wrangle_report

July 13, 2019

1 Wrangle Report

1.0.1 Table Of Content

- Introduction
- Gathering Data
- Assessing Data
- Cleaning Data

Introduction

Data wrangling is process of getting data ready for analysis and visualization. It is an iterative process with 3 stages. We can iterate through gathering,assessing and cleaning to get good quality data for analysis. So let's start with importing necessary libraries.

We are provided with three main sources to gather data, But i'll be using an external API to fetch some data for my wrangling.

1.0.2 Gathering Data

In this step i gathered data from three given sources and three different methods. First source provided was a direct download link for our main data Enhanced Twitter Archive. I downloaded it using browser.

Second source of data was Image Predictions. which we were told to download using python. We used python's `requests` package to fetch data from udacity's servers and then saved it locally using python's file I/O functions.

Third source of our data was in form an API. I used Twitter's developer API to retrieve full information on tweets usind ids we're provided. First i had to create account on twitter developers. Then i generated Access token from Dashboard and saved them locally to access keys in my project.

Fetching data from API required and programming interface. I'm using python so I installed `Tweepy`, A python package to play around with Twitter API. Then I wrote code to fetch the data and saved it into `tweet_json.txt` ato access it later.

I also used Dog API to fetch all of the dog breeds which can be later used in cleaning process.

1.0.3 Assessing Data

Assessing data means investigating the data and finding issues with data. Assessing can be done in two ways. Programatically and visually. Data issues can be devided into two types. One is Quality issues and other is Tidiness issues.

Assesing visually means opening dataset in some software and looking for problems in it. We can find issues like inaccurate data, incomplete data and tidiness issues visually data.

Assessing data programatically means using in built functions from python and finding issues. We can find issues like missing data and incorrect data types by assessing data programatically.

Issues in data are devided into quality issues and tidiness issues. Quality issues are related to quality of data like missing data, incorrect data types and incomplete data.

Tidiness issues are related with structure of data. You can read more about tidiness in here.

1.0.4 Cleaning

Cleaning data means fixing the issues we found during assessment. Cleaning data is mainly done programatically, but some data contains outliers which can not be fixed prgramatically so sometimes we need to fix them personally.

Cleaning is done in 3 steps. First we define our problems which means we take our issue and describe how we're going to fix that issue.

Second step is coding. We just code to fix the issue we described in our Define step. In this step i used python and pandas to fix data programatically. And third step is testing the solution. We check if our data issue is fixed or not.