# UNIVERSIDADE FEDERAL DE OURO PRETO INSTITUTO DE CIÊNCIAS EXATAS E APLICADAS - ICEA

## Relatorio TCC

TCC<sub>1</sub>

# Filipe Nunes Soares | 18.2.8084

#### Antes de incrementar as threads

O processo estava demorando muito e não chegava perto de parar, depois de 45.278s equivalentes a 12 horas o programa só havia encontrado 47 mil links e rastreado 25 mil

```
Total Links: 47445 | Rastreados: 25947 | Erros: 6637 | Tempo de execução: 45278.25s
NOVO LINK SERÁ RASTREADO !!
Link: http://www.repositorio.ufop.br/jspui/handle/123456789/9914
```

Os links rastreados continham PDF, links dos rádios, páginas de notícias como a revista curinga, pegando páginas extremamente antigas.

#### Após incrementar as threads

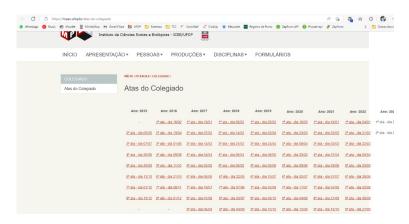
Foi adicionado 10 threads para rodar em simultâneo sem forçar muito o computador, tendo disponibilidade de adicionar mais futuramente, com este mecanismo o processo foi agilizado.

```
Total Links3: 46619 | Rastreados: 115473 | Erros: 1070 | Tempo de execução: 60648.53s | PDF: 12099 | Radio: 285

NOVO LINK SERÁ RASTREADO !!

Link: https://www.jornalismo.ufop.br/revistacuringa/?attachment_id=19#comment-38903
```

Chegando a encontrar em média de 466.190 links em 16 horas e rastreando 115.473 links onde destes 12 mil eram PDFs. Como podemos ver os PDFs assumem uma grande parte de toda a análise. O que originou uma quantidade exuberante de links foram páginas de notícias, boletins .... Todos com datas extremamente antigas que ninguém acessa mais pois estão completamente desatualizados mas que ainda fazem parte do domínio. Acredito que podemos encontrar uma maneira de retirar estes links para reduzirmos a nossa amostra.



# UNIVERSIDADE FEDERAL DE OURO PRETO INSTITUTO DE CIÊNCIAS EXATAS E APLICADAS - ICEA

#### TCC<sub>1</sub>

### • Utilização de arquivos

Um mecanismo que estou implementando é a utilização de arquivos para não se perder os links que já foram rastreados, pois até o momento, todos os links estão ficando em cache e assim que termina a execução do programa, todos são perdidos e temos que coletar todos de novo.

Entretanto com a utilização de arquivos não teremos mais que começar do zero, já aproveitando links que foram rastreados. Um problema sobre a utilização de arquivos são páginas que foram alteradas entre a execução do programa, elas não serão novamente rastreadas. Entretanto com este mecanismo poderemos chegar até o final do processo sem ter que manter o computador ligado sem parar por horas.

## Outros problemas

- A revista coringa tem páginas que possuem comentários e essas páginas são extremamente grandes e pesadas, demorando muito o processamento delas. Ex:
  - https://www.jornalismo.ufop.br/revistacuringa/?attachment\_id=19#comment-39141
- Novamente o calendário, para quase todos os subdomínios da UFOP, vão ter calendários com os eventos, trazendo eventos de vários anos atrás.
   <a href="https://ppgmqmg.ufop.br/calendar/week/2023-02?type=month&amp;amp%3Bamp%3Bamp%3Bamp%3Bmonth=2023-02&amp;amp%3Bamp%3Bamp%3Bmini=2023-03&amp;amp%3Bamp%3Bmini=2023-03&amp;amp%3Bmonth=2023-02&amp;month=2023-02</a>

C