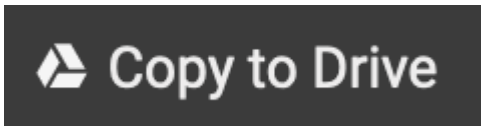
Instructions

Please make a copy of this file before continuing. You can click on the "Copy to Drive" button above. It looks like this:



1 Setup

First, we will have to install a couple of packages. `pandoc` will help us deal with Word files and avoid the encoding issues some people faced. As we are now working on a remote Notebook, we will need to downlaod the spacy models we want to use every time we use Colab. This is the same as what you did in your local Jupyter notebooks in the past.

```
pip install pypandoc --quiet
```

```
!python -m spacy download en_core_web_lg --quiet
```

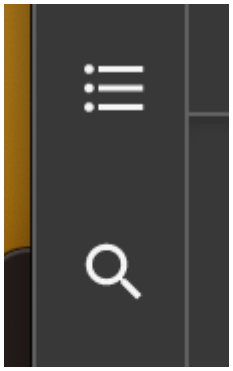
```
2023-10-17 04:14:26.839191: I tensorflow/core/platform/cpu_feature_guard.c
To enable the following instructions: AVX2 FMA, in other operations, rebui
2023-10-17 04:14:27.980879: W tensorflow/compiler/tf2tensorrt/utils/py_uti
     ━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━━ 587.7/587.7 MB 1.7 MB/s eta (
✔ Download and installation successful
You can now load the package via spacy.load('en_core_web_lg')
```

2 Loading data

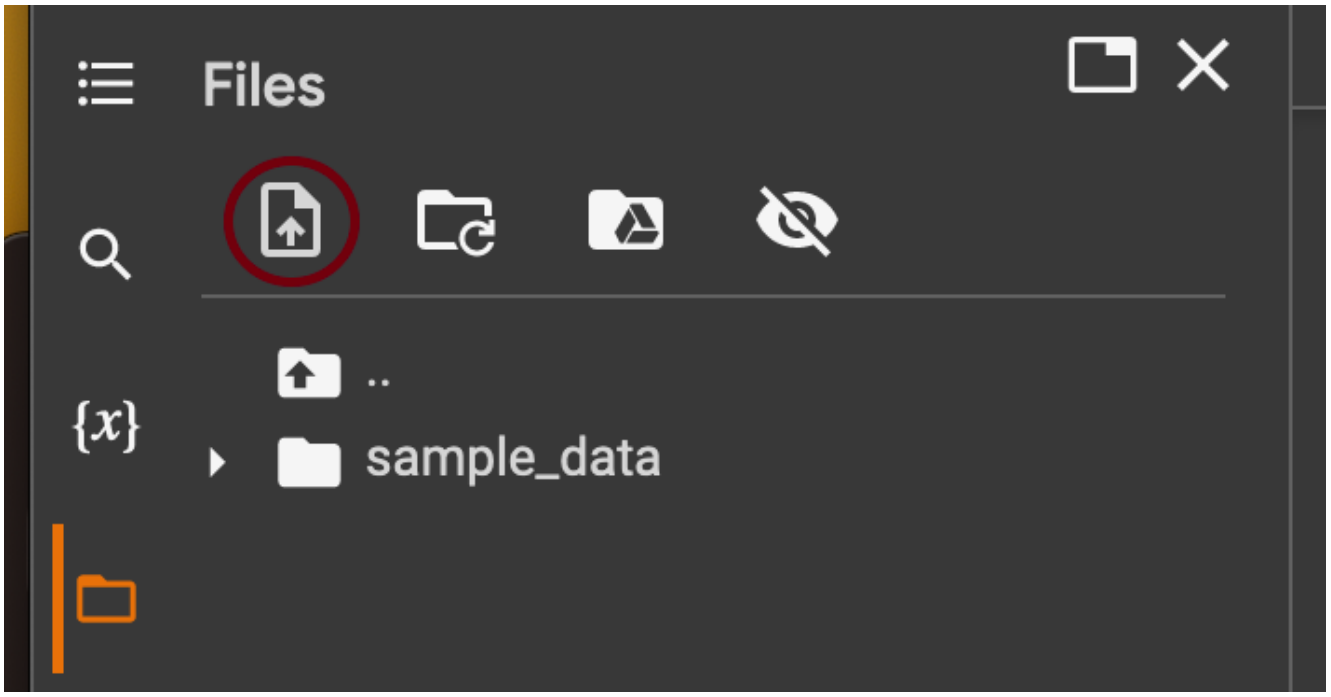
As stated in worksheet 3.2, each of you must select an academic essay that you have written (or some other piece of writing you have done).

You can upload files to Google colab in this way:

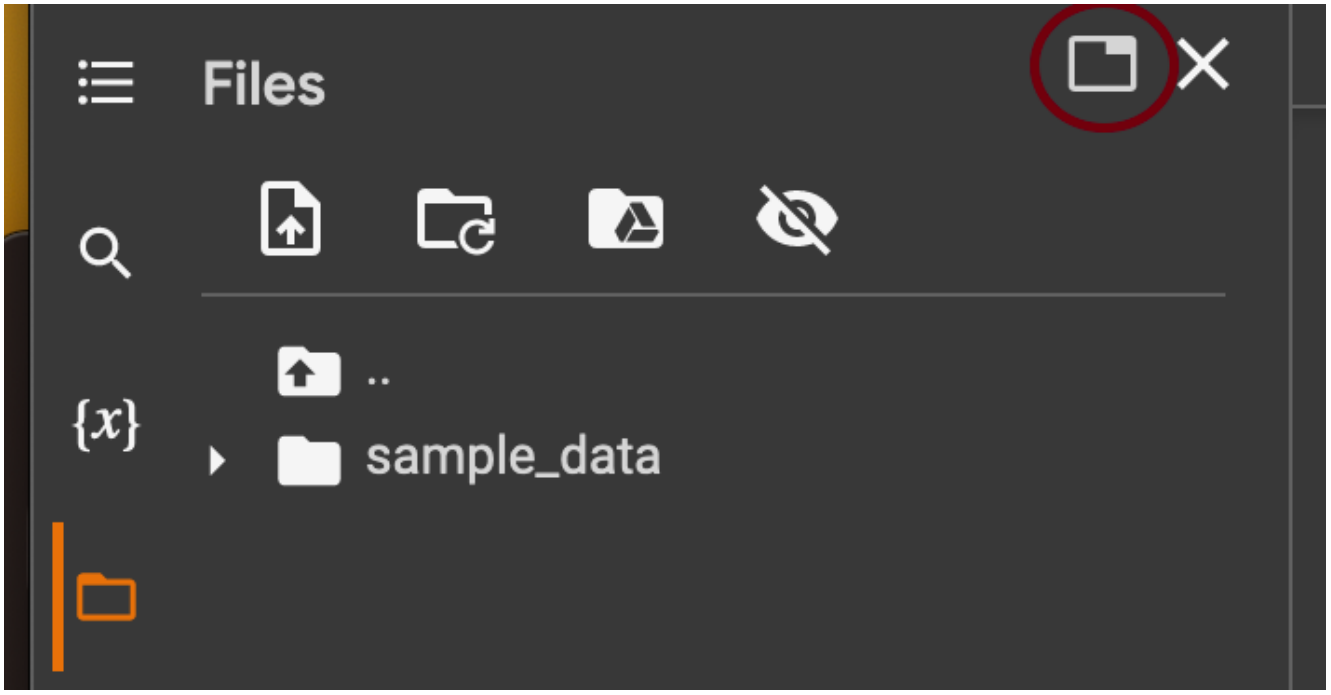
1) Click on the Files button on the left of the notebook.



2) Click upload and select a file.



3) If you want to organize your uploads into subfolders, you can click on the "add a new folder" button.



3 Loading modules

Then we will import `pydoc`, `files` from Google Colab to upload files from our computer, as well as `spacy` and its `Matcher` class.

```
import pypandoc
from google.colab import files
import spacy
from spacy.matcher import Matcher
```

4 Converting the data

This is an additional step we didn't see before. We will convert our word file to a txt file to avoid encoding errors. Instead of "sample.docx" use the name of the file you uploaded in step #2.

```
pypandoc.convert_file('sample.docx', 'plain', outputfile="sample.txt")
```

5 Analysis. This is the same as what we say in week 5.

```
nlp  = spacy.load("en_core_web_lg")
```

Make sure you replace "sample.txt" with the name you gave to your output file (unless you didn't change this).

```
with open("Midterm Essay.txt") as f:
    text = f.read()
```

```
doc = nlp(text)
```

```
matcher = Matcher(nlp.vocab)
```

Ok, so now it is time for you to add your own code! What is the ratio of nouns-per-verbs in your essay?

```
count_noun = 0
count_verb = 0
for token in doc:
  if token.pos_ == "NOUN":
    count_noun += 1
  if token.pos_ == "VERB":
```

```
    if token.pos_ == "VERB":
      count_verb += 1
print("Noun per Verb:", count_noun/count_verb)
```

```
Noun per Verb: 1.945578231292517
```

Now check with your friends. Who amongst you had the highest ratio of nouns-per-verbs? Each person in the team must submit your own Jupyter notebook to Canvas. This file must include the code, and a markdown cell with your group's answer to the question. Upload the notebook as an HTML file. Make sure everyone of you understands all the steps required to complete this worksheet.