Transcriptor Case Study

- language: python 3.9
- framework: apiflask

Open ai implementation

- I have tried different prompts to get correct answers in desired format, my last prompt is:

```
'content': '''Find if there is a homonymic word in sentence, if
     example: { "word": "high", "possible_word": "hi", "index"
   there can be no error in sentence, in that case return inde
```

- I limited answer with 150 tokens, since we do not need continuous streaming for these kind of short answer I declare stream as false
- This implementation is not restricted to the given similar words, since gpt infers the homophones easily.

Bert implementation

- Bert relies on masking the homophone word and returning the possible words instead of the word.
- I stored the homophones of each word in a dictionary for simplicity.
- If there is a homophone of the word in the returned possible words, I replace the MASK with the possible word.
- If there is no homophones in the returned value, I simply return the original sence.

Possible Improvements

- I split the sentence with space but it has bugs in upper case and commas. I have searched for a better tokenization method, Using tokenizer API from

huggingface can be enough

- This implementation does not cover multiple homophones error in bert implementation, I simply request from user to send the text until there is no modification left for the sentence.

Example request

```
curl -X POST \
  http://localhost:80 \
  -H 'Content-Type: application/json' \
  -d '{
    "text": "Stand by me.",
    "choice": "bert"
}'

returns:
{text: "Stand by me.", modified: "false"}
```

Deployment

- I have used AWS auto scaling group to handle high workloads and moreover spot instances besides ec2 to minimize cost.
- I used application load balancer associated with the global accelerator. Current static IPs:

```
http://15.197.134.218
http://35.71.171.16
```

Improvements on Deployment

- I did not focus on network security, and used default VPC server. VPC configured with public bastion host and private ec2 machines may be better in terms of security.
- I have not used IaC in the case study, It could be better to use github action to continuous development.
- Using EKS instead of autoscaling group can be more effective but it will propably cost more than autoscaler.