FINA 5240: FinTech Analytics Assignment 2 Solution

Halis Sak

Due date: September 15, 2020

Question. We will work with Financial Times data (see "train.csv") for this homework assignment.

a) Read "train.csv" data to a `Python` dataframe named "df". Create a new column named "year" for "df" including only year information of "date" column. ("date" consists of strings and the first four characters is holding year information. Thus, we can simply get the first four characters and then convert this to an integer using `int()` function.) And please show that "year" column consist of only two unique values; 1998 and 2002.

b) What is the number of articles that were published in 2002 and has label of 1 in our dataset? (We are looking for number of rows for which "year" column is equal to 2002 and "label" is equal to 1.)

c) In part d) of homework assignment 1, we counted the number words in each article and put this into a list using the code below

```
>>> numbofWordsList = []
>>> for news in df.news:
      words = [word for word in news.split(" ") if word != ""]
      numbofWordsList.append(len(words))
```

We can simply create a new column "word_count" holding word counts as follows.

```
>>> df["word_count"] = numbofWordsList
```

Please compute the average word counts for articles of 1998 and 2002. And draw word count frequencies for articles of 1998 and 2002 seperately. The figures that I got are as follows. Hint. You can use `hist` function of `matplotlib.pyplot` library.