

## ✓ Various Steps in NLP

We will be using a Python library called NLTK (Natural Language Toolkit).

NLTK is a powerful open source tool that provides a set of methods and algorithms to perform a wide range of NLP tasks, including tokenizing, parts-of-speech tagging, stemming, lemmatization, and more.

## ✓ Tokenization

Tokenization refers to the procedure of splitting a sentence into its constituent parts—the words and punctuation that it is made up of.

NLTK provides a method called `word_tokenize()`, which tokenizes given text into words.

It actually separates the text into different words based on punctuation and spaces between words.

[+ Code](#)[+ Text](#)

Import the necessary libraries and download the different types of NLTK data

```
from nltk import word_tokenize, download
download(['punkt', 'averaged_perceptron_tagger', 'stopwords'])
```

```
➔ [nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
[nltk_data] Downloading package averaged_perceptron_tagger to
[nltk_data]   /root/nltk_data...
[nltk_data]   Unzipping taggers/averaged_perceptron_tagger.zip.
[nltk_data] Downloading package stopwords to /root/nltk_data...
```

```
[nltk_data] Unzipping corpora/stopwords.zip.
True
```

We need to add a sentence as input to the `word_tokenize()` method so that it performs its job.

```
def get_tokens(sentence):
    words = word_tokenize(sentence)
    return words
```

```
print(get_tokens("I am reading NLP Fundamentals."))
```

```
↩ ['I', 'am', 'reading', 'NLP', 'Fundamentals', '.']
```

## ✓ PoS Tagging

PoS refers to parts of speech.

PoS tagging refers to the process of tagging words within sentences with their respective PoS.

import the necessary libraries

```
from nltk import word_tokenize, pos_tag
```

Using `word_tokenize()` method, find the tokens in the sentence.

Please follow our [blog](#) to see more information about new features, tips and tricks, and featured notebooks such as [Analyzing a Bank Failure with Colab](#).

### 2024-06-18

- Inline AI completions are now available to free tier users.
- Reduced latency for LSP and terminal connections
- Improved quality of inline completions
- Visual improvements to switch controls across Colab
- Various bug fixes, performance and a11y improvements to the user secrets panel
- Improved tooltip UX behavior
- Improved behavior when copying data from Google Sheets and pasting in Colab
- Scroll to cell fixes for single tabbed view and jump to cell command
- Improved tab header behavior
- A11y improvements for notebook-focused cells
- Python package upgrades
  - torch 2.2.1 -> 2.3.0
  - torchaudio 2.2.1 -> 2.3.0
  - torchvision 0.17.1 -> 0.18.0
  - torchtext 0.17.1 -> 0.18.0
  - google-cloud-aiplatform 1.51.0 -> 1.56.0
  - bigframes 1.5.0 -> 1.8.0
  - regex 2023.12.25 -> 2024.5.15

### 2024-05-13

- Code actions are now supported to automatically improve and refactor code. Code actions can be triggered by the keyboard shortcut "Ctrl/⌘ + ."
- Python package upgrades
  - bigframes 1.0.0 -> 1.5.0

```
def get_tokens(sentence):
    words = word_tokenize(sentence)
    return words
```

```
words = get_tokens("I am reading NLP Fundamentals")
print(words)
```

```
→ ['I', 'am', 'reading', 'NLP', 'Fundamentals']
```

Use the pos\_tag() method.

```
words12 = "I am reading NLP Fundamentals"
print(pos_tag(words12.split()))
```

```
→ [('I', 'PRP'), ('am', 'VBP'), ('reading', 'VBG'), ('NLP', 'NNP'), ('Fundament
```



```
def get_pos(words):
    return pos_tag(words)
get_pos(words)
```

```
→ [('I', 'PRP'),
    ('am', 'VBP'),
    ('reading', 'VBG'),
    ('NLP', 'NNP'),
    ('Fundamentals', 'NNS')]
```

PRP stands for personal pronoun.

VBP stands for verb present.

VGB stands for verb gerund.

NNP stands for proper noun singular

- google-cloud-aiplatform 1.47.0 -> 1.51.0
- jax[tpu] 0.4.23 -> 0.4.26
- Python package inclusions
  - cudf 24.4.1

## 2024-04-15

- TPU v2 runtime is now available
- L4 runtime is now available for paid users
- New distributed fine-tuning Gemma tutorial on TPUs ([GitHub](#))
- Symbol rename is now supported with keyboard shortcut F2
- Fixed bug causing inability to re-upload deleted files
- Fixed breaking bug in colabtools %upload\_files\_async
- Added syntax highlighting to %%writefile cells
- Cuda dependencies that come with Torch are cached for faster downloads for packages that require Torch and its dependencies ([GitHub issue](#))
- Python package upgrades
  - bigframes 0.24.0 -> 1.0.0
  - duckdb 0.9.2 -> 0.10.1
  - google-cloud-aiplatform 1.43.0 -> 1.47.0
  - jax 0.4.23 -> 0.4.26

## 2024-03-13

- Fixed bug that sometimes caused UserSecrets to move / disappear
- Improved messaging for mounting drive in an unsupported environment ([GitHub issue](#))
- Python package upgrades
  - torch 2.1.0 -> 2.2.1
  - torchaudio 2.1.0 -> 2.2.1
  - torchvision 0.16.0 -> 0.17.1

NNS stands for noun plural.

## ✓ Stop Word Removal

Stop words are the most frequently occurring words in any language.

They are just used to support the construction of sentences and do not contribute anything to the semantics of a sentence.

Removing them will help us clean our data, making its analysis much more efficient.

Import the necessary libraries

```
from nltk import download
download('stopwords')
from nltk import word_tokenize
from nltk.corpus import stopwords
```

```
➡ [nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data] Package stopwords is already up-to-date!
```

In order to check the list of stop words provided for English, we pass it as a parameter to the words() function.

```
stop_words = stopwords.words('english')
```

```
print(stop_words)
```

```
➡ ['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're"]
```

- torchtext 0.16.0 -> 0.17.1
- PyMC 5.7.2 -> 5.10.4
- BigFrames 0.21.0 -> 0.24.0
- google-cloud-aiplatform 1.42.1 -> 1.43.0
- tornado 6.3.2 -> 6.3.3

### 2024-02-21

- Try out Gemma on [Colab](#)!
- Allow unicode in form text inputs
- Display documentation and link to source when displaying functions
- Display image-like ndarrays as images
- Improved UX around quick charts and execution error suggestions
- Released Marketplace image for the month of February ([GitHub issue](#))
- Python package upgrades
  - bigframes 0.19.2 -> 0.21.0
  - regex 2023.6.3 -> 2023.12.25
  - spacy 3.6.1 -> 3.7.4
  - beautifulsoup4 4.11.2 -> 4.12.3
  - tensorflow-probability 0.22.0 -> 0.23.0
  - google-cloud-language 2.9.1 -> 2.13.1
  - google-cloud-aiplatform 1.39.0 -> 1.42.1
  - transformers 4.35.2 -> 4.37.2
  - pyarrow 10.0.1 -> 14.0.2

### 2024-01-29

- New [Kaggle Notebooks <> Colab updates](#)! Now you can:
  - Import directly from Colab without having to download/re-upload
  - Upload via link, by pasting Google Drive or Colab URLs
  - Export & run Kaggle Notebooks on Colab with 1 click

To remove the stop words from a sentence-

Assign a string to the sentence variable.

Tokenize it into words using the `word_tokenize()` method.

```
sentence = "I am learning Python. It is one of the \"\nmost popular programming languages\"\nsentence_words = word_tokenize(sentence)
```

```
print(sentence_words)
```

```
['I', 'am', 'learning', 'Python', '.', 'It', 'is', 'one', 'of', 'the', 'most']
```

To remove the stop words, we need to loop through each word in the sentence, check whether there are any stop words, and then finally combine them to form a complete sentence.

```
def remove_stop_words(sentence_words, stop_words):
    return ' '.join([word for word in sentence_words if \
                     word not in stop_words])
```

```
print(remove_stop_words(sentence_words, stop_words))
```

```
I learning Python . It one popular programming languages
```

Add your own stop words to the stop word list

- Try these notebooks that talk to Gemini:
  - [Gemini and Stable Diffusion](#)
  - [Learning with Gemini and ChatGPT](#)
  - [Talk to Gemini with Google's Speech to Text API](#)
  - [Sell lemonade with Gemini and Sheets](#)
  - [Generate images with Gemini and Vertex](#)
- Python package upgrades
  - google-cloud-aiplatform 1.38.1 -> 1.39.0
  - bigframes 0.18.0 -> 0.19.2
  - polars 0.17.3 -> 0.20.2
  - gdown 4.6.6 -> 4.7.3 ([GitHub issue](#))
  - tensorflow-hub 0.15.0 -> 0.16.0
  - flax 0.7.5 -> 0.8.0
- Python package inclusions
  - sentencepiece 0.1.99

## 2024-01-08

- Avoid nested scrollbars for large outputs by using `google.colab.output.no_vertical_scroll()` [Example notebook](#)
- Fix [bug](#) where downloading models from Hugging Face could freeze
- Python package upgrades
  - huggingface-hub 0.19.4 -> 0.20.2
  - bigframes 0.17.0 -> 0.18.0

## 2023-12-18

- Expanded access to AI coding has arrived in Colab across 175 locales for all tiers of Colab users
- Improvements to display of ML-based inline completions (for eligible Pro/Pro+ users)
- Started a series of [notebooks](#) highlighting Gemini API capabilities
- Enable ⌘/Ctrl+L to select the full line in an editor

```
stop_words.extend(['I', 'It', 'one'])
print(remove_stop_words(sentence_words, stop_words))
```

→ learning Python . popular programming languages

## ✓ Text Normalization

Unsupported Cell Type. Double-Click to inspect/edit the content.

There are various ways of normalizing text-

1. spelling correction
2. stemming
3. lemmatization

Unsupported Cell Type. Double-Click to inspect/edit the content.

Assign a string to the sentence variable

```
sentence = "I visited the US from the UK on 22-10-18"
```

Replace-

1. "US" with "United States"
2. "UK" with "United Kingdom"
3. "18" with "2018"

- Fixed [bug](#) where we weren't correctly formatting output from multiple execution results
- Python package upgrades
  - CUDA 11.8 to CUDA 12.2
  - tensorflow 2.14.0 -> 2.15.0
  - tensorboard 2.14.0 -> 2.15.0
  - keras 2.14.0 -> 2.15.0
  - Nvidia drivers 525.105.17 -> 535.104.05
  - tensorflow-gcs-config 2.14.0 -> 2.15.0
  - bigframes 0.13.0 -> 0.17.0
  - geemap 0.28.2 -> 0.29.6
  - pyarrow 9.0.0 -> 10.0.1
  - google-generativeai 0.2.2 -> 0.3.1
  - jax 0.4.20 -> 0.4.23
  - jaxlib 0.4.20 -> 0.4.23
- Python package inclusions
  - kagglehub 0.1.4
  - google-cloud-aiplatform 1.38.1

### 2023-11-27

- Removed warning when calling await to make it render as code
- Added "Run selection" to the cell context menu
- Added highlighting for the %%python cell magic
- Launched AI coding features for Pro/Pro+ users in more locales
- Python package upgrades
  - bigframes 0.12.0 -> 0.13.0
- Python package inclusions
  - transformers 4.35.2
  - google-generativeai 0.2.2

### 2023-11-08

- Launched Secrets, for safe storage of private keys on Colab ([tweet](#))

To do so, use the `replace()` function and store the updated output in the "normalized\_sentence" variable.

```
def normalize(text):
    return text.replace("US", "United States")\
.replace("UK", "United Kingdom")\
.replace("-18", "-2018")
```

Check whether the text has been normalized

```
normalized_sentence = normalize(sentence)
print(normalized_sentence)
```

 I visited the United States from the United Kingdom on 22-10-2018

## ✓ Spelling Correction

Spelling correction is one of the most important tasks in any NLP project.

Unsupported Cell Type. Double-Click to inspect/edit the content.

Unsupported Cell Type. Double-Click to inspect/edit the content.

Import the necessary libraries

```
pip install autocorrect
```

- Fixed issue where TensorBoard would not load ([#3990](#))
- Python package upgrades
  - lightgbm 4.0.0 -> 4.1.0
  - bigframes 0.10.0 -> 0.12.0
  - bokeh 3.2.2 -> 3.3.0
  - duckdb 0.8.1 -> 0.9.1
  - numba 0.56.4 -> 0.58.1
  - tweepy 4.13.0 -> 4.14.0
  - jax 0.4.16 -> 0.4.20
  - jaxlib 0.4.16 -> 0.4.20

## 2023-10-23

- Updated the **Open notebook** dialog for better usability and support for smaller screen sizes
- Added smart paste support for data from Google Sheets for R notebooks
- Enabled showing release notes in a tab
- Launched AI coding features for Pro/Pro+ users in Australia AU Canada CA India IN and Japan JP ([tweet](#))
- Python package upgrades
  - earthengine-api 0.1.357 -> 0.1.375
  - flax 0.7.2 -> 0.7.4
  - geemap 0.27.4 -> 0.28.2
  - jax 0.4.14 -> 0.4.16
  - jaxlib 0.4.14 -> 0.4.16
  - keras 2.13.1 -> 2.14.0
  - tensorboard 2.13.0 -> 2.14.1
  - tensorflow 2.13.0 -> 2.14.0
  - tensorflow-gcs-config 2.13.0 -> 2.14.0
  - tensorflow-hub 0.14.0 -> 0.15.0
  - tensorflow-probability 0.20.1 -> 0.22.0
  - torch 2.0.1 -> 2.1.0
  - torchaudio 2.0.2 -> 2.1.0
  - torchtext 0.15.2 -> 0.16.0
  - torchvision 0.15.2 -> 0.16.0
  - xgboost 1.7.6 -> 2.0.0

```

Collecting autocorrect
  Downloading autocorrect-2.6.1.tar.gz (622 kB)
    622.8/622.8 kB 8.8 MB/s eta 0:0
  Preparing metadata (setup.py) ... done
Building wheels for collected packages: autocorrect
  Building wheel for autocorrect (setup.py) ... done
  Created wheel for autocorrect: filename=autocorrect-2.6.1-py3-none-any.whl
  Stored in directory: /root/.cache/pip/wheels/b5/7b/6d/b76b29ce11ff8e2521c8c
Successfully built autocorrect
Installing collected packages: autocorrect
Successfully installed autocorrect-2.6.1

```

```

from nltk import word_tokenize
from autocorrect import Speller

```

Unsupported Cell Type. Double-Click to inspect/edit the content.

```

spell = Speller(lang='en')
spell('Natureal')

```

```

'Natural'

```

Unsupported Cell Type. Double-Click to inspect/edit the content.

```

sentence = word_tokenize("Ntural Luanguage Processin deals with "\
"the art of extracting insights from "\
"Natural Languaes")
print(sentence)

```

```

['Ntural', 'Luanguage', 'Processin', 'deals', 'with', 'the', 'art', 'of', 'ex

```

- Python package inclusions
  - bigframes 0.10.0
  - malloy 2023.1056

## 2023-09-22

- Added the ability to scope an AI generated suggestion to a specific Pandas dataframe ([tweet](#))
- Added Colab link previews to Docs ([tweet](#))
- Added smart paste support for data from Google Sheets
- Increased font size of dropdowns in interactive forms
- Improved rendering of the notebook when printing
- Python package upgrades
  - tensorflow 2.12.0 -> 2.13.0
  - tensorboard 2.12.3 -> 2.13.0
  - keras 2.12.0 -> 2.13.1
  - tensorflow-gcs-config 2.12.0 -> 2.13.
  - scipy 1.10.1 -> 1.11.2
  - cython 0.29.6 -> 3.0.2
- Python package inclusions
  - geemap 0.26.0

## 2023-08-18

- Added "Change runtime type" to the menu in the connection button
- Improved auto-reconnection to an already running notebook ([#3764](#))
- Increased the specs of our highmem machines for Pro users
- Fixed add-apt-repository command on Ubuntu 22.04 runtime ([#3867](#))
- Python package upgrades
  - bokeh 2.4.3 -> 3.2.2
  - cmake 3.25.2 -> 3.27.2
  - cryptography 3.4.8 -> 41.0.3
  - dask 2022.12.1 -> 2023.8.0
  - distributed 2022.12.1 -> 2023.8.0



Unsupported Cell Type. Double-Click to inspect/edit the content.

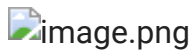
```
def correct_spelling(tokens):
    sentence_corrected = ' '.join([spell(word) \
for word in tokens])
    return sentence_corrected
print(correct_spelling(sentence))
```

➡ Natural Language Processing deals with the art of extracting insights from Na



## ✓ Stemming

Unsupported Cell Type. Double-Click to inspect/edit the content.



Unsupported Cell Type. Double-Click to inspect/edit the content.

Import the necessary libraries

```
from nltk import stem
```

Pass the words as parameters to the stem() method.

- earthengine-api 0.1.358 -> 0.1.364
- flax 0.7.0 -> 0.7.2
- ipython-sql 0.4.0 -> 0.5.0
- jax 0.4.13 -> 0.4.14
- jaxlib 0.4.13 -> 0.4.14
- lightgbm 3.3.5 -> 4.0.0
- mkl 2019.0 -> 2023.2.0
- notebook 6.4.8 -> 6.5.5
- numpy 1.22.4 -> 1.23.5
- opencv-python 4.7.0.72 -> 4.8.0.76
- pillow 8.4.0 -> 9.4.0
- plotly 5.13.1 -> 5.15.0
- prettytable 0.7.2 -> 3.8.0
- pytensor 2.10.1 -> 2.14.2
- spacy 3.5.4 -> 3.6.1
- statsmodels 0.13.5 -> 0.14.0
- xarray 2022.12.0 -> 2023.7.0
- Python package inclusions
  - PyDrive2 1.6.3

## 2023-07-21

- Launched auto-plotting for dataframes, available using the chart button that shows up alongside datatables ([post](#))



- Added a menu to the table of contents to support running a section or collapsing/expanding sections ([post](#))

```
def get_stems(word,stemmer):
    return stemmer.stem(word)
porterStem = stem.PorterStemmer()
get_stems("production",porterStem)
```

```
→ 'product'
```

```
get_stems("coming",porterStem)
```

```
→ 'come'
```

```
get_stems("firing",porterStem)
```

```
→ 'fire'
```

```
get_stems("battling",porterStem)
```

```
→ 'battl'
```

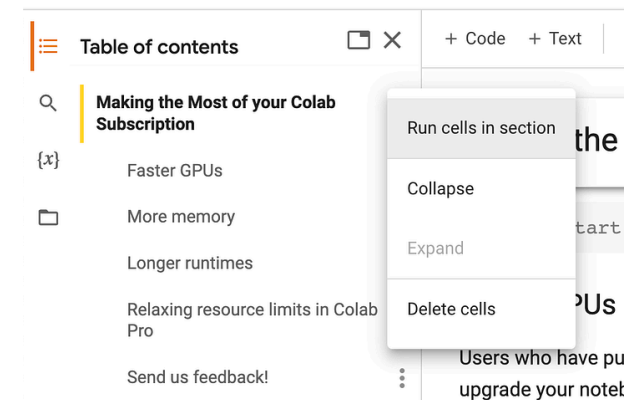
```
stemmer = stem.SnowballStemmer("english")
get_stems("battling",stemmer)
```

```
→ 'battl'
```

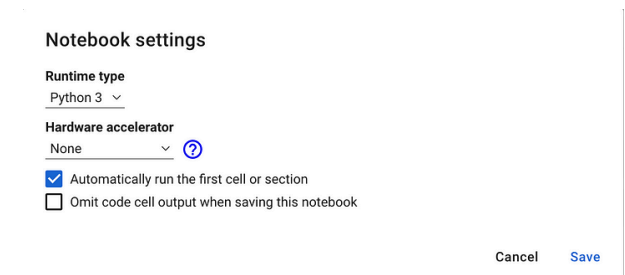
## ✓ Lemmatization

Unsupported Cell Type. Double-Click to inspect/edit the content.

Unsupported Cell Type. Double-Click to inspect/edit the content.



- Added an option to automatically run the first cell or section, available under Edit -> Notebook settings ([post](#))



- Launched Pro/Pro+ to Algeria, Argentina, Chile, Ecuador, Egypt, Ghana, Kenya, Malaysia, Nepal, Nigeria, Peru, Rwanda, Saudi Arabia, South Africa, Sri Lanka, Tunisia, and Ukraine ([tweet](#))
- Added a command, "Toggle tab moves focus" for toggling tab trapping in the editor (Tools -> Command palette, "Toggle tab moves focus")
- Fixed issue where `files.upload()` was sometimes returning an incorrect filename ([#1550](#))
- Fixed f-string syntax highlighting bug ([#3802](#))
- Disabled ambiguous characters highlighting for commonly used LaTeX characters ([#3648](#))
- Upgraded Ubuntu from 20.04 LTS to [22.04 LTS](#)
- Updated the Colab Marketplace VM image

Unsupported Cell Type. Double-Click to inspect/edit the content.

## Import the necessary libraries

```
from nltk import download
download('wordnet')
from nltk.stem.wordnet import WordNetLemmatizer
```


 [nltk\_data] Downloading package wordnet to /root/nltk\_data...

Create an object of the WordNetLemmatizer class.

```
lemmatizer = WordNetLemmatizer()
```

Bring the word to its proper form by using the lemmatize() method of the WordNetLemmatizer class.


```
def get_lemma(word):
    return lemmatizer.lemmatize(word)
get_lemma('products')
```

 'product'

```
get_lemma('production')
```

 'production'

```
get_lemma('coming')
```

 'coming'

- Python package upgrades:
  - autograd 1.6.1 -> 1.6.2
  - drivefs 76.0 -> 77.0
  - flax 0.6.11 -> 0.7.0
  - earthengine-api 0.1.357 -> 0.1.358
  - GDAL 3.3.2->3.4.3
  - google-cloud-bigquery-storage 2.20.0 -> 2.22.2
  - gspread-dataframe 3.0.8 -> 3.3.1
  - holidays 0.27.1 -> 0.29
  - jax 0.4.10 -> jax 0.4.13
  - jaxlib 0.4.10 -> jax 0.4.13
  - jupyterlab-widgets 3.0.7 -> 3.0.8
  - nbformat 5.9.0 -> 5.9.1
  - opencv-python-headless 4.7.0.72 -> 4.8.0.74
  - pygame 2.4.0 -> 2.5.0
  - spacy 3.5.3 -> 3.5.4
  - SQLAlchemy 2.0.16 -> 2.0.19
  - tabulate 0.8.10 -> 0.9.0
  - tensorflow-hub 0.13.0 -> 0.14.0

## 2023-06-23

- Launched AI coding features to subscribed users starting with Pro+ users in the US ([tweet](#), [post](#))
- Added the Kernel Selector in the Notebook Settings ([tweet](#))
- Fixed double space trimming issue in markdown [#3766](#)
- Fixed run button indicator not always centered [#3609](#)
- Fixed inconsistencies for automatic indentation on multi-line [#3697](#)
- Upgraded Python from 3.10.11 to 3.10.12
- Python package updates:
  - duckdb 0.7.1 -> 0.8.1
  - earthengine-api 0.1.350 -> 0.1.357
  - flax 0.6.9 -> 0.6.11
  - google-cloud-bigquery 3.9.0 -> 3.10.0

## ✓ Named Entity Recognition (NER)

Unsupported Cell Type. Double-Click to inspect/edit the content.

Import the necessary libraries

```
from nltk import download
from nltk import pos_tag
from nltk import ne_chunk
from nltk import word_tokenize
download('maxent_ne_chunker')
download('words')
```

```
↳ [nltk_data] Downloading package maxent_ne_chunker to
[nltk_data] /root/nltk_data...
[nltk_data] Unzipping chunkers/maxent_ne_chunker.zip.
[nltk_data] Downloading package words to /root/nltk_data...
[nltk_data] Unzipping corpora/words.zip.
True
```

Declare the **sentence** variable and assign it a string.

```
sentence = "We are reading a book published by Packt "\
"which is based out of Birmingham."
```

Find the named entities from the preceding text

- google-cloud-bigquery-storage 2.19.1 -> 2.20.0
- grpcio 1.54.0 -> 1.56.0
- holidays 0.25 -> 0.27.1
- nbformat 5.8.0 -> 5.9.0
- prophet 1.1.3 -> 1.1.4
- pydata-google-auth 1.7.0 -> 1.8.0
- spacy 3.5.2 -> 3.5.3
- tensorboard 2.12.2 -> 2.12.3
- xgboost 1.7.5 -> 1.7.6
- Python package inclusions:
  - gcsfs 2023.6.0
  - geopandas 0.13.2
  - google-cloud-bigquery-connection 1.12.0
  - google-cloud-functions 1.13.0
  - grpc-google-iam-v1 0.12.6
  - multidict 6.0.4
  - tensorboard-data-server 0.7.1

### 2023-06-02

- Released the new site [colab.google](https://colab.google)
- Published Colab's Docker runtime image to [us-docker.pkg.dev/colab-images/public/runtime](https://docker.pkg.dev/colab-images/public/runtime) ([tweet](#) [instructions](#))
- Launched support for Google children accounts ([tweet](#))
- Launched DagsHub integration ([tweet](#), [post](#))
- Upgraded to Monaco Editor Version 0.37.1
- Fixed various Vim keybinding bugs
- Fixed issue where the N and P letters sometimes couldn't be typed ([#3664](#))
- Fixed rendering support for compositional inputs ([#3660](#), [#3679](#))
- Fixed lag in notebooks with lots of cells ([#3676](#))
- Improved support for R by adding a Runtime type notebook setting (Edit -> Notebook settings)
- Improved documentation for connecting to a local runtime (Connect -> Connect to a local runtime)
- Python package updates:

```
def get_ner(text):
    i = ne_chunk(pos_tag(word_tokenize(text)), binary=True)
    return [a for a in i if len(a)==1]
get_ner(sentence)
```

```
→ [Tree('NE', [('Packt', 'NNP')]), Tree('NE', [('Birmingham', 'NNP')])]
```

## ✓ Activity: Preprocessing of Raw Text

Follow these steps to implement this activity:

1. Import the necessary libraries.
2. Load the text corpus to a variable.
3. Apply the tokenization process to the text corpus and print the first 20 tokens.
4. Apply spelling correction on each token and print the initial 20 corrected tokens as well as the corrected text corpus.
5. Apply PoS tags to each of the corrected tokens and print them.
6. Remove stop words from the corrected token list and print the initial 20 tokens.
7. Apply stemming and lemmatization to the corrected token list and then print the initial 20 tokens.
8. Detect the sentence boundaries in the given text corpus and print the total number of sentences.

Start coding or generate with AI.

- holidays 0.23 -> 0.25
- jax 0.4.8 -> 0.4.10
- jaxlib 0.4.8 -> 0.4.10
- pip 23.0.1 -> 23.1.2
- tensorflow-probability 0.19.0 -> 0.20.1
- torch 2.0.0 -> 2.0.1
- torchaudio 2.0.1 -> 2.0.2
- torchdata 0.6.0 -> 0.6.1
- torchtext 0.15.1 -> 0.15.2
- torchvision 0.15.1 -> 0.15.2
- tornado 6.2 -> 6.3.1

## 2023-05-05

- Released GPU type selection for paid users, allowing them to choose a preferred NVidia GPU
- Upgraded R from 4.2.3 to 4.3.0
- Upgraded Python from 3.9.16 to 3.10.11
- Python package updates:
  - attrs 22.2.0 -> attrs 23.1.0
  - earthengine-api 0.1.349 -> earthengine-api 0.1.350
  - flax 0.6.8 -> 0.6.9
  - grpcio 1.53.0 -> 1.54.0
  - nbclient 0.7.3 -> 0.7.4
  - tensorflow-datasets 4.8.3 -> 4.9.2
  - termcolor 2.2.0 -> 2.3.0
  - zict 2.2.0 -> 3.0.0

## 2023-04-14

- Python package updates:
  - google-api-python-client 2.70.0 -> 2.84.0
  - google-auth-oauthlib 0.4.6 -> 1.0.0
  - google-cloud-bigquery 3.4.2 -> 3.9.0
  - google-cloud-datastore 2.11.1 -> 2.15.1
  - google-cloud-firestore 2.7.3 -> 2.11.0
  - google-cloud-language 2.6.1 -> 2.9.1
  - google-cloud-storage 2.7.0 -> 2.8.0
  - google-cloud-translate 3.8.4 -> 3.11.1
  - networkx 3.0 -> 3.1

- notebook 6.3.0 -> 6.4.8
- jax 0.4.7 -> 0.4.8
- pandas 1.4.4 -> 1.5.3
- spacy 3.5.1 -> 3.5.2
- SQLAlchemy 1.4.47 -> 2.0.9
- xgboost 1.7.4 -> 1.7.5

## 2023-03-31

- Improve bash ! syntax highlighting ([GitHub issue](#))
- Fix bug where VIM keybindings weren't working in the file editor
- Upgraded R from 4.2.2 to 4.2.3
- Python package updates:
  - arviz 0.12.1 -> 0.15.1
  - astropy 4.3.1 -> 5.2.2
  - dopamine-rl 1.0.5 -> 4.0.6
  - gensim 3.6.0 -> 4.3.1
  - ipykernel 5.3.4 -> 5.5.6
  - ipython 7.9.0 -> 7.34.0
  - jax 0.4.4 -> 0.4.7
  - jaxlib 0.4.4 -> 0.4.7
  - jupyter\_core 5.2.0 -> 5.3.0
  - keras 2.11.0 -> 2.12.0
  - lightgbm 2.2.3 -> 3.3.5
  - matplotlib 3.5.3 -> 3.7.1
  - nltk 3.7 -> 3.8.1
  - opencv-python 4.6.0.66 -> 4.7.0.72
  - plotly 5.5.0 -> 5.13.1
  - pymc 4.1.4 -> 5.1.2
  - seaborn 0.11.2 -> 0.12.2
  - spacy 3.4.4 -> 3.5.1
  - sympy 1.7.1 -> 1.11.1
  - tensorboard 2.11.2 -> 2.12.0
  - tensorflow 2.11.0 -> 2.12.0
  - tensorflow-estimator 2.11.0 -> 2.12.0
  - tensorflow-hub 0.12.0 -> 0.13.0
  - torch 1.13.1 -> 2.0.0
  - torchaudio 0.13.1 -> 2.0.1
  - torchtext 0.14.1 -> 0.15.1

- torchvision 0.14.1 -> 0.15.1

## 2023-03-10

- Added the [Colab editor shortcuts](#) example notebook
- Fixed triggering of @-mention and email autocomplete for large comments ([GitHub issue](#))
- Added View Resources to the Runtime menu
- Made file viewer images fit the view by default, resizing to original size on click
- When in VIM mode, enable copy as well as allowing propagation to monaco-vim to escape visual mode ([GitHub issue](#))
- Upgraded CUDA 11.6.2 -> 11.8.0 and cuDNN 8.4.0.27 -> 8.7.0.84
- Upgraded Nvidia drivers 525.78.01 -> 530.30.02
- Upgraded Python 3.8.10 -> 3.9.16
- Python package updates:
  - beautifulsoup4 4.6.3 -> 4.9.3
  - bokeh 2.3.3 -> 2.4.3
  - debugpy 1.0.0 -> 1.6.6
  - Flask 1.1.4 -> 2.2.3
  - jax 0.3.25 -> 0.4.4
  - jaxlib 0.3.25 -> 0.4.4
  - Jinja2 2.11.3 -> 3.1.2
  - matplotlib 3.2.2 -> 3.5.3
  - nbconvert 5.6.1 -> 6.5.4
  - pandas 1.3.5 -> 1.4.4
  - pandas-datareader 0.9.0 -> 0.10.0
  - pandas-profiling 1.4.1 -> 3.2.0
  - Pillow 7.1.2 -> 8.4.0
  - plotnine 0.8.0 -> 0.10.1
  - scikit-image 0.18.3 -> 0.19.3
  - scikit-learn 1.0.2 -> 1.2.2
  - scipy 1.7.3 -> 1.10.1
  - setuptools 57.4.0 -> 63.4.3
  - sklearn-pandas 1.8.0 -> 2.2.0
  - statsmodels 0.12.2 -> 0.13.5
  - urllib3 1.24.3 -> 1.26.14
  - Werkzeug 1.0.1 -> 2.2.3

- wrapt 1.14.1 -> 1.15.0
- xgboost 0.90 -> 1.7.4
- xlrd 1.2.0 -> 2.0.1

## 2023-02-17

- Show graphs of RAM and disk usage in notebook toolbar
- Copy cell links directly to the clipboard instead of showing a dialog when clicking on the link icon in the cell toolbar
- Updated the [Colab Marketplace VM image](#)
- Upgraded CUDA to 11.6.2 and cuDNN to 8.4.0.27
- Python package updates:
  - tensorflow 2.9.2 -> 2.11.0
  - tensorboard 2.9.1 -> 2.11.2
  - keras 2.9.0 -> 2.11.0
  - tensorflow-estimator 2.9.0 -> 2.11.0
  - tensorflow-probability 0.17.0 -> 0.19.0
  - tensorflow-gcs-config 2.9.0 -> 2.11.0
  - earthengine-api 0.1.339 -> 0.1.341
  - flatbuffers 1.12 -> 23.1.21
  - platformdirs 2.6.2 -> 3.0.0
  - pydata-google-auth 1.6.0 -> 1.7.0
  - python-utils 3.4.5 -> 3.5.2
  - tenacity 8.1.0 -> 8.2.1
  - tiff file 2023.1.23.1 -> 2023.2.3
  - notebook 5.7.16 -> 6.3.0
  - tornado 6.0.4 -> 6.2
  - aiohttp 3.8.3 -> 3.8.4
  - charset-normalizer 2.1.1 -> 3.0.1
  - fastai 2.7.0 -> 2.7.1
  - soundfile 0.11.0 -> 0.12.1
  - typing-extensions 4.4.0 -> 4.5.0
  - widgetsnbextension 3.6.1 -> 3.6.2
  - pydantic 1.10.4 -> 1.10.5
  - zipp 3.12.0 -> 3.13.0
  - numpy 1.21.6 -> 1.22.4
  - drivefs 66.0 -> 69.0
  - gdal 3.0.4 -> 3.3.2 [GitHub issue](#)



- Added libudunits2-dev for smoother R package installs [GitHub issue](#)

## 2023-02-03

- Improved tooltips for pandas series to show common statistics about the series object
- Made the forms dropdown behave like an autocomplete box when it allows input
- Updated the nvidia driver from 460.32.03 to 510.47.03
- Python package updates:
  - absl-py 1.3.0 -> 1.4.0
  - bleach 5.0.1 -> 6.0.0
  - cachetools 5.2.1 -> 5.3.0
  - cmdstanpy 1.0.8 -> 1.1.0
  - dnspython 2.2.1 -> 2.3.0
  - fsspec 2022.11.0 -> 2023.1.0
  - google-cloud-bigquery-storage 2.17.0 -> 2.18.1
  - holidays 0.18 -> 0.19
  - jupyter-core 5.1.3 -> 5.2.0
  - packaging 21.3 -> 23.0
  - prometheus-client 0.15.0 -> 0.16.0
  - pyct 0.4.8 -> 0.5.0
  - pydata-google-auth 1.5.0 -> 1.6.0
  - python-slugify 7.0.0 -> 8.0.0
  - sqlalchemy 1.4.46 -> 2.0.0
  - tensorflow-io-gcs-filesystem 0.29.0 -> 0.30.0
  - tiff file 2022.10.10 -> 2023.1.23.1
  - zipp 3.11.0 -> 3.12.0
  - Pinned sqlalchemy to version 1.4.46

## 2023-01-12

- Added support for @-mention and email autocomplete in comments
- Improved errors when GitHub notebooks can't be loaded
- Increased color contrast for colors used for syntax highlighting in the code editor

- Added terminal access for custom GCE VM runtimes
- Upgraded Ubuntu from 18.04 LTS to 20.04 LTS ([GitHub issue](#))
- Python package updates:
  - GDAL 2.2.2 -> 2.2.3.
  - NumPy from 1.21.5 to 1.21.6.
  - attrs 22.1.0 -> 22.2.0
  - chardet 3.0.4 -> 4.0.0
  - cloudpickle 1.6.0 -> 2.2.0
  - filelock 3.8.2 -> 3.9.0
  - google-api-core 2.8.2 -> 2.11.0
  - google-api-python-client 1.12.11 -> 2.70.0
  - google-auth-httpplib2 0.0.3 -> 0.1.0
  - google-cloud-bigquery 3.3.5 -> 3.4.1
  - google-cloud-datastore 2.9.0 -> 2.11.0
  - google-cloud-firestore 2.7.2 -> 2.7.3
  - google-cloud-storage 2.5.0 -> 2.7.0
  - holidays 0.17.2 -> holidays 0.18
  - importlib-metadata 5.2.0 -> 6.0.0
  - networkx 2.8.8 -> 3.0
  - opencv-python-headless 4.6.0.66 -> 4.7.0.68
  - pip 21.1.3 -> 22.04
  - pip-tools 6.2.0 -> 6.6.2
  - prettytable 3.5.0 -> 3.6.0
  - requests 2.23.0 -> 2.25.1
  - termcolor 2.1.1 -> 2.2.0
  - torch 1.13.0 -> 1.13.1
  - torchaudio 0.13.0 -> 0.13.1
  - torchtext 0.14.0 -> 0.14.1
  - torchvision 0.14.0 -> 0.14.1

## 2022-12-06

- Made fallback runtime version available until mid-December ([GitHub issue](#))
- Upgraded to Python 3.8 ([GitHub issue](#))
- Python package updates:
  - jax from 0.3.23 to 0.3.25, jaxlib from 0.3.22 to 0.3.25

- pyarrow from 6.0.1 to 9.0.0
- torch from 1.12.1 to 1.13.0
- torchaudio from 0.12.1 to 0.13.0
- torchvision from 0.13.1 to 0.14.0
- torchtext from 0.13.1 to 0.14.0
- xldr from 1.1.0 to 1.2.0
- DriveFS from 62.0.1 to 66.0.3
- Made styling of markdown tables in outputs match markdown tables in text cells
- Improved formatting for empty interactive table rows
- Fixed syntax highlighting for variables with names that contain Python keywords ([GitHub issue](#))

## 2022-11-11

- Added more dark editor themes for Monaco (when in dark mode, "Editor colorization" appears as an option in the Editor tab of the Tools → Settings dialog)
- Fixed bug where collapsed forms were deleted on mobile [GitHub issue](#)
- Python package updates:
  - rpy2 from 3.4.0 to 3.5.5 ([GitHub issue](#))
  - notebook from 5.5.0 to 5.7.16
  - tornado from 5.1.1 to 6.0.4
  - tensorflow\_probability from 0.16.0 to 0.17.0
  - pandas-gbq from 0.13.3 to 0.17.9
  - protobuf from 3.17.3 to 3.19.6
  - google-api-core[grpc] from 1.31.5 to 2.8.2
  - google-cloud-bigquery from 1.21.0 to 3.3.5
  - google-cloud-core from 1.0.1 to 2.3.2
  - google-cloud-datastore from 1.8.0 to 2.9.0
  - google-cloud-firestore from 1.7.0 to 2.7.2
  - google-cloud-language from 1.2.0 to 2.6.1
  - google-cloud-storage from 1.18.0 to 2.5.0
  - google-cloud-translate from 1.5.0 to 3.8.4

## 2022-10-21

- Launched a single-click way to get from BigQuery to Colab to further explore query results ([announcement](#))
- Launched [Pro, Pro+, and Pay As You Go](#) to 19 additional countries: Austria, Belgium, Bulgaria, Croatia, Cyprus, Czechia, Denmark, Estonia, Finland, Greece, Hungary, Latvia, Lithuania, Norway, Portugal, Romania, Slovakia, Slovenia, and Sweden ([tweet](#))
- Updated jax from 0.3.17 to 0.3.23, jaxlib from 0.3.15 to 0.3.22, TensorFlow from 2.8.2 to 2.9.2, CUDA from 11.1 to 11.2, and cuDNN from 8.0 to 8.1 ([backend-info](#))
- Added a readonly option to [drive.mount](#)
- Fixed bug where Xarray was not working ([GitHub issue](#))
- Modified Markdown parsing to ignore block quote symbol within MathJax ([GitHub issue](#))

## 2022-09-30

- Launched [Pay As You Go](#), allowing premium GPU access without requiring a subscription
- Added vim and tcclib to our runtime image
- Fixed bug where open files were closed on kernel disconnect ([GitHub issue](#))
- Fixed bug where the play button/execution indicator was not clickable when scrolled into the cell output ([GitHub issue](#))
- Updated the styling for form titles so that they avoid obscuring the code editor
- Created a GitHub repo, [backend-info](#), with the latest apt-list.txt and pip-freeze.txt files for the Colab runtime ([GitHub issue](#))
- Added [files.upload\\_file\(filename\)](#) to upload a file from the browser to the runtime with a specified filename

## 2022-09-16

- Upgraded pymc from 3.11.0 to 4.1.4, jax from 0.3.14 to 0.3.17, jaxlib from 0.3.14 to 0.3.15, fsspec from

2022.8.1 to 2022.8.2

- Modified our save flow to avoid persisting Drive filenames as titles in notebook JSON
- Updated our [Terms of Service](#)
- Modified the Jump to Cell command to locate the cursor at the end of the command palette input (Jump to cell in Tools → Command palette in a notebook with section headings)
- Updated the styling of the Drive notebook comment UI
- Added support for terminating your runtime from code: `python from google.colab import runtime runtime.unassign()`
- Added regex filter support to the Recent notebooks dialog
- Inline `google.colab.files.upload` JS to fix `files.upload()` not working ([GitHub issue](#))

## 2022-08-26

- Upgraded PyYAML from 3.13 to 6.0 ([GitHub issue](#)), drivefs from 61.0.3 to 62.0.1
- Upgraded TensorFlow from 2.8.2 to 2.9.1 and ipywidgets from 7.7.1 to 8.0.1 but rolled both back due to a number of user reports ([GitHub issue](#), [GitHub issue](#))
- Stop persisting inferred titles in notebook JSON ([GitHub issue](#))
- Fix bug in background execution which affected some Pro+ users ([GitHub issue](#))
- Fix bug where `Download as .py` incorrectly handled text cells ending in a double quote
- Fix bug for Pro and Pro+ users where we weren't honoring the preference (Tools → Settings) to use a temporary scratch notebook as the default landing page
- Provide undo/redo for scratch cells
- When writing ipynb files, serialize empty multiline strings as `[ ]` for better consistency with JupyterLab

## 2022-08-11

- Upgraded ipython from 5.5.0 to 7.9.0, fbprophet 0.7 to prophet 1.1, tensorflow-datasets from 4.0.1 to 4.6.0, drivefs from 60.0.2 to 61.0.3, pytorch from 1.12.0 to 1.12.1, numba from 0.51 to 0.56, and lxml from 4.2.0 to 4.9.1
- Loosened our requests version requirement ([GitHub issue](#))
- Removed support for TensorFlow 1
- Added Help → Report Drive abuse for Drive notebooks
- Fixed indentation for Python lines ending in [  
• Modified styling of tables in Markdown to left-align them rather than centering them
- Fixed special character replacement when copying interactive tables as Markdown
- Fixed ansi 8-bit color parsing ([GitHub issue](#))
- Configured logging to preempt transitive imports and other loading from implicitly configuring the root logger
- Modified forms to use a value of None instead of causing a parse error when clearing raw and numeric-typed form fields

## 2022-07-22

- Update scipy from 1.4.1 to 1.7.3, drivefs from 59.0.3 to 60.0.2, pytorch from 1.11 to 1.12, jax & jaxlib from 0.3.8 to 0.3.14, opencv-python from 4.1.2.30 to 4.6.0.66, spaCy from 3.3.1 to 3.4.0, and dlib from 19.18.0 to 19.24.0
- Fix Open in tab doc link which was rendering incorrectly ([GitHub issue](#))
- Add a preference for the default tab orientation to the Site section of the settings menu under Tools → Settings
- Show a warning for USE\_AUTH\_EPHEM usage when running authenticate\_user on a TPU runtime ([code](#))

## 2022-07-01

- Add a preference for code font to the settings menu under Tools → Settings
- Update drivefs from 58.0.3 to 59.0.3 and spacy from 2.2.4 to 3.3.1
- Allow [display\\_data](#) and [execute\\_result](#) text outputs to wrap, matching behavior of JupyterLab (does not affect stream outputs/print statements).
- Improve LSP handling of some magics, esp. %%writefile ([GitHub issue](#)).
- Add a [FAQ entry](#) about the mount Drive button behavior and include link buttons for each FAQ entry
- Fix bug where the notebook was sometimes hidden behind other tabs on load when in single pane view.
- Fix issue with inconsistent scrolling when an editor is in multi-select mode.
- Fix bug where clicking on a link in a form would navigate away from the notebook
- Show a confirmation dialog before performing Replace all from the Find and replace pane.

## 2022-06-10

- Update drivefs from 57.0.5 to 58.0.3 and tensorflow from 2.8.0 to 2.8.2
- Support more than 100 repos in the GitHub repo selector shown in the open dialog and the clone to GitHub dialog
- Show full notebook names on hover in the open dialog
- Improve the color contrast for links, buttons, and the ipywidgets.Accordion widget in dark mode

## 2022-05-20

- Support URL params for linking to some common pref settings: [force\\_theme=dark](#), [force\\_corgi\\_mode=1](#), [force\\_font\\_size=14](#). Params forced by URL are not persisted unless saved using Tools → Settings.
- Add a class markdown-google-sans to allow Markdown to render in Google Sans

- Update monaco-vim from 0.1.19 to 0.3.4
- Update drivefs from 55.0.3 to 57.0.5, jax from 0.3.4 to 0.3.8, and jaxlib from 0.3.2 to 0.3.7

## 2022-04-29

- Added 🦀 mode (under Miscellaneous in Tools → Settings)
- Added "Disconnect and delete runtime" option to the menu next to the Connect button
- Improved rendering of filter options in an interactive table
- Added git-lfs to the base image
- Updated torch from 1.10.0 to 1.11.0, jupyter-core from 4.9.2 to 4.10.0, and cmake from 3.12.0 to 3.22.3
- Added more details to our [FAQ](#) about unsupported uses (using proxies, downloading torrents, etc.)
- Fixed [issue](#) with apt-get dependencies

## 2022-04-15

- Add an option in the file browser to show hidden files.
- Upgrade gdown from 4.2.0 to 4.4.0, google-api-core[grpc] from 1.26.0 to 1.31.5, and pytz from 2018.4 to 2022.1

## 2022-03-25

- Launched [Pro/Pro+](#) to 12 additional countries: Australia, Bangladesh, Colombia, Hong Kong, Indonesia, Mexico, New Zealand, Pakistan, Philippines, Singapore, Taiwan, and Vietnam
- Added