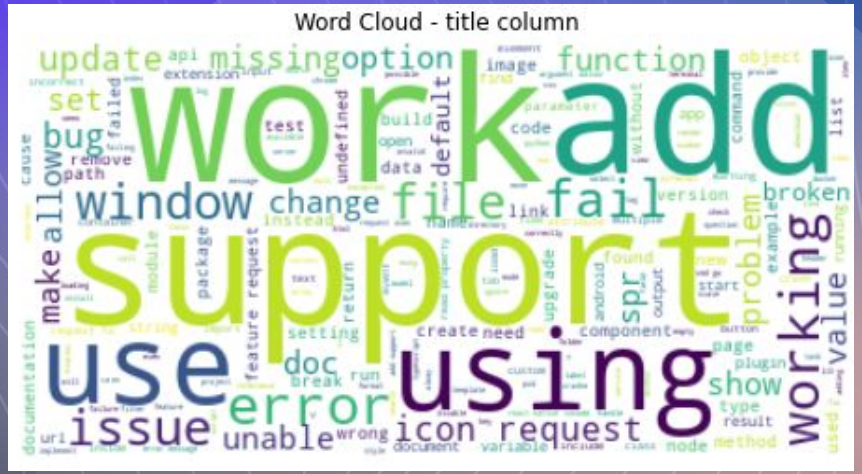
Preprocessed Sample Issue

in the challenge **link**, if an extra character/characters is/are added in the code between the `</sup>` and `</time>` tags, the challenge still passes. the buggy code looks like this: **code** note the `randomweirdcharacters` part, we want a test to disallow it. refer to the discussion on pr #35553

03 Exploratory Data Analysis

EDA Findings





04

Modelling 

Attention Is All You Need

Zero-Shot Learning

BERT

Bi-directional Encoder Representations from Transformer

`Bert-base-cased`

`BertForMaskedLM`

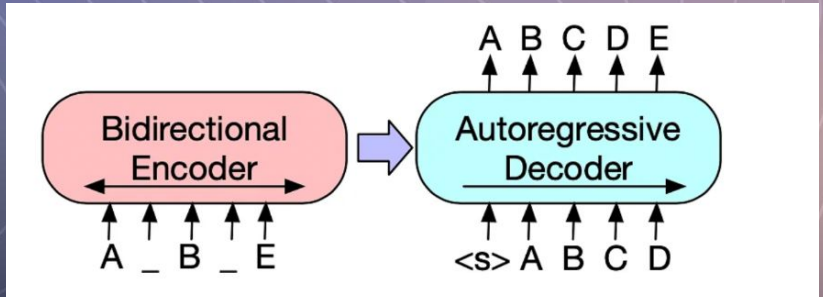
BART

Bidirectional and Auto-Regressive Transformer

- **modified BERT**
- **bi-directional encoder of BERT**
- **autoregressive decoder of GPT2**

`sshleifer/distilbart-cnn-6-6`

`BartForConditionalGeneration`



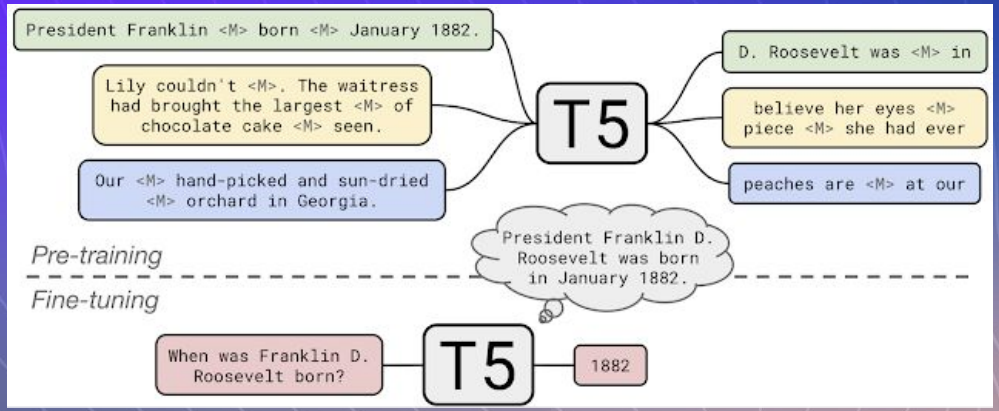
T5

Text-to-Text Transfer Transformer

`t5-small`

`T5ForConditionalGeneration`

`T5Tokenizer`



05 Evaluation

Zero-Shot Learning Results

	Precision Exp#1	Precision Exp#2
BART	41.67	18.18
BERT	35.71	11.27
T5	23.81	23.53

Comparing Different Summaries

Original Title

```
add test to disallow extra characters in applied
accessibility: standardize times with the html5
datetime attribute
```

Generated Sample Title

BART:

```
The buggy code looks like this: code note the
'randomweirdcharacters'
```

T5:

```
we want a test to disallow it. refer to the
discussion on pr #3
```

Disadvantages of BLEU:

- Doesn't consider meaning
- Doesn't map well to human judgements

Conclusion

- Fine tune models on iTape dataset => Amazon Sagemaker
- Different Evalution Method

Resources

- Songqiang Chen, Xiaoyuan Xie, Bangguo Yin, Yuanxiang Ji, Lin Chen, and Baowen Xu. 2020. Stay professional and efficient: Automatically generate titles for your bug reports. In 2020 35th IEEE/ACM International Conference on Automated Software Engineering (ASE). IEEE, 385–397.
- Zhang, T. *et al.* (2022) “ITiger: An automatic issue title generation tool,” *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering* [Preprint]. Available at: https://doi.org/10.1145/3540250.3558934.
- https://github.com/nlp-with-transformers/notebooks/blob/main/06_summarization.ipynb

Thanks :)

Do you have any questions?

CREDITS: This presentation template was created by Slidesgo, and includes icons by Flaticon and infographics & images by Freepik