TASK FOR CLASS

TASK 1: NLTK

Load all the documents in inaugural Corpus using the corpus reader class.

i) Obtain the words from the Trump's speech
ii) Find the total number of words in Trump's 2017 speech.
iii) Find the total number of distinct words in the same speech
iv) Find the average word type length of the same speech.

Code for obtaining words in Trump's speech:

```
from nltk.corpus import inaugural

# Get the file ID for Trump's speech in 2017
trump_speech_id = [file_id for file_id in inaugural.fileids() if "2017" in file_id and "Trump" in file_id][0]

# Load Trump's speech
trump_speech_words = inaugural.words(trump_speech_id)
len(trump_speech_words)
```

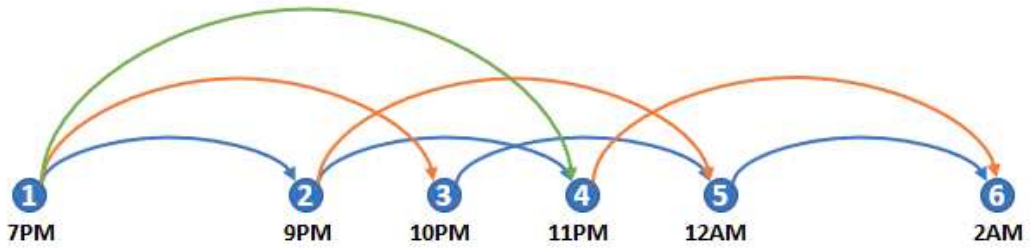
TASK 2: defaultdict, counter, and NLTK

i) Convert all the words identified above into lower case.
ii) Construct a frequency distribution over the lowercased words.
iii) Print out the words sorted by descending frequency along with the frequency.
iv) Write a function that finds the 50 most frequently occurring words in the speech that are not stop words.
v) Find out how many times the words 'india' and 'america' were used in the speech.

TASK 3: Networkx

Simplexville College campus shuttle bus begins running at 7:00pm and continues until 2:00am. Several drivers will be used, but only one should be on duty at any time. If a shift starts at or before 9:00pm, a regular driver can be obtained for a 4-hour shift at a cost of $50. Otherwise, part-time drivers need to be used. Several part-time drivers can work 3-hour shifts at $40, and the rest are limited to 2-hour shifts at $30. The college's goal is to schedule drivers in a way that minimizes the total cost of staffing the shuttle bus.

1. Formulate this problem as a graph problem. [**Hint.** Think of it as a weighted-directed graph where traversing a node can give a staffing schedule with edge weights as the cost of shift.]
2. Find all possible staffing schedules (from 7 PM to 2 AM) and print the staffing cost.
3. Find the shortest path length identifying the least staffing cost and the shortest path giving the schedule.



Each Node represents the hour. Node 1 is the source node and Node 6 is the sink node.

Each Blue arc represents the part time drivers doing 2 hour shift. Cost of each Blue arc is $ 30

Each Orange arc represents the part time drivers doing 3 hour shift. Cost of each Orange arc is $ 40

Each Green arc represents the regular drivers doing 4 hour shift. Cost of each Green arc is $ 50