

Tutorial para instalação e utilização básica do doccano



Autor: Messias Gomes da Silva

Índice

Introdução.....	3
Sobre o doccano.....	3
Pré-requisitos.....	3
Criação do ambiente virtual.....	4
Instalação do doccano.....	5
Iniciando o doccano.....	5
Finalizando o doccano.....	6
Utilização básica.....	7
Criar projeto.....	7
Importar dataset.....	8
Criar labels.....	10
Exportar labels.....	11
Importar labels.....	11
Fazer anotações.....	12
Exemplos de anotações.....	15
Exportar datasets.....	16
Apagar datasets.....	17
Como proceder para fazer as anotações.....	17
Preparando a ferramenta.....	18
Realizando as anotações.....	18

Introdução

Este tutorial cobrirá a instalação da ferramenta doccano e instruções para uma utilização básica a fim de possibilitar anotação de textos diversos que serão utilizados para criação de um corpus para tarefas de Reconhecimento de Entidades Nomeadas.

Sobre o doccano

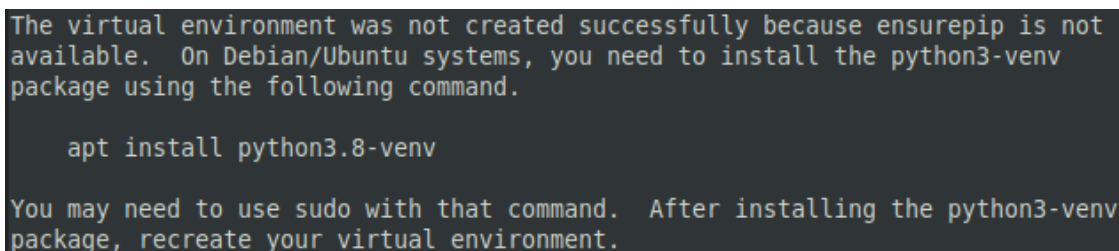
É uma ferramenta de anotação de texto de código aberto para humanos. Ele fornece recursos de anotação para classificação de texto, rotulagem de sequência e sequência para tarefas de sequência. Portanto, você pode criar dados rotulados para análise de sentimento, reconhecimento de entidade nomeada, resumo de texto e assim por diante. Basta criar um projeto, fazer upload dos dados e começar a anotar. Você pode construir um conjunto de dados em horas.

Traduzido de: <https://github.com/doccano/doccano>

Pré-requisitos

Esses são os pré-requisitos para a instalação do doccano utilizando o pip (ferramenta para instalação de pacotes do python) e o venv (ferramenta para criação de ambientes virtuais python):

1. Python 3 (de preferência a última versão). Instruções para instalação no windows: <https://python.org.br/instalacao-windows/>. Obs.: No Linux, geralmente já vem instalado por padrão.
2. Para criação de ambientes virtuais python no Linux é preciso instalar o **pythonX.Y-venv**, conforme tela de erro ao tentar criar o ambiente com o comando **python3 -m venv <nome do ambiente>**



```
The virtual environment was not created successfully because ensurepip is not
available.  On Debian/Ubuntu systems, you need to install the python3-venv
package using the following command.

apt install python3.8-venv

You may need to use sudo with that command.  After installing the python3-venv
package, recreate your virtual environment.
```

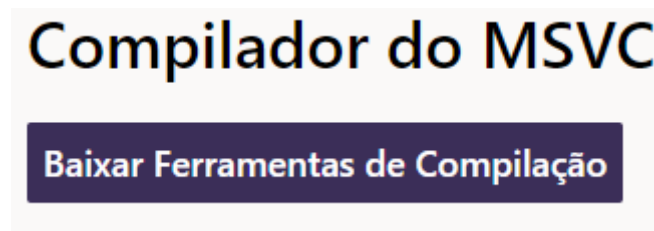
Sendo assim, já instale o pacote **pythonX.Y-venv** onde X.Y é a versão do python. Neste tutorial foi utilizado o comando: **apt install python3.8-venv**. Este comando deve ser adaptado de acordo com a distribuição Linux e a versão do python que está instalada no computador.

3. Para instalação do doccano no Windows, é requerido o “**Microsoft C++ Build Tools**”. Se não estiver instalado, aparecerá uma tela de erro conforme a abaixo:

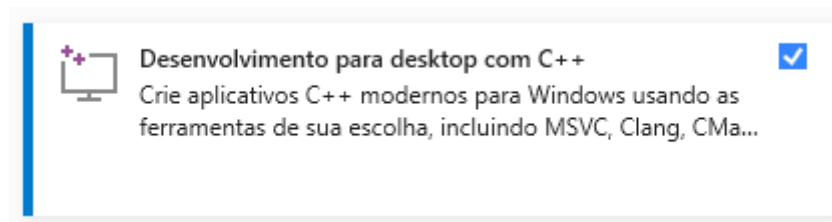
```
raise distutils.errors.DistutilsPlatformError(
distutils.errors.DistutilsPlatformError: Microsoft Visual C++ 14.0 or greater is required. Get it with "Microsoft C++
Build Tools": https://visualstudio.microsoft.com/visual-cpp-build-tools/
-----
ERROR: Failed building wheel for scikit-learn
Failed to build scikit-learn
```

Logo, será preciso instalar. Siga os passos:

a) Baixe o instalador conforme indicado na tela de erro: <https://visualstudio.microsoft.com/pt-br/visual-cpp-build-tools/>. Depois da página carregada, clique em **“Baixar Ferramentas de Compilação”**



b) Execute o instalador que foi baixado e deixe marcada somente a opção **“Desenvolvimento para desktop com C++”** e clique em instalar.



Criação do ambiente virtual

É interessante criar uma separação do ambiente para evitar conflitos entre versões de bibliotecas e pacotes de outros projetos python existentes no computador. Dessa forma, a instalação será realizada utilizando-se um ambiente virtual separado somente para o doccano. Portanto, é preciso criar tal ambiente antes da instalação.

- Para criar o ambiente siga os passos abaixo:

- Crie uma pasta para a instalação do doccano num local de sua preferência (Sugestão: chame a pasta de doccano);
- Abra um prompt de comando do Windows (ou terminal do Linux);
- Entre na pasta do doccano, **criada no passo a:**
cd <caminho completo da pasta criada>
- Crie o ambiente (neste exemplo o nome do ambiente é **doc**):
Windows: `py -m venv doc`
Linux: `python3 -m venv doc`

- e) Ative o ambiente:
Windows: `.\doc\Scripts\activate`
Linux: `source doc/bin/activate`

Obs.: Ao ativar aparecerá o nome do ambiente criado na frente do prompt de comando.
Exemplo no Windows: `(doc) C:\teste\doccano>`

- f) Instale o pacote wheel (será utilizado para construir o doccano):
`pip install wheel` (Se for Windows, pode ser que este pacote já esteja instalado)

Obs.: Mantenha o prompt (ou terminal Linux) aberto para a instalação do doccano!

Fonte: <https://packaging.python.org/guides/installing-using-pip-and-virtual-environments/#creating-a-virtual-environment>

Instalação do doccano

- Depois do ambiente criado e de dentro do ambiente ativado no prompt, rode os comandos abaixo para instalação do doccano:

```
pip install doccano
```

```
doccano init
```

```
doccano createuser --username admin --password pass@123_doc
```

Se tudo correr bem, o doccano estará instalado e com um usuário para acesso.

Obs.: Caso aconteça algum erro nesse processo, verifique as saídas no prompt para mais detalhes, solucione o problema e repita o processo de instalação.

- Saia do ambiente virtual:
`deactivate`

- Feche o prompt de comando (ou terminal Linux).

Fonte: <https://github.com/doccano/doccano>

Iniciando o doccano

Toda vez que for utilizar o doccano será necessário executar os passos abaixo para subir a aplicação (caso já não esteja em execução). A aplicação ficará em execução enquanto os prompts abertos não forem finalizados.

- a) Abra um prompt de comando do Windows (ou terminal do Linux):
- Entre na pasta do doccano, que foi criada anteriormente:
`cd <caminho completo da pasta criada>`

- Ative o ambiente criado:
Windows: `.\doc\Scripts\activate`
Linux: `source doc/bin/activate`

- Rode o comando:
`doccano webserver --port 8000`

- Mantenha o terminal aberto, só feche quando for encerrar o doccano.

b) Abra um **outro** prompt de comando do Windows (ou terminal do Linux):

- Entre na pasta do doccano, que foi criada anteriormente:
`cd <caminho completo da pasta criada>`

- Ative o ambiente criado:
Windows: `.\doc\Scripts\activate`
Linux: `source doc/bin/activate`

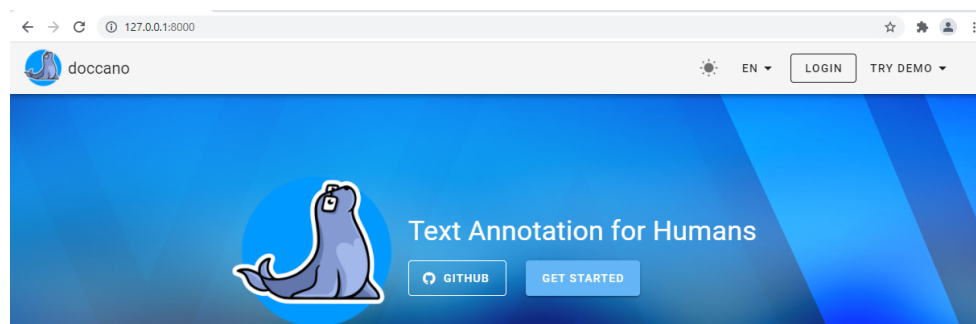
- Rode o comando:
`doccano task`

- Mantenha o terminal aberto, só feche quando for encerrar o doccano.

c) Abra um navegador e acesse o endereço:

<http://127.0.0.1:8000/>

Se tudo estiver ok, aparecerá a interface WEB do doccano (Para efetuar *login* utilize o usuário **admin** e senha **pass@123_doc**):



Finalizando o doccano

Para finalizar o sistema, **feche a aba do navegador onde o doccano está rodando** e execute os passos abaixo em cada um dos dois terminais abertos:

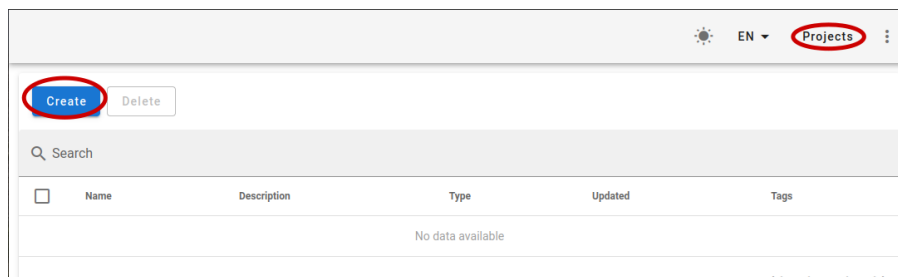
- Dê um `CTRL+C`
- Dê o comando `deactivate` (para desativar o ambiente virtual python)
- Feche o terminal.

Utilização básica

Para utilizar o doccano, depois de iniciá-lo, utilize um navegador e acesse o endereço <http://127.0.0.1:8000/>, faça login com o usuário **admin** e senha **pass@123_doc**. A seguir algumas ações que devem ser feitas para possibilitar as anotações de textos na ferramenta.


Criar projeto

Para realizar as anotações é necessário criar um projeto. Clique em **Projects** e depois em **Create**.




Preencha os dados do projeto, conforme tela abaixo, e clique em **Save**. Obs.: Nesse exemplo, foi escolhido um tipo de projeto (campo *Project Type*) específico para anotação de entidades nomeadas.


Add Project

 Project name

REN

 Description

Projeto de REN

 Project type

Sequence Labeling

☐ Allow overlapping entity

Tokyo National Museum

•LOC

o

ORG

I

LOC

f

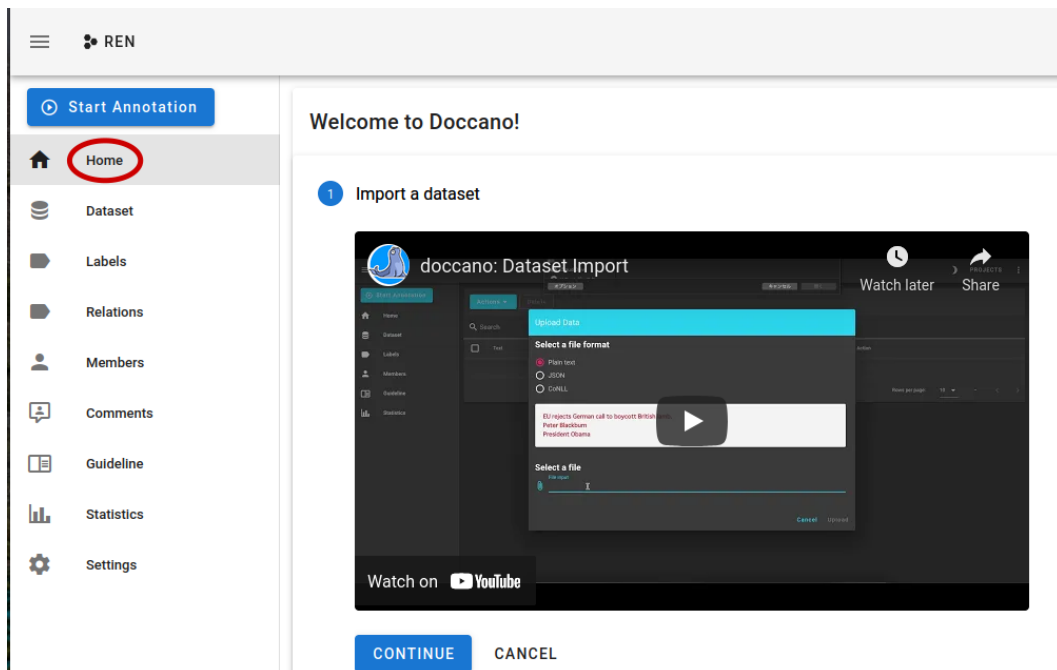
LOC

☐ Count [grapheme clusters](#) as one character☐ Randomize document order☐ Share annotations across all users

Cancel

Save

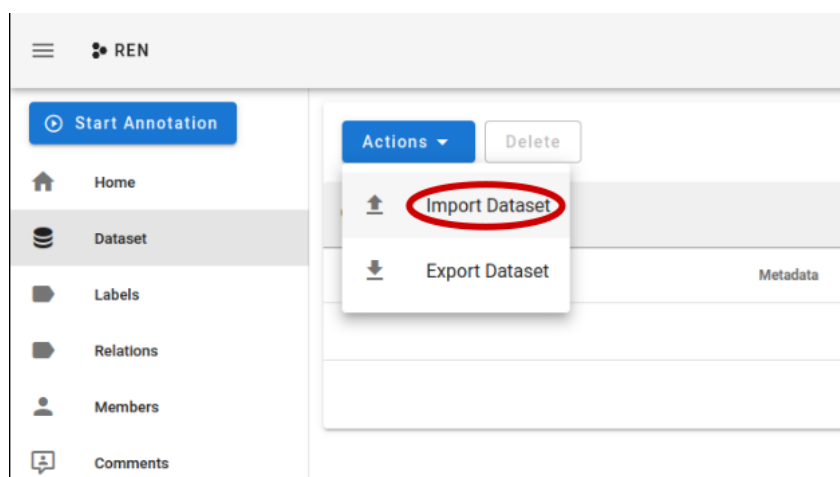
Depois do projeto criado, o sistema entra automaticamente nele. Na **Home** do projeto é possível assistir vídeos que ensinam os passos básicos para interação com o sistema. Apesar disso, este tutorial cobrirá como fazer essa interação básica.



Importar dataset

Após a criação do projeto é necessário importar um *dataset* que será utilizado para fazer as anotações das entidades.

Clique em **Dataset**, **Actions** e escolha **Import Dataset**.

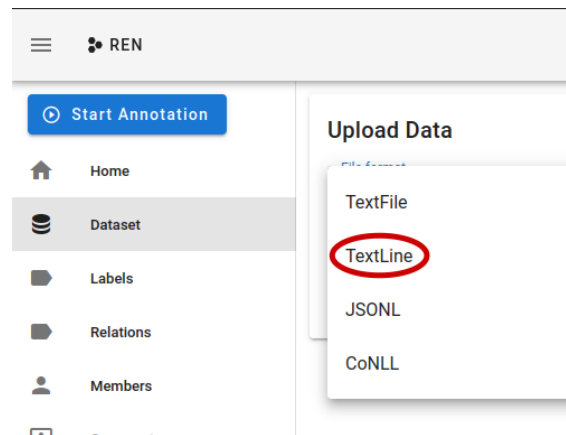


Clique na seta para baixo e escolha o formato do arquivo como **TextLine**.

Upload Data

File format

Ingest

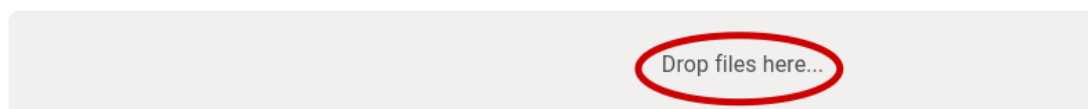


Selecione a codificação utf_8 (pode ser que o arquivo possua outra codificação, mas geralmente vem como UTF-8).

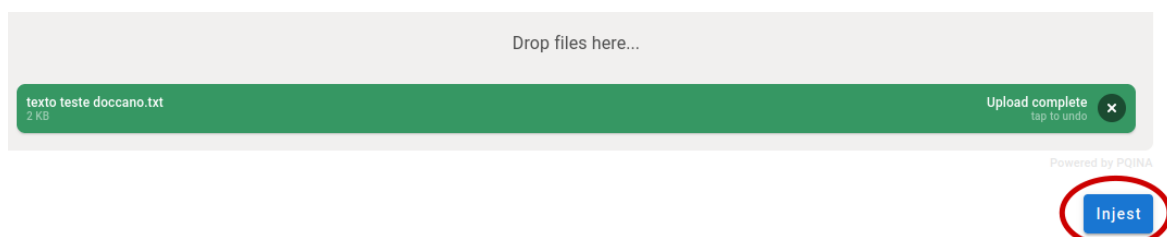
Encoding

utf_8

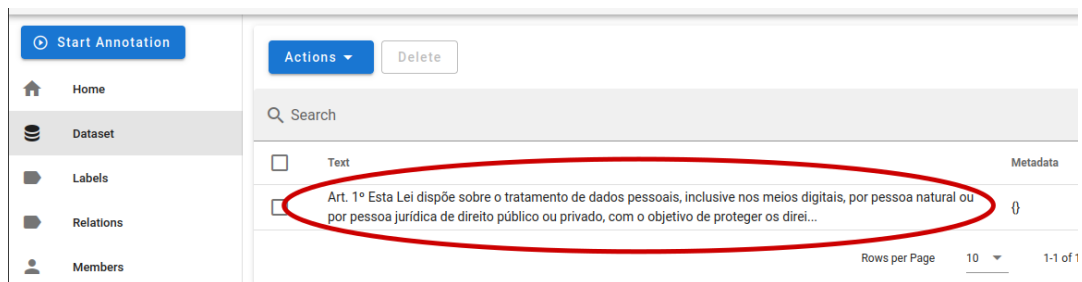
Clique em **Drop files here...** (ou arraste o arquivo e solte em cima do campo) e escolha o arquivo que contém o *dataset*.



Quando a carga do arquivo estiver completa, clique em **Ingest**.

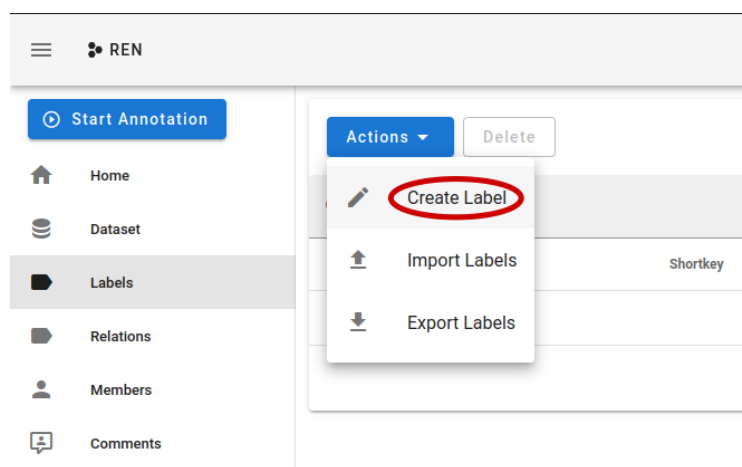


Clique em **Dataset** e confira se o arquivo carregado aparece corretamente.

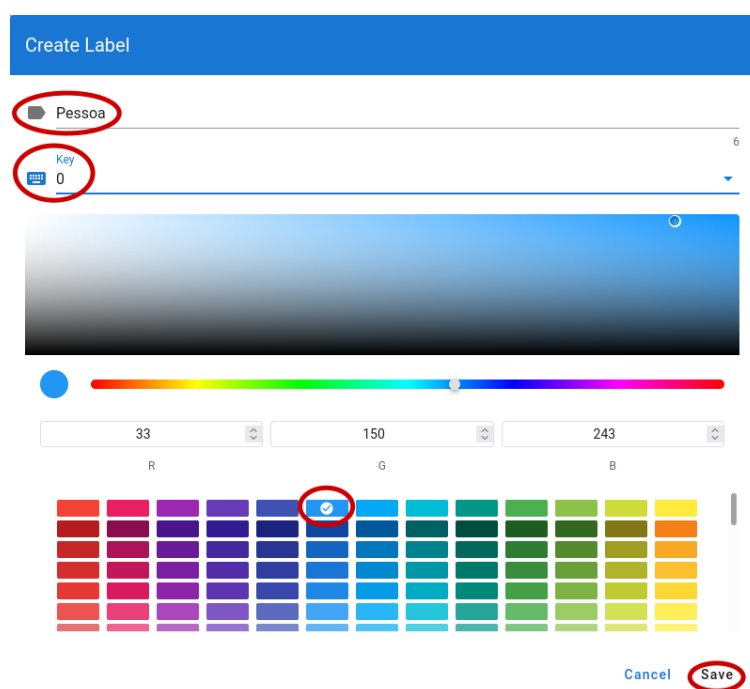


Criar labels

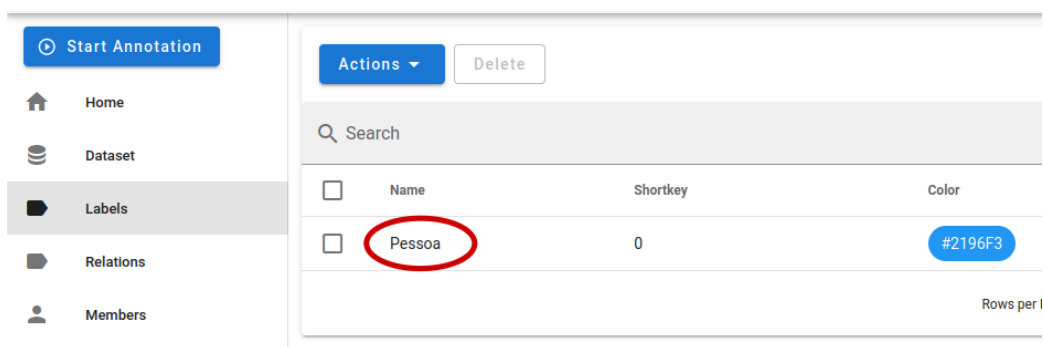
Para criar os *labels* que serão utilizados nas anotações vá em **Labels**, **Actions** e selecione **Create Label**.



Dê um nome para o label (nesse exemplo foi dado o nome Pessoa), escolha um atalho (**Key**), escolha uma cor e clique em **Save**.



Certifique-se que o label foi criado, caso não apareça, dê um **F5** para atualizar a página

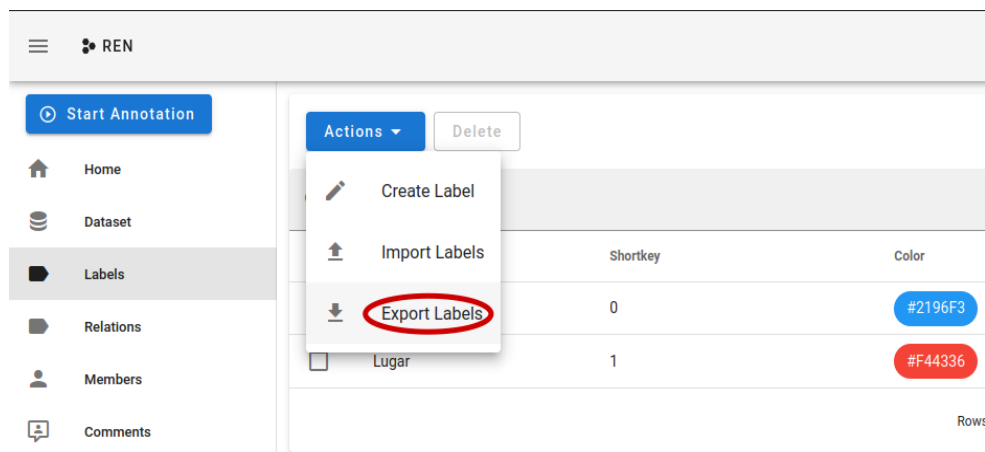


Exportar labels

Há a possibilidade de exportar os *labels* criados. Dessa forma é possível fazer um backup dos *labels* e importá-los caso haja necessidade. Outra possibilidade é exportar para importar em outra instalação da ferramenta.

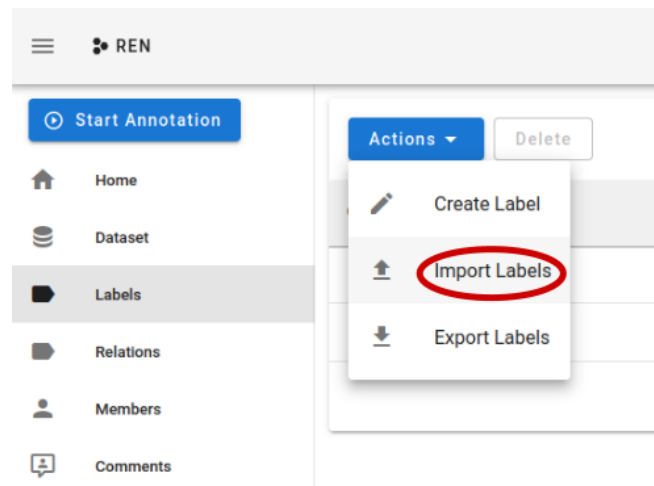
Para exportar vá em **Labels**, **Actions** e clique em **Export Labels**. O arquivo será salvo no formato JSON (<https://www.json.org/json-en.html>).

Obs.: Por padrão o arquivo será salvo na pasta que estiver configurada para receber os downloads feitos através do navegador.

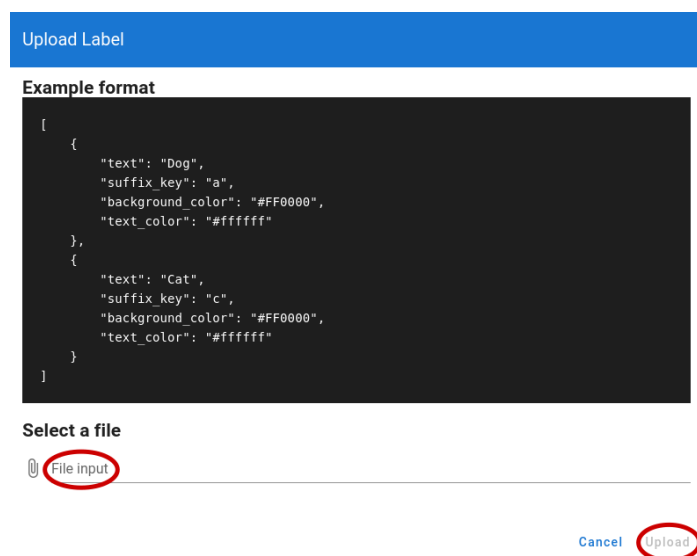


Importar labels

Para importar os *labels* vá em **Labels**, **Actions** e clique em **Import Labels**.

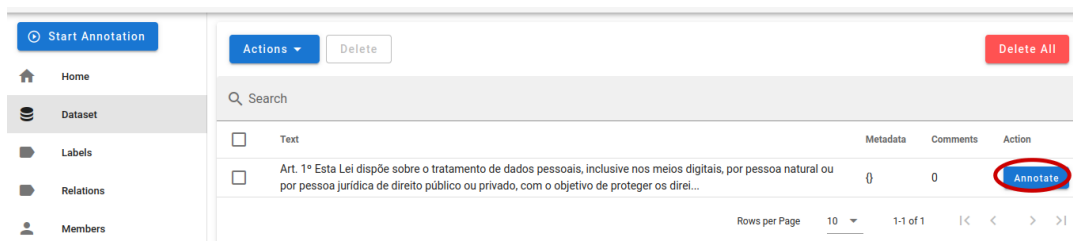


Clique no campo **File Input**, escolha o arquivo JSON com os *labels* e clique em **Upload**.

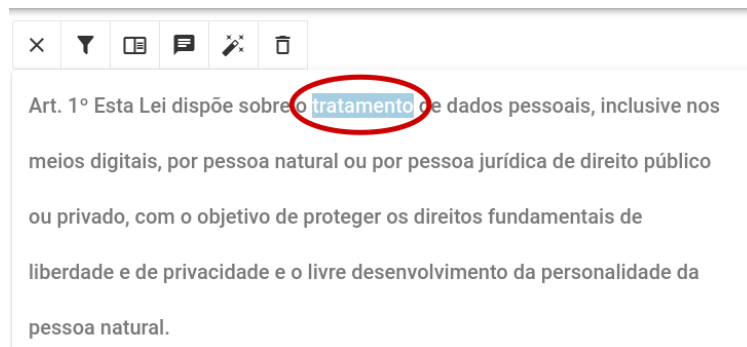


Fazer anotações

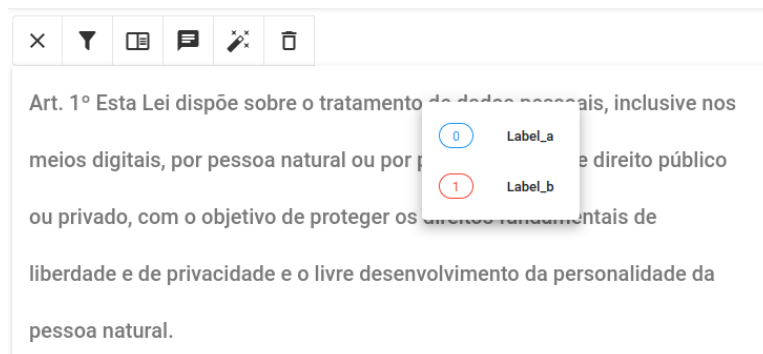
Para fazer anotações vá em **Dataset**, escolha em qual *dataset* as anotações serão feitas e clique em **Annotate**.



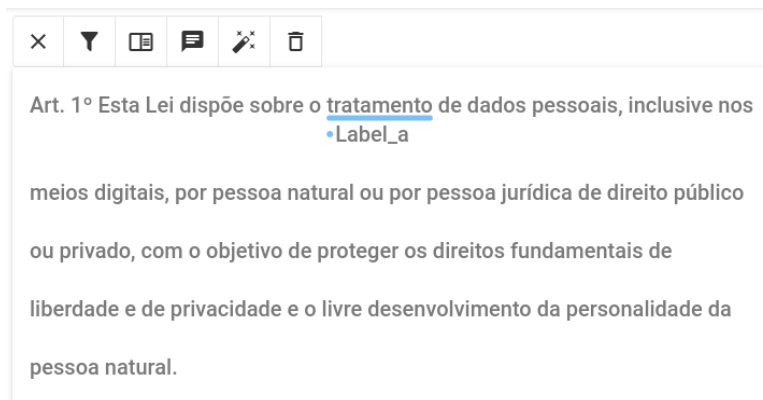
Um editor será aberto para que as anotações sejam feitas. Para anotar uma palavra basta **selecioná-la** ou dar um **clique duplo** em cima dela. No exemplo a seguir foi escolhido selecionar a palavra.



Logo em seguida, após a seleção ou clique duplo, aparecerá um menu com os *labels* para que seja feita a anotação.



Escolha um label de acordo com a classe da palavra. Obs.: O label pode ser escolhido utilizando o atalho indicado na frente dele, por exemplo o label **Label_a** tem como atalho a tecla **0**.



Caso deseje anotar uma entidade que seja composta por mais de uma palavra, por exemplo um nome próprio composto, basta selecionar as duas ou mais palavras de uma só vez.

Art. 1º Esta Lei dispõe sobre o tratamento de dados pessoais, inclusive nos meios digitais, por pessoa natural ou por pessoa jurídica de direito público ou privado, com o objetivo de proteger os direitos fundamentais de liberdade e de privacidade e o livre desenvolvimento da personalidade da pessoa natural.

Art. 1º Esta Lei dispõe sobre o tratamento de dados pessoais, inclusive nos meios digitais, por pessoa natural ou por pessoa jurídica de direito público
•Label_b
ou privado, com o objetivo de proteger os direitos fundamentais de liberdade e de privacidade e o livre desenvolvimento da personalidade da pessoa natural.

Obs.: Se as palavras que serão selecionadas estiverem em linhas separadas, há um *bug* no sistema que não deixa anotá-las juntas. Para contornar isso, **diminua o zoom** do navegador até que as palavras a serem anotadas **fiquem na mesma linha**. Depois de anotado, o zoom pode ser restaurado para o valor normal que a anotação não será perdida.

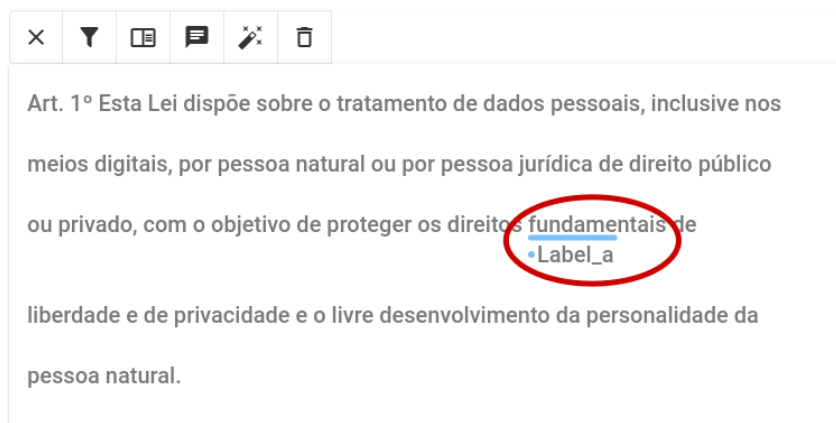
Para trocar um label de uma palavra, clique em cima do label e escolha outro.

Art. 1º Esta Lei dispõe sobre o tratamento de dados pessoais, inclusive nos meios digitais, por pessoa natural ou por pessoa jurídica de direito público
•Label_a
ou privado, com o objetivo de proteger os direitos funda
liberdade e de privacidade e o livre desenvolvimento da personalidade da
pessoa natural.

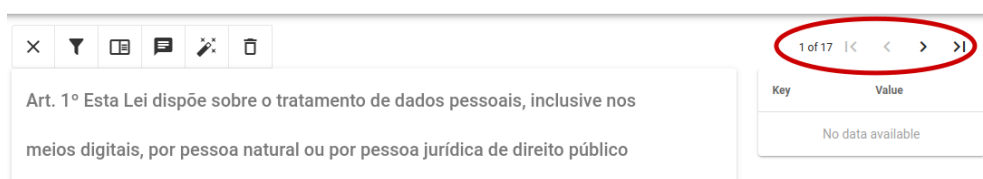
Para **apagar um label atribuído a uma entidade**, clique com o botão direito do mouse em cima do label que o mesmo será apagado imediatamente.

ATENÇÃO: Muito cuidado para não selecionar **somente um pedaço da palavra**! Se acontecer isso, será necessário apagar o label e refazer a anotação.

No exemplo abaixo, a anotação está **incorreta**. Deveria ser na palavra **fundamentais** porém foi somente em parte da palavra, que nesse caso foi **fundame**.

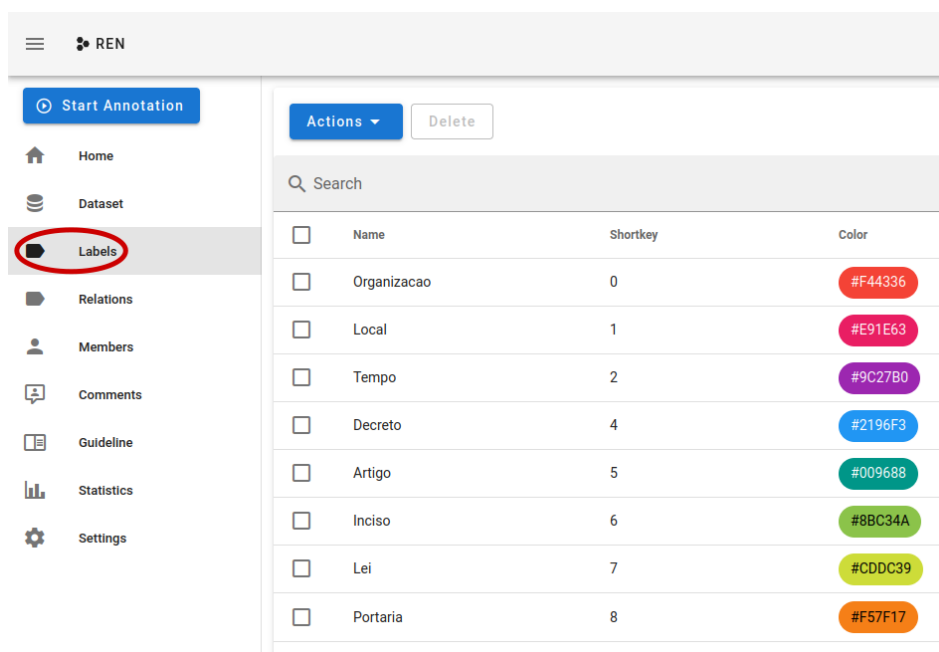


Depois de anotar um documento ou frase, caso tenha mais para ser anotado, é preciso navegar nos controles do editor para o próximo documento que receberá a anotação.



Exemplos de anotações

A seguir alguns exemplos de anotações utilizando os *labels* reais que foram importados no sistema. Pode-se verificar no menu **Labels** quais são as opções de *labels* existentes:



As anotações serão feitas com base nesses *labels*. Abaixo um exemplo de texto genérico anotado:

Neste arquivo , criado em 21/01/2022 às 19:51hs , tem alguns exemplos de entidades anotadas .

•Tempo

Se um token contendo hora (como 20:30hs , por exemplo) estiver distante da data (27/01/2022 ou 27 de janeiro de 2022 , etc.) deve ser anotado separadamente .

No caso de organização podem ser anotados nomes de empresas , órgãos , etc. tais como: Supermercados

•Organizacao

Perim, Ministério da Educação .

•Organizacao

Vitória-ES, Serra-ES e Praça Costa Pereira são exemplos de locais .

•Local •Local •Local

Já os exemplos de Decreto, Artigo, Inciso, Lei e Portaria podem ser encontrados abaixo:


Decreto Estadual 154/2022 , Artigo 8º , Art. 1º ,

•Decreto •Artigo •Artigo

inciso III , Lei 4.320/1964 e Portaria nº 512 .

•Inciso •Lei •Portaria

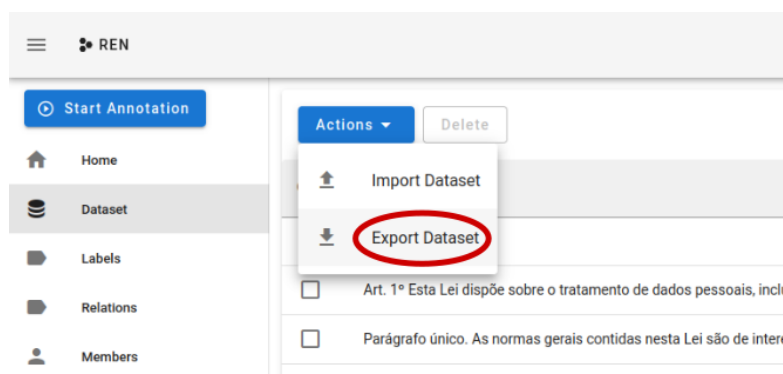
□



Exportar datasets

Depois de anotar os *datasets* é possível exportá-los para que sejam carregados em outra instalação do sistema ou utilizado para alimentar *scripts* de PLN (processamento de linguagem natural), etc.

Vá em **Dataset**, **Actions** e escolha **Export Dataset**.



Escolha o formato JSONL.

Export Data

Select a file format

☒ JSONL

Select a file name

☐ Export only approved documents

Cancel Export

Clique em **Export**.

Export Data

Select a file format

☒ JSONL

```
{ "text": "EU rejects German call to boycott British lamb.", "label": [ [0, 2, "ORG"], [2, 10, "PERSON"] ] }, { "text": "Peter Blackburn", "label": [ [0, 15, "PERSON"] ] }, { "text": "President Obama", "label": [ [10, 15, "PERSON"] ] }
```

Select a file name

☐ Export only approved documents

Cancel

Export

Obs.: Por padrão o arquivo será salvo na pasta que estiver configurada para receber os downloads feitos através do navegador.

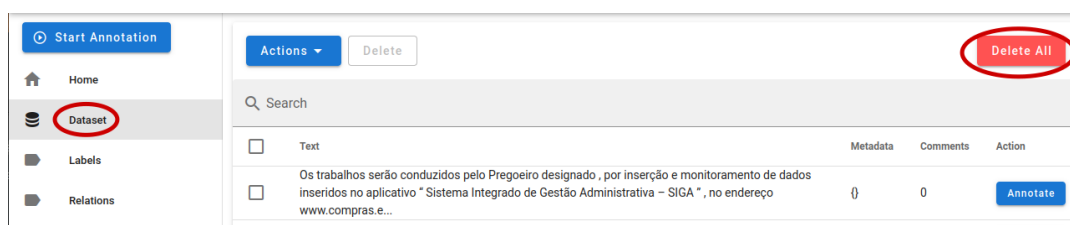
Será gerado um arquivo .zip com um nome aleatório contendo dois arquivos:

admin.jsonl – Contém as anotações realizadas pelo usuário. Neste exemplo o usuário foi o **admin**.

unknown.jsonl – Contém as sentenças que não receberam nenhuma anotação.

Apagar datasets

Para apagar todos os *datasets* do projeto e deixá-lo limpo, vá em **Datasets** e clique em **Delete All**



Responda **Yes** para confirmar a ação.

Delete All Documents

Are you sure you want to delete all documents from this project?

Cancel

Yes

Como proceder para fazer as anotações

Esta parte do tutorial é específica para instruir a atividade de anotação. Todas as funções realizadas na ferramenta já foram explicadas anteriormente no tutorial e serão somente referenciadas aqui. Caso tenha alguma dúvida, faça a releitura da respectiva seção no tutorial. Se a dúvida persistir, por favor, coloque-a no grupo de WhatsApp “**Anotação Corpus NER**” e assim que possível será sanada.

Preparando a ferramenta

Antes de iniciar as anotações, na ferramenta recém-instalada, será necessário executar os seguintes passos:

1. Acesse a aplicação ([Iniciando o doccano](#));
2. Crie um projeto ([Criar projeto](#)). Dê o nome **REN** ao projeto;
3. Importe os *labels* ([Importar labels](#)) através do arquivo chamado **label_config.json**, que foi disponibilizado.

Obs.: Estes passos devem ser executados **somente** uma vez. O intuito é preparar a ferramenta para o uso!

Realizando as anotações

Em todos os domingos, durante um prazo necessário para anotar um conjunto de documentos, serão enviados arquivos (por e-mail) com o seguinte padrão de nome:

s<n>_<nome>.txt onde:

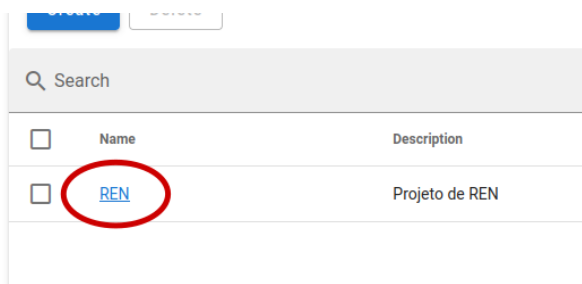
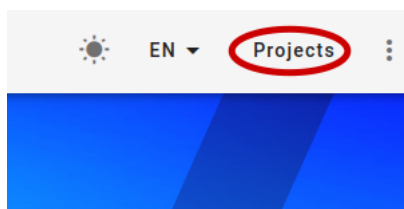
n = Número da semana em que o arquivo foi enviado

nome = primeiro nome do anotador(a) – minúsculo, sem acentos, cedilhas e caracteres especiais

Um exemplo de nome de arquivo para a anotadora Maria de Lourdes, referente à primeira semana de anotação seria: **s1_maria.txt**

Após receber estes arquivos, faça o seguinte:

1. Baixe o arquivo;
2. Acesse a aplicação ([Iniciando o doccano](#));
3. Caso não abra automaticamente no projeto, clique em **Projects** e escolha o projeto **REN**;



4. **IMPORTANTE:** Se ainda **não tiver terminado** de fazer a anotação do arquivo recebido na **semana anterior**, **termine** a anotação dos *datasets* e **execute a exportação** antes de executar este passo. Apague ([Apagar datasets](#)) todos os *datasets* que constam no projeto

REN. Esta ação é específica para este projeto e evita que os arquivos recebidos sejam misturados dentro da ferramenta e haja duplicidade na exportação de arquivos;

5. Importe o arquivo ([Importar dataset](#)). **Obs.:** Só importe o arquivo se tiver certeza que todos os *datasets* foram apagados, conforme o **passo 4**. Caso tenha importado sem apagar, execute o passo 4; e execute o passo 5 novamente;

6. Faça as anotações ([Fazer anotações](#));

Dica: As anotações não precisam ser feitas todas de um vez, caso necessário, feche a aplicação ([Finalizando o doccano](#)) e recomece a anotação numa outra oportunidade (**a partir daqui – passo 6**). Não se preocupe, o doccano salva, automaticamente, tudo o que foi anotado!

7. Exporte o arquivo anotado ([Exportar datasets](#)). **Obs.:** Será gerado um arquivo .zip com um nome aleatório (Ex. **3f4e2eba-6cc7-493c-8dbf-bccf271e8099.zip**);
8. Renomeie este arquivo, mantendo a extensão zip, para:

s<n>_<nome>_anotado.zip onde:

n = Número da semana em que o arquivo foi enviado

nome = primeiro nome do anotador(a) – minúsculo, sem acentos, cedilhas e caracteres especiais

Um exemplo de arquivo gerado por João na décima semana seria: **s10_joao_anotado.zip**

Atenção: Em hipótese alguma, altere o **conteúdo** do arquivo .zip exportado! Caso faça isso, este arquivo se tornará inválido e deverá ser apagado e gerado novamente.

9. Responda ao e-mail, onde o arquivo .txt foi recebido, anexando o arquivo .zip que foi exportado e devidamente renomeado.

Nota: O prazo para a entrega do arquivo anotado é de 7 dias corridos. Devendo ser entregue até o domingo posterior ao recebimento do arquivo txt. Cada arquivo conterà 100 sentenças (frases). Foram feitos alguns ensaios e o tempo médio para anotar um arquivo com esta quantidade de sentenças foi de aproximadamente 30 minutos ininterruptos.

A cada e-mail recebido com um arquivo .txt para ser anotado, é necessário **repetir** os passos de 1 a 9 desta seção.