# Week 6: Evals, HPC3 & AWS

Alfonso Vieyra, Denise Wei, Raha Ghazi, Tom Gao

Evaluations

- Began preliminary evaluation of models
  - Started with bart-large-cnn for summarization
- Poor scoring
- Used pipeline wrapper to load model
  - Maybe loading model directly and tweaking the prompting will yield better results?
- Took a little over 3 hours to generate summary for 2000 rows

...

BART ROUGE Scores

Density Plot for ROUGE-1

Density Plot for ROUGE-2
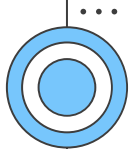
Density Plot for ROUGE-L
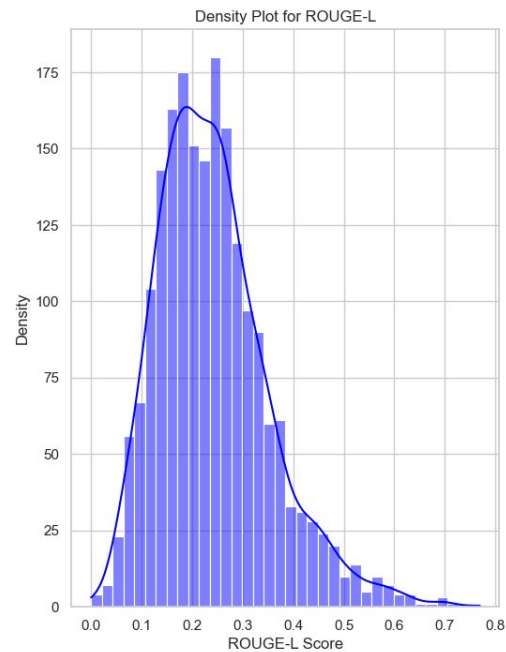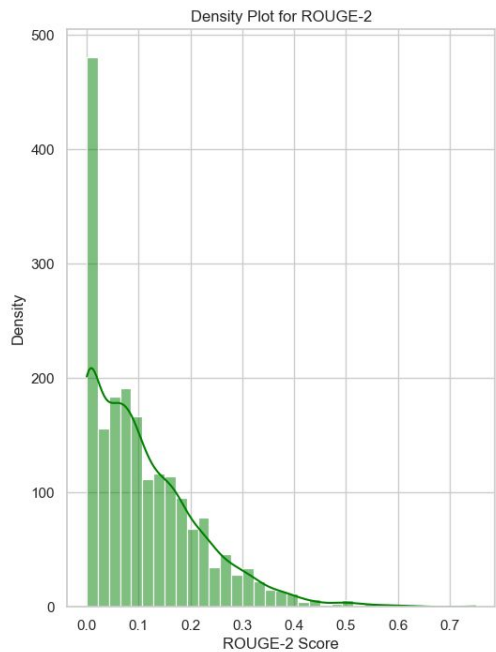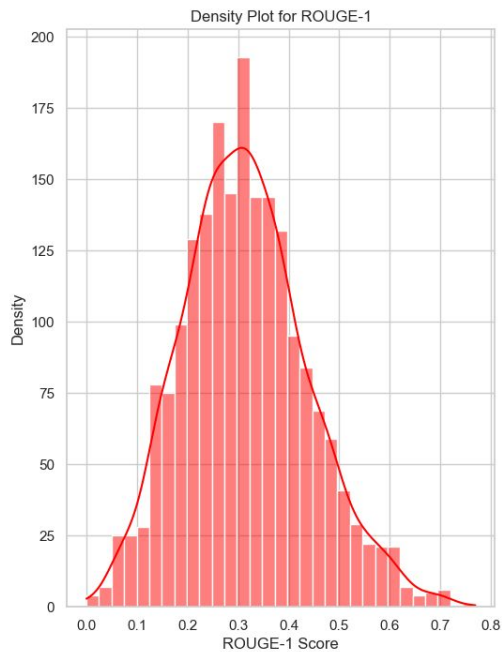
BART METEOR Scores

Density Plot for METEOR
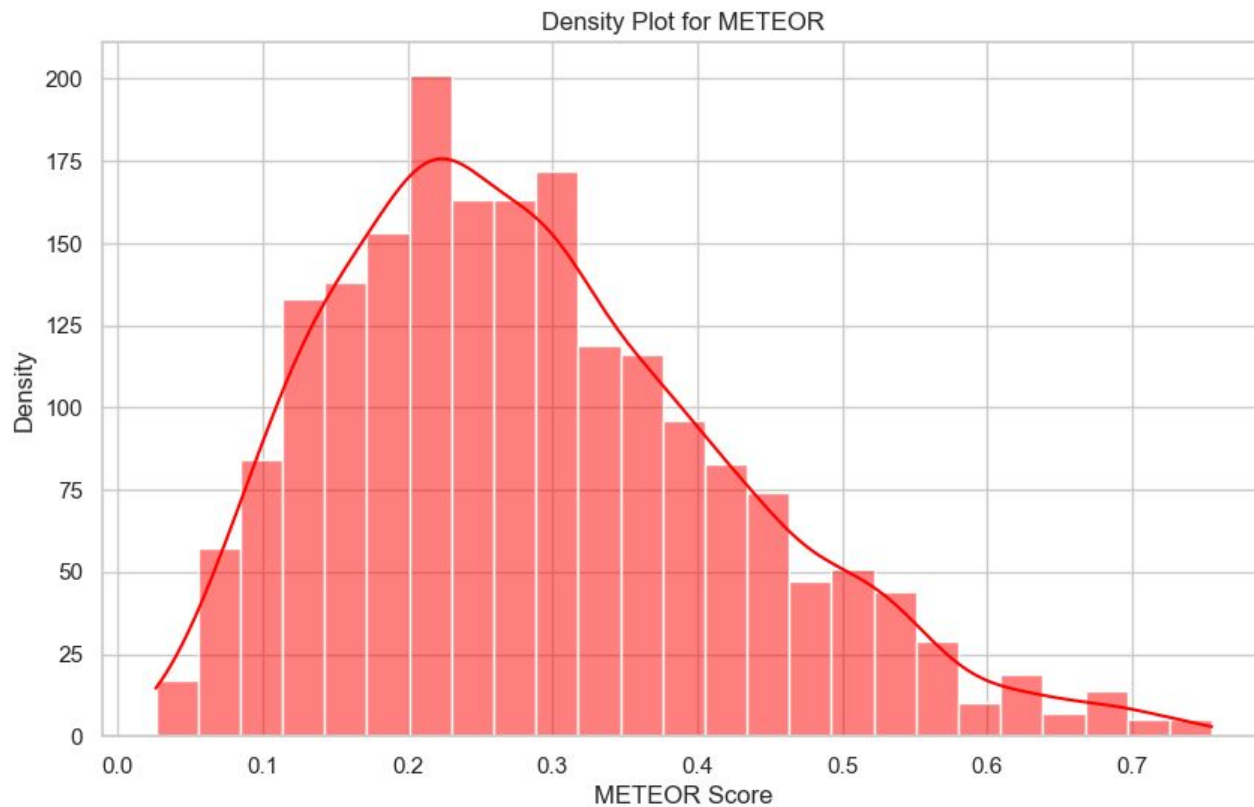
HPC3

- Set up environment has been temporarily resolved (Trying to resolve issue with PyTorch).

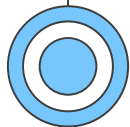- Encountering issues with resource allocation (i.e. A100 and CPU cores for running LLM's)

...

AWS Relational Database Service

- Leveraging the cloud to store ALL the data (Training, Validation, Test).

- Data is now remotely accessible from any device

- Changes to data can be enacted immediately for all users (i.e. no need to push, pull data files via github!)

...

Example: Structure

Dialogues

dialogue_id	dialogue_text	actual_summ...	actual_senti...	data_set
1	"#Person1#: Hi, Mr. Smith. I'm Doctor H...	"Mr. Smith's getting a ch...	"positive"	"training"

Models

model_id	model_name
1	Mystral

Summaries

summary_id	dialogue_id	model_id	generated_sum...	rouge_score	meteor_score
1	1	1	"Mr. Smith's getting a check-up..."	0.9	0.95

Sentiments

sentiment_id	dialogue_id	model_id	generated_sentim...
1	1	1	"positive"

Sentiment_evaluation

evaluation_id	model_id	roc_score
1	1	0.85