1. Scrape this medium article (<https://towardsdatascience.com/data-science-and-machine-learning-interview-questions-3f6207cf040b>) and create a word cloud to visualize frequently used words

(Note: Select all paragraphs and join them together as a single string and pass it to python word cloud package)

1. Using amazon reviews data set do the following
   1. create a corpus using sklearn package.
   2. Create the Document Term Matrix using unigrams
   3. Identify top 25 unigrams and filter their respective columns from DTM
   4. Using these 25 columns, compute correlation matrix (if possible visualize the same using seaborn heatmap function).
   5. Using correlation matrix, Identify the pair of unigrams which has the highest correlation value
2. Using the above Document Term Matrix
   1. Use K-means clustering to cluster the reviews in to four groups
   2. Plot a bar chart to display no. of documents under each group
   3. Create Word cloud for each group
3. Using imdb movies reviews data set (imdb\_sentiment.csv: <https://bit.ly/2S2yXEd>), create a supervised model to predict the sentiment of user reviews.
   1. Create a word cloud using the user reviews
   2. Use random\_state=100 while splitting data in to training (80%) and testing(20%).
   3. Comment if the target variable contains balanced or imbalanced classes (In target variable: 0 = negative, 1 = positive sentiment)
   4. Build various supervised classification models using the training data set
   5. Predict and compute accuracy of the model on test data set
   6. Identify which model is best for sentiment prediction
4. Identify sentiment for the imdb user reviews using vader package. Comment the performance of supervised vs unsupervised sentiment prediction.
5. Apply topic modelling to cluster news based on their headlines. Using abcnews.csv (<https://github.com/skathirmani/datasets/raw/master/abcnews.csv.zip> file, randomly select 10,000 rows. Using multiple iterations, choose appropriate number of topics. Visualize number of documents under each cluster using a bar chart