Neural Modeling and Probabilistic Sequencing in Movie Script Generation

Ryan Krawczyk

Natural Language Processing · Boston College

Project Overview

Characteristics

- ❑ Model-view-controller web framework
- ❑ Regex search, Markov chains, RNN text generation

Input	Output
genre, title, author, character names, start sentence	100 lines of movie script

Inspiration

- ❑ Accomplishment in artificial intelligence field
- ❑ Capability to engage in human activity, create "art"

Relevance

- ❑ Education on computer science, form of entertainment

Objective

- ❑ Generate text distinguishable as movie script, containing readable English and logical flow

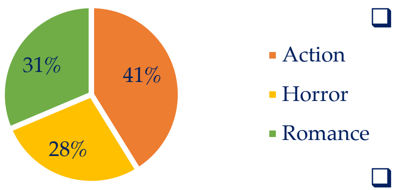
Processing & Training Data

U.S. Social Security Database

- ❑ Lexicon of 94,000 most common first names
- ❑ Sentences with names ignored (character conflict)

IMDb

- ❑ 153 raw text movie scripts from Datasets API
- ❑ Movie scripts manually annotated as action, romance, or horror

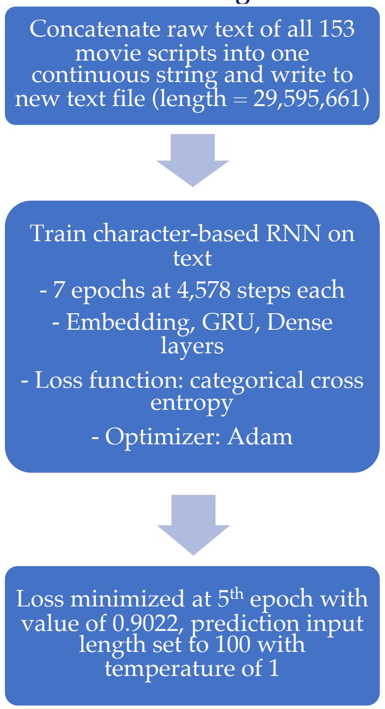


Normalization

- ❑ Sentences with numbers, URLs, or named entities excluded from generation to maintain informational consistency in scripts
- ❑ No random names, dates, etc.

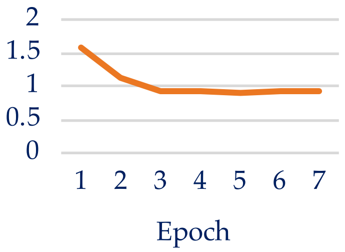
Core Logic

Training

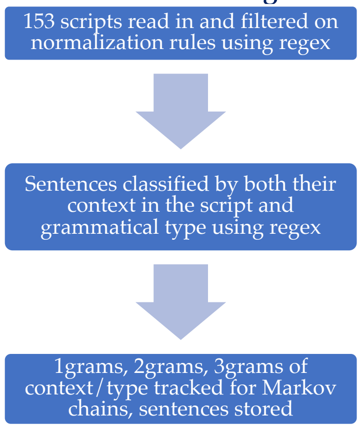

Concatenate raw text of all 153 movie scripts into one continuous string and write to new text file (length = 29,595,661)

Train character-based RNN on text
- 7 epochs at 4,578 steps each
- Embedding, GRU, Dense layers
- Loss function: categorical cross entropy
- Optimizer: Adam

Loss minimized at 5th epoch with value of 0.9022, prediction input length set to 100 with temperature of 1



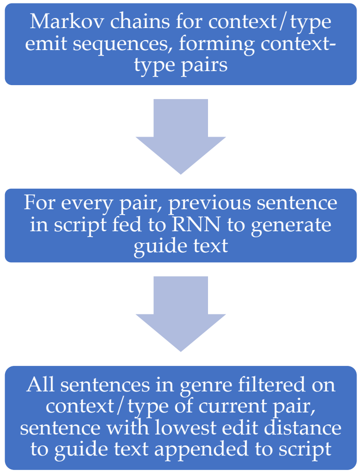
Text Processing


153 scripts read in and filtered on normalization rules using regex

Sentences classified by both their context in the script and grammatical type using regex

1grams, 2grams, 3grams of context/type tracked for Markov chains, sentences stored

Text Generation


Markov chains for context/type emit sequences, forming context-type pairs

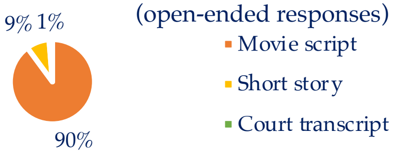
For every pair, previous sentence in script fed to RNN to generate guide text

All sentences in genre filtered on context/type of current pair, sentence with lowest edit distance to guide text appended to script

Text Intrusion Results

Question 1: What is this text?

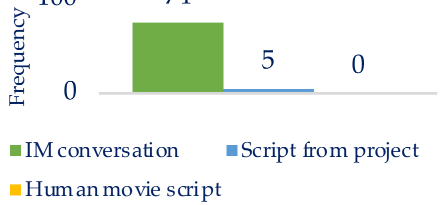
(open-ended responses)



Structure and content of project output recognizable by human eye as movie script

Question 2: Which of these is not a movie script?

(select 1 of 3 options)



Quality of generated text on par with human script and has characteristics distinguishable from other form of dialogue

Sample Script

1

RESUME

It's a human toe. Frightened, he's wielding a SCISSORS, poised to strike.

2

JEFF is tattered and smeared with vermillion ink.

THE DOCTOR:
"Single someone out"

3

MEGAN:
"Okay?"

JEFF sits at his desk, dark circles beneath his eyes.

4

VICTOR:
"Stop that!"

Trying to follow, he COLLIDES With some furniture.

Conclusion & Future Work

Successful Experiment

- ❑ Test subjects able to reliably identify output as movie script and associate it with human scripts

Looking Ahead

- ❑ More robust software for differentiating dialogue from narration and tagging named entities
- ❑ Generation with sequence-to-sequence modeling
- ❑ Deep learning structures to develop 5-stage plot

Libraries implemented: en_core_web_sm | nltk | numpy | pandas | spacy | setuptools | tensorflow

Github repository: https://github.com/ryankrawz/pocket-movie