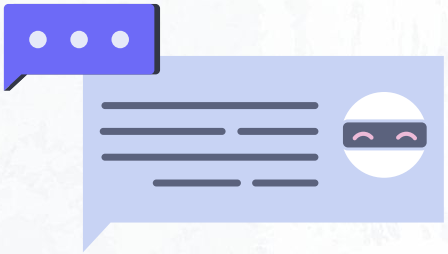
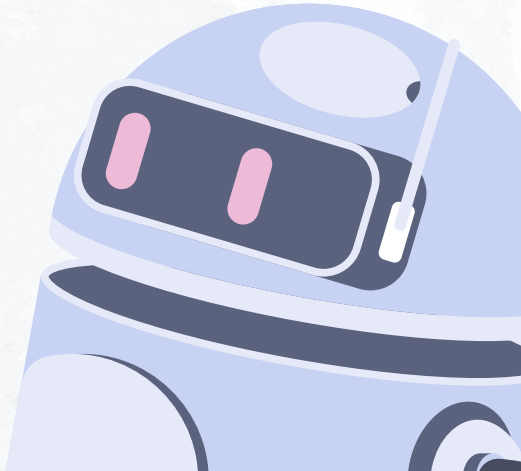
Generisanje Regular Show epizoda





Sadržaj

01 →

Opis problema

(AI)

Opis problema

Regular Show je emitovan u 8 sezona do 2017. godine. Cilj ovog projekta je da generiše nove epizode na osnovu ulaznog prompt-a.

Korišćen je openAI transformer model (GPT-2) fine-tune-ovan na postojećim epizodama Regular Show-a.

02 →

Skup podataka

(AI)

Skup podataka

Skup podataka je scrape-ovan sa Regular Show Wiki korišćenjem biblioteke urllib i regex-a. Podaci su procesirani korišćenjem regex-a tako da su uklonjeni html elementi i transkripti epizoda su snimljeni u tekstualne datoteke.

Skup za treniranje sadrži 210 (80%) epizoda, a skup za evaluaciju 52 epizode (20%). Prilikom učitavanja sa krajeva transkripata se uklanjaju prazni znakovi i preskaču se transkripti koji nemaju sadržaj.

03 →

Fine-tuning modela

(AI)

Fine-tuning modela

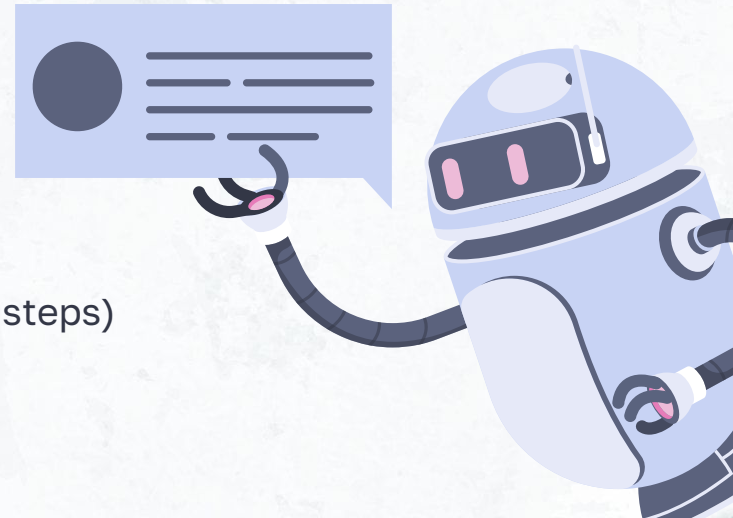
Korišćen je pretrained GPT-2 model koji je pomoću torch i huggingface biblioteka treniran sa raznim parametrima. Za trening je korišćena klasa Trainer iz modula transformers u kombinaciji sa TrainingArguments klasom.

Za većinu parametara su korišćene podrazumevane vrednosti:

- Learning rate = 5e-5
- Scheduling type = linear
- Weight decay = 0

Kombinacije parametara treniranih modela:

- Epochs = 3, Batch size = 4
- Epochs = 3, Batch size = 2 (8 gradient accumulation steps)
- Epochs = 4, Batch size = 4



04 →

Evaluacija modela

(AI)

Evaluacija modela

Modeli su evaluirani uz pomoć Trainer i TrainerArguments klasa slično procesu fine-tuninga. Za batch size je uzeta vrednost 2, dok su ostali argumenti ostali na podrazumevanim vrednostima.

Za metriku je korišćen ROUGE set, specifično: rouge 1, rouge 2 i rouge l. Set podataka za evaluaciju sadrži 20% nasumičnih epizoda iz nasumičnih sezona. Rezultati evaluacija modela ispunjavaju predviđanja, i pretpostavku da na kvalitet fine-tuninga utiče ne samo broj epoha nego i batch size kao i gradient accumulation steps.

Kod modela čiji trening je izveden u 4 epohe se može primetiti nedostatak kreativnosti (moguće zbog overfitting-a). Model koji je treniran sa povećanim korakom akumulacije gradijenta je postao nepredvidljiv i slabo povezan sa tematikom serije. Trenutno najbolja verzija modela je nastala treningom u 3 epohe i sa batch size 4, ali je za pronalazak zaista optimalnih vrednosti potrebno još istraživanja.

Evaluacija modela

(E - epochs, B - batch size)

Model	*Evaluation loss*	*Rouge 1*	*Rouge 2*	*Rouge L*
GPT-2	2.74742	0.53905	0.14282	0.32777
E4 B4	2.54160	0.56071	0.15077	0.34351
E3 B4	2.54791	0.55847	0.14941	0.34162
E3 B2	2.58894	0.55627	0.14641	0.33779

Hvala na pažnji!

SV31-2021 Gašić Dimitrije

SV54-2021 Ivanov Maša