プロジェクト第２回

NLPイントロダクション

環境構築

形態素解析をするために

- 形態素解析とは

  テキストを辞書に基づいて品詞などを判別すること

文字列	読み	原形	品詞の種類	活用の種類	活用形
お待ち	オマチ	お待ち	名詞-サ変接続		
し	シ	する	動詞-自立	サ変・スル	連用形
て	テ	て	助詞-接続助詞		
おり	オリ	おる	動詞-非自立	五段・ラ行	連用形
ます	マス	ます	助動詞	特殊・マス	基本形
。	。	。	記号-句点		

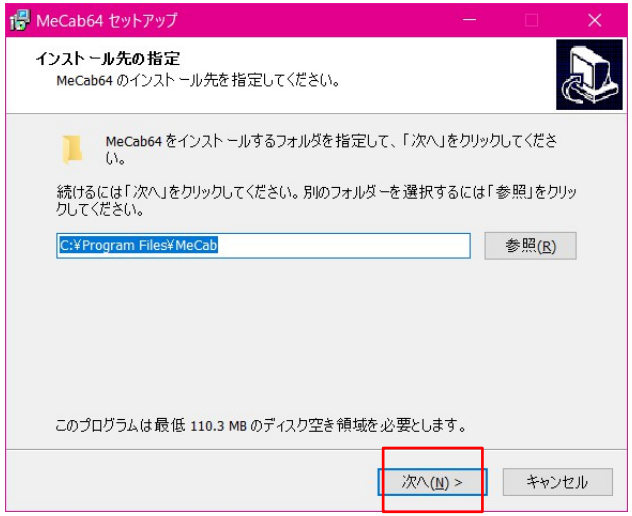
- 形態素解析器MeCabを使用する

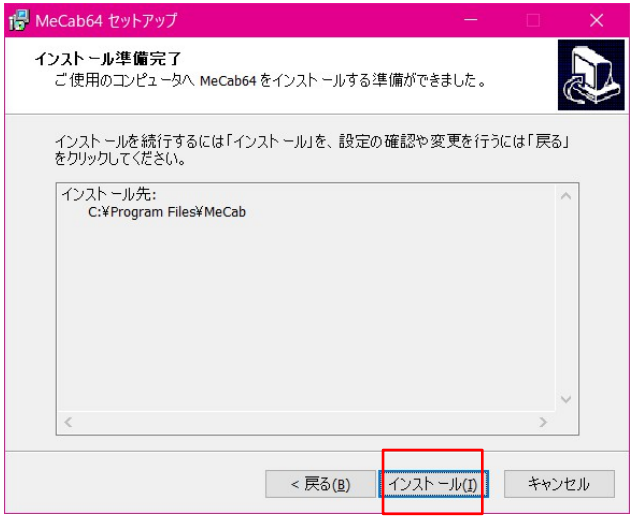
インストールする

- https://github.com/ikegami-yukino/mecab/releases/tag/v0.996

にアクセスする。mecab-0.996-64.exeを選択してダウンロードする。

ダウンロード後、そのファイルを実行する。





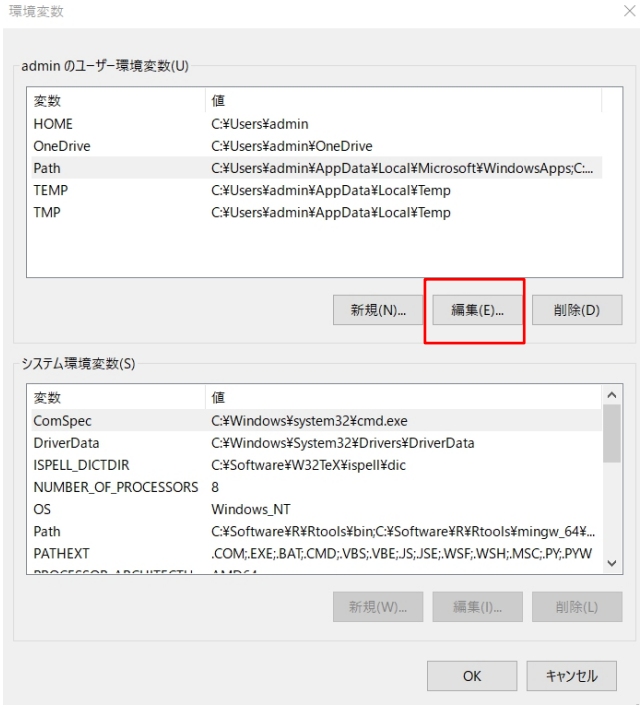
環境変数の設定

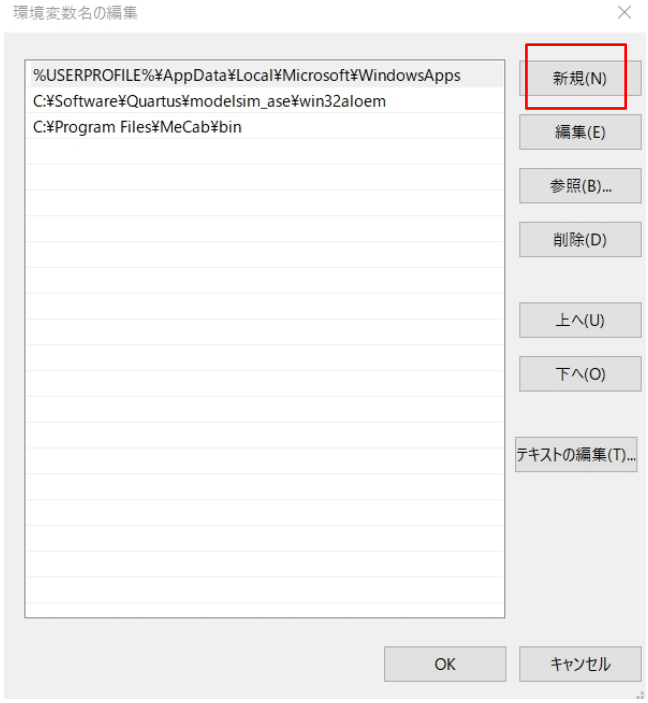
- Cortanaで「環境変数」と検索して、”環境変数の編集”で”path”を選択し、”編集”の”新規”で

  C:¥Program Files¥MeCab¥bin

  を追加する。

  OKで終了





pythonにMeCabを導入

- Cortana（左下の検索のこと）でcmdと検索する。そして、コマンドプロンプトを右クリックして「管理者として実行」を選択する。

pip install mecab-python-windows

または

python –m pip install mecab-python-windows

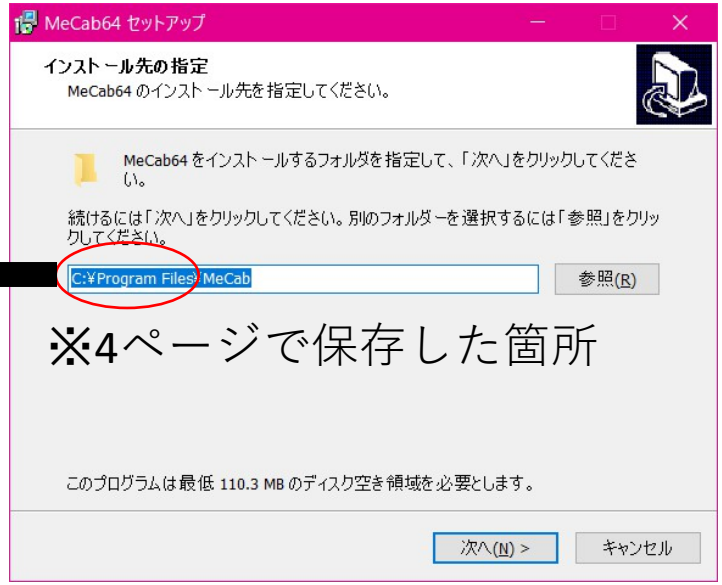
エラーが出た場合

- https://github.com/ikegami-yukino/mecab/releases/tag/v0.996
  にアクセスする。Source code(gz)を選択する。

解凍後、setup.pyの中身を変更する。（次へ）

setup.py

```
version="0.996",
py_modules=["MeCab"],
ext_modules=[
    Extension("_MeCab",
    ["MeCab_wrap.cxx",],
    include_dirs=[r"～～～¥MeCab¥sdk"]
    library_dirs=[r"～～～¥MeCab¥sdk"]
    libraries=["libmecab"]
```



そのmecabフォルダの中にpythonがあるのでpythonの中身までコマンドプロンプト（管理者）からcdコマンドで移動する。

移動後、setup.pyがあることを確認する。

管理者コマンドプロンプトで

python setup.py build

python setup.py install

を入力する。

Pythonで実行を確認する

- IDLEを起動して以下のコードを実行する

```
>>> import MeCab
>>> m=MeCab.Tagger("-Ochasen")
>>> print(m.parse("すももももももものうち"))
```

- エラーが出なければOK

NLPの基本

Ex2_1. 文字列のスライス

ex=“abcdefg”

print(ex[1:4]) # 1番目から4番目まで
print(ex[:4])　# 0番目から4番目まで
print(ex[3:])　# 3番目から最後まで
print(ex[::-2]) # 末尾から

Ex2_2. 文字列のステップ

ex=“abcdefg”

print(ex[::2]) # 最初から最後まで2つおきに
print(ex[1:5:2]) # 1番目から4番目まで2つおきに

演習1. 文字列の逆順

文字列”abcdefg”の文字を逆順にせよ.

Ex2_3. 文字列の連結

ex1=“abc”

ex2=“def”

print(ex1+ex2)

演習2. ホウセイダイガク

“ホセダガ”+“ウイイク”の文字を先頭から交互に連結せよ.

Ex2_4. 文字列の分割

ex=“Hello,world”

print(ex.split(“,”))

演習3. 円周率

“Now I need a drink, alcoholic of course, after the heavy lectures involving quantum mechanics.”という文を単語に分解し、各単語の（アルファベット）の文字数を先頭から出現順に並べたリストを作成せよ.

ヒント：

“AAA”.replace(“A”, “B”)を実行

演習4. 元素記号

“Hi He Lied Because Boron Could Not Oxidize Fluorine. New Nations Might Also Sign Peace Security Clause. Arthur King Can.”という文を単語に分解し、1,5,6,7,8,9,15,16,19番目の単語は先頭の1文、それ以外の単語は先頭に2文字を取り出し、取り出した文字列から単語の位置（先頭から何番目の単語か）への連想配列（辞書型またはマップ型）を作成せよ．（マグネシウムはMiで表す）

周期＼族	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	1 H 水素 Hydrogen 1.00798																	2 He ヘリウム Helium 4.0026
2	3 Li リチウム Lithium 6.968	4 Be ベリリウム Beryllium 9.01218											5 B 硼(ホウ)素 Boron 10.814	6 C 炭素 Carbon 12.0106	7 N 窒素 Nitrogen 14.0069	8 O 酸素 Oxygen 15.9994	9 F 弗(フッ)素 Fluorine 18.9984	10 Ne ネオン Neon 20.1797
3	11 Na ナトリウム Sodium 22.9898	12 Mg マグネシウム Magnesium 24.306											13 Al アルミニウム Aluminum 26.9815	14 Si 珪(ケイ)素 Silicon 28.085	15 P 燐(リン) Phosphorus 30.9738	16 S 硫黄 Sulfur 32.068	17 Cl 塩素 Chlorine 35.452	18 Ar アルゴン Argon 39.948
4	19 K カリウム Potassium 39.0983	20 Ca カルシウム Calcium 40.078	21 Sc スカンジウム Scandium 44.9559	22 Ti チタン Titanium 47.867	23 V バナジウム Vanadium 50.9415	24 Cr クロム Chromium 51.9961	25 Mn マンガン Manganese 54.938	26 Fe 鉄 Iron 55.845	27 Co コバルト Cobalt 58.9332	28 Ni ニッケル Nickel 58.6934	29 Cu 銅 Copper 63.546	30 Zn 亜鉛 Zinc 65.38	31 Ga ガリウム Gallium 69.723	32 Ge ゲルマニウム Germanium 72.630	33 As 砒(ヒ)素 Arsenic 74.9216	34 Se セレン Selenium 78.971	35 Br 臭素 Bromine 79.904	36 Kr クリプトン Krypton 83.798

演習5. テンプレート文生成

引数x, y, zを受け取り「x時のyはz」という文字列を返す関数を作成せよ。さらに、x=12, y=“気温”, z=22.4として、実行結果を確認せよ。

演習6. Typoglycemia

スペースで区切られた単語列に対して、各単語の先頭と末尾の文字は残し、それ以外の文字の順序をランダムに並び替えるプログラムを作成せよ。ただし、長さが４以下の単語は並び替えないこととする。適当な英語の文を与え、その実行結果を確認せよ。