Ｗｏｒｄクラウドを利用し、課題提出の際に一番多く使われた言葉を探す

経済科学部総合経済学科１年八幡怜花

はじめに

Wordクラウドを利用した学習を行いたいと思い、簡単に手に入りやすくかつ長すぎない記録となった時に課題が思い浮かんだため調査した。この調査を通して私が作った文章の傾向、癖のようなものを明らかにし、今後の課題提出の際多くの語彙を使い大学生らしい文章を生成するのに役立てたい。

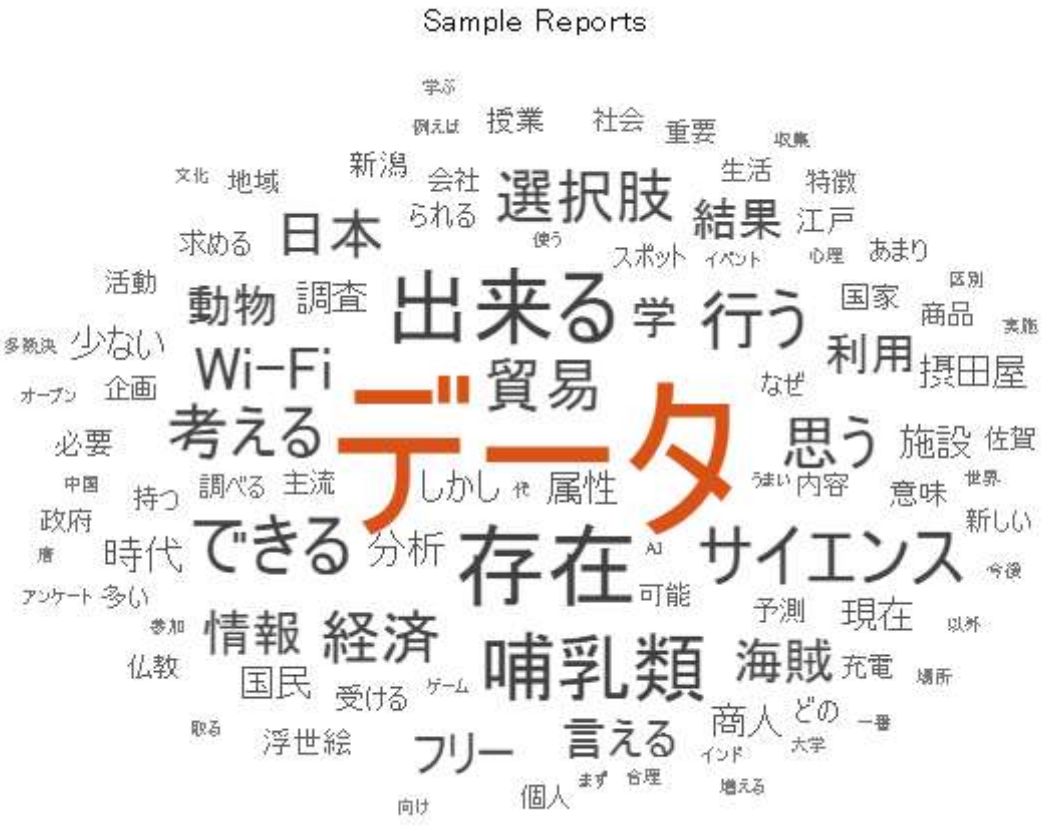
データの取得

学務情報システムより第3ターム授業の課題で既に提出済み、または提出受付中であるものを利用した。調査結果を分かりやすくするために日本語での提出の場合のみに絞った。その他Matlabのヘルプページを活用し、以下のように調査した。

過程

学務情報システムに提出した課題をtext analiytics boxを使ってWordクラウド化し、どの単語がよく使われているのかを確かめる。

```
T= readtable('sample01.csv','TextType','string');
str = T.Var1;
str(1:29);
figure;
wordcloud(str);
title("Sample Reports")
```



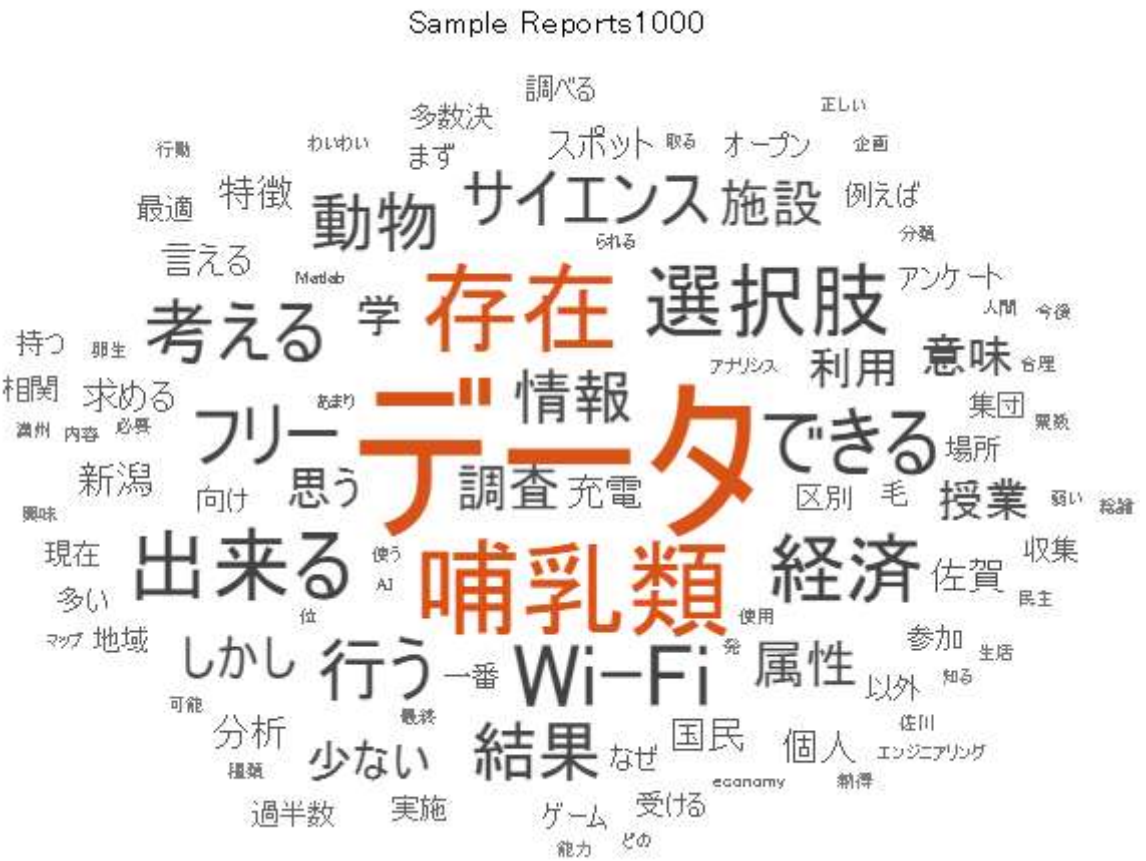
名詞で出ているのはデータサイエンス総論なら「データ」、人文社会科学なら「海賊」「貿易」という風にそれぞれの授業でキーワードになっていた言葉が多かった。授業の内容問わずによく使われる動詞だと「出来る（できる）」「思う」「考える」「言える」あたりがよく使う言葉だと判明した。

「データ」「サイエンス」という言葉が多かった理由は、今季課題最大文字数の1000文字の課題が出されたのはデータサイエンス総論だけだったからだと考察。文字数が多ければ多いほど同じ単語を使う傾向があると思い、1000文字の課題に絞ってみた。

```
U = readtable('1000sample.csv','TextType','string');
```

```
"現在経済というと財・サービスとお金のやり取りの事を指す。しかしeconomyの語源のオイコノミアには家政と
"どちらも個人の合理性を求めた結果、最適解を求めることが出来ない場合を示している。人間が合理的に動く前
"経済は個人的な感情や考えを排除した、合理的な賢い人間たちの集団によって形成されるものだと考えていた。
"ロビンソンは欲望を満足させることを目的としている。そのために限りある行動可能時間や材料の分量を最適に
"佐賀県が実施した「佐川わいわいWi-Fiマップ」とは佐賀県内のフリーWi-Fiスポット・スマートフォンなどの
"選択肢が２つであり、それぞれにメリットデメリットが存在し絶対的に正しい選択肢が存在しない時に多数決に
"資料内で得られるデータによると哺乳類に分類される動物は全て胎生であり、卵生の動物は哺乳類以外であっ
"近代以前では身分や宗教、村などの社会集団を一つの単位として秩序を維持していた。近代になると今まで集団
```

```
str = U.Var1;
str(1:11);
figure;
wordcloud(str);
title("Sample Reports1000")
```



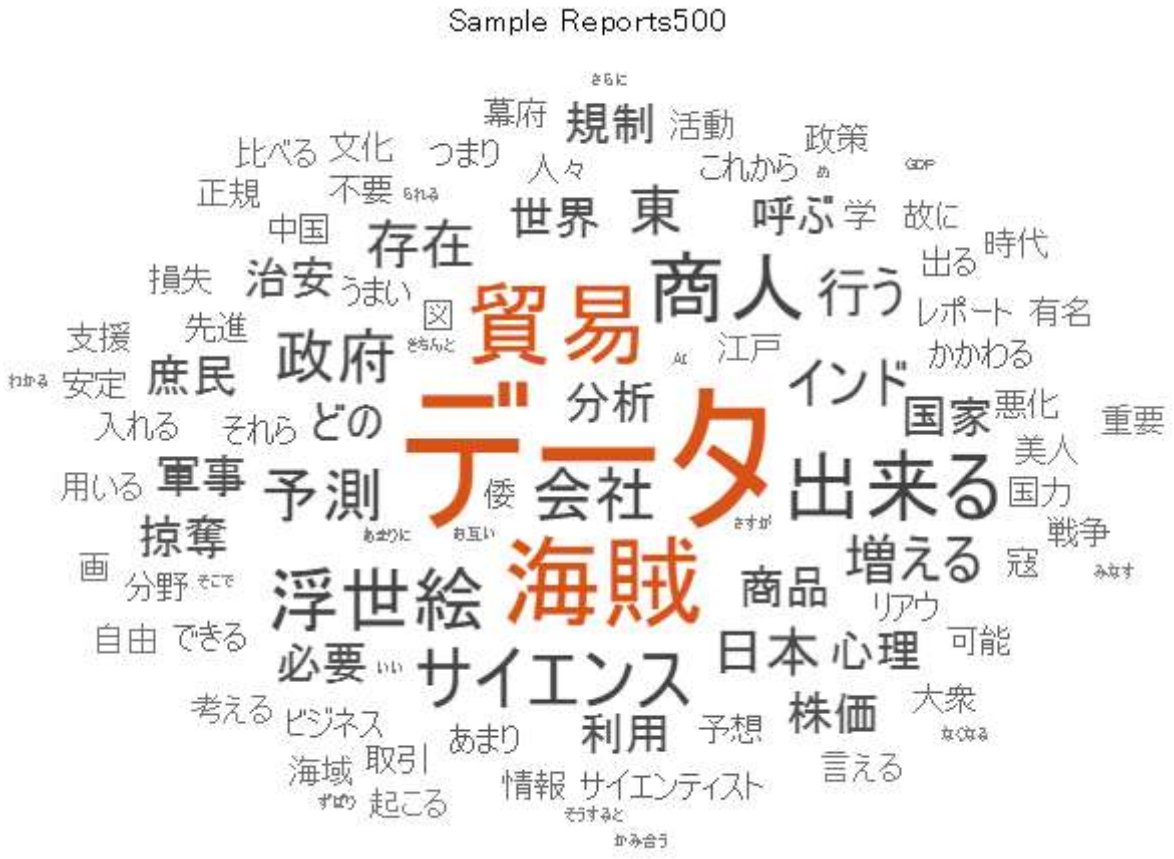
考察通りこちらでも「データ」が主に使われた単語となった。あと「学」については、○○学という言葉を使った際に前半の言葉と分離されてしまったため多くなったと考えている。

次に500文字に絞ってみる。

```
V = readtable('500sample.csv','TextType','string');
```

"日本がアメリカや中国などと比べてデータサイエンティストの数が少ないのは、ずばりデータサイエンス自体
"将来のことは誰にも予測することが出来ない。故に過去のデータを参考にして、これから起こりうることを予
"私のレポートには図が一つも存在しなかったので、プレゼンテーションでは図を適度用いて聴衆が退屈しない
"戦時期で貿易船が攻撃される危険性があり、貿易による収入が安定しない時には商人が海賊を支援する。海賊
"戦争によって内陸部にある中央政府の支配力が沿海部まで行き届かなくなり、治安が悪化する。そうすると国
"東アジア海域において、明王朝では海禁政策を行っており朝貢貿易しか認めていなかった。それにもかかわら
"浮世絵は江戸時代を代表する美術品である。菱川師宣の「見返り美人」、葛飾北斎の「冨嶽三十六景」が有名

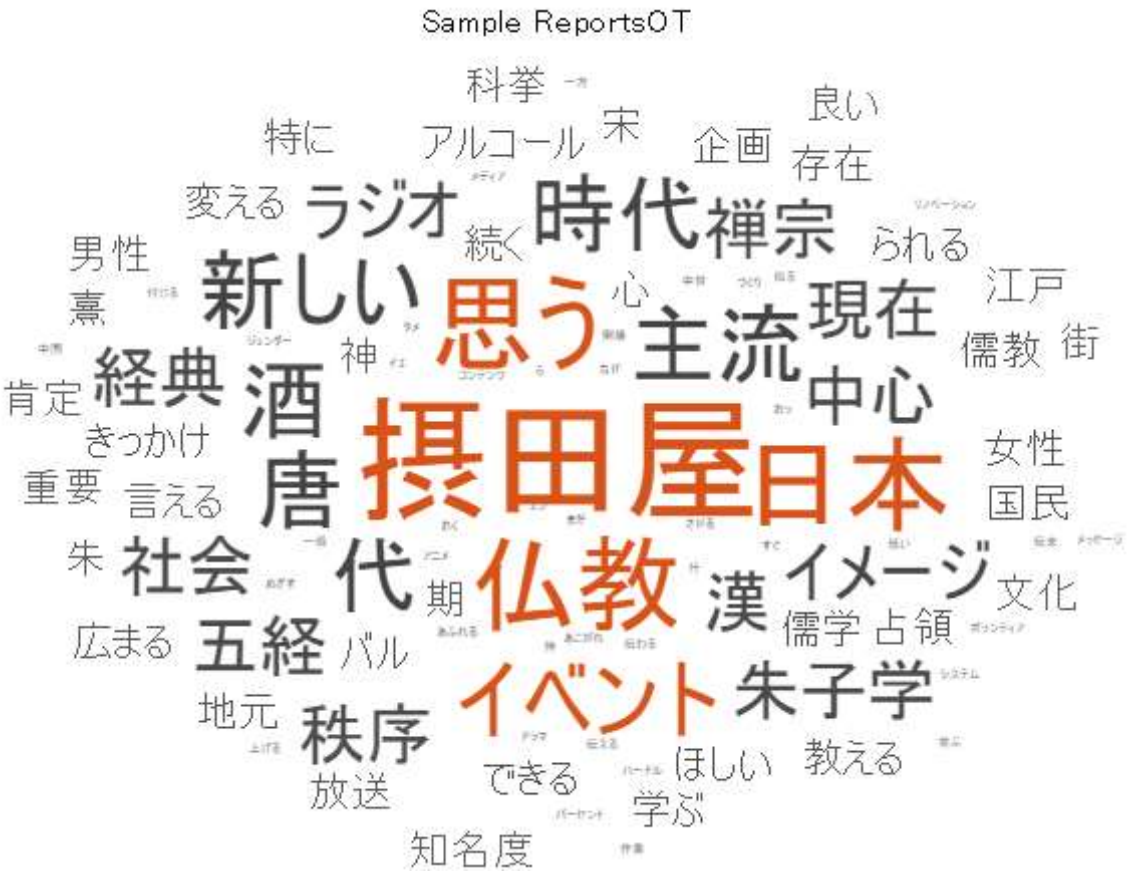
```
str = V.Var1;
str(1:7);
figure;
wordcloud(str);
title("Sample Reports500")
```



思っていたのとは異なり、今回もまた「データ」が一番多いという結果になった。複数の講義内容が混ざっているので単語数としては先ほどの調査より多いことがわかる。

最後に残りの100～300文字を調査する。

```
 W= readtable('3~100sample.csv','TextType','string');
str = W.Var1;
str(1:11);
figure;
wordcloud(str);
title("Sample ReportsOT")
```



摂田屋に野外学習する授業があったので「摂田屋」が多いという結果になった。動詞については今までは中央にあった「出来る」よりも、「思う」を頻繁に使っていることがわかる。300文字以下の時に「思う」を多く使う傾向があるのだろうか。また前回の7つの文と比べてこちらは11文であったのに対し語彙数は少ないので同じ言葉を繰り返している率が高いことが分かった。

結論

最初は「思う」「考える」「わかる」などの動詞が全体を通して頻繁に使われていると想定していたので、名詞が中央に集まっているのが意外だった。今回の調査では学情での提出のみ、つまりWordファイルでの提出の場合はカウントしていないのでその結果も含めればより詳しく傾向がわかるかもしれない。

参考文献

ヘルプとサポート - MATLAB & Simulink - MathWorks 日本

感想

初めてのMatlab利用だったのでかなり苦戦し、結果的に雑で拙い研究となってしまいました。本格的なレポート提出の経験がまだ少なく、まとめ方自体もよくわかっていない状態なので今後のためにアドバイスをもらえればありがたいです。