

Projeto TCC: pré processamento, execução e pós processamento da aplicação do aprendizado de máquina para detecção de anomalias nos dados referentes aos gastos das prefeituras.

A estrutura geral do projeto é bem simples, como mostra a Figura 1, ela é formada por três diretórios principais: “dados”, “Preprocessamento” e “tools”. A pasta “dados” contém todos os *scripts* (arquivos .bat), cópias dos arquivos bytecode java (arquivos .class), dados de entradas (inputs), dados de saídas (outputs gerados), entre outros arquivos. O diretório “Preprocessamento” contém todos os *sources* do código escrito em Java. Isto é, trata-se da pasta do projeto java criado na IDE Eclipse. Já a pasta “tools” contém a ferramenta Elki (arquivo Jar).

Figura 1 - Estrutura de diretórios do projeto



Name	Date modified	Type	Size
dados	11/12/2017 11:41	File folder	
Preprocessamento	05/03/2017 11:23	File folder	
tools	17/04/2017 22:34	File folder	
Readme	04/03/2017 15:41	Google document	1 KB

Logo, na necessidade de alteração dos códigos fontes, a pasta “Preprocessamento” deve ser levada em conta. Entretanto, para execução dos scripts e análise dos dados, o diretório “dados” que deve ser utilizado. Na Figura 2, é possível ver todos subdiretórios e arquivos contidos nesta pasta. A pasta “helper” e “model” contém apenas arquivos .class (bytecode java). A pasta “INPUT” contém os arquivos de entrada de dados. Em “originals” estão os arquivos de dados originais baixados. Os diretórios “OUTPUT” e “OUTPUT_ELKI” são onde as saídas da execução dos scripts e da ferramenta elki são encontradas, respectivamente. A planilha denominada “IFGF_Prefeituras_2013_2014_2015” contém justamente o índice Firjan de todas prefeituras dentre os anos 2013 a 2015. Este arquivo pode ser útil ao trazer informações relevantes da qualidade da gestão fiscal das prefeituras que podem ser analisadas junto aos resultados.

Figura 2 - Estrutura da pasta dados

Name	Date modified	Type	Size
helper	15/05/2017 21:54	File folder	
INPUT	11/12/2017 11:45	File folder	
model	15/05/2017 21:54	File folder	
originals	11/12/2017 11:44	File folder	
OUTPUT	11/12/2017 11:36	File folder	
OUTPUT_ELKI	11/12/2017 11:42	File folder	
IFGF_Prefeituras_2013_2014_2015	19/05/2017 19:42	Planilha OpenDoc...	367 KB
Main.class	11/12/2017 11:10	CLASS File	5 KB
Preprocessamento.class	16/05/2017 22:29	CLASS File	21 KB
RUN_Multienio_part_1	11/12/2017 10:53	Windows Batch File	2 KB
RUN_Multienio_part_2	11/12/2017 10:32	Windows Batch File	1 KB
RUN_part_1	11/12/2017 11:12	Windows Batch File	2 KB
RUN_part_2	11/12/2017 11:34	Windows Batch File	1 KB

Ao todo são 4 arquivos .bat, sendo dois para processamento de dois ou mais anos/arquivos (RUN_multienio_part_1 e 2) e dois para processamento de um único arquivo por vez ("RUN_Part_1" e "RUN_Part_2"). A única diferença entre os scripts com "multienio" dos outros é a variável que armazena o parâmetro do ano a ser processado. No "multienio" são passados dois parâmetros: um para indicar o início e outro o fim da sequência de anos / arquivos a serem processados.

Na Figura 3, é possível verificar os arquivos originais dos anos 2014 e 2015. Esta pastas e arquivos apenas foram mantidos por questão de organização do projeto. Entretanto, não são necessários após serem copiados para a pasta "INPUT".

O download desses e outros arquivos pode ser feito em: https://siconfi.tesouro.gov.br/siconfi/pages/public/consulta_finbra/finbra_list.jsf e em <http://tesouro.fazenda.gov.br/contas-anuais> para os dados dos anos anteriores a 2013.

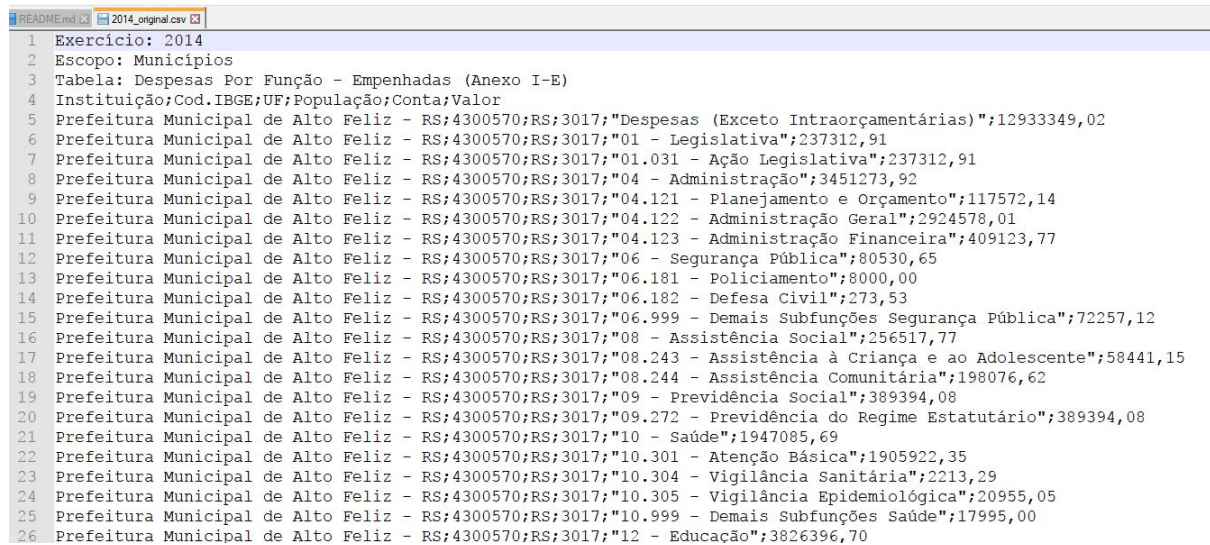
Figura 3 - Pasta com os dados originais

This PC > HD Data (D:) > REPOSITORY > TCC > PROJETO > dados > originals			
Name	Date modified	Type	Size
2014_original	24/11/2016 20:39	Planilha OpenOffi...	21.763 KB
2015_original	24/11/2016 20:57	Planilha OpenOffi...	21.531 KB

A Figura 4, mostra um dos arquivos originais abertos em um editor de texto. Vale ressaltar que, aparentemente, ao baixar esses dados do portal siconfi citado acima, algumas diferenças podem ser notadas. Por exemplo, uma nova coluna foi inserida

classificando o tipo da despesa. Com isso, algumas alterações e precauções devem ser tomadas para evitar a duplicação ou até mesmo erros no processamento.

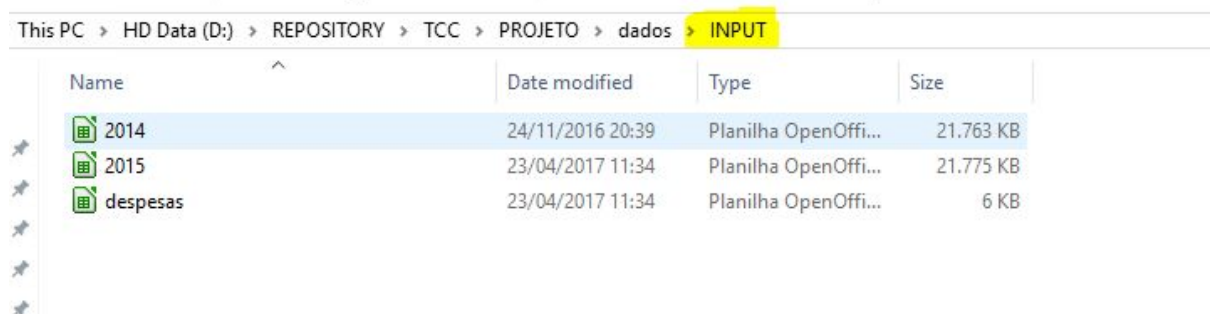
Figura 4 - Arquivo original aberto com editor de texto



Exercício	Escopo	Tabela	Despesas Por Função - Empenhadas (Anexo I-E)	Instituição;Cod.IBGE;UF;População;Conta;Valor
2014	Municípios	Despesas Por Função - Empenhadas (Anexo I-E)		
Prefeitura Municipal de Alto Feliz	RS;4300570;RS;3017	"Despesas (Exceto Intraorçamentárias)"	12933349,02	
Prefeitura Municipal de Alto Feliz	RS;4300570;RS;3017	"01 - Legislativa"	237312,91	
Prefeitura Municipal de Alto Feliz	RS;4300570;RS;3017	"01.031 - Ação Legislativa"	237312,91	
Prefeitura Municipal de Alto Feliz	RS;4300570;RS;3017	"04 - Administração"	3451273,92	
Prefeitura Municipal de Alto Feliz	RS;4300570;RS;3017	"04.121 - Planejamento e Orçamento"	117572,14	
Prefeitura Municipal de Alto Feliz	RS;4300570;RS;3017	"04.122 - Administração Geral"	2924578,01	
Prefeitura Municipal de Alto Feliz	RS;4300570;RS;3017	"04.123 - Administração Financeira"	409123,77	
Prefeitura Municipal de Alto Feliz	RS;4300570;RS;3017	"06 - Segurança Pública"	80530,65	
Prefeitura Municipal de Alto Feliz	RS;4300570;RS;3017	"06.181 - Policiamento"	8000,00	
Prefeitura Municipal de Alto Feliz	RS;4300570;RS;3017	"06.182 - Defesa Civil"	273,53	
Prefeitura Municipal de Alto Feliz	RS;4300570;RS;3017	"06.999 - Demais Subfunções Segurança Pública"	72257,12	
Prefeitura Municipal de Alto Feliz	RS;4300570;RS;3017	"08 - Assistência Social"	256517,77	
Prefeitura Municipal de Alto Feliz	RS;4300570;RS;3017	"08.243 - Assistência à Criança e ao Adolescente"	58441,15	
Prefeitura Municipal de Alto Feliz	RS;4300570;RS;3017	"08.244 - Assistência Comunitária"	198076,62	
Prefeitura Municipal de Alto Feliz	RS;4300570;RS;3017	"09 - Previdência Social"	389394,08	
Prefeitura Municipal de Alto Feliz	RS;4300570;RS;3017	"09.272 - Previdência do Regime Estatutário"	389394,08	
Prefeitura Municipal de Alto Feliz	RS;4300570;RS;3017	"10 - Saúde"	1947085,69	
Prefeitura Municipal de Alto Feliz	RS;4300570;RS;3017	"10.301 - Atenção Básica"	1905922,35	
Prefeitura Municipal de Alto Feliz	RS;4300570;RS;3017	"10.304 - Vigilância Sanitária"	2213,29	
Prefeitura Municipal de Alto Feliz	RS;4300570;RS;3017	"10.305 - Vigilância Epidemiológica"	20955,05	
Prefeitura Municipal de Alto Feliz	RS;4300570;RS;3017	"10.999 - Demais Subfunções Saúde"	17995,00	
Prefeitura Municipal de Alto Feliz	RS;4300570;RS;3017	"12 - Educação"	3826396,70	

Em “INPUT” como mostra a Figura 5, estão os mesmos arquivos originais, porém renomeados para somente o ano que representam. Manter somente o ano como nome do arquivo é essencial para rodar os scripts. O arquivo “despesas” foi criado a partir da extração das categorias/ funções encontradas nos gastos, e é utilizado pelos algoritmos chamados nos scripts.

Figura 5 - Pasta com os arquivos de entrada para processamento



Name	Date modified	Type	Size
2014	24/11/2016 20:39	Planilha OpenOffi...	21.763 KB
2015	23/04/2017 11:34	Planilha OpenOffi...	21.775 KB
despesas	23/04/2017 11:34	Planilha OpenOffi...	6 KB

Na Figura 6, estão destacados as variáveis do primeiro script que devem ser alteradas, caso necessário, antes de executar esse script. A variável “YEAR” indica o arquivo a ser processado, e nas duas linhas com o “java jar” estão todos os parâmetros passados para o elki. Isto é, na necessidade de alteração do algoritmo (LOF, COF, SVM, etc) ou de seus parâmetros (valor de k, agrupamentos, etc) essas linhas deverão ser alteradas.

Figura 6 - Primeiro script (RUN_part_1.bat) aberto com editor de texto

```

1 xcopy ..\Preprocessamento\bin\helper /I /Y helper
2 xcopy ..\Preprocessamento\bin\model /I /Y model
3 copy ..\Preprocessamento\bin\Main.class Main.class
4 copy ..\Preprocessamento\bin\Preprocessamento.class Preprocessamento.class
5
6
7
8 SET YEAR=2014
9 SET TYPE=ano
10 SET DIR=ANO_%YEAR%
11 SET ORIGINAL_FILE=%YEAR%.csv
12
13 SET FILE_REL=_REL_NOR.csv
14 SET FILE_SUA=_SUA_NOR.csv
15 SET FILE_ELKI=lof-outlier_order.txt
16 SET FILE_ELKI_SUA=ELKI_SUA_NOR.txt
17 SET FILE_ELKI_REL=ELKI_REL_NOR.txt
18 SET OUTPUT_ELKI=OUTPUT_ELKI\%DIR%\
19 RMDIR "%cd%\OUTPUT\%DIR%" /S /Q
20
21 md "%cd%\OUTPUT\%DIR%"
22 if not exist "%cd%\OUTPUT_ELKI%" md "%cd%\OUTPUT_ELKI%"
23 del /q "%cd%\OUTPUT_ELKI%*"
24 ::java Main %cd% [ano ou bienio] [XXXX ou XXXX-XXXX] [first ou last] > %OUTPUT%
25
26 java Main %cd% %DIR% %TYPE% %YEAR% first %ORIGINAL_FILE%
27
28 java -jar ../tools/elki.jar KDDCLIApplication -verbose -verbose -dbc.in "%cd%\C...
29
30 rename "%cd%\OUTPUT_ELKI%\FILE_ELKI%" %FILE_ELKI_REL%
31
32 java -jar ../tools/elki.jar KDDCLIApplication -verbose -verbose -dbc.in "%cd%\C...
33
34 rename "%cd%\OUTPUT_ELKI%\FILE_ELKI%" %FILE_ELKI_SUA%
35
36 java Main %cd% %DIR% %TYPE% %YEAR% last %ORIGINAL_FILE%
37
38
39 pause
40

```

Ao executar o primeiro script “*_part_1”, a pasta do ano (ou anos) processado será criada em “OUTPUT” como mostra a Figura 7.

Figura 7 - Pasta OUTPUT após executar o primeiro script (part_1.bat)

This PC > HD Data (D:) > REPOSITORY > TCC > PROJETO > dados > OUTPUT			
Name	Date modified	Type	
Analizados	11/12/2017 11:14	File folder	
ANO_2014	11/12/2017 12:00	File folder	

Dentro desta nova pasta estarão os três arquivos gerados: O primeiro, como mostra a Figura 8, é o arquivo “_REL_NOR” que possui os dados normalizados por meio da relação entre o valor da população e as despesas e também, normalizados pela média e desvio padrão. O segundo arquivo, “_SCORED”, possui os dados brutos com as pontuações do algoritmo utilizado nas duas últimas colunas, para ambas abordagens de normalização. O último, “_SUA_NOR”, contém os dados suavizados por logaritmo decimal e normalizados por média e desvio padrão.

Figura 8 - Arquivos gerados ao executar o primeiro script (part_1.bat)

This PC > HD Data (D:) > REPOSITORY > TCC > PROJETO > dados > OUTPUT > ANO_2014				
Name	Date modified	Type	Size	
_REL_NOR	11/12/2017 11:59	Planilha OpenOffi...	869 KB	
_SCORED	11/12/2017 12:00	Planilha OpenOffi...	1.378 KB	
_SUA_NOR	11/12/2017 11:59	Planilha OpenOffi...	942 KB	

O segundo script (*_part_2), como mostra a Figura 9, é utilizado para executar o algoritmo “explicador”. Isto é, o algoritmo que busca os centróides das cidades de população similar e verifica as diferenças de gastos.

Figura 9 - Segundo script (RUN_part_2.bat) aberto com editor de texto

```

1 xcopy ..\Preprocessamento\bin\helper /I /Y helper
2 xcopy ..\Preprocessamento\bin\model /I /Y model
3 copy ..\Preprocessamento\bin\Main.class Main.class
4 copy ..\Preprocessamento\bin\Preprocessamento.class Preprocessamento.class
5
6
7 SET TYPE=ano
8
9
10
11 SET YEAR=2014
12 SET ORIGINAL_FILE=%YEAR%.csv
13 SET DIR=ANO %YEAR%
14 set COL_RANGE=3-32
15 SET SCORE_COL=34
16 SET RANKING=10
17 SET K=20
18 set MIN=0
19 SET MAX=1000
20 SET INPUT=OUTPUT\%DIR%\_SCORED.csv
21 set OUTPUT_TEMP=OUTPUT\%DIR%\_SCORED_TEMP.csv
22 set OUTPUT_RANKED=OUTPUT\%DIR%\_SCORED_RANKED.csv
23
24 java Preprocessamento --centroid %K% 2 0 %COL_RANGE% %SCORE_COL% %MIN% %MAX% < %INPUT% > %OUTPUT_TEMP%
25
26 :::java Preprocessamento --ranking rankingSize colRange < anomalies_suav_full.csv > anomalies_suav_rank
27 java Preprocessamento --ranking %RANKING% %COL_RANGE% < %OUTPUT_TEMP% > %OUTPUT_RANKED%
28
29 del "%cd%\%OUTPUT_TEMP%"
30
31 pause
32
33
34

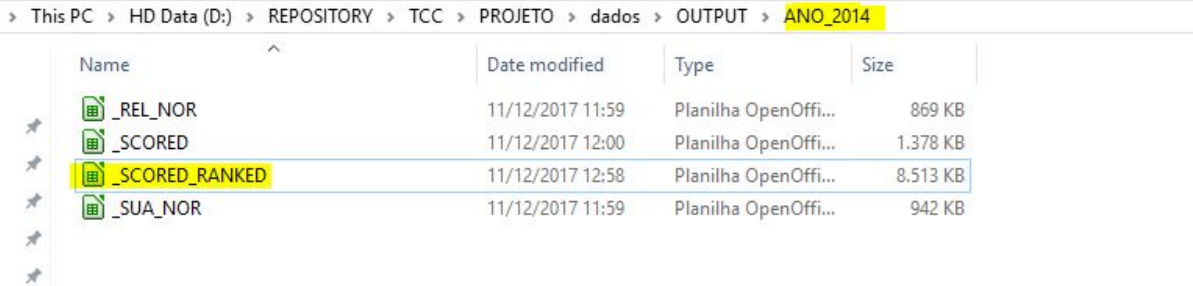
```





Os parâmetros / variáveis importantes neste arquivo são: “YEAR” que indica o arquivo a ser processado, “SCORE_COL” que é a coluna da pontuação a ser utilizada (como explicado acima, são geradas 2 colunas), “K” é o número de cidades

a serem consideradas para gerar o centróide e “MIN” é o valor da pontuação mínima a ser levado em consideração.

A Figura 10 mostra o arquivo de saída da execução do segundo script. Ou seja, o arquivo que contém os dados brutos + as diferenças de valores e porcentagens de influências dos gastos das cidades que foram responsáveis pelas suas pontuações de anormalidade.

Figura 10 - Arquivo gerado ao executar o segundo script (part_2.bat)



> This PC > HD Data (D:) > REPOSITORY > TCC > PROJETO > dados > OUTPUT > ANO_2014				
Name	Date modified	Type	Size	
 _REL_NOR	11/12/2017 11:59	Planilha OpenOffi...	869 KB	
 _SCORED	11/12/2017 12:00	Planilha OpenOffi...	1.378 KB	
 _SCORED_RANKED	11/12/2017 12:58	Planilha OpenOffi...	8.513 KB	
 _SUA_NOR	11/12/2017 11:59	Planilha OpenOffi...	942 KB	