Categorizing Search Strings via Deep Learning

ClassiPy

Presenter: Samuel Fu, Jack Shi, Justin Won

Overall Goal

- [x] Create general search model
- [x] Train with Deep Learning Neural Network
- [x] Categorize by key word descriptions
- [x] Return with high accuracy (>95%)

Real Life Applications

- Search Engine
- Online Shopping Assistance
- Social Media Interpreter

Current Capabilities

HCL: Search String Categorization

- Categorized over 11,000 search strings into 63 categories
- Trained deep learning neural network model
- Built front-end and back-end framework
- Predicts categories using real-time trained model
- Returns results with extremely high accuracy and precision (**>87%**)
- Reduced standard deviation and variance in model

Demonstration

ClassiPy-lookup

lookup: Submit

Search String: Katniss ignores Haymitch's advice and grabs some of the supplies placed around the Cornucopia, a structure at the starting point, and narrowly escapes death. Twelve tributes die in the initial melee, and only eleven, including the Careers, survive the first day
Search Result: DVDs and Movies
Training Accuracy: 94%
Testing Accuracy: 85%

GO BACK

OUR PROCESS

How we approached this problem

Step 1:
Clean Data

- Stripped all non-alphabetic characters
- Set to lower case
- Identified and standardize incorrect category IDs
- Removed low frequency key terms

OUR PROCESS

How we approached this problem

Step 2:
Machine Learning
Natural Language Processing

- 65% training data
- 15% validation data
- 20% test data
- One-hot encoding -> Bag of Words method
- Transformed data using vectorizer

OUR PROCESS

How we approached this problem

Step 3:
Neural Network

- Chose between single-layer, multi-layer, Convolutional Neural Network, Unigram SVM, Recurrent Neural Network
- 4 layers
- Dense layer
- Dropout layer
- Dropout layer
- Output layer
- Dropout is used to prevent overfitting

OUR PROCESS

How we approached this problem

Step 4:
Loss Function

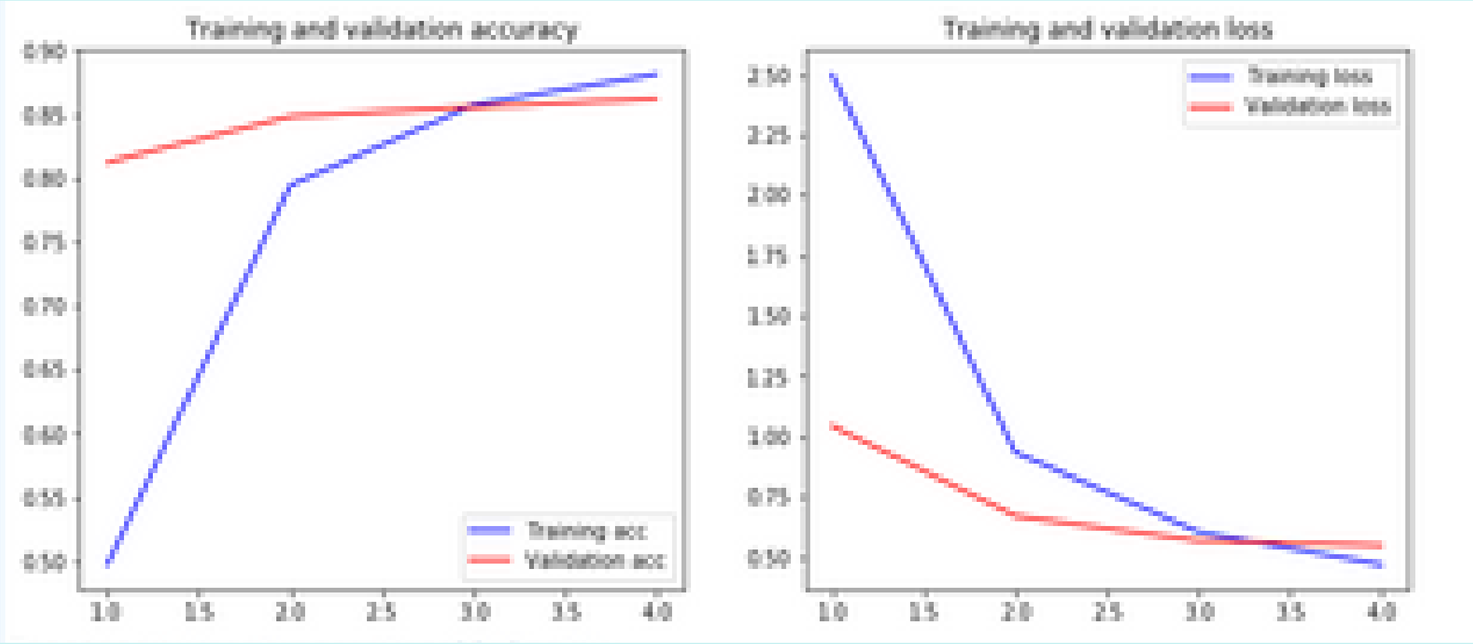
- Sparse Categorical Cross Entropy
- A search string can be only one category
- No regression among the categories
- Minimizes total cost associated with the weights

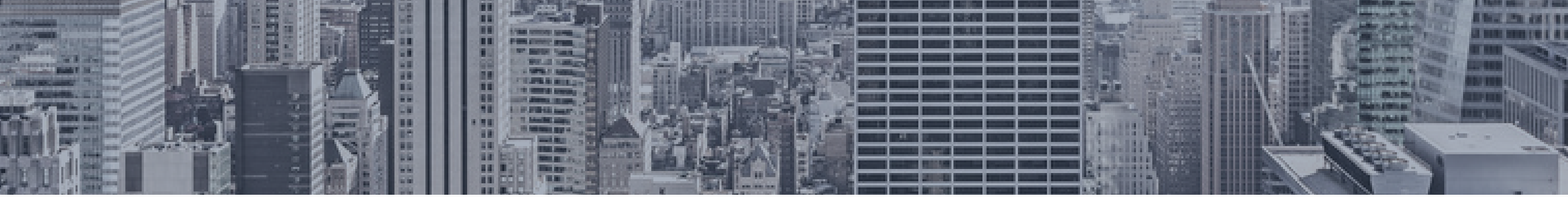
OUR PROCESS

How we approached this problem

Step 5:
Evaluate Model

- **Accuracy of Model:**
- **Variance:**
- **K-Fold Cross-Validation: Training 96%, Testing 85%)**
- **Graph here**





Thank You