TARTU ÜLIKOOL

Aksel Martin Muru

Rasmus Lille

E5: QUICKDRAW-ANALYSIS

CRISP-DM Start and Work Plan

Introduction to Data Science (LTAT.02.002) Group Project

Instructors:

Markus Haug

Victor Cabral Pinheiro

Tartu 2024

BUSINESS UNDERSTANDING

Identifying the Business Goals

Background

Data science project for students of University of Tartu to learn how to and deepen their understanding of working on a data science project. This subject of doodle analysis is a useful subject to dedicate oneself to, because doodles are a common mode of human expression, through which one could predict things about the drawer.

Business goals

Our main goal (1) is to create a solution for predicting objects from input drawings. In case of this goal's failure or unexpectedly fast success, pivot to predicting the subject's nationality through the doodle (2). If the next goal succeeds immediately or is deemed α failure, pivot the solution to displaying an average doodle representation of an object (3).

Business success criteria

The solution should be able to predict from input doodles what the doodles depict with a 75% accuracy, and/or predict the nationality of the input doodle's author with a 75% accuracy, and/or display an average representation of a doodled object.

Assessing the Situation

Inventory of resources

The project is planned taking into account the availability of the following resources.

- Available data
    - Google's Quick! Draw dataset.
- Available personnel:
    - two data science students
    - instructors who can help if problems arise.
- Available hardware
    - ~2 modern laptops.

- Available software
    - Python with data science libraries
    - Git (our repository https://github.com/Virteso/IDS-Quickdraw-Analysis)

Requirements, assumptions, and constraints

The project is under a time limit.

- Time limit: up until December 9th.
- Legal and security obligations: none, as the dataset is public.
- The work can be considered acceptable if it shows the students understanding of the learned techniques and materials and their capabilities of using them.

Risks and contingencies

We strive to account for issues that could arise and prepare solutions.

Risk	Contingency Plan
Computing takes more time than estimated for any one goal	Conclude goal failure, continue with next
Required data amount for reaching a goal is absurd	Conclude goal failure, continue with next

Terminology

It is useful to define some terminology that this project uses either exclusively or with a specific meaning to the goals and tasks at hand.

- The dataset: Google's Quick! Draw dataset

Costs and benefits

This is largely irrelevant to the current project.

Costs: 0€ (each person is responsible for their own costs), but it still requires some time put into it.

Benefits: 0€ but a max 30 points for the data science course and the connections we made along the way.

Defining machine learning goals

Machine learning goals

According to the business goals:

- Train a model to:
    - predict objects from input doodles
    - and/or predict nationality, region or doodler from input doodles
    - interpolate the average doodle-depiction of an object

Machine learning success criteria

Each model is successful if it

1. meets the business goals
2. works with available computing power

DATA UNDERSTANDING

Gathering data

Outline data requirements

The project requires data for representing each drawing, and a classification on what the doodle represents. For later goals, nationality or region data classification of each doodle might be required. We plan to use preprocessed “simplified” data that represents drawings in a more uniform manner.

Verify data availability

The data is freely available through Google Cloud. The possibility of downloading it was verified through Node.js API and web browser by downloading a single simplified drawing representation. Verified the required data’s existence through inspecting one source file visually.

Define selection criteria

Data should be contained in a specific file format (either JSON or specialised binary representation). This project is interested mainly in the data fields representing the drawing data, the drawing's description, and later possibly the location of the drawing's author.

Data collection conclusion

It was verified that the simplified data is accessible through the Google Cloud API and through a usual web browser file download function. It was verified that the data contains required fields. Caution should be taken, however, with the large storage space (and thus also memory) requirements of the data.

Describing data

The data is spread out into different source file formats in the Google Cloud. Each data source file has fields for the following:

- the drawing itself, represented as a nested array of bytes.
- a word describing the drawing (what it represents)
- the timestamp of its creation
- its "key_id" (unknown how it is specified)
- country code of the author
- a field for whether the Google Quick! Draw machine learning model predicted its word when it was drawn.

The data, as one can confirm visually, contains fields required for this project.

Exploring data

The following was found out about the data through a closer inspection.

- each drawing is represented through an array of strokes
    - each stroke contains following arrays:
        - the x-coordinates of a stroke through time (bytes)
        - the y-coordinates of a stroke through time (bytes)
- the country code is represented with two letters, for example "US", "EE".
    - for the second goal, it should be kept in mind that "US" citizens make up a majority of the submitted doodlers.

- the name of each drawing can contain capitalisation and spaces
- there are about 100 to 200 thousand drawings for each category

Verifying data quality

The data is accessible. We found the following quality issues:

- Some drawings are incomprehensible scribbles. This can theoretically skew training. We should use each drawing's "recognised" field to see whether Google's neural network recognised it, and if so, there is more of a chance that the doodle is something drawn in earnest.
- Some datasets have written words on the doodles sometimes while the "recognised" flag is true, which can again skew the results, so if we detect that a dataset contains words instead of drawings often we should skip it.
- People from the US have submitted more drawings overall, which can affect specific goals.

PLANNING THE PROJECT

Tasks

Task 1

Write a program to process data into machine-learnable format, for example replacing categorical values with numerical and maybe converting strokes into pixel data - testing required on which is more useful. Python libraries learned throughout the course should suffice.

Time: 7h each member

Deadline: 02.12

Task 2

Test ML algorithms (random forest, neural nets, linear regressions) on processed data to see what results the class predictions give at this stage. Adjust processing and algorithms if results are subpar. Adjust algorithm to work with multiple classes.

Time: 10h each member

Deadline: 05.12

Task 3

Train more data and try to reach 75% input prediction correctness with multiple classes. Try to write functions to automate testing as much as possible.

Time: 4h Rasmus 2h Martin

Deadline: 06.12

Task 4

IF previous task succeeds, attempt implementing predicting other things from data, like country code - goal 2 - and if this succeeds, attempt goal 3 as well.

Otherwise, try harder to reach the first goal.

Time: 6h each regardless of success

Deadline: 07.12

Task 5

Create poster using previous data gathered and to explain our work so far.

Time: 3h Rasmus 5h Martin

Deadline: 09.12