# Forming a Correlation Matrix

C ANISH

***Abstract*—This is a document explaining a method to find the Correlation Matrix.**

Download all python codes from

```
svn co https://github.com/chakki1234/
    Winter_intern/tree/main/correlation_matrix/
    codes
```

A video explaining the process documented can be found through the link

```
https://drive.google.com/file/d/1
    GKquFbnmw6De4ltrgvEE7SbDHoYYe6nG/
    view?usp=sharing
```

## 1 Solution

1.1. *aigiri.txt* contains a stothram in Telugu. The file is read and the variable *stothram_txt* is a string which contains the stothram.

```
stothram_file = open('aigiri.txt', 'r')
stothram_txt = stothram_file.read()
stothram_file.close()
```

1.2. *aigiri.txt* contains 21 slokams. Each slokam is added as an element to the empty list *slokams*. Before adding to the list each slokam is processed to remove escape sequences and the character $'|'$.

```
slokams = []
for sloka in stothram_txt.split('||'):
  if not sloka.replace('␣', '').isnumeric() and
      not sloka == '\n':
    slokams.append( sloka.replace('|', '').
        replace('\n', '') )
```

1.3. A empty ddictonary is initilized called *correlation_dict*. Each slokam from the list *slokams* is split into a list of words. A for loops runs through the list of words and each character is appened to the list *temp*. A key with the slokam number is created in the $correlation_dict$ and *temp* is intilized as its value.

```
correlation_dict = {}

for index, sloka in enumerate(slokams):
  temp = []
  for words in sloka.split('␣'):
    for char in words:
      temp.append(ord(char))
  correlation_dict[ str(index + 1) ] = np.
      array(temp)
```

1.4. The number of characters in each slokam is different. Hence the length of each list in the dictonary *correlation_dict* is different. To form a correlation matrix it is necessary that the length of all the lists are equal. The maximum length is found and all the list that have length less than the maximum length, $0's$ are added at the end to make it equal to the maximum length.

```
length = [ len(char_array) for char_array in
    correlation_dict.values() ]
max_length = max(length)

for key in correlation_dict:
  correlation_dict[key] = np.pad(
      correlation_dict[key], (0, max_length
      - len(correlation_dict[key])), 'constant
      ')
```

1.5. A pandas DataFrame object is created by passing *correlation_dict* as an argument and an predefind method is called to produce the correlation matrix.

```
df = pd.DataFrame( correlation_dict )
corrMatrix = df.corr()
print (corrMatrix)
```