**Image Detection – Simpson's Characters**

**Assignment 4 – Data Sets - Sprint 2**

# Data Acquisition

## Overview:

This dataset is an image dataset about the Simpson’s characters from Kaggle [1], it has 53.7k views and 7274 downloads. The Simpson image dataset has 20 folders for each character and 400 to 2000 pictures in each folder. Images are frames from Simpson’s cartoon series.

## Field Descriptions:

There are three files in Simpson’s characters dataset. They are:

File simpson-set.tar.gz: This is an image dataset: 20 folders (one for each character) with 400-2000 pictures in each folder.

File simpson-test-set.zip.: Preview of the image dataset

File annotation.txt: Annotation file for bounding boxes for each character

Below is some additional information:

URL: <https://www.kaggle.com/alexattia/the-simpsons-characters-dataset>

Title: The Simpsons Characters Data

Authors: alexiattia

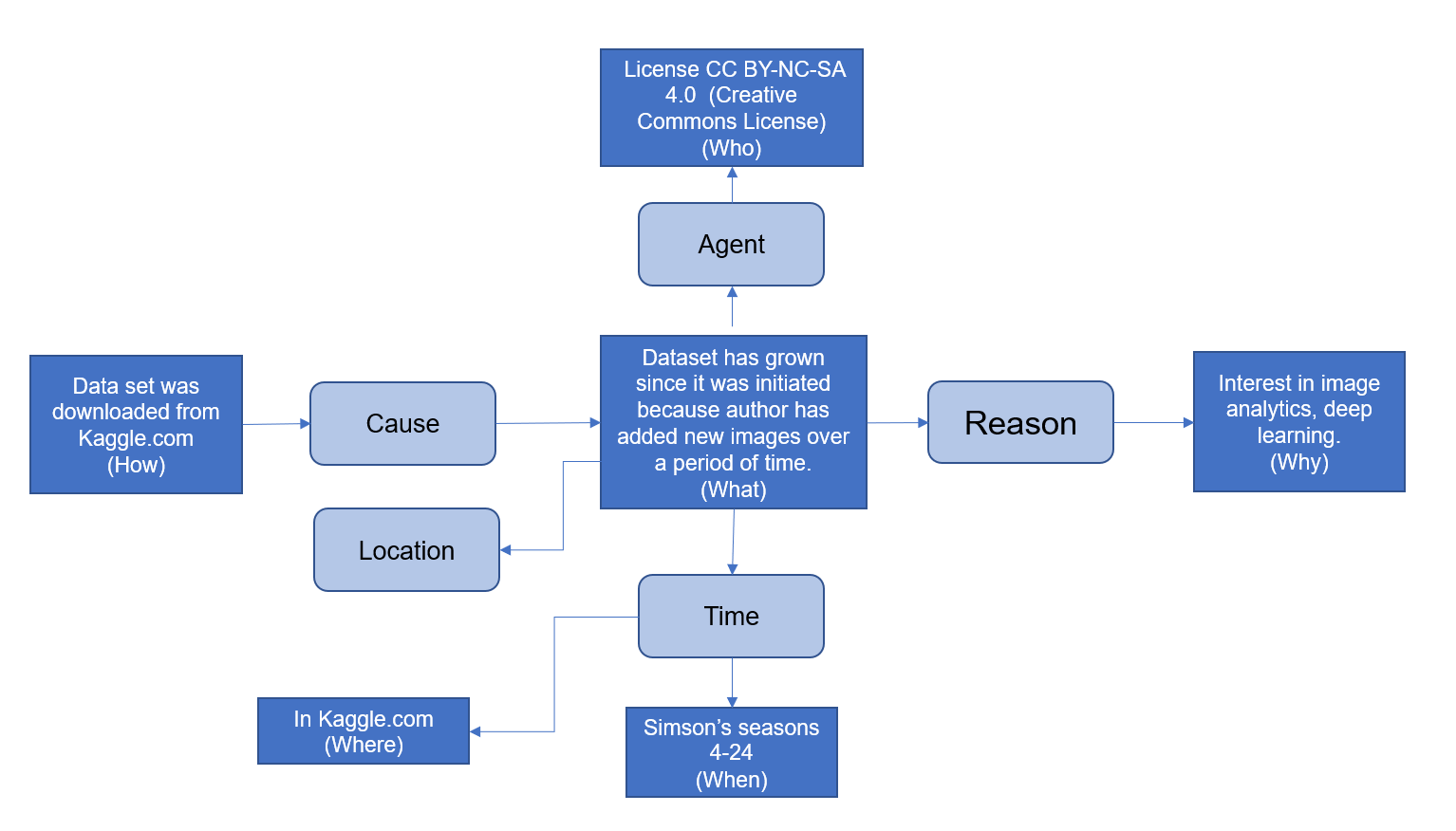
Publication Date: 2017-06-15

Text: image data

Tags: Deep Learning, Image data

## Data Context:

The dataset is downloaded from Kaggle, which is an image dataset about the Simpsons 4 to 24 seasons. This data set is a public legal data set with its own file license number "CC BY-NC-SA 4.0". Based on the legitimacy and openness of this data set, we chose it. However, the most important reason is that our team is interested in deep learning and image analysis.



## Data Conditioning

For existing data provided in Kaggle:

Use existing functions in Keras, PyTorch libraries to read images and preprocess them for classification and detection.

For extra/new data (if required):

We can label the data from the movie and we can get the cropped pictures. And then we wrote classify for each image though algorithm. When preprocessing those images, the first step is to adjust the sample size. using the function to adjust all images into the same size. The labels for the various characters are then converted from names to Numbers. Furthermore, the data set is divided into training set and test set.

## Data Quality Assessment:

* Completeness:
  + Characters being classified are when subject image count is greater than 100.
* Uniqueness:
  + Image set consists of mixed images (as depicted in previous slides). Total count of images makes it impossible to visually identify uniqueness.
* Accuracy:
  + Not determinable at this time. This can be determined after algorithms have been executed.
* Atomicity:
  + Training set contains images that are classified and grouped by character.
  + Image level annotation to box characters is also provided.
* Conformity:
  + The dataset provided conforms to one required to run image classification models.
* Overall Quality:
  + Good

## Other Data Sources

No other data sources considered at this time. As we start working on our stretch goals, this may be required.

# References

|  |  |
| --- | --- |
| [1] | Kaggle.com, "The Simpsons Characters Data," kaggle, 15 June 2017. [Online]. Available: <https://www.kaggle.com/alexattia/the-simpsons-characters-dataset>. [Accessed 8 June 2019]. |