Lab2_problem2

October 23, 2024

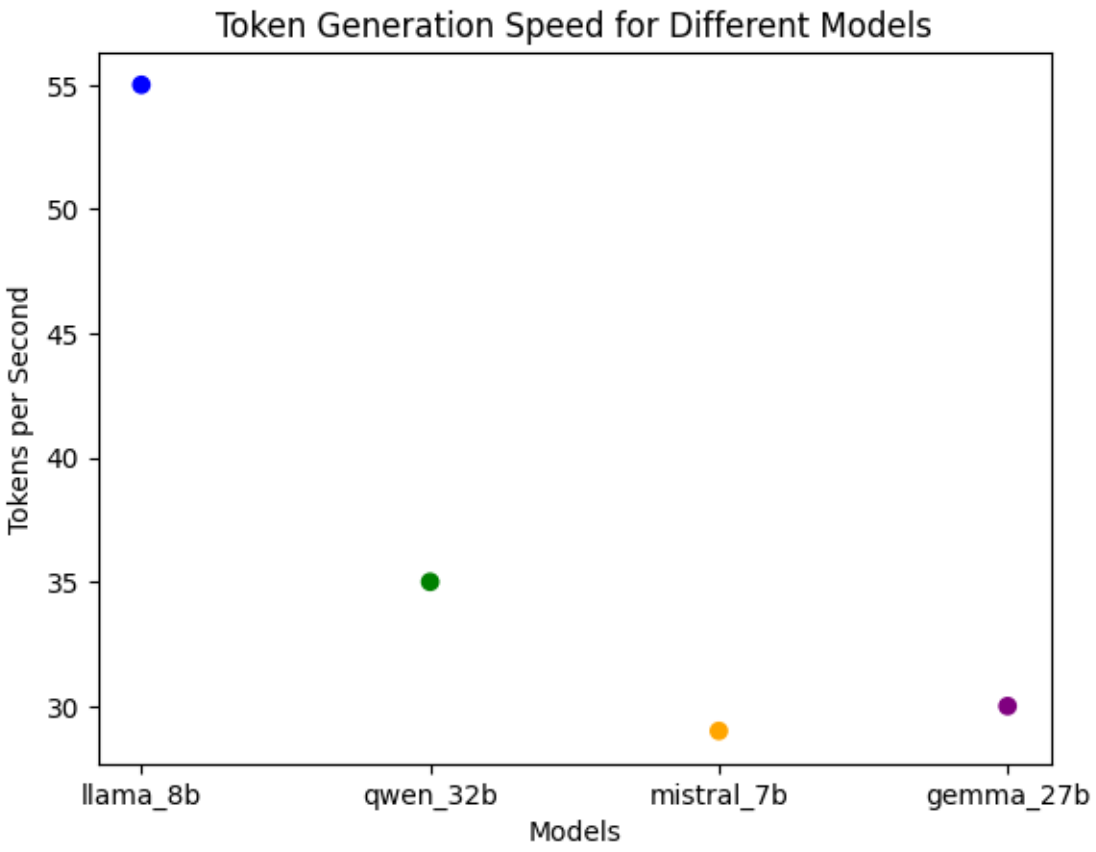
0.1 Problem 2

0.1.1 part A

[2]:
```
import matplotlib.pyplot as plt
```

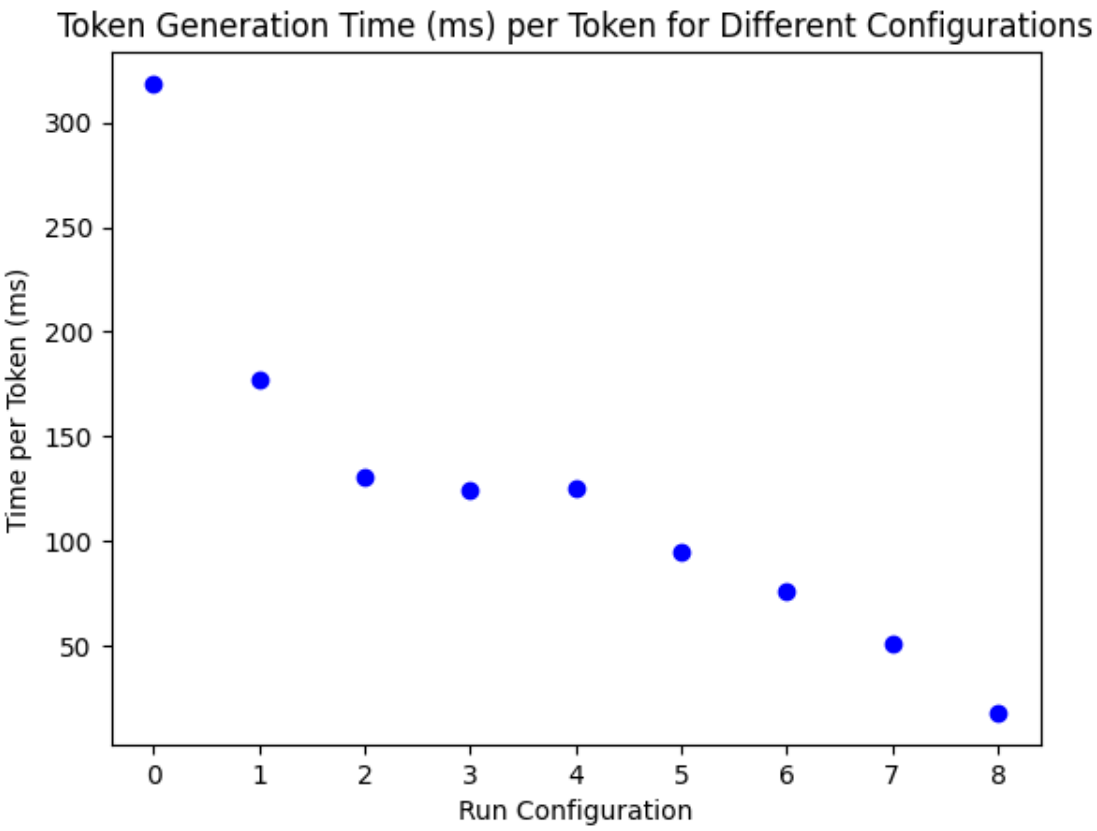
[16]:
```
import matplotlib.cm as cm
```

[23]:
```
models = ["llama_8b", "qwen_32b", "mistral_7b" ,"gemma_27b"]
token_speed = [55, 35, 29, 30],


colors = ["blue", "green", "orange", "purple"]

# for bar chart
#plt.bar(models, token_speed, color=colors)

#for scatter chart
plt.scatter(models, token_speed,color=colors)

plt.xlabel('Models')
plt.ylabel('Tokens per Second')
plt.title('Token Generation Speed for Different Models')
plt.show()
```



0.1.2 Part B

[26]:
```
runs = range(9)
tokens_per_second = [3.14, 5.64, 7.64, 8.02, 7.96, 10.56, 13.10, 19.40, 55.09]

time_per_token = [(1 / speed) * 1000 for speed in tokens_per_second]

runs = range(len(tokens_per_second))


plt.scatter(runs, time_per_token, color='blue')


plt.xlabel('Run Configuration')
plt.ylabel('Time per Token (ms)')
plt.title('Token Generation Time (ms) per Token for Different Configurations')

plt.show()
```


Token Generation Time (ms) per Token for Different Configurations
Time per Token (ms)
300
250
200
150
100
50
0
1
2
3
4
5
6
7
8
Run Configuration