515 HOMEWORK 4:
KICKSTARTER CLASSIFICATION

Your assignment is to create a tool that trains several machine learning models to perform the task of classifying Kickstarter posts. Kickstarter is an online platform on which individuals or teams can pitch their business ideas, and they collect funding from many online users ("crowdfunding"). In this dataset, some posts met their funding goals, while others did not. The goal of the assignment is to predict which Kickstarter posts will be successful based on their text.

The dataset

The dataset is available at **kickstarter_data_full.csv** or, if preferred, online at https://dgoldberg.sdsu.edu/515/kickstarter_data_full.csv. The data is formatted as a CSV. An example of the formatting is below. Note that there may be some special characters in the file, which can cause an error when opening it. You can resolve this by specifying the encoding: `open("kickstarter_data_full.csv", "r", encoding = "latin-1")`.

Num	Title	Blurb	Status	Requested	Description
1	Clean Power For Nomads	An ethnographic project designed to assess the energy needs of Mongolian nomads…	Funding Unsuccessful	6500	Essentially, my project is an initial ethnographic study of the lives of Mongolian nomads...
2	High Noon California, a feature film	It's 2040; A Drug Cartel occupies the US southland. Troops with pocket transportals,flying personal hovercraft, try to capture El Capo.	Cancelled by Creator	335000	Thank you for watching our zero dollar ($-0-) budget Kickstarter Video. Now you can help us make an excellent feature film…

Machine learning analysis

The "x" variable will be the description, which is the long post describing the project concept. Using the techniques discussed in class, convert the text into a machine learning-usable format. Consider and remove stop words. In addition, since there are many unique words in the dataset, consider using your vectorizer's `max_features` setting, which limits to vectorizer to the top-ranking set of words. For example, using a `max_features = 1000` setting would use the top 1,000 words.

Generate your "x" data in two different ways:

- Use the raw data in the CSV file without changing the text.
- Autocorrect the text data, fixing any typos that may have been made in the description (does this make a difference?).

Optionally, you could consider whether other variables could be useful predictors. Some suggestions are the amount requested (provided in the CSV), the length of the post, the sentiment scores of the post, and the readability of the post (see the previous homework assignment).

The dependent or "y" variable will be the status; consider this a 1 of the status is "Funding Successful" and a 0 in all other cases.

Train decision tree, k-nearest neighbors, and neural network machine learning models. You may choose an appropriate training/test split. Report the accuracy values from all three machine learning models. Report accuracy values from both ways of generating your "x" data for comparison.

The printout of your code may be brief. Simply report the metrics mentioned above.

Last considerations

Some considerations as you write your code:

Consider the possibility that the dataset cannot be loaded properly (the CSV does not exist, etc.). Handle this case gracefully in your code and print an appropriate error message.

When training your neural network model, you may see a warning message "Maximum iterations reached and the optimization hasn't converged yet." This message means that the neural network model would have preferred to work with a larger dataset, but it does not actually cause an error. However, *optionally*, if you would like to turn this message off, then you can do so using the following:

```
import warnings
warnings.filterwarnings("ignore")
```

Ensure that your output is crisp, professional, and well-formatted. For example, ensure that you have used spaces appropriately and checked your spelling.

Adding comments in your code is encouraged. At minimum, please use a comment at the start of your code to describe its basic functionality. In addition, for any, write appropriate docstrings. Ensure that your code would be as understandable as possible for a programmer working with your code for the first time.