文章からワードクラウドを作成して頻出単語を視覚的に確認する

【重要】以下の欄をすべて埋めてから応募してください

氏名：[redacted]

所属学部学科(専攻)：[redacted]

email：[redacted]

【考察記述欄】作業をすべて終えてから考察を自由に記述してください。実行結果の考察だけでなく、データを集めたりプログラムを実行するにあたって苦労した点や、こんなデータを収集、解析してみたい、プログラムを改造してこんな解析や予測もしてみたいというアイデアなどでも構いません。また、実際にプログラムを改造するなど特別に工夫したことがあれば記載してください。

－－－記述欄はじめ

結果及び考察

今回私は，企業に期待される大学は，どのような特徴があるのかを可視化することを目的とした．大学の指標には「偏差値」というものがあるが，一方で企業の大学への評価は，必ずしも偏差値が絶対的な評価にはなっていないと感じたためである．今回利用したランキングは NIKKEI STYLE(https://style.nikkei.com/article/DGXKZO72472380R00C21A6TCN000/)に掲載されたものを利用した．このランキングでは，確かに上位は旧帝大が占めるものの，その順番は偏差値の高さとは一致していない．解析にはこのランキング上位５校に関して，その大学の理念，及び中期戦略うち，全体を俯瞰する部分用いた．ここでこれらを分析対象として選んだのは，学校理念が最もその大学の運営や教育方針を示すと考えたためである．また，中期戦略を含めることで，最近力を入れている部分にも焦点と当てようと考えた．しかしながら，中期戦略全てを解析した場合，具体性が強くなり，それぞれの大学を俯瞰した分析が困難になると考えたため，中期戦略全てではなく，初めに述べられる全体を俯瞰した部分のみを対象とした．また，同じ新潟にありながら，企業期待度が高いとされている長岡技術科学大学も対象として分析を行った．

以下の解析結果は，先頭が新潟大学，以降ランキングの 1～5 位までの大学が並び，最後に長岡技術科学大学の結果を示している．結果を見ると，新潟大学は中心に「社会」という単語がきており，「研究」よりも多く使用されていることが分かった．一方で，ランキング上位校はすべて「研究」が「社会」よりも上位に表れた．また，長岡科学技術大学も，やや単語は違うが「技術」という，「研究」と比較的近い単語が中心に表れている．この結果を見るに，企業は「社会に即応用できる結果」よりも，「研究を行える人材」を求めている現状が可視化できたと考える．最近は基礎研究よりも応用研究の方が評価されやすいという意見があるが，それは必ずしも社会先行の研究ではないと考えられる．企業が注目するのは，研究に重きを置き，その先に社会への応用を考える大学なのではないかという結論に至った．

今回作業を行った感想

今回の作業で苦労した点は，データの内容をどこまで揃えるかという点です．今回複数大学の理念を収集する中で，各大学ごとに当然形式が違いますし，中期戦略なのか，それとは別のものなのかの判断が難しいものも存在しました．収集時に文量などのバランスはできるだけ揃えたつもりですが，それらの判断に私の主観が入る余地があったように感じます．この主観が結果にどれほど反映されているのかは，今回評価できず，不安

な点として残してしまいました．また，文脈によっても同じ単語でも扱いが変わる場合があると感じました．例えば同じ「社会」という単語でも，使われている文が肯定文か，否定文かで，その扱いは変えるべきだと考えます．今回の手法ではこれが行えないので，今後の改善点になると感じております．加えて，今回は理念から各大学の全体を読み取るということを行いましたが，これは因果関係の原因に当たると考えます．より精度よく各大学を分析するには，因果関係の結果に当たる，各大学の研究成果に関する分析も必要だと感じました．今後の解析対象として，最近 5 年間の各大学研究動向とすることに関心を抱きました．

記述欄おわり---

(ここからプログラムが始まります)

最近良く見かけるワードクラウドは、文章の中から単語を出現頻度に応じた大きさで可視化する手法です。MATLAB のテキストマイニング機能を利用して文章データから簡単にワードクラウドを作成できます。

身の回りにあるテキストデータを使ってワードクラウドを作成してみてください。よく使われている単語を視覚で確認することで、思わずおおっ！と声が出そうになる驚きや気づきがあるかもしれません。

事前準備:

- テキストファイルを用意してください。英語でも日本語でも大丈夫です。自分が毎日つけている日記やメモ書き、ブログを開設していたらブログ記事でも良いでしょう。インターネット上で公開されているパブリックドメインの文章、演説や小説などでは話し手や書き手の癖が垣間見えたりするかもしれません。ファイル形式がテキスト形式ではない場合は、Windows メモ帳を開いて、該当の文章をすべてコピーしてメモ帳に張り付けてから、テキスト形式(.txt)のファイルとして保存してください。

プログラムの実行:

下のステップ 1 を読んでプログラムを書き換えてから、このプログラムを実行してください。

Tip: メニューに実行ボタンが見当たらない時は、[ライブエディター]タブを選択してください。

プログラム

ステップ 1: テキストファイルの読み込み

以下の作業を済ませてからプログラムを実行してください。

- 事前準備で用意したテキストファイルをこのプログラムと同じ場所に保存してください。
- MATLAB の「現在のフォルダー」をこのプログラムとテキストファイルが保存されているフォルダーに変更してください。
- 下のプログラムの filename = の後のファイル名を実際に読み込むファイル名に書き換えてください。

同梱のサンプルファイル"hakusho.txt"は令和 2 年度版情報通信白書「はじめに」の本文です。

出典：「令和 2 年度版情報通信白書」(総務省) 「はじめに」 https://www.soumu.go.jp/johotsusintokei/whitepaper/ja/r02/pdf/n1000000.pdf

令和 2 年度版情報通信白書の二次利用について(二次利用可能)：https://www.soumu.go.jp/main_content/000700124.pdf

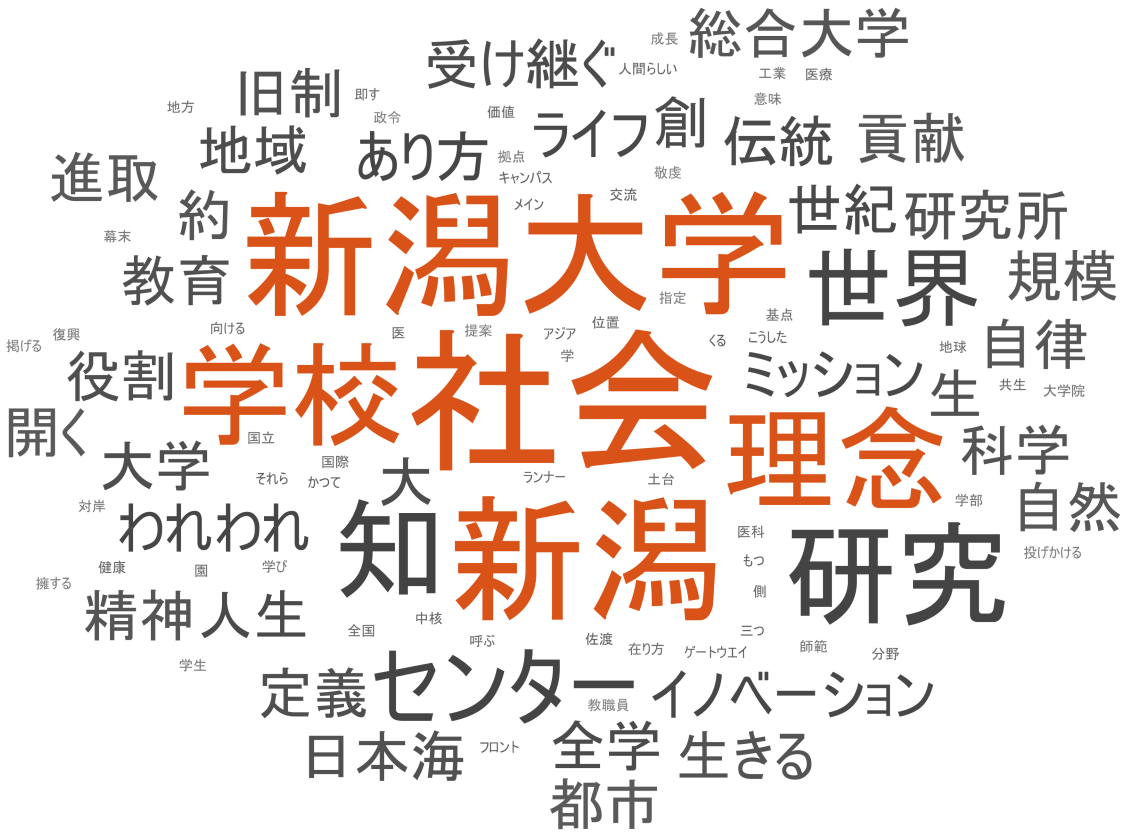
```
% 初期化します
clear

% 実際に読み込むファイル名に書き換えてください
filename_1 = ['Niigata.txt'];
filename_2 = ['Hokkaido_1.txt'];
filename_3 = ['Kyoto_2.txt'];
filename_4 = ['Tohoku_3.txt'];
filename_5 = ['Tokyo_4.txt'];
filename_6 = ['Nagoya_5.txt'];
filenameGib = ['Nagaoka.txt'];
% 上で指定したファイルを文字列として変数 text に格納します
text_N = extractFileText(filename_1);
text_H = extractFileText(filename_2);
text_K = extractFileText(filename_3);
text_T = extractFileText(filename_4);
text_To = extractFileText(filename_5);
text_Nago = extractFileText(filename_6);
text_Na = extractFileText(filenameGib);
```

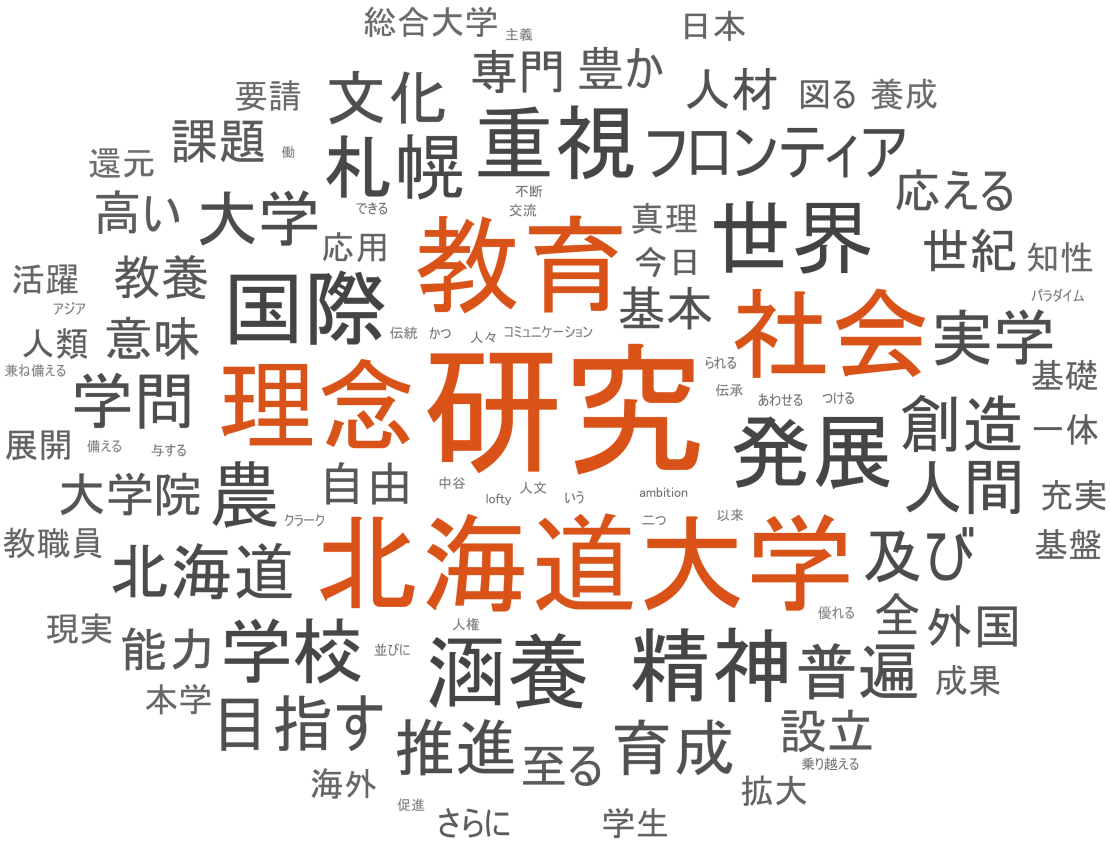
ステップ 2: ワードクラウドの作成

MATLAB では文章データから自動的に単語を抽出、カウントしてワードクラウドを作成してくれます。

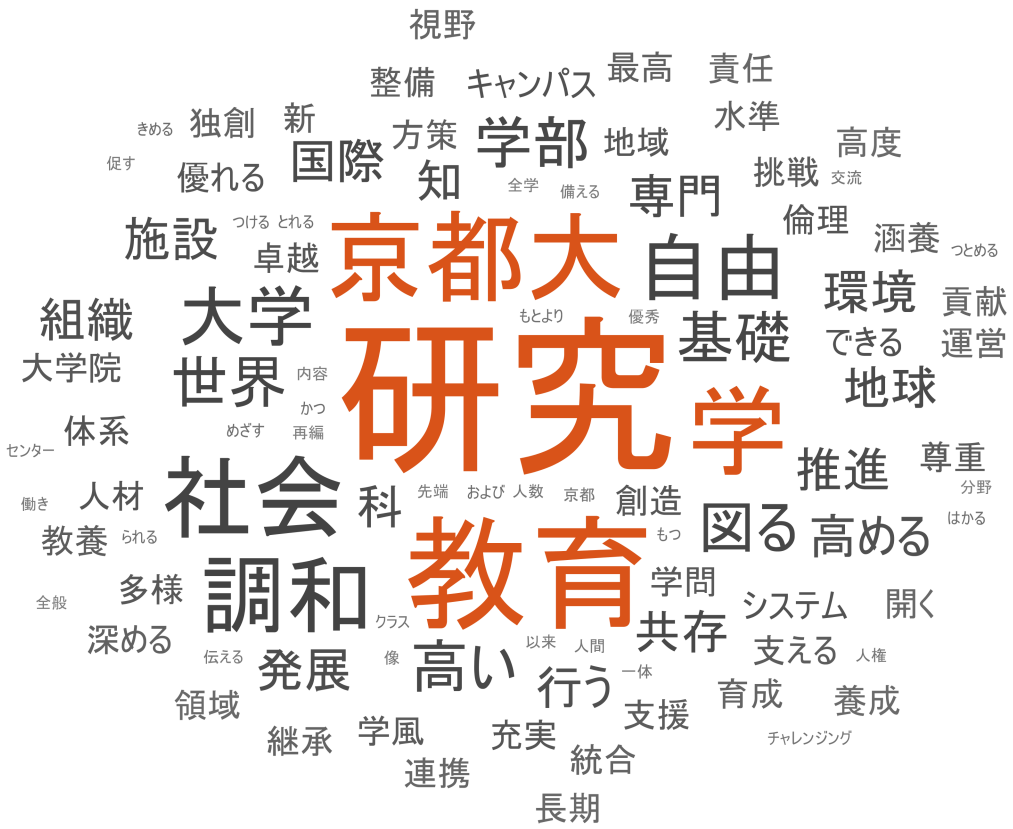
```
% ワードクラウドを作成します
wordcloud(text_N);
```



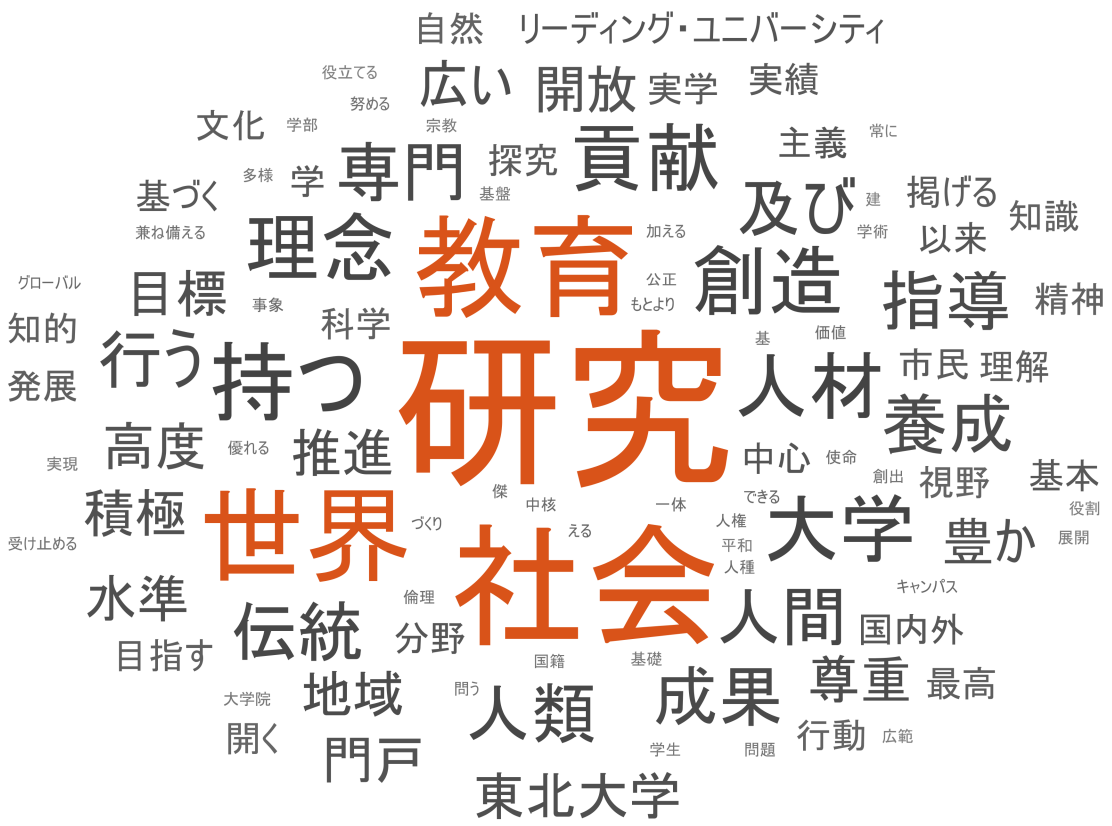

```
wordcloud(text_H);
```



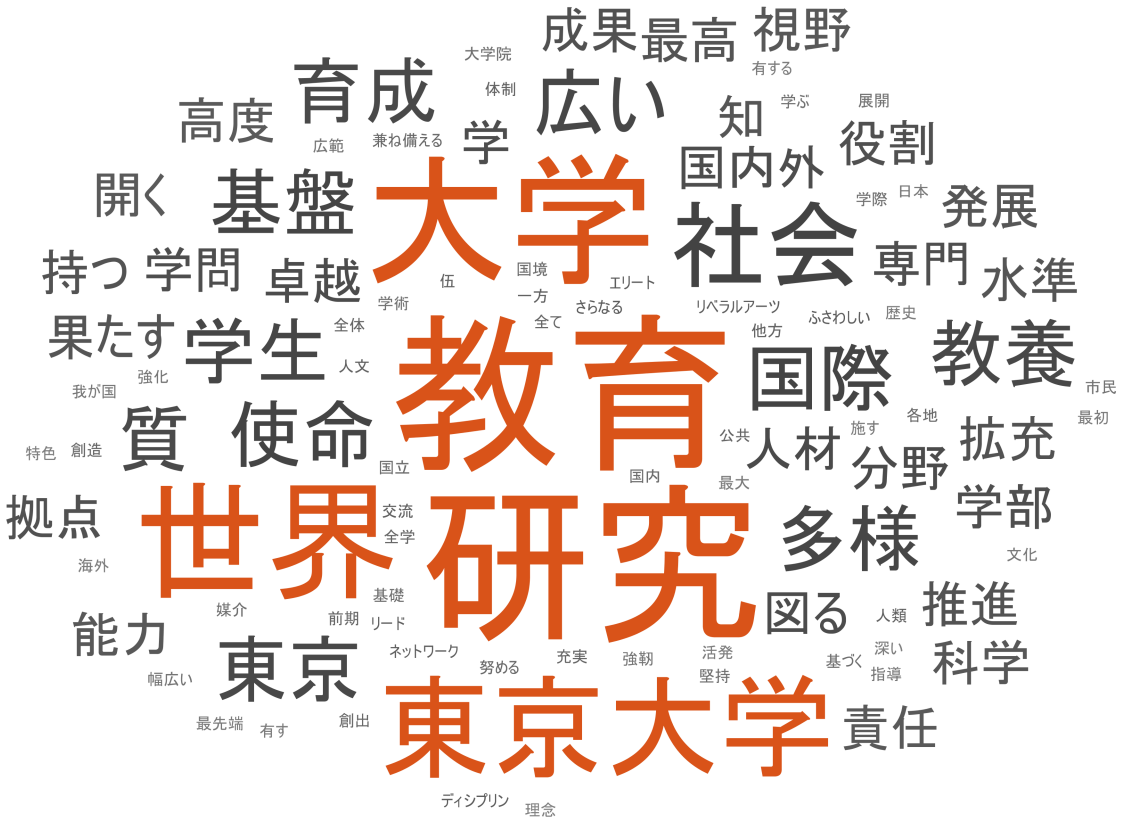
```
wordcloud(text_K);
```



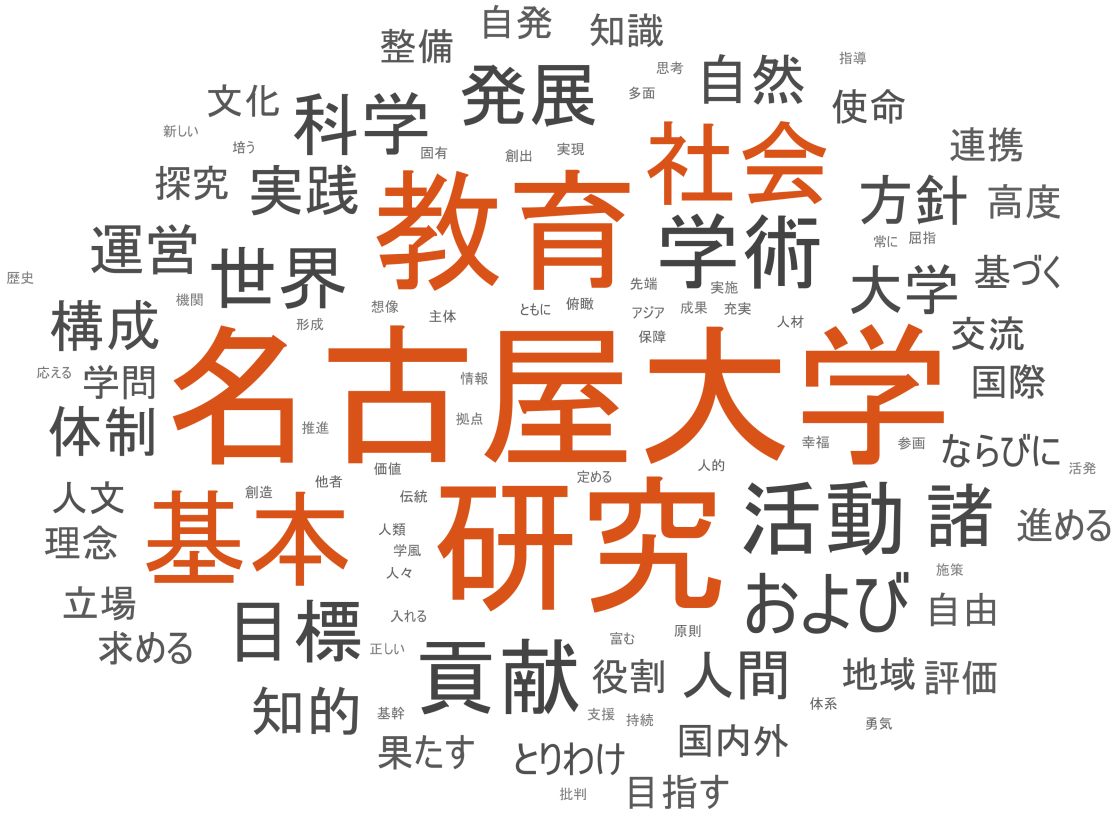

```
wordcloud(text_T);
```




```
wordcloud(text_To);
```



```
wordcloud(text_Nago);
```




```
wordcloud(text_Na);
```

考察:

実行結果に対して、冒頭の考察記述欄に考察を自由に記述してみましょう。

保存(レポート化):

ここまでの作業お疲れさまでした。最後に作業の結果を残しておきましょう。

このプログラムは実行結果とともにレポート出力できます。

[ライブエディター]タブの[エクスポート]ボタンを押して、PDF を選んでエクスポートしてください。

保存された PDF 形式のファイルをコンテスト窓口に送付してください。

Copyright 2021 The MathWorks, Inc.