Attempt all.

Text = “The Indian state of Kerala has been devastated by severe floods. More than 350 people have died, while more than a million have been evacuated to over 4,000 relief camps. Tens of thousands remain stranded.

The crisis is a timely reminder that climate change is expected to increase the frequency and magnitude of severe flooding across the world. Although no single flood can be linked directly to climate change, basic physics attests to the fact that a warmer world and atmosphere will hold more water, which will result in more intense and extreme rainfall.

The monsoon season usually brings heavy rains but this year Kerala has seen 42% more rain than would be expected, with more than 2,300mm of rain across the region since the beginning of June, and over 700mm in August alone.

These are similar levels seen during Hurricane Harvey, that hit Houston in August 2017, when more than 1,500mm of rain fell during one storm. Tropical cyclones and hurricanes, such as Harvey, are expected to increase in strength by up to 10% with a 2℃ rise in global temperature. Under climate change the probability of such extreme rainfall is also predicted to grow by up to sixfold towards the end of the century. The rivers and drainage systems of Kerala have been unable to cope with such large volumes of water and this has resulted in flash flooding.

Much of that water would normally be slowed down by trees or other natural obstacles. Yet over the past 40 years Kerala has lost nearly half its forest cover, an area of 9,000 km², just under the size of Greater London, while the state’s urban areas keep growing. This means that less rainfall is being intercepted, and more water is rapidly running into overflowing streams and rivers.”

1. Summarize this paragraph onto 3 sentences. Feel free to use either a count vectorizer or a tf-idf matrix.
2. Would it be possible to extract only the places from this paragraph? Use either a POS tagger with the noun entity or a NER (Named Entity recognition).
3. What is the overall polarity of the para? Also identify its subjectivity.
4. What is the paragraph about? Identify the top 5 words and figure out what the topic of discussion is.
5. training = [

('Tom Holland is a terrible spiderman.','pos'),

('a terrible Javert (Russell Crowe) ruined Les Miserables for me...','pos'),

('The Dark Knight Rises is the greatest superhero movie ever!','neg'),

('Fantastic Four should have never been made.','pos'),

('Wes Anderson is my favorite director!','neg'),

('Captain America 2 is pretty awesome.','neg'),

('Let\s pretend "Batman and Robin" never happened..','pos'),

]

testing = [

('Superman was never an interesting character.','pos'),

('Fantastic Mr Fox is an awesome film!','neg'),

('Dragonball Evolution is simply terrible!!','pos')]

Classify “the weather is terrible” into pos or neg. Also calculate the accuracy based on the testing set of your classifier.

1. Can you parse any news article of your choice using urllib/lxml? Please do mention the news article link in your code as a comment.
2. N-grams are defined as the combination of N keywords together. How many bi-grams can be generated from given sentence:

“Batman is a fictional superhero appearing in American comic books published by DC Comics.”

1. Solve the equation according to the sentence “I am planning to visit New Delhi to attend Analytics Vidhya Delhi Hackathon”.

A = (# of words with Noun as the part of speech tag)  
B = (# of words with Verb as the part of speech tag)  
C = (# of words with frequency count greater than one)

What are the correct values of A, B, and C?

1. In a corpus of N documents, one document is randomly picked. The document contains a total of T terms and the term “data” appears K times.

What is the correct value for the product of TF (term frequency) and IDF (inverse-document-frequency), if the term “data” appears in approximately one-third of the total documents?

A) KT \* Log(3)  
B) K \* Log(3) / T  
C) T \* Log(3) / K  
D) Log(3) / KT

1. What is the right order for a text classification model components
2. Text cleaning
3. Text annotation
4. Gradient descent
5. Model tuning
6. Text to predictors