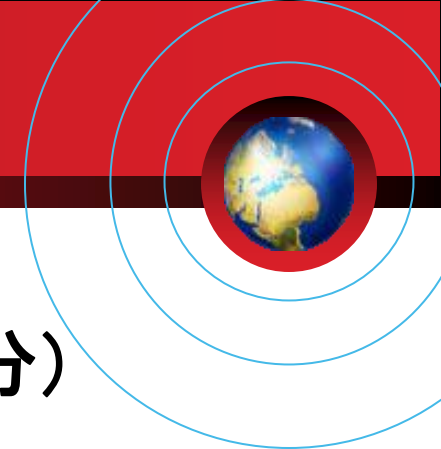


Introduction
プログラミングの基礎
Pythonの基礎



永田亮(甲南大学／理研)
川崎義史(東京大学)
内田諭(九州大学)

このパートの内容

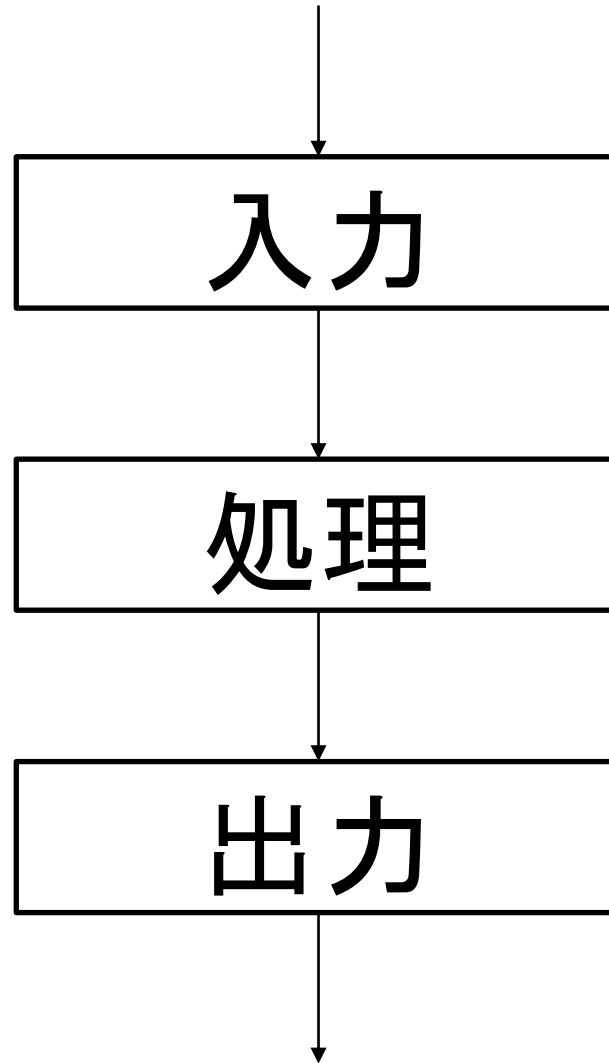
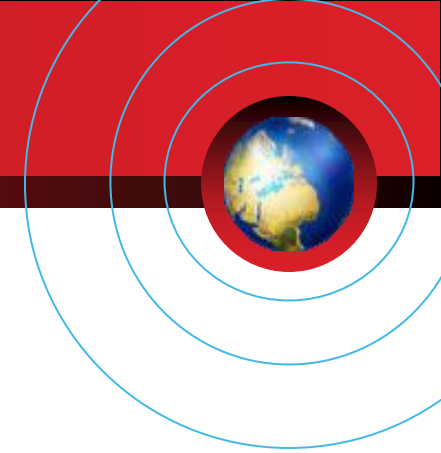


- **導入：プログラミングとは何か（5分）**
 - 重要な三要素：**入力**，処理，**出力**
- **基礎：プログラミング準備と基礎（85分）**
 - Google Colabの準備
 - 簡単なプログラミング（文字列処理）

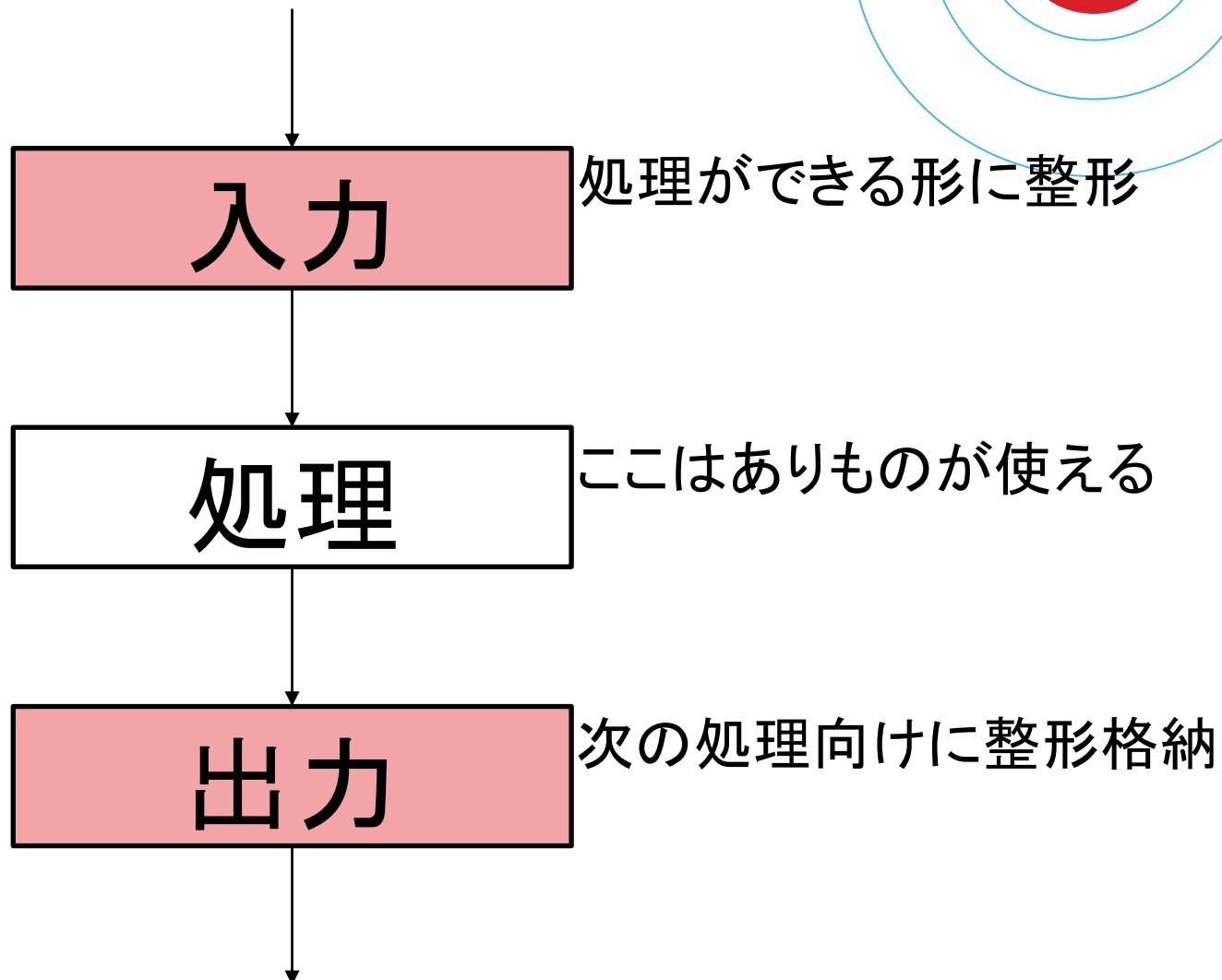
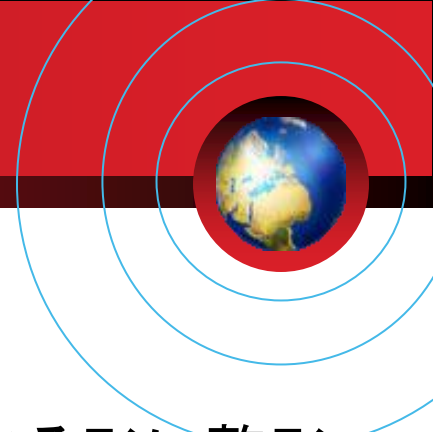
導入：プログラミングとはなにか？



プログラムの三大要素

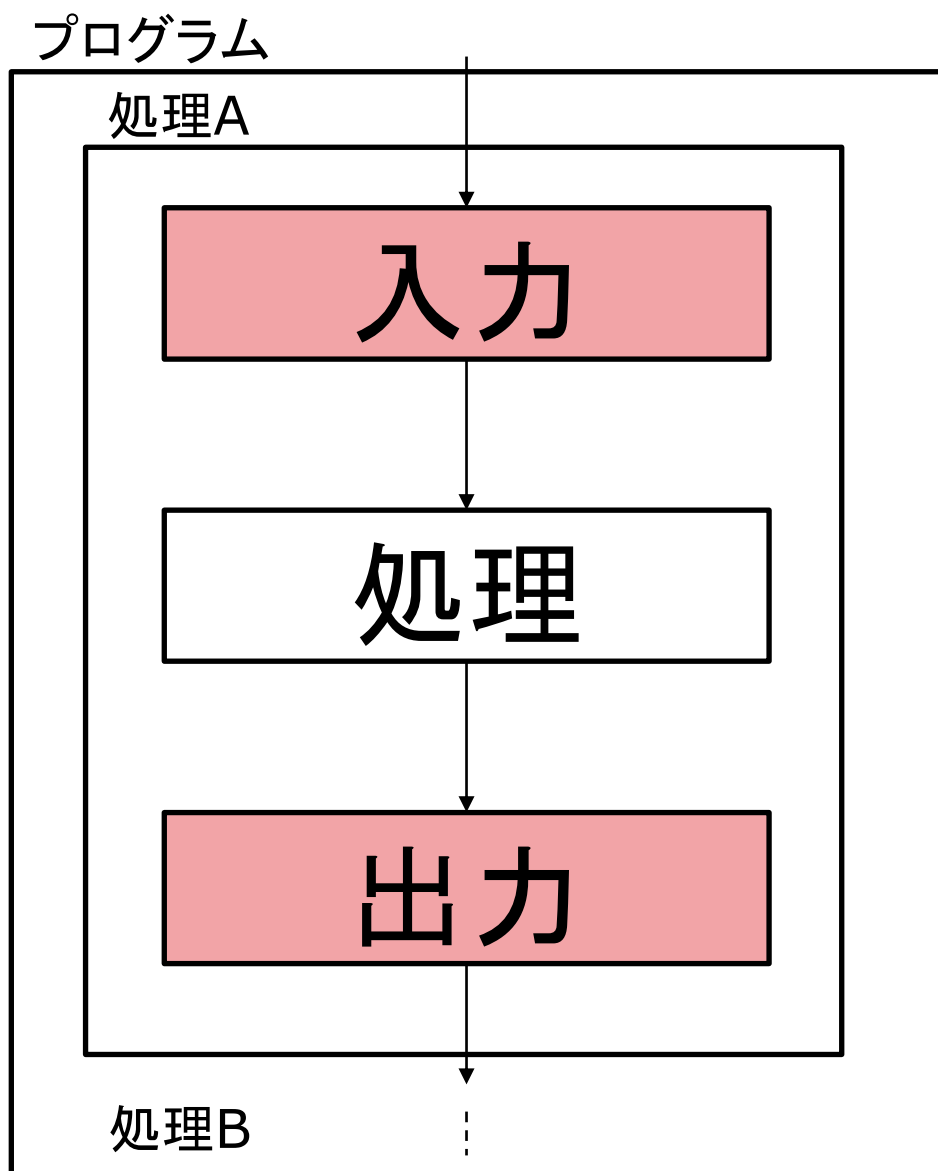
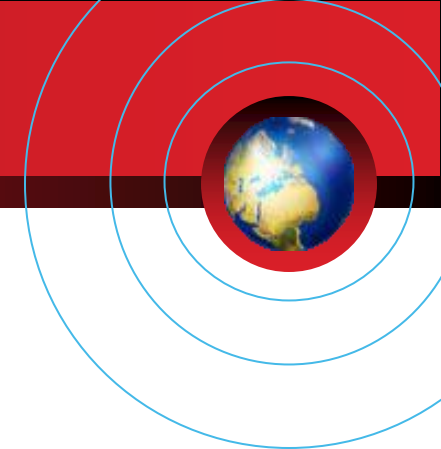


プログラムの三大要素



今回重要なのはこの二つ

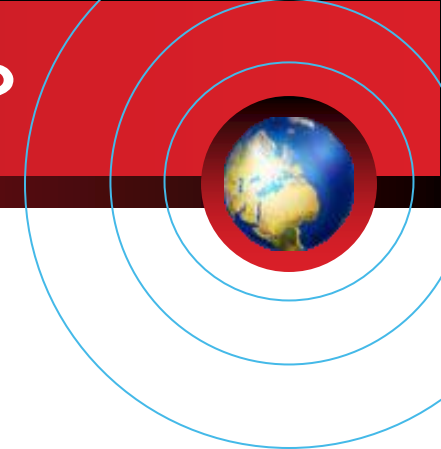
プログラム内も同じ



環境準備: Google Colab



Google Colaboratory (Colab)とは？

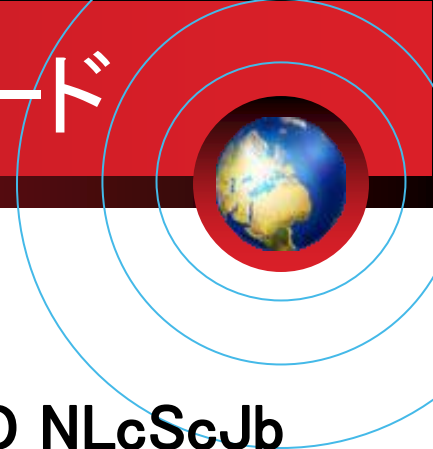


■ Google Colabの概要

- オンラインのPythonの実行環境
- Jupyter Notebookという方式
 - 「プログラム」と「実行結果」を一画面で確認可
- 無料（GPUも使える！）
（1200円/月～の有料版はさらに強力）
- numpy, pandas, scipyなど標準的なライブラリはデフォルトでインストール済み



ファイルのダウンロードとアップロード



- Step 1: 次のURLからファイルをダウンロード

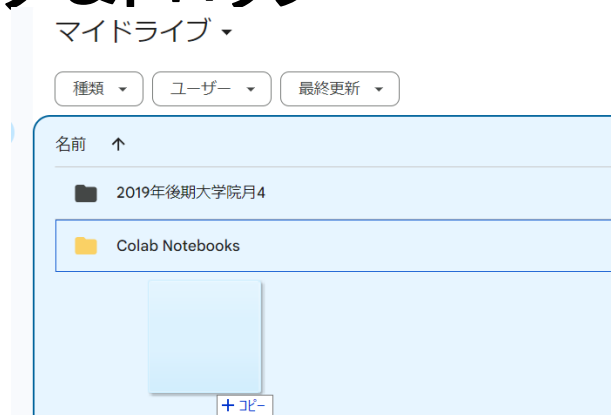
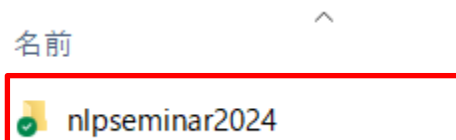
https://drive.google.com/drive/folders/1KI7xebHkD_NLcScJb4QxUNU9rO5S_sMi?usp=sharing

- Step 2: zipファイルの解凍

Windows: 右クリック→すべて展開(あるいは解凍ソフトを実行)

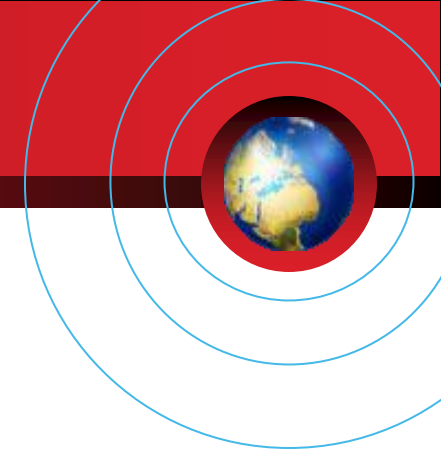
Mac: ファイルをダブルクリックする

- Step 3: フォルダをGoogle Driveにドラッグ & ドロップ



- Step 4: ファイルをダブルクリックする

アップロードしたファイルの確認



マイドライブ > nlpseminar2024 ▾















種類 ▾

ユーザー ▾

最終更新 ▾

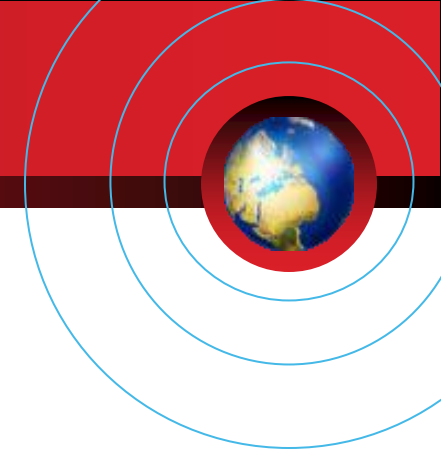
名前 ↑

オーナー

 __pycache__	 自分
 _4_count_words.ipynb	 自分
 _5_search_similar_word_usage_from_nltk_corpus_by_bert_vec.ipynb	 自分
 _tokenize_count_words.ipynb	 自分
 1_read_data-0.ipynb	 自分
 1_read_data-1.ipynb	 自分
 2_split_sentences2words_output_per_line.ipynb	 自分

ファイルをダブルクリックすればコードが開く

Google Colabの画面



1_read_data-0.ipynb ☆

ファイル 編集 表示 挿入 ランタイム ツール ヘルプ 保存しています

ファイル

- drive
- MyDrive
- sample_data

プログラムの実行ボタン

準備: モジュールの導入 (Googleドライブを使用するモジュール)

```
1 import google.colab.drive
```

準備: ドライブ使用準備 (ドライブのマウント)

```
[2] 1 google.colab.drive.mount('/content/drive/')
```

Mounted at /content/drive/

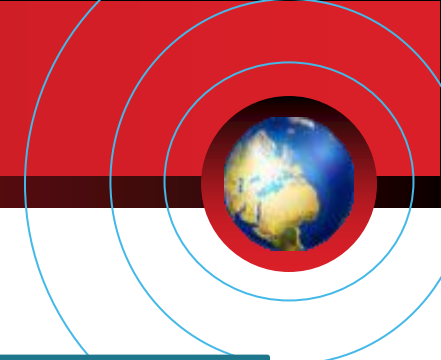
メイン処理: ファイルの読み込みと内容の出力

```
[ ] 1 with open('/content/drive/MyDrive/nlpseminar2024/simple_corpus.txt') as f:
    2     for data_line in f:
    3         print(data_line)
```

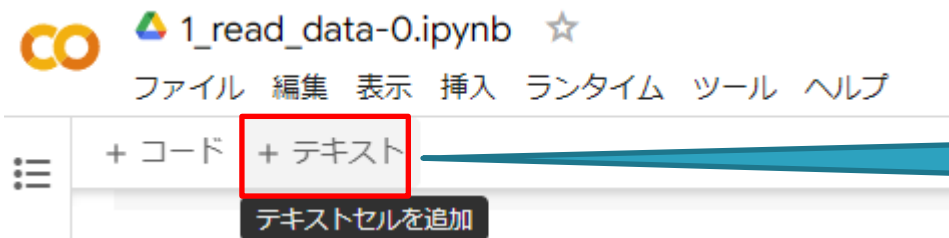
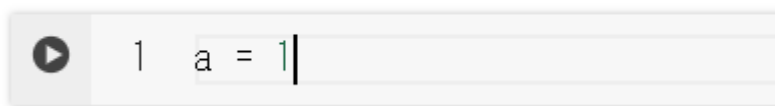
He drew some money from the bank.

プログラムの実行結果

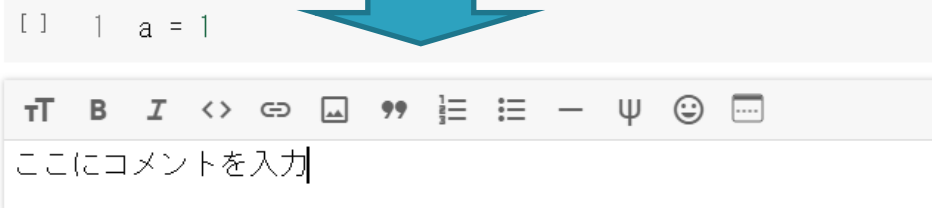
コードとテキストの入力



新しくコードを入力する



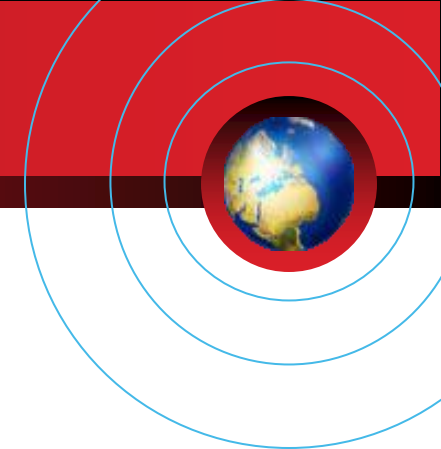
新しくテキストを入力する



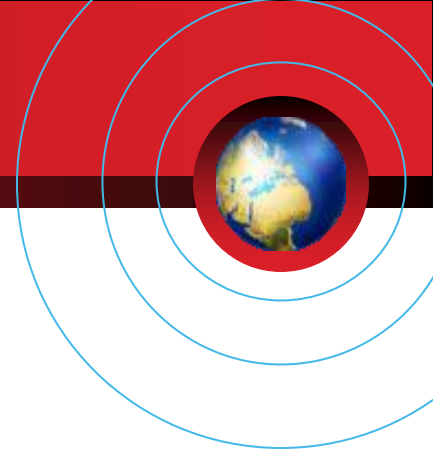
基礎: Pythonプログラミング



基礎のメニュー

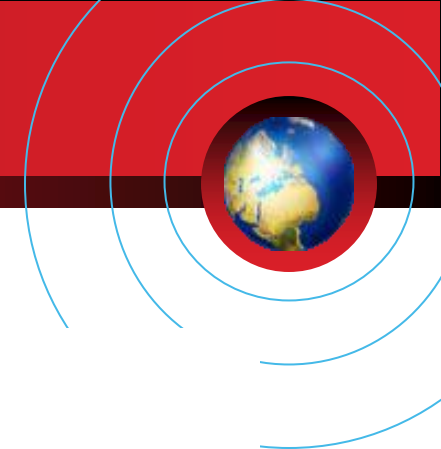


- ファイル読み込みと出力
- 簡単な文字列処理
- 単語長ヒストグラムの作成
- (単語頻度のカウント)



- ファイル読み込み
 - ファイルのオープン
 - 一行読み込み
 - (処理いろいろ)
- 出力: 画面に一行ずつ出力

Pythonでの入出力プログラム例



1_read_data-0.ipynb

準備：モジュールの導入（Googleドライブを使用するモジュール）

```
[1] 1 # #以降はコメントとして無視される（実行されない）
    2 import google.colab.drive # モジュールの読み込み
```

準備：ドライブ使用準備（ドライブのマウント）

```
[2] 1 google.colab.drive.mount('/content/drive/')
```

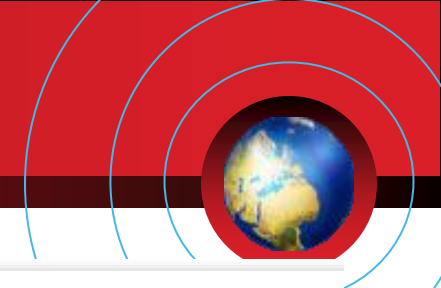
⇄ Mounted at /content/drive/

メイン処理：ファイルの読み込みと内容の出力

ファイルをオープンし、fとして使用

```
▶ 1 with open('/content/drive/MyDrive/nlpseminar2024/simple_corpus.txt') as f:
    2     for data_line in f: 変数(任意の名前)
    3         print(data_line)
```


演習：変数名を変えてみましょう



準備：モジュールの導入（Googleドライブを使用するモジュール）

```
[ ] 1 import google.colab.drive
```

準備：ドライブ使用準備（ドライブのマウント）

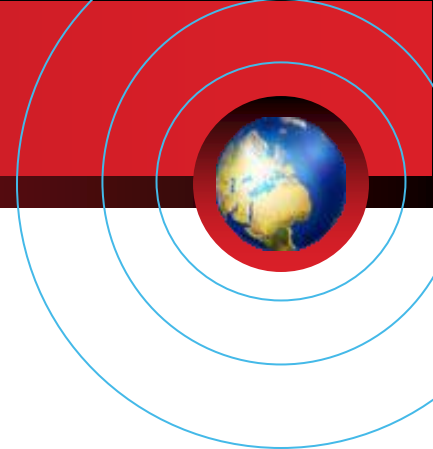
```
[ ] 1 google.colab.drive.mount('/content/drive/')
```

⇄ Mounted at /content/drive/

メイン処理：ファイルの読み込みと内容の出力

```
[ ] 1 with open('/content/drive/MyDrive/nlpseminar2024/simple_corpus.txt') as f:  
  2     for data_line in f:  
  3         print(data_line)
```

余分な空行(改行コード)が！



```
[3] 1 with open('/content/drive/MyDrive/nlpseminar2024/simple_corpus.txt') as f:  
    2   for data_line in f:  
    3       print(data_line)
```

⇒ He drew some money from the bank.

John got a bank transfer form to make a bank transfer.

My father had worked as a bank clerk for a long time.

Someone raided a bank.

It functions as a data bank.

They walked along the river bank.

The city stands on the right bank of the Saine.

There is a sand bank between the two towns.

They sat down against the bank by the wayside.

The heavy rain broke the bank.

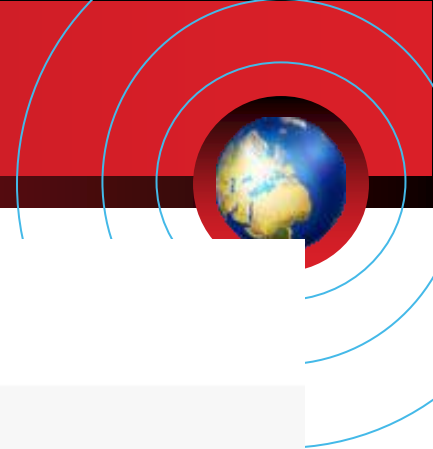
}

演習：行末の改行コードを除去してみよう



🔍 python 改行コード 削除

余分な改行コードの除去



準備：モジュールの導入



✓
0秒

```
[1] 1 import google.colab.drive
```



準備：ドライブ使用準備（ドライブのマウント）



✓
24秒

```
[2] 1 google.colab.drive.mount('/content/drive/')
```

⇄ Mounted at /content/drive/

メイン処理：ファイルの読み込みと内容の出力

✓
0秒



```
1 with open('/content/drive/MyDrive/nlpseminar2024/simple_corpus.txt') as f:  
2     for data_line in f:  
3         data_line = data_line.rstrip()  
4         print(data_line)
```

ここまでは常套句（コピーして使用）

基礎 : a step further
ちよつと文字列処理

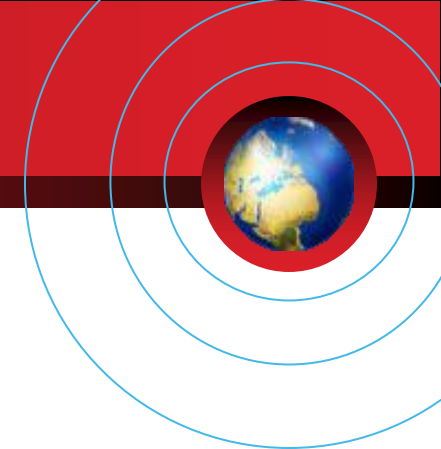


文(文字列)を単語(文字列)に分割



- ファイル読み込み
 - ファイルのオープン
 - 一行読み込み(文)
- 単語分割(様々な方法があるが今回はこれ)
 - 空白で単語に分割
 - 分割結果を格納(リスト)
- 出力:一文中の単語を一行ずつ画面に出力

演習：文(文字列)を分割



🔍 python 文字列 空白で分割

文(文字列)を単語(文字列)に分割



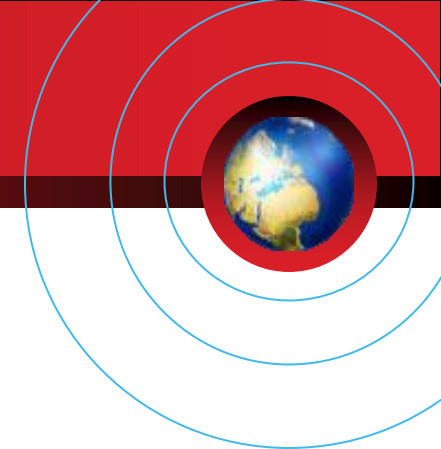
2_split_sentences2words.ipynb

(前の部分は省略)

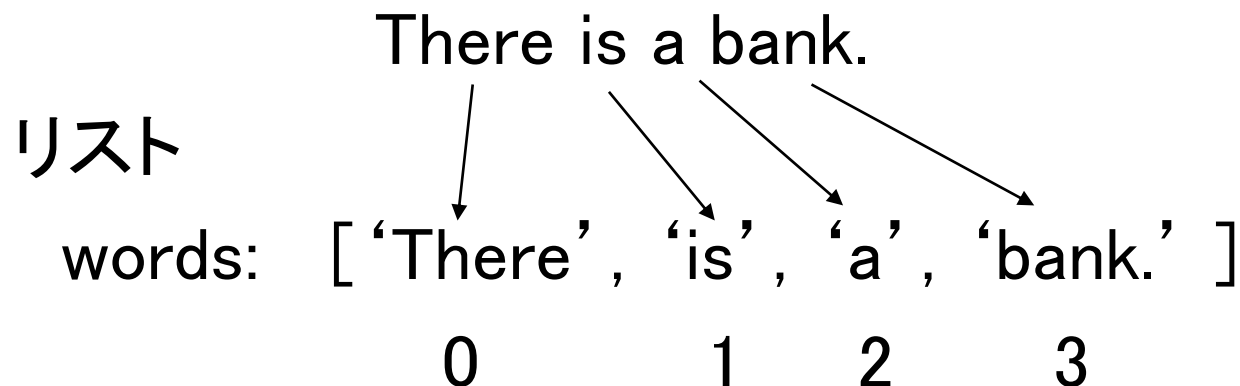
```
[3] 1 with open('/content/drive/MyDrive/nlpseminar2024/simple_corpus.txt') as f:
    2     for data_line in f:
    3         data_line = data_line.rstrip()
    4         words = data_line.split(' ')
    5         print(words)
```

wordsはリストという変数の一種.

リストのイメージ



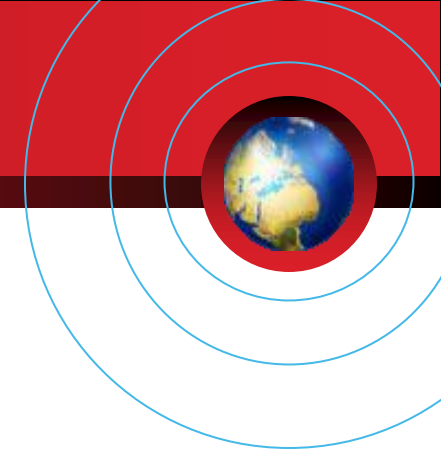
■ 複数の値を保持可能



任意の名前が付けられます(中身を表すように)

値(単語)だけでなく順番の情報も保持

文を単語に分割(出力変更)



- ファイル読み込み
 - ファイルのオープン
 - 一行読み込み(文)
- 単語分割
 - 空白で単語に分割
- 出力: **一単語一行ずつ画面に出力**

文(文字列)を単語(文字列)に分割



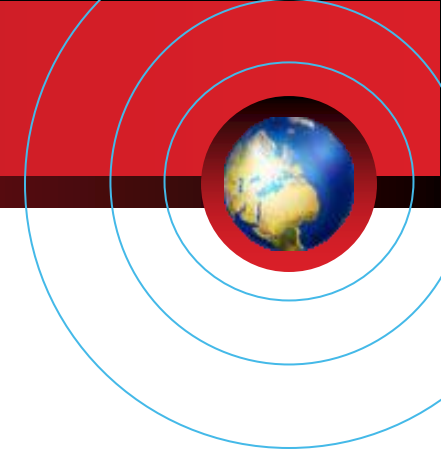
2_split_sentences2words_output_per_line.ipynb

(前の部分は省略)

```
1 with open(' /content/drive/MyDrive/dl/data/simple_corpus.txt' ) as f:
2     for data_line in f:
3         data_line = data_line.rstrip()
4         words = data_line.split(' ')
5         for word in words:
6             print(word)
```

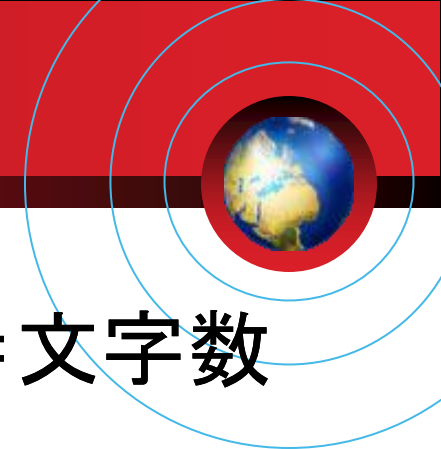
リストの先頭からひとつずつ値を取り出す常套句

演習：特定の単語の出力



- 文頭(先頭)の単語
- 2番目の単語
- 文頭(先頭)～9番目の単語(先頭から10個)

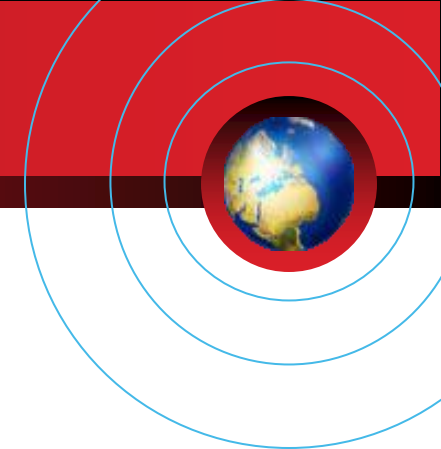
演習：単語とその長さを出力



- ヒント：単語の長さ＝文字列の長さ＝文字数

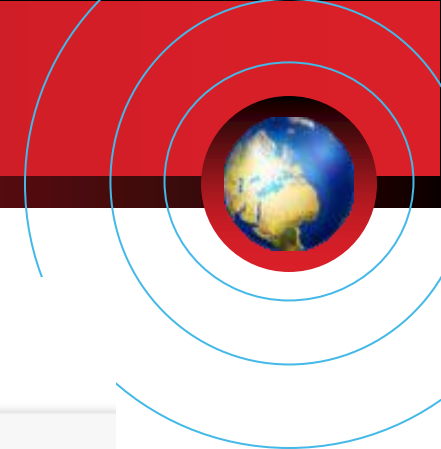
🔍 python 文字列 長さ

単語長のヒストグラムの作成



- ファイル読み込み
- 単語分割
- 単語長のカウント
 - 単語の取り出し
 - 単語長の算出
 - 単語長の保存(リストで)
使用するリストの準備: `word_lengths = []`
- ヒストグラムの作成(ありものを利用)
- 出力: ヒストグラム(図)

単語長のヒストグラムの作成



3_make_word_length_hist.ipynb

メイン処理：ファイルの読み込み単語の長さ求め保存

```
1 word_lengths = []
2 with open('/content/drive/MyDrive/nlpseminar2024/simple_corpus.txt') as f:
3     for data_line in f:
4         data_line = data_line.rstrip()
5         words = data_line.split(' ') # 入力文を単語に分割
6         for word in words:
7             l = len(word) # len()で文字列の長さ (文字数)
8             word_lengths.append(l) # word_lengthsに長さlを格納
9
10 print(word_lengths)
```

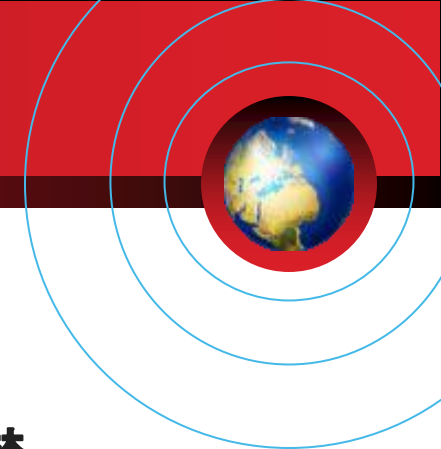
ヒストグラムを描画

```
1 import matplotlib.pyplot as plt # 描画用モジュールの読み込み
2
3 # ヒストグラムの描画
4 plt.hist(word_lengths, bins=20) # 区間は20個
5 plt.show() # 画面表示
```

演習: ヒストグラムの見た目を変えてみよう



スケジュール(再掲)



■ 午前

10:00 -- 11:30 導入と基礎: Pythonの基礎

11:30 -- 13:00 (お昼休憩)

■ 12:30 -- 12:55 再度入館手続きが必要

■ 午後

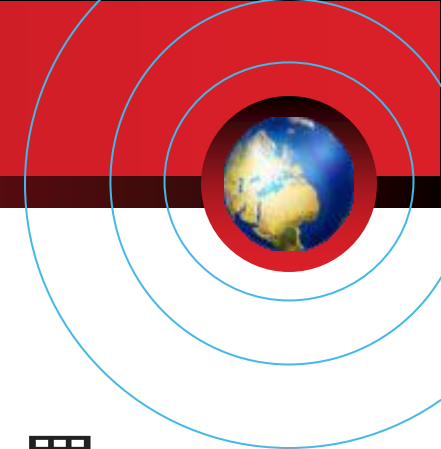
13:00 -- 13:50 深層学習入門

14:00 -- 15:10 応用1: word2vecベースの言語分析

15:20 -- 16:50 応用2: 言語モデルベースの言語分析

16:50 -- 17:00 クロージング

会場周辺案内(再掲)



- **注意:再入館は一度のみ可能**
 - 12:30 -- 12:55 3Fに受付デスク設置
- **建物内(再入館手続き不要)**
 - 3F: コンビニ
 - 3F: スターバックス
 - 3F: イタリア料理店
- **建物外(再入館手続き必要)**
 - 大丸(東京駅徒歩10分ほど. お弁当売り場あり)