2013年1月14日

進捗状況

前回ご指摘があった以下の課題をやりました．

・宿題

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | CommitComment | Create | Delete | Follow | ... |
| Aさん | 2 | 15 | 6 | 9 |  |
| Bさん |  |  |  |  |  |
| Cさん |  |  |  |  |  |

というような表を作って、

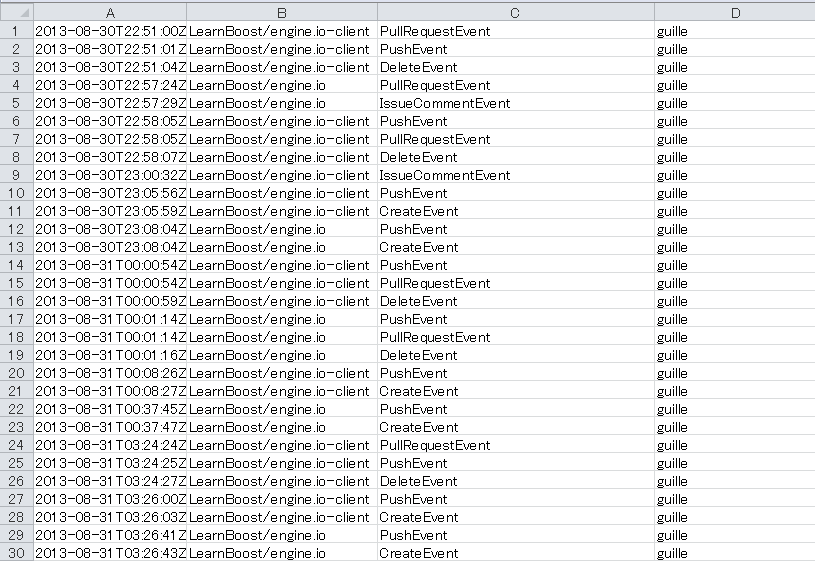
* 主成分分析
* 因子分析
* 自己組織化マップ
* クラスター分析

で分析し、結果を解釈してみる。

・卒業論文の執筆

・表を作る．

六人それぞれ全員の活動ログを一から調べ，その中でLearnBoostの結果のみを取り出しました．



図GuiileのAPIから直接得られた生のままのGithub上での個人のログ

そこからプロジェクトをLearnBoost関連に絞り，エクセルのピポットテーブルに表示しました．

|  |  |
| --- | --- |
| guille | (すべて) |
|  |  |
| **行ラベル** | **データの個数 / LearnBoost** |
| **CommitCommentEvent** | **2** |
| **CreateEvent** | **12** |
| **DeleteEvent** | **5** |
| **IssueCommentEvent** | **97** |
| **IssuesEvent** | **17** |
| **PullRequestEvent** | **48** |
| **PullRequestReviewCommentEvent** | **5** |
| **PushEvent** | **39** |
| **総計** | **225** |

図APIで得られたデータからLearnBoostのみを抽出したデータをピポットテーブルで表示

この作業をメンバ六人全員分やり，以下の表のように表してみました．

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | guille | retrofox | TooTallNate | Matthew Mueller | stambizzle | visionmedia |
| CommitCommentEvent | 2 | 0 | 0 | 0 | 0 | 0 |
| CreateEvent | 12 | 6 | 0 | 0 | 0 | 0 |
| DeleteEvent | 5 | 2 | 0 | 0 | 0 | 0 |
| IssueCommentEvent | 97 | 0 | 2 | 0 | 0 | 0 |
| IssuesEvent | 17 | 0 | 1 | 0 | 0 | 0 |
| PullRequestEvent | 48 | 5 | 1 | 0 | 0 | 0 |
| PullRequestReviewCommentEvent | 5 | 0 | 1 | 0 | 0 | 0 |
| PushEvent | 39 | 10 | 1 | 0 | 0 | 0 |
| WatchEvent | 0 | 0 | 0 | 0 | 0 | 0 |
| ForkEvent | 0 | 0 | 0 | 0 | 0 | 0 |
| GollumEvent | 0 | 0 | 0 | 0 | 0 | 0 |

「収集されたデータ」に保存してあります．

上記のデータを使い各解析を行おうとしましたが，Ｒのデータの読み込みがどうしてもできなかったです．

色々自分なりに調べ，エクセルからCSV形式にしたのですが，

> x <- read.table("C:/R/data.csv")

以下にエラー file(file, "rt") : コネクションを開くことができません

追加情報: 警告メッセージ:

In file(file, "rt") :

ファイル 'C:/R/data.txt' を開くことができません: No such file or dire

と出てしまいます．作業ディレクトリは指定したのですが，どうもRは読み込んでくれなくて困っています．．

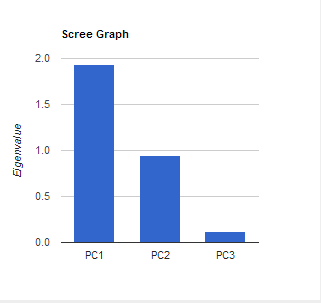
データの収集は上記のように終わったのですが，解析はまだ手詰まりな感じです・・・

なにもせずに終わりたくはなかったので，Rを使わず，フリーソフトなどを使い解析してみました．

・主成分分析

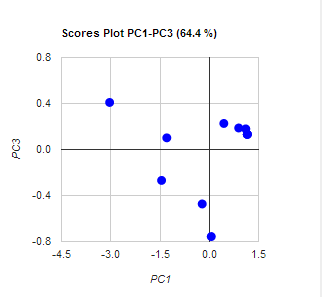
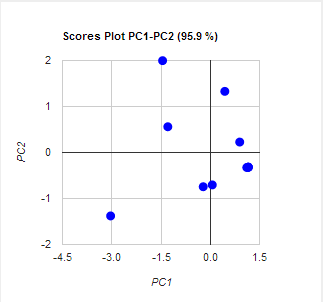
簡易的に主成分分析を行うことが出来るツールEasy PCAを使用しました．

このツールを使い，先に収集して出たデータを読み込み，グラフ化しました．



図の棒グラフは Scree Graph といって、各主成分が持つ情報量の大きさを示していて，

主成分は，情報量の多い方から第一主成分(PC1)，第二主成分(PC2)，第三主成分(PC3)…となります．Scree Graph から、第何主成分までが情報量を多く持つかを確認．



このようにそれぞれ、第一主成分vs第二主成分、第一主成分vs第三主成分、第二主成分vs第三主成分をプロットしたもので，これらの図を見てクラスターが形成されているかどうかが分かる．

・卒業論文の執筆

今まで進捗に挙げてきた事柄を論文に反映させている最中です．

・次の講義，異文化理解は期末テストでして，お昼は欠席します．最後のテスト頑張ってきます．

Rでの主成分分析

Excelでデータをつくり（1行に1データ）、c:/cit/sample.csvに保存する。

setwd("c:/cit")

myData <- read.csv("sample.csv")

head(myData)

#biplotの結果を見やすくするために、行に名前（店名）を付けておく

row.names(myData) <- myData$name

myData

#主成分分析（全体を括弧で囲むと結果が表示される）

myResult <-prcomp(myData[,-1]) #1列目（name）は不要

#結果の概要

summary(myResult)

#主成分（固有ベクトル）

myResult$rotation

#主成分の図示

barplot(sort(myResult$rotation[,1]))

barplot(sort(myResult$rotation[,2]))

#主成分スコア

myResult$x

#主成分スコアの図示

biplot(myResult)

宿題

清水・久保のプロジェクトで（25個）、3つのグラフを描いてみる。

> barplot(sort(myResult$rotation[,1]))

> barplot(sort(myResult$rotation[,2]))

> biplot(myResult)

主成分分析について数ページ書いておく。（豊田秀樹『データマイニング入門』を参考に）