Universität Passau
Fakultät für Informatik und Mathematik

CLASSIFICATION OF VISUALIZATIONS IN SCIENTIFIC LITERATURE

Masterarbeit zur Erlangung des akademischen Grades
Master of Science (M.Sc.)

am Lehrstuhl für ... oder an der Professur für ...
der Fakultät für Informatik und Mathematik
der Universität Passau

Name:	Arnold Azeem
Matrikelnummer:	79176
Fachbereich:	Informatik
Studiengang:	Master Informatik
Schwerpunkt:	TODO
Studienjahrgang:	TODO
Erstprüfer:	Prof. Dr. Christin Siefert
Zweitprüfer:	Prof. Dr. Michael Granitzer

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abstract

Einleitung

Motivation

Abbildung 1.1: Describe this picture.

CREATING PLOTS

The dataset consists of images of plots which are automatically created and labeled. Scripts are created in Python, Matlab,R language and Java. This is done to make our model very robust since plots used in scientific papers could be created in any of the above programming languages. The dummy data used in creating these plots are all in csv format. All the csv files are put into one folder and each is read at a time. For each csv file the script reads the first column and discards it if it contains any string values except in its header. Its saves this column into an array as the x-axis and saves the next column into an array as the y-axis if it does not also contain string values in the column. This two arrays are then used to create the plots, their headers serve as the label of their axis and the name of the file serve as the title. The previous y-axis becomes the new x-axis and the next column becomes the y-axis and the process continues till the columns in the csv file are all used. So depending on the number of numeric columns in the csv file many plots are created and labeled.

Scatter plot

For constructing the plots for the scatter plots I looked through a dataset from this paper[]. The dataset contained 555 scatter plots and looking through them I realized the markers used to plot where mainly zeros, triangles,squares,dots and x's. So instead of just plotting all the markers randomly i plotted those mentioned above with a higher probability than the rest of the markers. Also the scatter plots consisted of positive and negative correlation and just a few zero correlation. I also noticed that some markers were filled and other were not filled. So for creating my dataset I plotted with some markers being filled and other unfilled and this was done randomly.

Dataset for Scatter plot

For plotting my scatter plots I used raw data correlated into csv files. Next I will describe the data used and their sources. This group of data was gotten from [vincentarel].
The first is the **loti.csv** file which contains temperature anomalies for the years 1880 to 2010 and this data is from GISS(Goddard Institute for Space Studies) Land-Ocean Temperature Index(LOTI) data. It has data for months of the year and averages of some combined months. Second raw data used is **USjudgerAtings.csv** contains the ratings of 43 state judges in the US Superior Court by some lawyers. The rating were given based on attributes like Diligence,integrity,demeanor etc. Third data used is the famous Iris dataset by Edgar Anderson This dataset gives the measurements in centimeters of the variables sepal length and width and petal length and width for 50 flowers of 3 different species of iris; setosa,versicolor and virginica. The next data Longley's Economic Regression Data which was good because it gave very good positive correlation scatter plots even though it was not as huge as the other dataset used. It consisted of 7 economical variables observed yearly from 1947 to 1962. The next dataset is called the quakes, which has information of locations of Earthquakes off Fiji. It has 1000 observations with 5 variables(longitude,latitude, depth,Richter Magnitude and number of stations reporting). The nuclear Power station Construction Data was also used. It contains information about the construction of 32 light water reactor plants in

the U.S.A in the 90's.
The second group of data was taken from [Datplot website] a website for a software called DatPlot for plotting raw data.

Literaturverzeichnis

[BDE-1, o.J.] “The Big List of D3.js Examples“, http://christopheviau.com/d3list/, 20.01.2015

Erklärung zur Masterarbeit

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

Passau, den <date>

<First Name, Last Name>