Text Mining Maniax for Python

Python による
日本語計量テキスト分析の基礎

後藤和智
（後藤和智事務所 OffLine）

Text Mining Maniax for Python

Python による
日本語計量テキスト分析の基礎

著：後藤和智（後藤和智事務所 OffLine）
発行：2022 年 12 月 31 日
（コミックマーケット 101）

はじめに

（執筆予定だったものを含めると）90 番目の同人誌になります、後藤和智です。

今回は、満を持して話題のプログラミング言語「Python」の解説書になります。弊サークルでは主としてフリーの統計ソフト「R」やフリーのテキストマイニングソフト「KH Coder」の解説書を多数書いてきましたが、数年前から話題になっていた Python については、2014 年に出した『R Maniax Advance』で少し解説をしただけで、それ以降触れることはありませんでした。

Python は、R に比べると、統計解析という面では劣りますが、ただ機械学習やディープラーニングなどの分野では R よりも優れているため、弊サークルのような統計解析を中心にやっている人はともかく、先進的な分野、少なくともエンジニアとして生活する人たちとしては必須のスキルになっているようです。

とはいえ、私自身はエンジニアではなく、あくまでも趣味の範囲として統計解析などを行っているのですが、ただ、せっかくだから触れてみようと思って数年前から構想は立てていました。しかし、私自身他の解説書のほうに興味が行ってしまい、何回か Python に触れたのですがその都度諦め、というのを繰り返したせいで、今の今まで刊行が遅れてしまったことを深くお詫びいたします。コミケでは「次は Python をやるかも」ということを何度もサークルカットに書いてきたのですが、結局今になってようやくの刊行というあたり、なんか意欲が落ちてきているなぁ……という気がしています。

世の中には R 以上に Python の解説書が多く、一部の解説書は無料で読めるものもあります。例えば東京大学の『Python プログラミング入門』や、京都大学の『プログラミング演習 Python』は、大学の教材であることもあり、体系的に学べます（詳しくは「@IT」の「無料で読める、東大／京大の「Python 教科書」電子書籍」https://atmarkit.itmedia.co.jp/ait/articles/2105/26/news025.html を参照されたい）。Python を学習する環境が整ってきているので、この機会に学んでみてはいかがでしょうか。

さて、弊サークルは最近はテキストマイニングを主にやっているサークルでもあるため、Python を用いた簡単なテキストマイニングを行う方法を解説します。とはいえ、まずはテキストマイニングに必要な Python の操作の基本を解説してから、文章を読み込み、そして解析する方法、そして簡単な分析を行う、という段階までの解説となります。R のテキストマイニング解説ではかなり深いところまで解説をしている感がありますが、初めての Python 本ということで、ご容赦いただけますと幸いです。

また、本書ではウェブスクレイピングという、インターネット上から情報を取得する方法も少しだけ解説します。テキストマイニングでは、ウェブ上に公開されている文章の解析も行うこともあるので、そのとっかかりの分野だけでも覚えてもらえると幸いです。

なお、本書では Windows で Python を使用することを前提に書かれております。Macintosh や Linux での使用には対応しておりませんので、何卒ご了承願います。

目　　次

第 1 章 Python の導入

1.1 はじめに

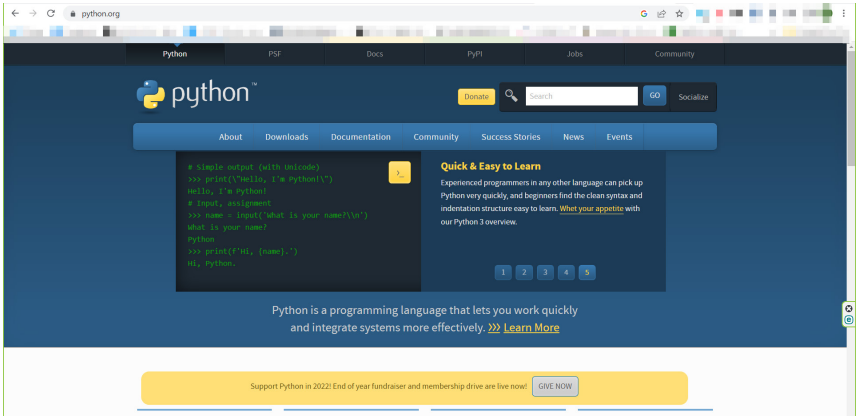
本書では Python でテキストマイニング（計量テキスト分析）を行う方法を解説しますが、まず、Python 含め、本書で必要なソフトの導入について解説します。

1.2 Python の導入

まずは Python がないと始まりませんので、Python の導入から解説します。Python はフリーソフトであり、公式サイト https://www.python.org/ からダウンロードできます。「downloads」をクリックすればインストーラーがダウンロードされますので、あとはインストーラーの指示に従ってインストールを行います。

ここで注意すべきなのは、Python のバージョンです。現在の Python には Python2 と Python3 というバージョンがあり、それぞれでプログラムの表記が大きく異なるようです。そして、Python2 は 2020 年でサービスが終了しているので、インストールする際は Python3 を選びましょう。本書では（ずいぶん前にインストールしてその後アップグレードを怠ったため）Python3.9.7 を使っていますが、2022 年 12 月 19 日現在の最新のバージョンは 3.11.1 なので、定期的に新しいものにしておきましょう。

なお、Windows で Python を使うときは、コマンドプロンプトで行います。また、詳細は後述しますが、Python はプログラムが書かれたファイルを別に作成して、それを読み込ませるという方法をとるので、テキストエディタも用意しておきましょう。もちろん、Windows に標準で搭載されているメモ帳でもいいのですが、弊サークルの同人誌で何回か登場している「秀丸エディタ」（有限会社サイトー企画。https://hide.maruo.co.jp/software/



Python の公式サイト

6

hidemaru.html）なら、Python で頻繁に使う「'」で囲まれた部分の色が変わったり、「import」などの一部コマンドが強調されたりと便利なので、他にも LaTeX などを使う機会がある方は導入をお勧めします。

1.3 MeCab の導入

次に、本書で使う形態素解析エンジンである、フリーソフト「MeCab」をインストールします。MeCab は文章を単語に分けて集計する形態素解析ソフトです。そもそも形態素解析というのは、文章を文字や単語などといった形態素に分けて解析するもので、MeCab は日本語形態素解析ソフトの中でも最も広く使われているものです。

とはいえ、MeCab 本体の開発は、2013 年に出た MeCab0.996 で終了しており、現在の Windows パソコンの主流である 64bit に対応しておりません。そのため、MeCab をインストールするときは、有志による 64bit 対応版を使う必要があります。

64bit 版の MeCab は、https://github.com/ikegami-yukino/mecab/releases から取得します（Google などで「MeCab 64bit」などと検索しても可）。そこから「mecab-64-0.996.2.exe」をダウンロードして、インストーラーの指示に従います。

ただし、Python で MeCab を使うときに注意すべきなのは、文字コードです。Windows の標準の文字コードは Shift-JIS なのに対して、Python は Unicode（UTF-8）を使っています。この文字コードの違いは、Windows で Python を使うときに常に意識しておかなければなりません。インストーラーでインストールすると、文字コードが確認されますので、必ず「UTF-8」を選びましょう。

MeCab は、そのままインストールすると Windows の「Program File」のフォルダに保存されますが、Python だけでなく KH Coder などで頻繁に MeCab を使う場合は、Program File のフォルダの中にあると分析対象に応じてカスタマイズをすることができません。そのため、C ドライブ直下に「usr」というフォルダを作り、さらにその中に「local」というフォルダを使って、そこにインストールすることを本書では推奨しています。

また、コマンドプロンプトなどで MeCab を使えるようにするために、Windows の設定を変える必要があります。Windows10 であれば、スタートメニューから設定を開き、「設定の検索」に「環境変数」と入力すると、「環境変数を編集」という項目が出るので、それをクリックし、「Path」を選択し編集して「c:\usr\local\mecab\bin」があることを確認し、なければ追加します。

1.4 ubuntu LTS と neologd の導入

前述の通り、MeCab は 2013 年に開発が終わっているため、新語や流行語などの多くに対応しておりません。それらをユーザーで辞書を作って適用させる方法もあるのですが、ここでは neologd という辞書を使う方法を解説します。neologd の導入については、『Twitter Analysis Maniax』でも解説したのですが、改めて説明します（詳しい方法は「いるかのボックス」の「もう少し簡単に KH Coder で新語に対応するために mecab-ipadic-NEologd を使う」

7

https://irukanobox.blogspot.com/2018/08/kh-codermecab-ipadic-neologd.html 参照)。

まず、Windows で neologd を導入するために、Windows で動く ubuntu LTS を導入する必要があります。ubuntu LTS は、Microsoft Store からでもダウンロードできます。ubuntu LTS を導入したら、まずはソフトのアップデートを行います（sudo は管理者として実行するときのコマンドです。なお、ubuntu を初めて起動したときにパスワードの設定が求められます)。

```
sudo apt update
```

これを行わないと、MeCab 含め多くのソフトがインストールされませんのでご注意ください。次に、ubuntu LTS に MeCab をインストールします。

```
sudo apt install mecab
```

MeCab をインストールしているときにエラーが出たら、同じように「sudo apt install」を使って、インストールされていないと指摘されたものをその都度インストールしましょう。次に、

```
git clone --depth 1 https://github.com/neologd/mecab-ipadic-neologd.git
```

と入力すると、mecab-ipadic-neologd が ubuntu にインストールされるので、これを Windows に移動する必要があります。MeCab で利用することも考え、MeCab の辞書フォルダに移すようにします。

```
cd mecab-ipadic-neologd
./bin/install-mecab-ipadic-neologd -n
echo `mecab-config --dicdir`"/mecab-ipadic-neologd"
```

2 行目で更新、3 行目で保存されているディレクトリの確認をしています。そして、

```
cp -a（neologd が保存されているディレクトリ） /mnt/c/usr/local/mecab/dic
```

と入力すれば、MeCab の dic フォルダに neologd がコピーされます。以降は、neologd が「c:\usr\local\mecab\dic」の中にある前提で解説を行います。

1.5 Python を起動してみる

Windows では、Python はコマンドプロンプトで操作します。また、Python は R のパッケージのように、様々なモジュールを使って機能を強化することができます。モジュールのインストールは、コマンドプロンプトで次のように入力します。

```
python -m pip install インストールしたいモジュール
```

なお、単に「pip install モジュール名」でも可能です。例えば、Python で MeCab を使う

8

ためのモジュールである「mecab-python3」をインストールするときは、次のように入力します。

```
pip install mecab-python3
```

Python は、プログラムを拡張子「.py」というファイルに入れて、それを起動させるという方法をとります。例えば、「sample.py」というファイルを使って Python を実行させるときは、次のように入力します。

```
python sample.py
```

コマンドプロンプトの基本的な操作方法については省略しますが、一点だけ述べると、「cd」というコマンドは、作業フォルダを移動するときに使います。例えば作業フォルダが C ドライブ直下の「usr」＞「local」＞「python」というところにあるときは、

```
cd c:¥usr¥local¥python
```

というように入力します。Python 以外で本書で使うコマンドはこれだけなので、これだけ覚えておけばいいでしょう。

9

第 2 章 Python の基礎

2.1 はじめに

本章では、四則演算などの基本的な方法を解説します。Python もプログラミング言語なので、まずは基本的な計算ができる必要があります。

前述の通り、Python を使うときは、あらかじめ「ファイル名 .py」というファイルを作って、それを起動させるという方法をとりますので、まずは基本的な書き方を学んでおきましょう。

2.2 四則演算

まず、次のようなファイルを作り、「sample.py」という名前で作業フォルダに保存します。

```
print(1)
print('hello,world')
print(1+1)
print(2-1)
print(2*3)
print(2**3)
print(9/2)
print(9//2)
print(9%2)
```

print というコマンドは、結果を表示させるためのもので、これがないと結果が表示されません。これを起動させると、次のような結果が出されます。

```
>python sample.py
1
hello,world
2
1
6
8
4.5
4
1
```

** は累乗、// は Excel の int 関数のように結果が整数になり、% は割ったときのあまりを示します。また、文字列を ' で囲めば文章を出力することだって可能です。ただし、日本語の文章を表示させたいときや、日本語を取り扱いたいときは、あらかじめ文字コードを

10

UTF-8 にしておく必要があります。そうしないとエラーになります（これに気付かなかったことが、私が Python 本の執筆に取りかかれなかったことでもあります）。

秀丸エディタの場合、右下に文字コードが表示されるので、文字コードが「日本語 (Shift-JIS)」になっていたら、「ファイル」の「エンコードの種類」から「Unicode(UTF-8)」を選択して、内容を保持したまま文字コードを変更します。

Python に既にインストールされているモジュール math を使えば、三角関数なども使えます。次のような「sample2.py」を作ってみましょう。

```
import math
print(math.sqrt(2)) # 平方根
print(math.pi) # 円周率
print(math.sin(math.pi / 6))
print(math.cos(math.pi / 6))
print(math.sqrt(3) / 2)
```

```
>python sample2.py
1.4142135623730951
3.141592653589793
0.49999999999999994
0.8660254037844387
0.8660254037844386
```

import はモジュールを適用させるためのコマンドであり、「モジュール名 . 関数名」と入力することでモジュール内の関数を使えます。また、例えば「import math as m」とすれば、「m.pi」と入力しても円周率を出すこともできます。

また、文字に数字を入れることもできます。次のような「sample3.py」を作ってみます。

```
import math
a = math.pi
print(math.cos(a/6))
print(math.cos(a/3))
print(math.cos(a/2))
print(math.cos(a*2/3))
print(math.cos(a*5/6))
print(math.cos(a))
```

```
>python sample3.py
0.8660254037844387
```


11