mi-7. 自然言語処理

（人工知能シリーズ）

https://www.kkaneko.jp/cc/mi/index.html

金子邦彦

CC BY NC SA

アウトライン

7-1 単語の切り出し，品詞の判定

7-2 HTML（Web ページ）のタグの除去，単語の切り出し

7-3 音声合成の例

自然言語処理

- **人間の言葉**（日本語や英語など）を，コンピュータが処理することを，**自然言語処理**という

- **人間の言葉**を理解できる能力をもつアプリやサービスの制作に役立つ

  （用途の例）

  インターネット検索，対話システム，音声合成，
  翻訳，かな漢字変換での予測変換，
  迷惑メールフィルタ

7-1 単語の切り出し，品詞の判定

自然言語処理の技術

- 文を単語に分割（**単語の切り出し**）
- 単語の種類（**品詞**）の判定
- 係り受けの関係などの分析（**構文解析**）
- **意味解析**

Windows などで，単語を切り出し，品詞を自動判定する機能を持つ mecab コマンド

白い雲と青い空が美しい

```
C:¥Users¥user>mecab
白い雲と青い空が美しい
白い    形容詞,自立,*,*,形容詞・アウオ段,基本形,白い,シロイ,シロイ
雲      名詞,一般,*,*,*,*,雲,クモ,クモ
と      助詞,並立助詞,*,*,*,*,と,ト,ト
青い    形容詞,自立,*,*,形容詞・アウオ段,基本形,青い,アオイ,アオイ
空      名詞,一般,*,*,*,*,空,ソラ,ソラ
が      助詞,格助詞,一般,*,*,*,が,ガ,ガ
美しい  形容詞,自立,*,*,形容詞・イ段,基本形,美しい,ウツクシイ,ウツク
シイ
EOS
```

単語に切り分けられ，単語ごとの**読み仮名**，**品詞**などが，得られる

mecab は Windows の標準機能ではない.
使えるようにするにはインストールが必要.

日本国民は、正当に選挙された国会における代表者を通じて行動し

```
日本国民は，正当に選挙された国会における代表者を通じて行動し
日本	名詞,固有名詞,地域,国,*,*,日本,ニッポン,ニッポン
国民	名詞,一般,*,*,*,*,国民,コクミン,コクミン
は	助詞,係助詞,*,*,*,*,は,ハ,ワ
，	記号,読点,*,*,*,*,，,，,，
正当	名詞,形容動詞語幹,*,*,*,*,正当,セイトウ,セイトー
に	助詞,副詞化,*,*,*,*,に,ニ,ニ
選挙	名詞,サ変接続,*,*,*,*,選挙,センキョ,センキョ
され	動詞,自立,*,*,サ変・スル,未然レル接続,する,サ,サ
れ	動詞,接尾,*,*,一段,連用形,れる,レ,レ
た	助動詞,*,*,*,特殊・タ,基本形,た,タ,タ
国会	名詞,一般,*,*,*,*,国会,コッカイ,コッカイ
における	助詞,格助詞,連語,*,*,*,における,ニオケル,ニオケル
代表	名詞,サ変接続,*,*,*,*,代表,ダイヒョウ,ダイヒョー
者	名詞,接尾,一般,*,*,*,者,シャ,シャ
を通じて	助詞,格助詞,連語,*,*,*,を通じて,ヲツウジテ,ヲツージテ

行動	名詞,サ変接続,*,*,*,*,行動,コウドウ,コードー
し	動詞,自立,*,*,サ変・スル,連用形,する,シ,シ
EOS
```

日本語の文章から，単語を切り出し，品詞を自動判定する Python プログラム

```
import sys
import MeCab
m = MeCab.Tagger("-Ochasen")
print(m.parse ("日本国民は、正当に選挙された国会における代表者を通じて行動し"))
```

```
In [2]: import sys
   ...: import MeCab
   ...: m = MeCab.Tagger("-Ochasen")
   ...: print(m.parse ("日本国民は、正当に選挙された国会における代表者を通じて行動し"))
日本 ニッポン    日本 名詞-固有名詞-地域-国
国民 コクミン    国民 名詞-一般
は   ハ   は    助詞-係助詞
、   、   、    記号-読点
正当 セイトウ    正当 名詞-形容動詞語幹
に   ニ   に    助詞-副詞化
選挙 センキョ    選挙 名詞-サ変接続
```

（前準備）MeCab のインストール，Python の mecab パッケージのインストール

まとめ

- **単語**　単一の意味のまとまり
- **品詞**　単語の種別

7-2 HTML（Web ページ）のタグの除去，単語の切り出し

Web ページの例
URL: https://www.kkaneko.jp

① Web ページのダウンロード（Python プログラム）

```
import urllib.request
r =  urllib.request.urlopen('https://www.kkaneko.jp')
html = r.read()
print(html.decode())
```

```
In [9]: import urllib.request
   ...: r =  urllib.request.urlopen('https://www.kkaneko.jp')
   ...: html = r.read()
   ...: print(html.decode())
<!DOCTYPE html>
<html lang="ja">
<head>
<meta content="text/html; charset=utf-8" http-equiv="Content-Type">
<meta content="text/javascript" http-equiv="Content-Script-Type">
<meta content="text/css" http-equiv="Content-Style-Type">
<meta content="width=device-width, initial-scale=1.0, maximum-scale=1.0, minimum-
scale=1.0" name="viewport">
<link href="https://www.kkaneko.jp/css/blue.css" rel="stylesheet" type="text/
css">
<link href="./index-j.html" rel="contents">
<link href="https://www.kkaneko.jp/sitemap-j.html" rel="index">
<title>トップページ(金子邦彦研究室)</title>
</head>
<body>
<div align="left"><a href="index-j.html">トップページ</a>
```

②①でダウンロードしたWebページからHTMLタグを取りのぞく（テキストとJavaScriptプログラムが残る）　（Pythonプログラム）


```
from bs4 import BeautifulSoup
soup = BeautifulSoup(html,'html5lib')
t = soup.get_text()
print(t)
```

```
In [10]: from bs4 import BeautifulSoup
    ...: soup = BeautifulSoup(html,'html5lib')
    ...: t = soup.get_text()
    ...: print(t)
```

```
トップページ（金子邦彦研究室）


トップページ




[サイトマップへ],
[サイト内検索へ],
[アクセスログへ],
[英語版へ]
```

③ ②の結果を切り分ける
※ 改行や空白文字を区切りとして切り分ける

```
tokens = [i for i in t.split()]
print(tokens)
```

```
In [11]: tokens = [i for i in t.split()]
    ...: print(tokens)
['トップページ(金子邦彦研究室)', 'トップページ', '[サイトマップへ],', '[サイト内検索へ],', '[アクセスログへ],', '[英語版へ]', 'トップページ(金子邦彦研究室)', 'サイト構成', 'データベース関連技術', 'データの扱い', 'インストール, 設定, 利用', 'プログラミング', '講義実習資料', 'サポートページ', '連絡先, 業績など', '金子邦彦研究室:', 'データベース、人工知能(AI)、データサイエンスの融合により不可能を可能にする', '【掲載しているトピックス】', 'データベース関連技術', 'データベース関連分野の技術を, すぐに, 手元のパソコン等で実験, 実施, 評価する手順等について.', '(データサイエンス, 人工知能, コンピュータグラフィックス, コンピュータビジョン等のデータベース周辺領域を網羅)', 'データの扱い', 'オープンデータ, データの下準備, データ解析等について.', 'インストール, 設定, 利用', 'さまざまなOSやアプリケーションのインストール(Windows, Linux), 種々の設定, Raspberry', 'Pi,', 'データベースシステム, 仮想マシン, 情報ネットワーク, サーバの運用保守などについて.', 'プログラミング', 'Octave,', 'Java,', 'JavaScript,', 'Ruby', 'などのプログラミング. それらのデータベースや', 'Web', '連携.', '講義実習資料', 'データベース, 人工知能, コンピュータグラフィックス,', 'プログラミング, パソコン活用などの授業資料(パワーポイントファイル, PDF', 'ファイル, 動画など)の公開版.', 'サポートページ', '授業, 研究室活動, 地域連携活動メンバーへの案内', '連絡先, 業績など', '金子邦彦研究室では、こんなことを行っている.', '【我々が目指すもの】',
```

④ ②の結果を，単語に切り分ける

```
import sys
import MeCab
m = MeCab.Tagger("-Ochasen")
a = m.parse(t)
words = [i.split()[0] for i in a.splitlines()]
print(words)
```

```
In [47]: import sys
    ...: import MeCab
    ...: m = MeCab.Tagger("-Ochasen")
    ...: a = m.parse(t)
    ...: words = [i.split()[0] for i in a.splitlines()]
    ...: print(words)
['トップページ', '(', '金子', '邦彦', '研究', '室', ')', 'トップページ', '[', 'サイト', 'マッ
プ', 'へ', '],', '[', 'サイト', '内', '検索', 'へ', '],', '[', 'アクセス', 'ログ', 'へ',
'],', '[', '英語', '版', 'へ', ']', 'トップページ', '(', '金子', '邦彦', '研究', '室',
')', 'サイト', '構成', 'データベース', '関連', '技術', 'データ', 'の', '扱い', 'インストール',
'，', '設定', '，', '利用', 'プログラミング', '講義', '実習', '資料', 'サポート', 'ページ',
'連絡', '先', '，', '業績', 'など', '金子', '邦彦', '研究', '室', ':', '記号-空白',
'データベース', '、', '人工', '知能', '(', 'AI', ')、', 'データ', 'サイエンス', 'の', '融合',
'により', '不可能', 'を', '可能', 'に', 'する', '【', '掲載', 'し', 'て', 'いる', 'トピック
ス', '】', 'データベース', '関連', '技術', 'データベース', '関連', '分野', 'の', '技術',
'を', '，', 'すぐ', 'に', '，', '手元', 'の', 'パソコン', '等', 'で', '実験', '，', '実施',
'，', '評価', 'する', '手順', '等', 'について', '．', '(', 'データ', 'サイエンス', '，', '人
工', '知能', '，', 'コンピュータグラフィックス', '，', 'コンピュータ', 'ビジョン', '等', 'の', 'データ
```

7-3 音声合成の例

音声合成の例

音声合成を，Python プログラムで行う

（前準備）

- マイクロソフトから，Windows 用の音声合成機能をダウンロード
- Python と pywin32 パッケージをインストール

Jupyter QtConsole

File Edit View Kernel Window Help

```
Jupyter QtConsole 4.7.4
Python 3.7.7 (tags/v3.7.7:d7c567b08f, Mar 10 2020, 10:41:24)
[MSC v.1900 64 bit (AMD64)]
Type 'copyright', 'credits' or 'license' for more information
IPython 7.15.0 -- An enhanced Interactive Python. Type '?' for
help.

In [1]: import win32com.client
   ...: tts = win32com.client.Dispatch("Sapi.SpVoice")
   ...: tts.Speak("日本国民は、正当に選挙された国会における代表者を通じ
て行動し、われらとわれらの子孫のために、諸国民との協和による成果と、わが国全土
にわたつて自由のもたらす恵沢を確保し、政府の行為によつて再び戦争の惨禍が起る
ことのないやうにすることを決意し、ここに主権が国民に存することを宣言し、この憲法を
確定する。そもそも国政は、国民の厳粛な信託によるものであつて、その権威は国民に
由来し、その権力は国民の代表者がこれを行使し、その福利は国民がこれを享受す
る。これは人類普遍の原理であり、この憲法は、かかる原理に基くものである。われら
は、これに反する一切の憲法、法令及び詔勅を排除する。");

In [2]:
```