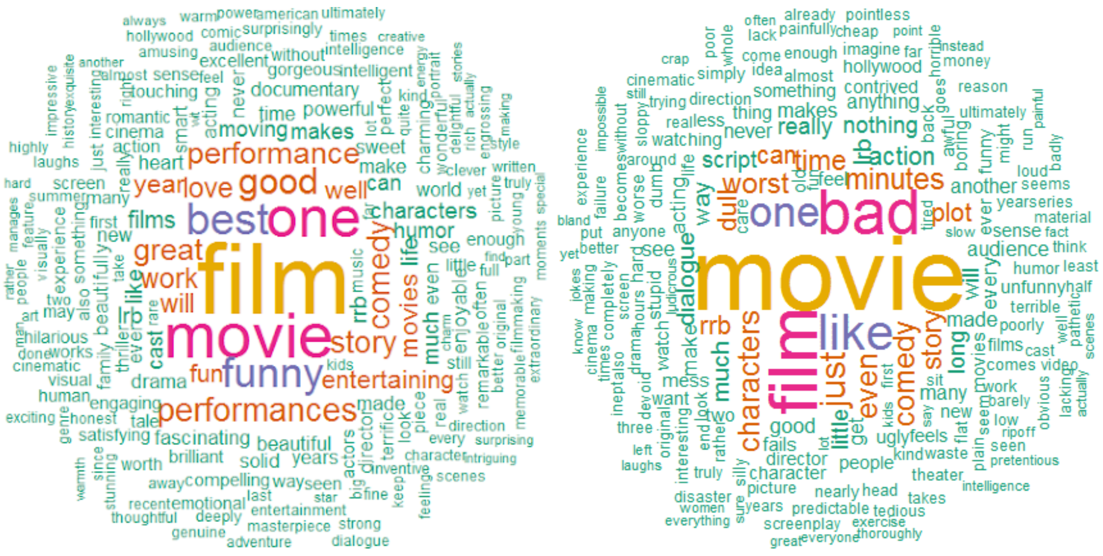
Using keras to classify movie reviews
Mathematical Programming with Python

MATH 2604: Advanced Scientific Computing 4
Spring 2025
Monday/Wednesday/Friday, 1:00-1:50pm
Room A202 Langley Hall

https://people.sc.fsu.edu/∼jburkardt/classes/python_2025/keras/keras.pdf



Words used in positive and negative movie reviews.

Classifying with keras

- We start with a set of movie reviews, each one classified as posiive or negative;
- We want to create a neural network able to examine new movie reviews and guess whether they are also positive or negative;
- We train it on some reviews, and then test it on the remainder.
- This is supervised learning - we want the network to “learn” how to classify reviews like we did.

1 What do we want?

When NetFlix recommends movies to you, it is using an automated system that has classified the movies in its library, and has also analyzed your own likes and dislikes. It then lists movies that you have not yet watched, but which most closely fit your apparent preferences.

We are going to experiment with a similar, but simpler task. We have collected the text of many movie reviews, and labeled them as *positive* or *negative*. How we did this is something of a mystery that would be very hard to express as a computer program. Some (but not all!) of our judgment could be explained simply by the choice of words in each review. If we encounter the words “miserable”, or “boring” or “stupid”, we are likely to be reading a negative review.

Now suppose we have a pile of movie reviews, and we have read each of them, and assigned them a rating of "positive" or "negative". Now imagine we hand the reviews and their ratings to a very intelligent Martian, who can see the words, but doesn't know what they mean. The Martian might notice that there is a rough relationship between the words and the classification: words like *great* and *thilling* may indicate a positive review, while *awful* would suggest the opposite. However, a bad movie can be called *a great disappointment*, so if we decided that *great* by itself was a sure indication of a positive review, we would soon notice some problems. To build a good classifier, we would have to review the data, make some tentative rules, test them, and then adjust our model to try to reduce the errors.

Finding patterns in data is something that the keras package does well. To deal with the movie review classification problem, we can use keras to build a neural network, apply it to our data to get a model, and then test the model to see how well it performs on new data.

The movie review dataset is a built-in feature of keras, and so we can take advantage of a large set of data which has already been labeled. This will allow us to create models and see how effective they are.

2 A reference for keras

You can find an introduction to keras in the book:

```
Francois Chollet,
Deep Learning with Python,
Second Edition,
Manning, 2021,
ISBN: 9781617296864
https://www.manning.com/books/deep-learning-with-python-second-edition
```

The IMDB movie review exercise is covered in section 4.1, *Classifying movie reviews: a binary classification example.*

3 Where can I run keras?

We have several options for running the movie review example with keras:

1. on your laptop, but you must install keras and some other libraries, see `https://keras.io/getting_started/`;
2. using Google Colab, which has the necessary libraries already, see `https://colab.research.google.com/`;

4 What is the IMDB data and how is it used?

The Internet Movie Database (IMDB) dataset consists of 50,000 movie reviews and 50,000 labels. Each review is labeled "0" if it was judged to be negative, or "1" if positive. These labels were supplied by humans who read the reviews.

Our goal is to come up with a procedure that can automatically produce a label for a movie review, and which will closely match the behavior of human readers. We are to do this by using a computer procedure which can see, but not understand, the words of the reviews. In other words, as far as we are concerned, these reviews could have been written in Polish, or Chinese, or Egyptian hieroglyphics.

A dictionary was created from all the words in all the reviews, and each word in that dictionary was given a numerical index. Then each review was used to generate a corresponding file of numbers, where each word was replaced by its index. Now, instead of English, each review can be transformed into a sequence of numbers, which refer to the words in the dictionary.

Thus the text of a movie review is now numeric. For technical reasons, we wish to consider only the 10,000 most common words, so each numeric file is modified to eliminate unusual words. When we actually process a movie review, we do one last step: we replace the file of numbers by a vector of length 10,000, where entry i of the vector is set to 1 if word[i] appeared at least once in the review. The reason for doing this is that the neural network needs to process vectors of a uniform size. We keep the neural network happy by making every movie review a 10,000 entry vector of 0's and 1's. We say that we have *vectorized* a review in this way.

We divide the data into three sets: training, validation, and testing data. We will build a model with the training data, and then use it to predict the labels on the validation data. The prediction failures will be used to adjust the model. We will do this adjustment a fixed number of times (perhaps 10 or 20 "epochs") and then declare the model ready for testing.

We now hold the model fixed, and try it out on the testing data. If the training procedure was done well, then the model should have good accuracy in predicting the labels for the testing data. If the model does poorly, then we must go back and adjust our model and repeat the entire process.

5 Peeking at the reviews

The movie reviews in the IMDB dataset are no longer humanly readable; they are just lists of numbers. However, it is possible to use the dataset to decode the numeric review back into a semi-readable version by replacing each number by the corresponding word. The file *imdb_decode.py* can be used this way, where the input argument specifies the index of the movie review you want to decode.

```
python3
from imdb_decode import imdb_decode
imdb_decode(7)

? lavish production values and solid performances in this straightforward adaption of jane ?
     satirical classic about the marriage game within and between the classes in ? 18th
    century england northam and paltrow are a ? mixture as friends who must pass through ?
    and lies to discover that they love each other good humor is a ? virtue which goes a
    long way towards explaining the ? of the aged source material which has been toned down
    a bit in its harsh ? i liked the look of the film and how shots were set up and i
    thought it didn't rely too much on ? of head shots like most other films of the 80s and
    90s do very good results
```

The question marks in the listing indicate unusual words that were not in the top 10,000 most common. You should be able to guess that this review is labeled 1 **positive**. We hope that our movie classifier will also be able to correctly label it.

6 keras on your laptop

You may find it tricky to install keras on your laptop. However, if you can get it set up, you may prefer to do your work there. The movie review is large, but not enormous, so it should run fairly quickly for you.

Installation information can be found on the keras website `keras.io`.

Mac and Linux users can try these install commands:

```
sudo pip install tensorflow
sudo pip install keras
```

while Windows users may try to install with:

```
pip install tensorflow
pip install keras
```

If your installs are successful, then you should download the file *imdb.py* from the lab website. Then you can execute it on your laptop with a command like:

```
python3 imdb.py
```

On Mac and Linux, you can save the "interesting" output to a separate text file:

```
python3 imdb.py > imdb.txt
```

7 Understanding the code

You will be working with the file *imdb.py*. The program can be thought of as having these parts:

1. Load the data, and prepare it for use;
2. Describe the model;
3. Create the model;
4. Use the model on training and validation data;
5. Evaluate the model on new test data;

We will not worry about how the data is loaded and prepared, except to note that keras will need to download the data from the Internet if you are running on your laptop. It will then rearrange and split the data so that it has the right shape for the neural network, and is divided into training, validation, and test sets.

The model is described as a sequential model, with two hidden layers, each with 16 hidden units, and a `relu` activation. The first layer expects an input vector of length 10,000; in other words, one of our movie reviews. Our output layer uses the `sigmoid` function, which returns a value between 0 and 1, the probability that the movie review is negative or positive.

```
model = models.Sequential()

model.add ( layers.Dense ( 16, activation = 'relu', input_shape = (word_num,) ) )
model.add ( layers.Dense ( 16, activation = 'relu' ) )
model.add ( layers.Dense ( 1, activation = 'sigmoid' ) )
```

The model is created by choosing an optimizer, loss function, and a metric. We have discussed the `rmsprop` optimizer in class. The `binary_crossentropy` loss function is a way of measuring the difference between two probability distributions.

```
model.compile (
  optimizer = 'rmsprop',
  loss = 'binary_crossentropy',
  metrics = ['accuracy'] )
```

Once the model is created, we apply it to the training and validation data, and save a report in a dictionary called `history`:

```
history = model.fit (
  partial_x_train,
  partial_y_train,
  epochs = 20,
  batch_size = 512,
  validation_data = (x_val, y_val) ))
```

We can use `history` to print out the final values of the validation loss and accuracy. The code prints these values like this:

```
  Model loss and accuracy on validation data:
    Final validation loss 0.7035827110290528
    Final validation accuracy 0.8651999831199646
```

Now that the model has been trained, we want to see how well it can handle new data. We carry out this experiment on the test data, for which we know the correct result, and we report how well our model does:

```
  results = model.evaluate ( x_test , y_test )

  for i in range ( len ( model.metrics_names ) ):
    print ( model.metrics_names[i], results[i] )
```

The test loss and test accuracy are printed out something like this:

```
  Model loss and accuracy on test data:
loss 0.3238627934074402
accuracy 0.87308
```

8 Choose an experiment

The validation accuracy and test accuracy measure how well our classifier performs on data for which it had not been trained. In the example code, after the 20th epoch, the validation accuracy was about 0.865 and the test accuracy was about 0.873.

The accuracy of our model is affected by the parameter choices that were made in the program. In each of the following experiments you are to vary one of the parameters in the model, and note the resulting validation and test accuracy for each parameter choice.

The list of experiments for you to choose from includes:

1. The example used the command `model.add(layers.Dense())` twice, specifying the `relu` activation function. Compare using `relu` versus using the `tanh` activation function.
2. The example used the command `model.add(layers.Dense())` twice, to set up two hidden layers. Compare using one, two, or three hidden layers.
3. The example used the command `model.add(layers.Dense())` with 16 units in the two hidden layers. Compare using 16, 32, or 64 hidden units in each layer.
4. The example used a `model.compile()` command in which the optimizer was `rmsprop`. Compare using `rmsprop` versus just one of the other optimizers on this list: `sgd`, `adagrad`, `adadelta`, `adamax`.
5. The example used a `model.compile()` command in which the loss function was `cross_entropy`. Compare the results for `cross_entropy` versus using `mse`.
6. The example used a `model.fit()` command in which 20 epochs of training were carried out. Compare your results using 5, 10, and 20 epochs.

9 Computing Assignment

Do **two** of the experiments from the above list. Write a short report in which you explain:

1. which two experiments you carried out;
2. where you did your experiments: laptop or Google Colab;
3. the validation and test accuracies for your several cases;

Example:

```
MATH 1900 Report for Joe User

I ran experiments #1 and #6.
I ran the experiments on my laptop.
Here are my tables:

Experiment #1:
  activation    validation    test

   relu         0.86          0.87
   tanh         0.73          0.68

Experiment #6:
  epochs        validation    test

    5           0.80          0.71
   10           0.82          0.75
   20           0.86          0.87
```