付録: 日本語文章の形態素解析：素人による入門

Author: 藤原 義久 yoshi.fujiwara@gmail.com (mailto:yoshi.fujiwara@gmail.com)
Data: ディレクトリ"data"以下

- 短いサンプル文章
- 日本国憲法

形態素解析のツール**mecab**

- mecab 本家: https://taku910.github.io/mecab/ (https://taku910.github.io/mecab/)
- mecab 自体のインストール
  - Windows: https://github.com/ikegami-yukino/mecab/releases/tag/v0.996 (https://github.com/ikegami-yukino/mecab/releases/tag/v0.996)
    「MeCab 0.996 64bit version (旧)」にあるインストーラ"mecab-0.996-64.exe"を使う
    注：環境変数 PATH にmecab をインストールしたパスを追加(例: C:\w10\mecab\bin)
    注：環境変数 MECABRC を新たに追加して設定(例: C:\w10\mecab\etc\mecabrc)
  - Linux: "mecab linux"でググるとUbuntu, CentOS などでのインストール方法が分かる
  - Mac: "mecab mac"でググる(未確認)
- python からmecab を使うパッケージのインストール
  jupyter notebook を新規に開いて以下を実行する(先頭の"!"に注意)
  - Windows:

```
!pip install mecab-python-windows
```

  - Linux:

```
!pip install mecab-python3
```

  - Mac: Linuxと同じ(?)

In [1]:
```
# !pip install mecab-python-windows  # Windows の場合のインストール
# !pip install mecab-python3  # Linux/Mac の場合のインストール
```

```
Collecting mecab-python3
  Downloading mecab_python3-1.0.1-cp37-cp37m-manylinux2010_x86_64.whl (3.5 M
B)
     |████████████████████████████████| 3.5 MB 3.6 MB/s eta 0:00:01
Installing collected packages: mecab-python3
Successfully installed mecab-python3-1.0.1
```

パッケージの読み込み

In [1]:
```
import MeCab
```

使い方

In [2]:
```
# tagger = MeCab.Tagger()  # Windows の場合
tagger = MeCab.Tagger("-r /etc/mecabrc")  # Linux の場合
```

In [3]:
```
print(tagger.parse("すももももももものうち"))
```

```
すもも	名詞,一般,*,*,*,*,すもも,スモモ,スモモ
も	助詞,係助詞,*,*,*,*,も,モ,モ
もも	名詞,一般,*,*,*,*,もも,モモ,モモ
も	助詞,係助詞,*,*,*,*,も,モ,モ
もも	名詞,一般,*,*,*,*,もも,モモ,モモ
の	助詞,連体化,*,*,*,*,の,ノ,ノ
うち	名詞,非自立,副詞可能,*,*,*,うち,ウチ,ウチ
EOS
```

In [4]:
```
# 文章：「今日はきれいな虹が出た。」
with open("data/sample1/01.txt", encoding="utf_8") as fin:
    s = fin.read()
    print(tagger.parse(s))
```

```
今日	名詞,副詞可能,*,*,*,*,今日,キョウ,キョー
は	助詞,係助詞,*,*,*,*,は,ハ,ワ
きれい	名詞,形容動詞語幹,*,*,*,*,きれい,キレイ,キレイ
な	助動詞,*,*,*,特殊・ダ,体言接続,だ,ナ,ナ
虹	名詞,一般,*,*,*,*,虹,ニジ,ニジ
が	助詞,格助詞,一般,*,*,*,が,ガ,ガ
出	動詞,自立,*,*,一段,連用形,出る,デ,デ
た	助動詞,*,*,*,特殊・タ,基本形,た,タ,タ
。	記号,句点,*,*,*,*,。,。,。
EOS
```

In [5]:
```
# 文章：日本国憲法前文(現代語)
with open("data/sample1/02.txt", encoding="utf_8") as fin:
    s = fin.read()
    print(tagger.parse(s))
```

```
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	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
、	記号,読点,*,*,*,*,、,、,、
われ	名詞,代名詞,一般,*,*,*,われ,ワレ,ワレ
ら	名詞,接尾,一般,*,*,*,ら,ラ,ラ
と	助詞,並立助詞,*,*,*,*,と,ト,ト
われ	名詞,代名詞,一般,*,*,*,われ,ワレ,ワレ
ら	名詞,接尾,一般,*,*,*,ら,ラ,ラ
の	助詞,連体化,*,*,*,*,の,ノ,ノ
子孫	名詞,一般,*,*,*,*,子孫,シソン,シソン
の	助詞,連体化,*,*,*,*,の,ノ,ノ
ため	名詞,非自立,副詞可能,*,*,*,ため,タメ,タメ
に	助詞,格助詞,一般,*,*,*,に,ニ,ニ
、	記号,読点,*,*,*,*,、,、,、
諸	接頭詞,名詞接続,*,*,*,*,諸,ショ,ショ
国民	名詞,一般,*,*,*,*,国民,コクミン,コクミン
と	助詞,格助詞,一般,*,*,*,と,ト,ト
の	助詞,連体化,*,*,*,*,の,ノ,ノ
協和	名詞,サ変接続,*,*,*,*,協和,キョウワ,キョーワ
による	助詞,格助詞,連語,*,*,*,による,ニヨル,ニヨル
成果	名詞,一般,*,*,*,*,成果,セイカ,セイカ
と	助詞,並立助詞,*,*,*,*,と,ト,ト
、	記号,読点,*,*,*,*,、,、,、
わが国	名詞,一般,*,*,*,*,わが国,ワガクニ,ワガクニ
全土	名詞,一般,*,*,*,*,全土,ゼンド,ゼンド
にわたって	助詞,格助詞,連語,*,*,*,にわたって,ニワタッテ,ニワタッテ
自由	名詞,形容動詞語幹,*,*,*,*,自由,ジユウ,ジユー
の	助詞,格助詞,一般,*,*,*,の,ノ,ノ
もたらす	動詞,自立,*,*,五段・サ行,基本形,もたらす,モタラス,モタラス
恵沢	名詞,一般,*,*,*,*,恵沢,ケイタク,ケイタク
を	助詞,格助詞,一般,*,*,*,を,ヲ,ヲ
確保	名詞,サ変接続,*,*,*,*,確保,カクホ,カクホ
し	動詞,自立,*,*,サ変・スル,連用形,する,シ,シ
、	記号,読点,*,*,*,*,、,、,、
政府	名詞,一般,*,*,*,*,政府,セイフ,セイフ
の	助詞,連体化,*,*,*,*,の,ノ,ノ
行為	名詞,サ変接続,*,*,*,*,行為,コウイ,コーイ
によって	助詞,格助詞,連語,*,*,*,によって,ニヨッテ,ニヨッテ
再び	副詞,助詞類接続,*,*,*,*,再び,フタタビ,フタタビ
戦争	名詞,サ変接続,*,*,*,*,戦争,センソウ,センソー
の	助詞,連体化,*,*,*,*,の,ノ,ノ
惨禍	名詞,一般,*,*,*,*,惨禍,サンカ,サンカ
が	助詞,格助詞,一般,*,*,*,が,ガ,ガ
起る	動詞,自立,*,*,五段・ラ行,基本形,起る,オコル,オコル
こと	名詞,非自立,一般,*,*,*,こと,コト,コト
の	助詞,格助詞,一般,*,*,*,の,ノ,ノ
ない	形容詞,自立,*,*,形容詞・アウオ段,基本形,ない,ナイ,ナイ
よう	名詞,非自立,助動詞語幹,*,*,*,よう,ヨウ,ヨー
に	助詞,格助詞,一般,*,*,*,に,ニ,ニ
する	動詞,自立,*,*,サ変・スル,基本形,する,スル,スル
こと	名詞,非自立,一般,*,*,*,こと,コト,コト
を	助詞,格助詞,一般,*,*,*,を,ヲ,ヲ
決意	名詞,サ変接続,*,*,*,*,決意,ケツイ,ケツイ
```

In [6]:
```
# 結果をくわしく見るには

with open("data/sample1/02.txt", encoding="utf-8") as fin:
    s = fin.read()
    node = tagger.parseToNode(s)

while node:
    print("%s\t%s" % (node.surface, node.feature))
    node = node.next
```

```
        BOS/EOS,*,*,*,*,*,*,*,*
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      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
、      記号,読点,*,*,*,*,、,、,、
われ    名詞,代名詞,一般,*,*,*,われ,ワレ,ワレ
ら      名詞,接尾,一般,*,*,*,ら,ラ,ラ
と      助詞,並立助詞,*,*,*,*,と,ト,ト
われ    名詞,代名詞,一般,*,*,*,われ,ワレ,ワレ
ら      名詞,接尾,一般,*,*,*,ら,ラ,ラ
の      助詞,連体化,*,*,*,*,の,ノ,ノ
子孫    名詞,一般,*,*,*,*,子孫,シソン,シソン
の      助詞,連体化,*,*,*,*,の,ノ,ノ
ため    名詞,非自立,副詞可能,*,*,*,ため,タメ,タメ
に      助詞,格助詞,一般,*,*,*,に,ニ,ニ
、      記号,読点,*,*,*,*,、,、,、
諸      接頭詞,名詞接続,*,*,*,*,諸,ショ,ショ
国民    名詞,一般,*,*,*,*,国民,コクミン,コクミン
と      助詞,格助詞,一般,*,*,*,と,ト,ト
の      助詞,連体化,*,*,*,*,の,ノ,ノ
協和    名詞,サ変接続,*,*,*,*,協和,キョウワ,キョーワ
による  助詞,格助詞,連語,*,*,*,による,ニヨル,ニヨル
成果    名詞,一般,*,*,*,*,成果,セイカ,セイカ
と      助詞,並立助詞,*,*,*,*,と,ト,ト
、      記号,読点,*,*,*,*,、,、,、
わが国  名詞,一般,*,*,*,*,わが国,ワガクニ,ワガクニ
全土    名詞,一般,*,*,*,*,全土,ゼンド,ゼンド
にわたって      助詞,格助詞,連語,*,*,*,にわたって,ニワタッテ,ニワタッテ
自由    名詞,形容動詞語幹,*,*,*,*,自由,ジユウ,ジユー
の      助詞,格助詞,一般,*,*,*,の,ノ,ノ
もたらす 動詞,自立,*,*,五段・サ行,基本形,もたらす,モタラス,モタラス
恵沢    名詞,一般,*,*,*,*,恵沢,ケイタク,ケイタク
を      助詞,格助詞,一般,*,*,*,を,ヲ,ヲ
確保    名詞,サ変接続,*,*,*,*,確保,カクホ,カクホ
し      動詞,自立,*,*,サ変・スル,連用形,する,シ,シ
、      記号,読点,*,*,*,*,、,、,、
政府    名詞,一般,*,*,*,*,政府,セイフ,セイフ
の      助詞,連体化,*,*,*,*,の,ノ,ノ
行為    名詞,サ変接続,*,*,*,*,行為,コウイ,コーイ
によって 助詞,格助詞,連語,*,*,*,によって,ニヨッテ,ニヨッテ
再び    副詞,助詞類接続,*,*,*,*,再び,フタタビ,フタタビ
戦争    名詞,サ変接続,*,*,*,*,戦争,センソウ,センソー
の      助詞,連体化,*,*,*,*,の,ノ,ノ
惨禍    名詞,一般,*,*,*,*,惨禍,サンカ,サンカ
が      助詞,格助詞,一般,*,*,*,が,ガ,ガ
起る    動詞,自立,*,*,五段・ラ行,基本形,起る,オコル,オコル
こと    名詞,非自立,一般,*,*,*,こと,コト,コト
の      助詞,格助詞,一般,*,*,*,の,ノ,ノ
ない    形容詞,自立,*,*,形容詞・アウオ段,基本形,ない,ナイ,ナイ
よう    名詞,非自立,助動詞語幹,*,*,*,よう,ヨウ,ヨー
に      助詞,格助詞,一般,*,*,*,に,ニ,ニ
する    動詞,自立,*,*,サ変・スル,基本形,する,スル,スル
こと    名詞,非自立,一般,*,*,*,こと,コト,コト
を      助詞,格助詞,一般,*,*,*,を,ヲ,ヲ
```

応用例：名詞だけを取り出して、それぞれの頻度を調べる

In [7]:
```
from collections import Counter
```

In [8]:
```
# 名詞だけを取り出して、それぞれの頻度を調べる
with open("data/sample1/02.txt", encoding="utf-8") as fin:
    s = fin.read()
    node = tagger.parseToNode(s)

l = []
while node:
    if node.feature.split(',')[0] == "名詞":
        l.append(node.surface)
    node = node.next

print(l)

freq = Counter(l)
for k,v in sorted(freq.items(), key=lambda x:x[1], reverse=True):
    print("%s\t%d" % (k,v))
```

['日本', '国民', '正当', '選挙', '国会', '代表', '者', '行動', 'われ', 'ら', 'われ', 'ら', '子孫', 'ため', '国民', '協和', '成果', 'わが国', '全土', '自由', '恵沢', '確保', '政府', '行為', '戦争', '惨禍', 'こと', 'よう', 'こと', '決意', 'ここ', '主権', '国民', 'こと', '宣言', '憲法', '確定', '国政', '国民', '厳粛', '信託', 'もの', '権威', '国民', '由来', '権力', '国民', '代表', '者', 'これ', '行使', '福利', '国民', 'これ', '享受', 'これ', '人類', '普遍', '原理', '憲法', '原理', 'もの', 'われ', 'ら', 'これ', '一切', '憲法', '法令', '詔勅', '排除', '日本', '国民', '恒久', '平和', '念願', '人間', '相互', '関係', '支配', '崇高', '理想', '自覚', 'の', '平和', '国民', '公正', '信義', '信頼', 'われ', 'ら', '安全', '生存', '保持', '決意', 'われ', 'ら', '平和', '維持', '専制', '隷従', '圧迫', '偏狭', '地上', '永遠', '除去', '国際', '社会', '名誉', '地位', 'われ', 'ら', '世界', '国民', '恐怖', '欠乏', '平和', 'うち', '生存', '権利', 'こと', '確認', 'われ', 'ら', '国家', '自国', 'こと', '専念', '他国', '無視', 'の', '政治', '道徳', '法則', '普遍', '的', 'もの', '法則', 'こと', '自国', '主権', '維持', '他国', '対等', '関係', '各国', '責務', '日本', '国民', '国家', '名誉', '全力', '崇高', '理想', '目的', '達成', 'こと']
国民　　11
われ　　7
ら　　7
こと　　7
これ　　4
平和　　4
日本　　3
憲法　　3
もの　　3
代表　　2
者　　2
決意　　2
主権　　2
普遍　　2
原理　　2
関係　　2
崇高　　2
理想　　2
の　　2
生存　　2
維持　　2
名誉　　2
国家　　2
自国　　2
他国　　2
法則　　2
正当　　1
選挙　　1
国会　　1
行動　　1
子孫　　1
ため　　1
協和　　1
成果　　1
わが国　　1
全土　　1
自由　　1
恵沢　　1
確保　　1
政府　　1
行為　　1
戦争　　1
惨禍　　1
よう　　1
ここ　　1
宣言　　1
確定　　1
国政　　1
厳粛　　1
信託　　1
権威　　1
由来　　1
権力　　1

応用例：名詞、動詞、形容詞、助詞だけを選んで文書の「ダイジェスト」を作る

In [9]:
```
# 名詞、動詞、形容詞、助詞だけを選んで文書の「ダイジェスト」を作る関数を定義

def digest_doc(filename):
    with open(filename, encoding="utf-8") as fin:
        s = fin.read()
        node = tagger.parseToNode(s)

    l = []
    while node:
        x = node.feature.split(',')[0]
        if x == "名詞" or x == "動詞" or x == "形容詞" or x == "助詞":
            l.append(node.feature.split(',')[6])  # 原形を用いる
        node = node.next

    return " ".join(l)
```

In [10]:
```
d = digest_doc("data/sample1/02.txt")
d
```

Out[10]:
```
'日本 国民 は 正当 に 選挙 する れる 国会 における 代表 者 を通じて 行動 する われ ら と われ ら の 子孫 の ため に 国民 と の 協和 による 成果 と わが国 全土 にわたって 自由 の もたらす 恵沢 を 確保 する 政府 の 行為 によって 戦争 の 惨禍 が 起る こと の ない よう に する こと を 決意 する ここ に 主権 が 国民 に 存する こと を 宣言 する 憲法 を 確定 する 国政 は 国民 の 厳粛 信託 による もの て 権威 は 国民 に 由来 する 権力 は 国民 の 代表 者 が これ を 行使 する 福利 は 国民 が これ を 享受 する これ は 人類 普遍 の 原理 憲法 は 原理 に 基づく もの われ ら は これ に 反する 一切 の 憲法 法令 詔勅 を 排除 する 日本 国民 は 恒久 の 平和 を 念願 する 人間 相互 の 関係 を 支配 する 崇高 理想 を 深い 自覚 する の て 平和 を 愛する 国民 の 公正 と 信義 に 信頼 する て われ ら の 安全 と 生存 を 保持 する と 決意 する われ ら は 平和 を 維持 する 専制 と 隷従 圧迫 と 偏狭 を 地上 から 永遠 に 除去 する と 努める て いる 国際 社会 において 名誉 ある 地位 を 占める と 思う われ ら は 世界 の 国民 が ひとしい 恐怖 と 欠乏 から 免れる 平和 の うち に 生存 する 権利 を 有する こと を 確認 する われ ら は いづ の 国家 も 自国 の こと のみ に 専念 する て 他国 を 無視 する て は なる の て 政治 道徳 の 法則 は 普遍 的 もの 法則 に 従う こと は 自国 の 主権 を 維持 する 他国 と 対等 関係 に 立つ と する 各国 の 責務 と 信ずる 日本 国民 は 国家 の 名誉 に かける 全力 を あげる て 崇高 理想 と 目的 を 達成 する こと を 誓う'
```

応用例：ディレクトリ以下のすべての文書について処理する

In [11]:
```
import glob
```

In [12]:
```
docs = []
for fn in glob.glob("data/sample2/*"):
    print(fn)
    d = digest_doc(fn)
    docs.append(d)
```

```
data/sample2/04.txt
data/sample2/03.txt
```

In [13]:
```
docs
```

Out[13]:
```
['国民 は すべて の 基本 的 人権 の 享有 を 妨げる られる 憲法 が 国民 に 保障 する 基本
的 人権 は 侵す こと の できる 永久 の 権利 として 現在 将来 の 国民 に 与える られる 憲
法 が 国民 に 保障 する 自由 権利 は 国民 の 不断 の 努力 によって これ を 保持 する ば
なる 国民 は これ を 濫用 する て は なる の て 公共 の 福祉 の ため に これ を 利用 す
る 責任 を 負う すべて 国民 は 個人 として 尊重 する れる 生命 自由 幸福 追求 に対する 国
民 の 権利 について は 公共 の 福祉 に 反する 限り 立法 その他 の 国政 の 上 で 最大 の
尊重 を 必要 と する すべて 国民 は 法 の 下 に 平等 て 人種 信条 性別 社会 的 身分 門地
により 政治 的 経済 的 社会 的 関係 において 差別 する れる 華族 その他 の 貴族 の 制度
は これ を 認める 栄誉 勲章 その他 の 栄典 の 授与 は 特権 も 伴う 栄典 の 授与 は これ
を 有す 将来 これ を 受ける 者 の 一代 に 限る 効力 を 有する',
 '日本 国民 は 正当 に 選挙 する れる 国会 における 代表 者 を通じて 行動 する われ ら と
われ ら の 子孫 の ため に 国民 と の 協和 による 成果 と わが国 全土 にわたって 自由 の
もたらす 恵沢 を 確保 する 政府 の 行為 によって 戦争 の 惨禍 が 起る こと の ない よう
に する こと を 決意 する ここ に 主権 が 国民 に 存する こと を 宣言 する 憲法 を 確定
する 国政 は 国民 の 厳粛 信託 による もの て 権威 は 国民 に 由来 する 権力 は 国民 の
代表 者 が これ を 行使 する 福利 は 国民 が これ を 享受 する これ は 人類 普遍 の 原理
憲法 は 原理 に 基づく もの われ ら は これ に 反する 一切 の 憲法 法令 詔勅 を 排除 する
日本 国民 は 恒久 の 平和 を 念願 する 人間 相互 の 関係 を 支配 する 崇高 理想 を 深い
自覚 する の て 平和 を 愛する 国民 の 公正 と 信義 に 信頼 する て われ ら の 安全 と
生存 を 保持 する と 決意 する われ ら は 平和 を 維持 する 専制 と 隷従 圧迫 と 偏狭 を
地上 から 永遠 に 除去 する と 努める て いる 国際 社会 において 名誉 ある 地位 を 占める
と 思う われ ら は 世界 の 国民 が ひとしい 恐怖 と 欠乏 から 免れる 平和 の うち に 生存
する 権利 を 有する こと を 確認 する われ ら は いづ の 国家 も 自国 の こと のみ に 専
念 する て 他国 を 無視 する て は なる の て 政治 道徳 の 法則 は 普遍 的 もの 法則 に
従う こと は 自国 の 主権 を 維持 する 他国 と 対等 関係 に 立つ と する 各国 の 責務 と
信ずる 日本 国民 は 国家 の 名誉 に かける 全力 を あげる て 崇高 理想 と 目的 を 達成 す
る こと を 誓う']
```

ディレクトリ以下のすべての文書について語とその頻度の表を作る

メモ：テキスト解析で、語とその頻度の表はterm-frequency matrix と呼ばれている
以下では機械学習の学習用パッケージ scikit-learn からテキスト解析のツールを用いる

In [14]:
```
import numpy as np
import pandas as pd
from sklearn.feature_extraction.text import CountVectorizer
```

In [15]:
```
count_vec = CountVectorizer()

x = count_vec.fit_transform(np.array(docs))
# 疎な行列として扱われている
# print(type(X))

td = x.toarray()  # term-document matrix

# 出現したすべての語のリスト
terms = count_vec.get_feature_names()
print(terms)

# term-frequency matrix の次元 = 文書数 * 全語数
print(td.shape)

# term-frequency の中身
print(td)
```

```
['あげる', 'ある', 'いづ', 'いる', 'うち', 'かける', 'から', 'ここ', 'こと', 'これ',
'すべて', 'する', 'その他', 'ため', 'できる', 'として', 'ない', 'なる', 'において', '
において', 'について', 'によって', 'により', 'による', 'にわたって', 'に対する', 'のみ
', 'ひとしい', 'もたらす', 'もの', 'よう', 'られる', 'れる', 'わが国', 'われ', 'を通
じて', '一代', '一切', '不断', '与える', '世界', '主権', '享受', '享有', '人権', '人
種', '人間', '人類', '他国', '代表', '伴う', '侵す', '保持', '保障', '信ずる', '信条
', '信義', '信託', '信頼', '個人', '偏狭', '免れる', '全力', '全土', '公共', '公正',
'利用', '制度', '努める', '努力', '効力', '勲章', '協和', '占める', '原理', '厳粛',
'反する', '受ける', '各国', '名誉', '国会', '国家', '国政', '国民', '国際', '圧迫',
'地上', '地位', '基づく', '基本', '妨げる', '子孫', '存する', '安全', '宣言', '対等',
'専制', '専念', '将来', '尊重', '崇高', '差別', '平和', '平等', '幸福', '従う', '必
要', '念願', '思う', '性別', '恐怖', '恒久', '恵沢', '惨禍', '愛する', '憲法', '成果
', '戦争', '授与', '排除', '支配', '政府', '政治', '日本', '普遍', '最大', '有す',
'有する', '栄典', '栄誉', '権利', '権力', '権威', '欠乏', '正当', '永久', '永遠', '
決意', '法令', '法則', '深い', '濫用', '無視', '特権', '現在', '理想', '生命', '生存
', '由来', '目的', '相互', '確保', '確定', '確認', '社会', '福利', '福祉', '立つ',
'立法', '経済', '維持', '自国', '自由', '自覚', '華族', '行使', '行動', '行為', '詔
勅', '認める', '誓う', '負う', '責任', '責務', '貴族', '起る', '身分', '追求', '道徳
', '達成', '選挙', '門地', '関係', '限り', '限る', '除去', '隷従']
(2, 187)
[[ 0  0  0  0  0  0  0  0  1  6  3  8  3  1  1  2  0  2  1  0  1  1  1  0
   0  1  0  0  0  0  0  2  2  0  0  0  1  0  1  1  0  0  0  1  2  1  0  0
   0  0  1  1  1  2  0  1  0  0  0  1  0  0  0  0  2  0  1  1  0  1  1  1
   0  0  0  0  1  1  0  0  0  0  1  9  0  0  0  0  0  2  1  0  0  0  0  0
   0  0  2  2  0  1  0  1  1  0  1  0  0  1  0  0  0  0  0  2  0  0  2  0
   0  0  1  0  0  1  1  1  2  1  3  0  0  0  0  1  0  0  0  0  0  1  0  1
   1  0  1  0  0  0  0  0  0  0  2  0  2  0  1  1  0  0  2  0  1  0  0  0
   0  1  0  1  1  0  1  0  1  1  0  0  0  1  1  1  1  0  0]
 [ 1  1  1  1  1  1  2  1  7  4  0 26  0  1  0  0  1  1  1  1  0  1  0  2
   1  0  1  1  1  3  1  0  1  1  7  1  0  1  0  0  1  2  1  0  0  0  1  1
   2  2  0  0  1  0  1  0  1  1  1  0  1  1  1  1  0  1  0  0  1  0  0  0
   1  1  2  1  1  0  1  2  1  2  1 11  1  1  1  1  1  0  0  1  1  1  1  1
   1  1  0  0  2  0  4  0  0  1  0  1  1  0  1  1  1  1  1  3  1  1  0  1
   1  1  1  3  2  0  0  1  0  0  1  1  1  1  1  0  1  2  1  2  1  0  1  0
   0  2  0  2  1  1  1  1  1  1  1  1  0  1  0  0  2  2  1  1  0  1  1  1
   1  0  1  0  0  1  0  1  0  0  1  1  1  0  2  0  0  1  1]]
```

In [16]:
```
# pandas のデータフレームに変換する

df_td = pd.DataFrame(data=td, columns=terms)
df_td
```

Out[16]:

	あげる	ある	いづ	いる	うち	かける	から	ここ	こと	これ	…	追求	道徳	達成	選挙	門地	関係	限り	限る	除去	隷従
0	0	0	0	0	0	0	0	0	1	6	…	1	0	0	0	1	1	1	1	0	0
1	1	1	1	1	1	1	2	1	7	4	…	0	1	1	1	0	2	0	0	1	1

2 rows × 187 columns

In [17]:
```
# 1番目の文書について、出現頻度によって語をソート
i = 0
df_td[i:i+1].sort_values(by=i, axis=1, ascending=False)
```

Out[17]:

	国民	する	これ	すべて	権利	その他	人権	自由	栄典	保障	…	国家	国会	名誉	各国	厳粛	原理	占める	協和	努める	隷従
0	9	8	6	3	3	3	2	2	2	2	…	0	0	0	0	0	0	0	0	0	0

1 rows × 187 columns

In [18]:
```
# 2番目の文書について、出現頻度によって語をソート
i = 1
df_td[i:i+1].sort_values(by=i, axis=1, ascending=False)
```

Out[18]:

	する	国民	こと	われ	平和	これ	日本	憲法	もの	生存	…	一代	特権	現在	制度	生命	幸福	平等	られる	努力	差別
1	26	11	7	7	4	4	3	3	3	2	…	0	0	0	0	0	0	0	0	0	0

1 rows × 187 columns

In [19]:
```
# 各文書について、頻度の合計を計算
df_td.sum(axis=1)
```

Out[19]:
```
0    119
1    205
dtype: int64
```

In [20]:
```
# 各文書について、語ごとの出現確率

freqs = np.array(td, np.float)
freq_sums = np.array(df_td.sum(axis=1), np.float).reshape(2,1)  # For numpy's broadcast

probs = freqs / freq_sums

for i in range(probs.shape[1]):
    print("%s\t%f\t%f" % (terms[i],probs[0,i],probs[1,i]))
```

あげる	0.000000	0.004878
ある	0.000000	0.004878
いづ	0.000000	0.004878
いる	0.000000	0.004878
うち	0.000000	0.004878
かける	0.000000	0.004878
から	0.000000	0.009756
ここ	0.000000	0.004878
こと	0.008403	0.034146
これ	0.050420	0.019512
すべて	0.025210	0.000000
する	0.067227	0.126829
その他	0.025210	0.000000
ため	0.008403	0.004878
できる	0.008403	0.000000
として	0.016807	0.000000
ない	0.000000	0.004878
なる	0.016807	0.004878
において	0.008403	0.004878
における	0.000000	0.004878
について	0.008403	0.000000
によって	0.008403	0.004878
により	0.008403	0.000000
による	0.000000	0.009756
にわたって	0.000000	0.004878
に対する	0.008403	0.000000
のみ	0.000000	0.004878
ひとしい	0.000000	0.004878
もたらす	0.000000	0.004878
もの	0.000000	0.014634
よう	0.000000	0.004878
られる	0.016807	0.000000
れる	0.016807	0.004878
わが国	0.000000	0.004878
われ	0.000000	0.034146
を通じて	0.000000	0.004878
一代	0.008403	0.000000
一切	0.000000	0.004878
不断	0.008403	0.000000
与える	0.008403	0.000000
世界	0.000000	0.004878
主権	0.000000	0.009756
享受	0.000000	0.004878
享有	0.008403	0.000000
人権	0.016807	0.000000
人種	0.008403	0.000000
人間	0.000000	0.004878
人類	0.000000	0.004878
他国	0.000000	0.009756
代表	0.000000	0.009756
伴う	0.008403	0.000000
侵す	0.008403	0.000000
保持	0.008403	0.004878
保障	0.016807	0.000000
信ずる	0.000000	0.004878
信条	0.008403	0.000000
信義	0.000000	0.004878
信託	0.000000	0.004878
信頼	0.000000	0.004878
個人	0.008403	0.000000
偏狭	0.000000	0.004878
免れる	0.000000	0.004878
全力	0.000000	0.004878
全土	0.000000	0.004878
公共	0.016807	0.000000
公正	0.000000	0.004878
利用	0.008403	0.000000
制度	0.008403	0.000000

In [21]:
```
# 各文書について，出現確率の合計は1になるはず
probs.sum(axis=1)
```

Out[21]:
```
array([1., 1.])
```