Atualizado em: {{date}} 02/10/2023

O repositório querido-diario-data-processing tem como objetivo gerar buscas mais assertivas para o usuário por meio do uso de técnicas de processamento de linguagem natural. O processo desse repositório pode ser referenciado a partir da imagem da Infraestrutura do Querido Diário no fluxograma_1.png. As partes referentes à indexação e extração do texto são responsabilidade desse repositório em específico. Afinal, para ter os documentos em formato de texto (.txt) disponíveis na plataforma é necessário que seja feito um processamento desse conteúdo (os PDFs coletados previamente pelo repositório querido-diario).

Esse é o objetivo principal mas não é o único, já que além da possibilidade da colaboração por meio do desenvolvimento, é também possível aplicar as técnicas de PLN em um *dataset* específico.

1. **Contribuindo no Desenvolvimento**
   Sempre fique ligado(a) ao documento de Contribuição, nele é possível verificar as exigências básicas como formatação *black*, configuração de ambiente seguro, detalhamento nas *issues* e *pull requests*.

Sabendo desses pontos, é necessário configurar o ambiente de trabalho. Existem três diferentes sistemas operacionais que são compatíveis com o ambiente desenvolvido: Linux (o padrão e raíz), Windows e Mac (os dois últimos ainda em testes). Vamos explorar cada um deles.

- **Linux**
  Se você já trabalha com Linux seguir as orientações de instalação contidas no repositório serão suficientes para instalar o ambiente.

Alguns possíveis problemas que talvez precisem de um cuidado a mais.

- **Windows**

1. Utilizando WSL
   O WSL é uma sigla para Subsistema de Windows para Linux, tradução de *Windows Subsystem for Linux*,

O sistema do Querido Diário foi totalmente desenvolvido para Linux e por isso algumas configurações não funcionam para Windows, sabendo disso uma das maneiras menos trabalhosas é configurar um subsistema para Linux, através de WSL.
https://www.alura.com.br/artigos/wsl-executar-programas-comandos-linux-no-windows?utm_term=&utm_campaign=%5BSearch%5D+%5BPerformance%5D+-

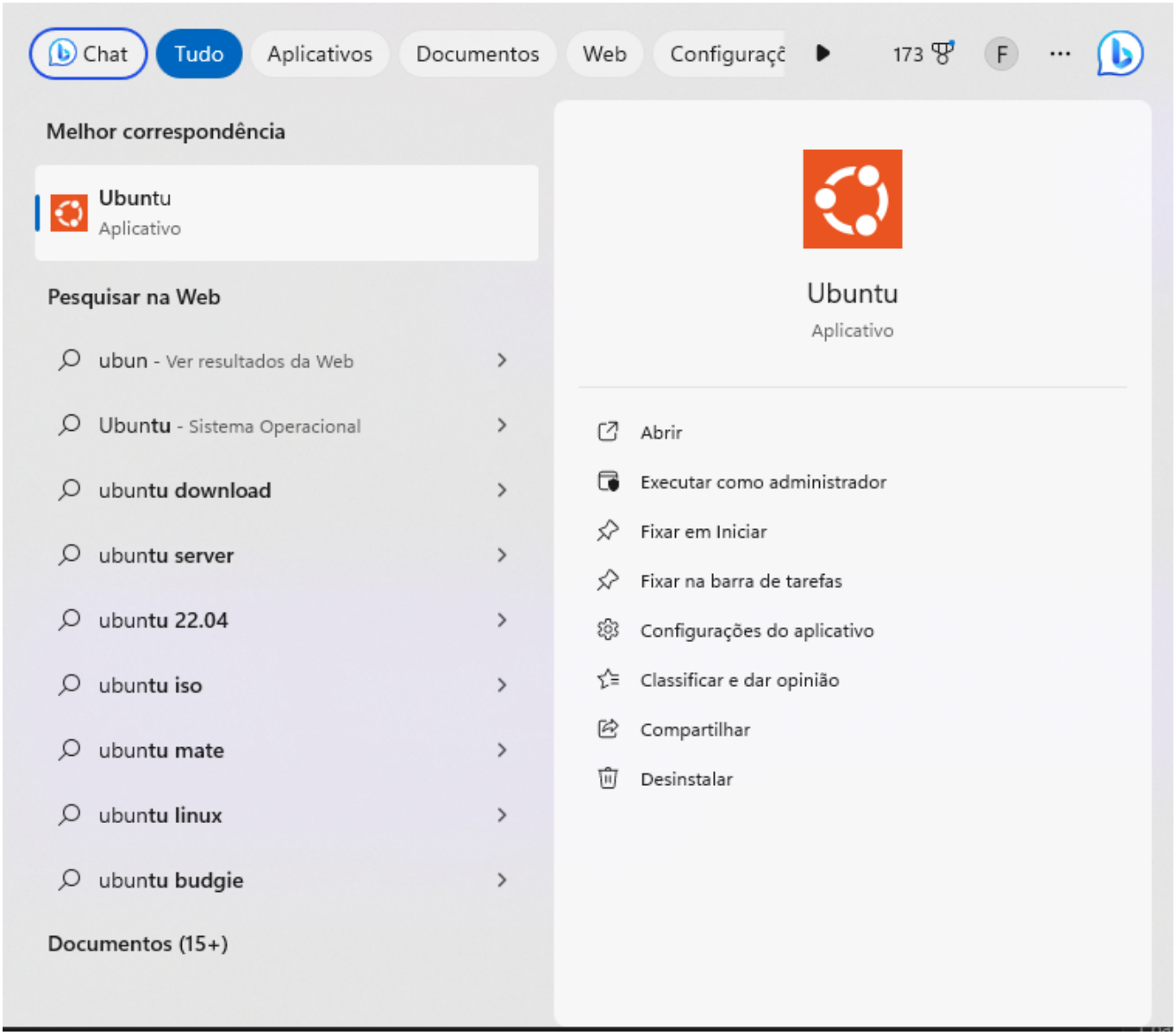
+Dynamic+Search+Ads+-
+Artigos+e+Conteúdos&utm_source=adwords&utm_medium=ppc&hsa_acc=7964138385&hsa_cam=11384329873&hsa_grp=111087461203&hsa_ad=662261334153&hsa_src=g&hsa_tgt=aud-456779235794:dsa-843358956400&hsa_kw=&hsa_mt=&hsa_net=adwords&hsa_ver=3&gclid=Cj0KCQjw06-oBhC6ARIsAGuzdw3Ok9G6C6iBPWLpHuicggkhdEvA_J-zz1fZbmtESuFHP17RmLyKZZgaApM4EALw_wcB

https://linuxize.com/post/how-to-install-pip-on-ubuntu-18.04/

```
sudo apt install python3-venv -y
sudo apt install python3.10-venv
python3 -m venv .venv
source .venv/bin/activate
```

Atenção: Recursos mais atuais do WSL exigem um sistema operacional Windows mais recentes (a partir do Windows 10)..

Ao iniciar uma nova máquina, já é possível acessá-la no menu iniciar do Windows. Por exemplo, caso tenha instalado o Ubuntu, pesquise assim:



A partir daí já será possível realizar o git clone de um repositório forked do qd-data-processing e então criar e iniciar um ambiente virtual:
\py3 -m venv .venv
\source .venv/Scripts/activate

Assim como está documentado no repositório.

Após essa etapa é necessário **conectar ao querido-diario** ao banco de dados gerados pelo repositório querido-diario o qual é responsável por extrair os diários oficiais.

Para fazer a conexão você precisará ter baixado e instalado tudo que for necessário no repositório querido-diario em outro lugar na sua máquina WSL. Deixe as pastas próximas uma da outra para facilitar seu trabalho. Abra uma outra máquina Ubuntu para iniciar o repositório querido-diario.

---- Mudar CSV para incluir ASSOCIAÇÕES (!!!)
----- Mudar settings.py em querido-diario > data_collection > gazette > settings.py

2. Utilizando somente Windows
   --- Ainda faltam configurações para que o make re-run compile.
   O que é necessário resolver:

- [ ] Habilitar pasta \mnt no Windows
- [ ] ...

Caso haja um erro com”pinned with ==” na hora de instalar os requerimentos, utilize o pip2 install e adicione um dos comandos abaixo:

```
--use-pep517
-no-deps
```

- **Mac**
  ...

1. **Contribuindo no geral**

- Como salvar os diários oficiais de forma local
  a) Utilizando somente o repositório querido-diario