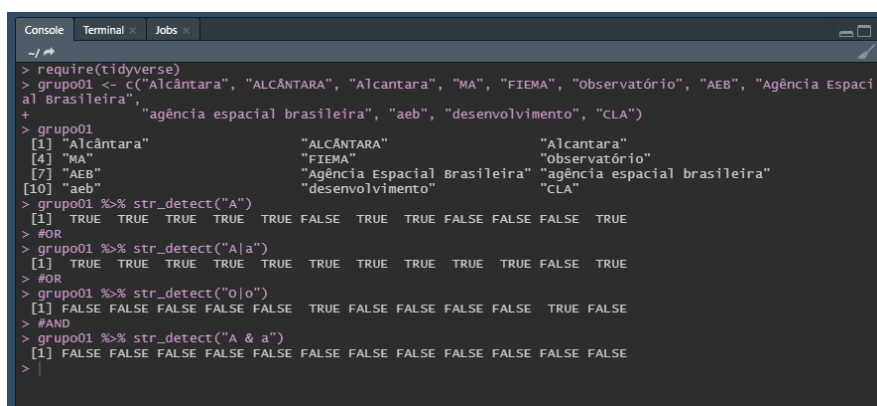


1 Resumo

Para o estudo de caso foi considerado que ao analisar um grupo de strings deve ser encontrado algum padrão da linguagem que está sendo utilizado. Dessa forma no momento da extração de dados é possível encontrar esses padrões e retirá-los do conjunto de dados que está sendo analisado. Para isso, foi utilizado a pacote tidyverse do R e com ele foram feitos alguns testes para verificar se o script é capaz de identificar esses padrões.

1.1 Detecção de Padrão

Para o primeiro caso foi criado um vetor de strings e sequencialmente foi verificado se é possível obter a detecção do padrão de texto, como teste foi considerado o padrão 'A'. No grupo01 foram inseridos strings que têm relação direta com o tema da AEB. Também foi levado em consideração que a linguagem R é case sensitive. Foi feito o teste utilizando também operador lógico para verificação.

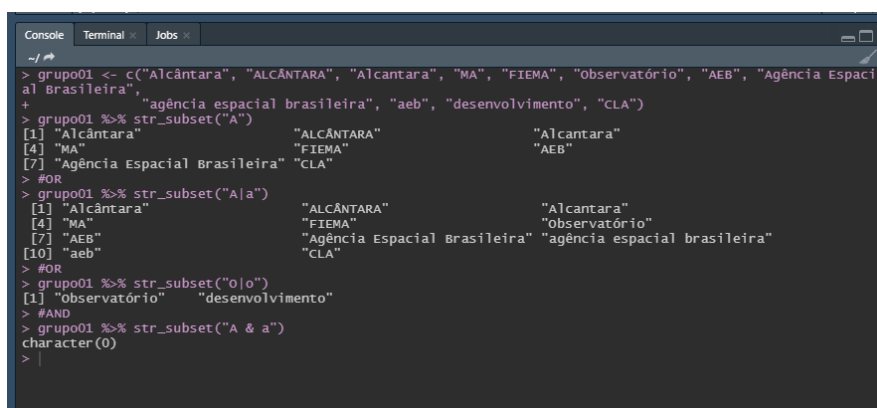


```
> require(tidyverse)
> grupo01 <- c("Alcantara", "ALCANTARA", "Alcantara", "MA", "FIEMA", "Observatório", "AEB", "Agência Espaci
al Brasileira", "agência espacial brasileira", "aeb", "desenvolvimento", "CLA")
+
> grupo01
[1] "Alcantara"          "ALCANTARA"          "Alcantara"
[4] "MA"                "FIEMA"              "Observatório"
[7] "AEB"               "Agência Espacial Brasileira" "agência espacial brasileira"
[10] "aeb"               "desenvolvimento"    "CLA"
> grupo01 %>% str_detect("A")
[1] TRUE TRUE TRUE TRUE FALSE TRUE TRUE FALSE FALSE FALSE TRUE
> #OR
> grupo01 %>% str_detect("A|a")
[1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE FALSE TRUE
> #OR
> grupo01 %>% str_detect("o|O")
[1] FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE TRUE FALSE
> #AND
> grupo01 %>% str_detect("A & a")
[1] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
> |
```

Abbildung 1: Detecção de padrão

1.2 Indexação

Para que seja possível realizar a localização a partir do índice é possível utilizar diretamente a função de subset que existe no pacote tidyverse, dessa forma é possível realizar a identificação do padrão através do índice.

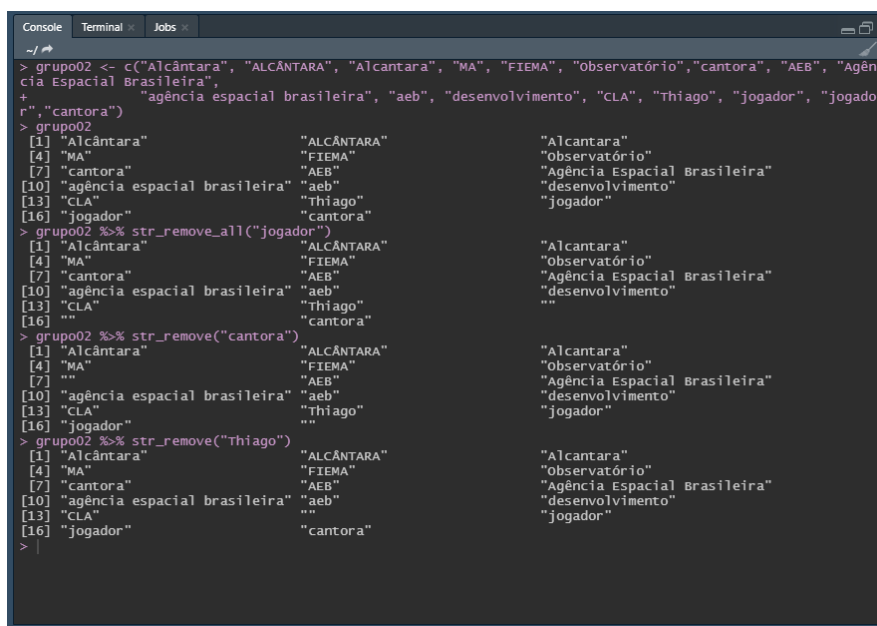


```
> grupo01 <- c("Alcantara", "ALCANTARA", "Alcantara", "MA", "FIEMA", "Observatório", "AEB", "Agência Espaci
al Brasileira", "agência espacial brasileira", "aeb", "desenvolvimento", "CLA")
+
> grupo01 %>% str_subset("A")
[1] "Alcantara"          "ALCANTARA"          "Alcantara"
[4] "MA"                "FIEMA"              "Observatório"
[7] "AEB"               "Agência Espacial Brasileira" "agência espacial brasileira"
[10] "aeb"               "desenvolvimento"    "CLA"
> #OR
> grupo01 %>% str_subset("A|a")
[1] "Alcantara"          "ALCANTARA"          "Alcantara"
[4] "MA"                "FIEMA"              "Observatório"
[7] "AEB"               "Agência Espacial Brasileira" "agência espacial brasileira"
[10] "aeb"               "desenvolvimento"    "CLA"
> #OR
> grupo01 %>% str_subset("o|O")
[1] "Observatório"      "desenvolvimento"
> #AND
> grupo01 %>% str_subset("A & a")
character(0)
> |
```

Abbildung 2: Indexação

1.3 Remoção de padrões

Nesse caso, o tratamento seria feito para realizar a retirada de outliers que se encontram no conjunto de dados que foi extraído. Para simulação foi criado um segundo vetor de strings que possui os três outliers que são mais encontrados na coleta do twitter do projeto para o PDI-CEA. Sendo eles 'cantora', 'jogador' e 'Thiago'. O que foi observado é que a função remove consegue detectar e retirar a palavra que foge do padrão do grupo de dados, no entanto ao realizar uma retirada de um segundo padrão em seguida ele retorna o padrão que havia sido removido anteriormente, causando um problema no processo de remoção do padrão.



```
> grupo02 <- c("Alcântara", "ALCÂNTARA", "Alcantara", "MA", "FIEMA", "Observatório", "cantora", "AEB", "Agência Espacial Brasileira", "agência espacial brasileira", "aeb", "desenvolvimento", "CLA", "Thiago", "jogador", "jogador", "cantora")
> grupo02
[1] "Alcântara"      "ALCÂNTARA"      "Alcantara"      "MA"             "FIEMA"          "Observatório"   "cantora"        "AEB"            "Agência Espacial Brasileira"
[10] "agência espacial brasileira" "aeb"            "desenvolvimento" "CLA"            "Thiago"         "jogador"        "jogador"        "cantora"
> grupo02 %>% str_remove_all("jogador")
[1] "Alcântara"      "ALCÂNTARA"      "Alcantara"      "MA"             "FIEMA"          "Observatório"   "cantora"        "AEB"            "Agência Espacial Brasileira"
[10] "agência espacial brasileira" "aeb"            "desenvolvimento" "CLA"            "Thiago"         ""              ""              ""
[16] ""              ""              "cantora"
> grupo02 %>% str_remove("cantora")
[1] "Alcântara"      "ALCÂNTARA"      "Alcantara"      "MA"             "FIEMA"          "Observatório"   ""              "AEB"            "Agência Espacial Brasileira"
[10] "agência espacial brasileira" "aeb"            "desenvolvimento" "CLA"            "Thiago"         ""              ""              "jogador"
[16] "jogador"
> grupo02 %>% str_remove("Thiago")
[1] "Alcântara"      "ALCÂNTARA"      "Alcantara"      "MA"             "FIEMA"          "Observatório"   ""              "AEB"            "Agência Espacial Brasileira"
[10] "agência espacial brasileira" "aeb"            "desenvolvimento" "CLA"            ""              ""              "jogador"
[16] "jogador"
>
```

Abbildung 3: Remoção de padrões

2 Conclusão

Diante do que foi apresentado, verifica-se a possibilidade de inserir algumas expressões regulares no script de coleta do projeto sobre o PDI-CEA. Visando ou na detecção de expressões que são de interesse do projeto e manter apenas ela no script ou da retirada de outliers.