 Logotipo, nome da empresa

Descrição gerada automaticamente

Manual de utilização da análise de logs de sites legados

INF2102 – Projeto Final de Programação

Professora: Clarisse

Aluno: Rodrigo Brito de Freitas Lima

Orientadora: Simone Diniz Junqueira Barbosa

# Introdução

Este Manual refere-se a análise de logs do site fluzao.xyz, alvo principal dos estudos em como melhorar a experiência do usuário em sistemas legados, algumas premissas tem que ficar bem explicadas para que o trabalho possa ser utilizado em qualquer outro trabalho relacionado.

1. Utilização da ferramenta Google Analytics 4 – extração de relatórios de navegação, páginas, títulos das páginas por data, títulos das páginas por data hora, acesso de usuários totais e acesso de usuários por cidade e estado, exceto os relatórios de acesso de usuários totais e por cidade todos os outros precisaram ser customizados/criados na ferramenta.
2. O site precisa estar com os títulos todos bem formatados de forma a permitir que a análise seja realizada nas páginas acessadas.
3. Um mapa do site precisa ser previamente criado com as páginas de acesso e seus códigos de forma a permitir a análise de páginas, deste mapa deve ser gerado o arquivo referencia.csv
4. Se for necessário dividir o sistema em grupos e subgrupos de forma a permitir a verificação dos acessos em diferentes áreas de atuação, validando assim a eficiência atual e em quais pontos devemos agir que todas as informações tenham relevância e sejam exploradas.

Os códigos de todo o sistema estão divididos em 2 partes, a primeira parte é particular e refere-se a todo o site [www.fluzao.xyz](http://www.fluzao.xyz) e não ficará disponível, a segunda parte refere-se a toda a análise dos logs, questionários e entrevistas e ficará disponível no git: <Preencher com o endereço quando existir>

# Pré processamento

Os relatórios do Analytics foram baixados em csv e sofrem algumas alterações para ficarem prontos para serem executados nos jupyter notebooks.

1. Devido a formatação de acentos e ao caractere de divisão dos campos é executada uma macro no Notepad++ que remove todos os acentos e troca o caractere de “,” para “;”, desta forma é possível a abertura no Excel sem problema.
2. Os campos de data são exportados em extenso, Ex: 20230920, por isso é necessário dentro do Excel a criação de fórmulas para transforma-los para o formato data.
3. Especificamente para o relatório de páginas de forma a dividir os acessos em grupos e subgrupos foram criadas fórmulas que quebram o caminho da página em 3 pedaços, sendo o primeiro referente ao grupo, o segundo referente ao subgrupo e o terceiro ao código acessado.
4. A página inicial neste processo fica com os campos em branco e para documentação o valor é editado manualmente com a palavra inicial em ambas as colunas de grupo e subgrupo, o código permanece em branco.
5. Para o caso específico desta análise é necessário manter um arquivo com os jogos disputados para análise de qual jogo é o mais acessado pelos usuários, isso foi feito manualmente porém é um detalhe muito específico deste sistema.

Ao final de todo esse processo são gerados os arquivos de entrada para a análise, são eles:

1. acesso\_cidades.csv
2. acesso\_estado.csv
3. acesso\_paginas.csv
4. acessos.csv
5. referencia.csv
6. telas\_acessadas.csv
7. Titulos\_data.csv
8. Titulos\_data\_hora.csv
9. Jogo.csv

# Arquivos de Entrada

1. **acesso\_cidades.csv**

|  |  |  |
| --- | --- | --- |
| **Colunas** | **Descrição** | **Origem** |
| Cidade | Cidade de onde o usuário acessou o site | Analytics |
| Data | Data por extenso do acesso | Analytics |
| Usuários | Número de usuários recorrentes acessando o site | Analytics |
| Novos Usuários | Número de novos usuários acessando o site | Analytics |
| Total | Total de usuários – Usuários + Novos Usuários | Excel |
| DT | Data formatada | Excel |

1. **acesso\_estado.csv**

|  |  |  |
| --- | --- | --- |
| **Colunas** | **Descrição** | **Origem** |
| Regiao | Estado de onde o usuário acessou o site | Analytics |
| Data | Data por extenso do acesso | Analytics |
| Usuário | Número de usuários recorrentes acessando o site | Analytics |
| Primeira Visita | Número de novos usuários acessando o site | Analytics |
| Total | Total de usuários | Analytics |
| Usuarios | Total de usuários | Analytics |
| DT | Data formatada | Excel |

1. **acesso\_paginas.csv**

|  |  |  |
| --- | --- | --- |
| **Colunas** | **Descrição** | **Origem** |
| Pagina | Página acessada | Analytics |
| Visualizações | Número de visualizações | Analytics |

1. **acessos.csv**

|  |  |  |
| --- | --- | --- |
| **Colunas** | **Descrição** | **Origem** |
| Data | Data do acesso | Analytics |
| Novos usuarios | Número de novos usuários | Analytics |
| Recorrentes | Número de usuários recorrentes acessando o site | Analytics |
| Total | Total de usuários | Excel |

1. **referencia.csv –** Este arquivo deve ser gerado uma vez e atualizado caso uma nova funcionalidade seja adicionada

|  |  |  |
| --- | --- | --- |
| **Colunas** | **Descrição** | **Origem** |
| Perguntas | Número da pergunta a que se refere a pesquisa | Mapa |
| Consulta | Nome da consulta | Mapa |
| Grupo | Grupo da consulta | Mapa |
| Subgrupo | Subgrupo da consulta | Mapa |
| Código | Código para acesso da consulta | Mapa |
| Pagina | Inicio do nome da página – referente ao grupo | Mapa |
| Pagina2 | Complemento do nome da página – subgrupo | Mapa |
| Visualizacao | Sempre zerado – será populado no primeiro processo | -- |
| Usuario | Sempre zerado – será populado no primeiro processo | -- |

1. **telas\_acessadas.csv**

|  |  |  |
| --- | --- | --- |
| **Colunas** | **Descrição** | **Origem** |
| Data | Data do acesso | Analytics |
| Visualizações | Número de visualizações | Analytics |
| Usuarios | Número de usuários recorrentes acessando o site | Analytics |
| Novos usuarios | Número de novos usuários | Analytics |
| Endereco | Página acessada | Excel |
| Página | Página acessada sem o código | Excel |
| Codigo | Codigo da pagina acessada | Excel |

1. **Titulos\_data.csv**

|  |  |  |
| --- | --- | --- |
| **Colunas** | **Descrição** | **Origem** |
| Titulo | Título da Página acessada | Analytics |
| Extenso | Data por extenso | Analytics |
| Sessões | Sessões que acessaram a página | Analytics |
| Usuarios | Número de usuários recorrentes acessando o site | Analytics |
| Novos usuarios | Número de novos usuários | Analytics |
| Data | Data do acesso formatada | Excel |

1. **Titulos\_data\_hora.csv**

|  |  |  |
| --- | --- | --- |
| **Colunas** | **Descrição** | **Origem** |
| Titulo | Título da Página acessada | Analytics |
| Extenso | Data por extenso | Analytics |
| Sessões | Sessões que acessaram a página | Analytics |
| Usuarios | Número de usuários recorrentes acessando o site | Analytics |
| Novos usuarios | Número de novos usuários | Analytics |
| Data | Data do acesso formatada | Excel |
| Hora | Hora do acesso | Excel |

1. **jogo.csv –** Este arquivo é específico deste sistema e fornece suporte para a identificação dos jogos mais acessados.

|  |  |  |
| --- | --- | --- |
| **Colunas** | **Descrição** | **Origem** |
| res\_flu | Resultado do Fluminense | BD |
| res\_adv | Resultado do Adversário | BD |
| Cod\_jogo | Código do jogo no BD | BD |
| Adversário | Nome do Adversário | BD |
| Data | Data do Jogo | BD |
| Agrupado | Formatação do título da página | BD |

# Processamento

A etapa de processamento foi realizada utilizando o Jupyter Notebook sendo acessado dentro do ambiente do Anaconda, as bibliotecas envolvidas no processo foram a numpy, Pandas, Plotly, matplotlib e warnings.

Para iniciar o processamento é necessário atualizar o arquivo e referência e depois realizar as diversas análises em cima dos logs, seguem abaixo o passo a passo e o link de acesso para cada jupyter notebook criado.

## Gera referência atualizada

Incluir link do Jupyter

O primeiro processo a ser executado é a geração da referência atualizada, ele consiste em pegar os arquivos de entrada referencia.csv e telas\_acessadas.csv com a finalidade de atualizar os campos Visualizacao e Usuario do arquivo de referencia gerando um outro arquivo chamada referencia\_atualizada.csv que será a base de todo o resto da análise de páginas acessadas.

O campo código da tabela é único no sistema e portanto é utilizado para atualizar os acessos de todas as páginas que contenham código, as páginas que não contém código porém tem os grupos e subgrupos bem determinados também é possível associar o código, fora isso páginas de link e gráfico também são atualizadas, as demais páginas são consideradas páginas de trânsito e não contém informações finais e não são importantes para o estudo e neste caso serão desconsideradas.

## Análise de Títulos das Páginas

Incluir link do Jupyter

Em qualquer momento de qualquer análise pode ser realizado um filtro por data limitando o tempo de análise, outro ponto importante é que o arquivo está listado por data e por isso quando queremos fazer uma análise do todo precisamos agrupar por título.

Inicialmente foi criada uma pesquisa por palavra chave, através de um array de palavras o programa deve isolar somente os títulos que tem alguma das palavras, após a geração deste novo dataset é necessário complementar com zero as datas que a palavra chave não foi acessada, isso foi feito para analisar a relevância das páginas no espaço de tempo estudado, o exemplo realizado foi colocar o nome de todos os adversários da série A e todos os adversários da Libertadores desde 03/08 e analisar como se comportam os acessos as páginas que contenham os nomes de cada clube.

A pesquisa por palavra chave permite fazer uma análise completa de qualquer formato visando estabelecer uma relação entre os objetos do sistema e sua quantidade de acessos ao longo do tempo, será gerado um gráfico de linha com o acompanhamento de cada palavra chave elencada.

Outra validação realizada é com relação as páginas mais visitadas em quantidade, neste quesito fazemos duas validações uma com todas as páginas e uma segunda removendo as páginas de entrada de cada seção, ou seja, excluímos a página inicial e páginas principais de consulta de cada grupo gerando os resultados consolidados e os resultados com o que de fato o usuário tem mais interesse em pesquisar, essa validação também foi quebrada por data onde é possível identificar a conexão das páginas mais acessadas contra fatos importantes que aconteceram no campo ou na imprensa esportiva, serão gerados gráficos de barra consolidados e por data com as páginas mais acessadas.

As análises a seguir são totalmente voltadas para os dados do sistema objeto do estudo tornando assim as definições bem específicas, a primeira análise realizada foi identificar historicamente quais os adversários são mais visitados independente de data, para isso o dataset foi filtrado pelo início ‘Fluminense X ‘, isso indica a visitação a página do confronto, depois é agrupado por título e visualizado em um gráfico de barras.

Seguindo a mesma linha da análise do confronto foi criado um gráfico de barras sobre os acessos as páginas de Técnicos do Fluminense, filtrando o dataset pela sentença ‘Tec – ‘.

Uma outra análise realizada diz respeito as páginas do sistema que contenham o ano listado, por isso foi criado um array com os anos desde 1902, fundação do Fluminense e depois filtrado o dataset, agrupado pelo título e visualizado em um gráfico de barras.

Análise dos títulos envolvendo uma palavra, filtrar pela palavra, agrupar por data e visualizar gráfico de barra com os títulos acessados e mostrar o detalhado por data, qualquer palavra que exista em um título do sistema.

## Análise de páginas acessadas

Incluir link do Jupyter

Com o arquivo referencia\_atualizada.csv conseguimos analisar as consultas no site por grupo e subgrupo, essa análise somente poderá ser realizada se o site/sistema tiver a possibilidade de ser quebrado em grupos e subgrupos.

Uma análise realizada foi a quantidade de páginas que tiveram uma determinada quantidade de visualizações, isso é interessante para identificar principalmente as páginas que nunca foram acessadas, com isso analisar o que precisa ser feito na interface para que todas as informações sejam acessadas.

Identificar as páginas mais acessadas, neste caso pegamos as 15 mais acessadas e geramos um gráfico de barras.

Análise sobre os Grupos de consulta, o dataset de referência será agrupado pela coluna Grupo e será gerado um gráfico de barra ordenado pelos Grupos mais acessados.

Análise sobre os Grupos e subgrupos de consulta, o dataset de referência será agrupado pela coluna Grupo e Subgrupo e será gerado um gráfico de barra ordenado pelos Grupos-Subgrupos mais acessados.

Análise das páginas que não tiveram acesso desde o início das medições, filtrar por 0 Visualizações, agrupar por grupo-subgrupo e gerar gráfico de barras para identificar onde devem ser focados os esforços para entender o porque as consultas estão sem acesso, perguntas no questionário e entrevista foram formuladas para buscar mais informações.

Para entender a profundidade desta falta de acesso foi analisado a quantidade de consultas com acesso e sem acesso dos grupos e subgrupos do sistema, foram criados gráficos de barra empilhada e barra lado a lado.

Análise dos jogos mais acessados, devido a uma limitação da análise de títulos, jogos com o mesmo placar tem o mesmo título e ocorreram em datas diferentes, sendo assim precisamos verificar os acessos pelas urls, nas urls temos o código do jogo e com isso é possível quantificar os acessos ao jogo correto, no final geramos um gráfico de barra com os dados consolidados e depois quebramos os acessos por data e geramos tantos relatórios quanto necessários com os acessos aos jogos por data.

## Análise de acessos

Incluir link do Jupyter

Utilizando os arquivos de acesso geral, acesso por cidade e acesso por estado serão gerados gráficos para estudar o fluxo de acessos diários, para os acessos gerais foi gerado um gráfico de linha com os usuários totais, novos e recorrentes, já os acessos por cidade e estado são gerados gráficos de barra com os 30 maiores níveis de acesso.

Continuando a análise filtrar o arquivo por uma cidade ou estado e gerar um gráfico de linha analisando o fluxo de acessos tentando relacionar o número de acessos com o local do próximo jogo ou do adversário.

## Análise de Títulos por data e Hora

Incluir link do Jupyter

Utilizando o arquivo de títulos acessados por data e hora serão gerados gráficos para analisar o impacto dos jogos no acesso as informações, ou seja, como se comporta o número de acessos durante a realização da partida, vale lembrar que as informações do site não são atualizadas em real time.

Foram propostas 3 análises, uma com as páginas mais acessadas dada uma palavra chave por data e hora via um gráfico de barras, com base neste mesmo dataset filtrado pela palavra chave ordenamos por data-hora e geramos um gráfico de linha com o número de acessos geral a qualquer página com a palavra chave.

Outra análise é sumarizar o número de acessos geral por data hora a fim de identificar como se comporta o número de acessos por data-hora e relacionar com a data hora da realização dos jogos.