2018/5/28

前回までのあらすじ

衆議院の動きの動きから、国会別に議員の名前をゲットすることができる

↓

エクセルでデータを成型する

データの区切り位置を選択→タブ、スペースで区切り、苗字、名前ごとにセルを分ける

検索で「理」「事」をすべて検索&消去

↓

5/29

野党の口汚さを調べる必要がある

末尾表現&動詞を調べる

野党同士を比較しないと意味がない

名詞だとその時のトピックが違うので意味がない

末尾表現は、「。」で調べてその前の分かち書き二つを取り出すなどの必要がある

更に発言数に対しての、正規化をする必要もある

6/5

データ数が少ないので、民主党政権の3年間と、阿部政治の3年間を比較することにした。

・民主党、与党時代

2009/9/16-2012/12/25年（平成21-24）

第172回国会で鳩山幸雄内閣が正式に発表された

～181回国会まで

・自民党与党時代

2012/12/26から

6/9  
パワーポイントの構成

1. 研究背景
2. 目的
3. 手法
4. 結果
5. 課題

使われている単語がどこで使われているか考えなくてはいけない

使っている人が何人いるか調べる必要がある

現在の手法

全員分スクレイピングする➡いっぺんに形態素解析する

手法A

ひとごとにスクレイピングする➡形態素解析する➡

手法B

調べてほしい単語をリストアップする

各単語ごとに使用している人がいれば数値を足していく

どの単語が何人使っているかわかる

ループをする前に形態素解析あ

ワードと一致している

+する

0だったら

語数が1個でもあれば＋１する

6月12日

初めてのゼミ発表

案の定ボコボコにされた。が、研究結果は面白かったとのこと

反省点としては

・有効数字が適当

・もしもお互いの発言数をそろえるのであれば、両者同じ基準にしてとる。

・末尾表現の判断基準がよくわかってない（句点の後ろからいくつか、とおもいきや「です」単体などもランクインしている）

・字が小さい（らしい）

6月19日

手法を変えることにした。

Globライブラリを用いることで、議員毎にtxtファイルを分別した。

具体的には、スクレイピングするときに議員毎にfor文を回すが、その際のtxtファイルの記名にformatメソッドを用いた

全体で何人使っているか調べるときにはglobライブラリでワイルドカード検索し、

読みだした。

このほうが誰がどの単語を使っているか、何人使っているかなどの判定がしやすい。

今後の研究の拡張性も高い。

6月20日

新しく統計を取ったmecabファイルがうまくいっていなかった。

各議員のスピーチが終わるごとにEOFが含まれてしまっていたことが原因だった。

検索にかけてCTRL+Dで一斉選択&削除を行った

6月23日

教授にぶちぶちに怒られた

「コーディングをちゃんとすれば、「です」は除ける  
はずです。もう少し丁寧な作業をしてください。  
現状のやり方はいい加減すぎて、研究としては  
全く通用しません。」

今回の末尾表現のルールは以下のようになっている。

句点、の前4語を取り出す。

その4語の中に読点がある場合は読点移行のみを取り上げるというものである。

また、全角スペースを取り除く必要もある。

この問題に対して、以下の様に取り組んだ。

正規表現を用いて、Vscodeで前処理をする

「.+ 」で全角スペース前までの任意の文字を削除する。これによって、全角スペースを取り除くことができる。

また、4語以下のどくてん以降のみ取得する方法についてだが、、、

まず、どくてん以下だけのtxtファイルを作成する。

それを読み込んで、形態素解析し、もう一度分類器にかけて4語以下にする

新しく作ったプログラムはこちら

Matubi2

6月26日

ナイーブベイズで判定するために、各議員の末尾表現のみのtxtファイルを作る。

しかし、

・句点が二個になってしまう「。。」

・〇なになに議員　、の部分が入ってしまう。

正規表現で、前処理するか、手作業でごりごりやるしかない。。。。

6月27日

re.subで難なく前処理完了した。

手法としては、txtファイルとラベルを両方、順番つけてリスト型に入れる。

↓

関数から出力させる

何らかの方法でテストデータとトレーニングデータを分ける。

Vectorizerで変換

学習させる

6月28日

ついに、windowsからMecabとipa辞書を使える兆しが見えた。

Bash on windowsを用いることで、DLが可能になる。

詳しくはipa辞書のgithubを参考にすれば大丈夫。

Dicにパスを通すことで出来るようになる。

と思いきや、unicodeescape error

これについて解明していきたい所存。

↑パスにシーケンス文字が使われているという警告r’’の形で指定すればよい

6月29日

反省】

・同じ処理をするならいっぺんにやる

（後から昔の事をやろうとすると、わからなくなる）

（よくわかんなくなった）

Git管理したほうが良さげ

自分がいまなんの処理をやっているかわからなくなってしまうので作業効率が低くなる可能性が高い。

Make\_line2を二回引用しているが、re,subがいるかいらないか分ける必要がある？

↑ちがうかも

文字化けを回避する方法としては、

R・ｗどちらもencodingを指定する必要がある。

file\_parsed

がラッパーに包まれている場合でも

In line for file\_parsed

を実行した場合、lineは文字列になる

split(): 指定の区切り文字で文字列を分割しリストとして返す

使用したファイル、

Delete

Mecab3,4

改善点

議員の名前がわかるファイル名にしないとだめ、

議員の人数がおかしい

とりあえず、前処理まではおわった。Docsに何も入っていないことがもんだい。