NATURAL LANGUAGE PROCESSING-SENTIMENTAL ANALYSIS

OVERVIEW:

- Purpose : is to build a prediction model which predicts if the review on the restaurant is positive or negative.
- Steps followed in this procedure:

1.Importing the dataset: Here we have used pandas library to import the restaurant review dataset.

2.preprocessing: this is done by removing unwanted information like stop words and normalizing using the method called stemmatization.

3.vectorization: tokenize the dataset, convert the textual data to numeric format . Here we have made use of bag of words model, i.e count vectorizer to create a matrix of features and reviews.

4.classification: we have used logistic regression as the classification algorithm to get the prediction.

STRINGS

In week 1, we learnt about string properties, operations on strings ,file handling and regular expressions.

- String representation: either in single, double or triple quotes.
- Properties: strings are immune to alterations.
- Slicing : accessing a part of the string based on indexing, wherein you include the start index, and exclude the element at the end index.
- String operations:

  membership property: in and not in (output is a Boolean value)

  String formatting: operator-> %

  Some other built-in string functions: let s be a string. s.islower(), s.isalpha(), s.find(), s.count(), s.replace(), s.split() are some of the operations that can be performed on it.
- file handling: open, read or write, close are the operations that can be performed on a file.
- Regular expressions: sequence of characters used for operations like string searching, pattern matching, and string manipulations. Python has a built-in package called *re* for this.which offers functions like search, findall, split and sub that operates on strings.
- Eg: “^” matches the start of the string, “w+” matches alphanumeric characters in a string, “s” creates spaces in a string.

LINGUISTIC ANNOTATIONS

Libraries used:

- 1.spaCy – helps to build applications that process and understand the large volume of text.
- 2. NLTK – has a suite of text processing libraries like tokenization, stemming, etc

Terminologies: doc(container), span(a slice from doc) , token(word), lexeme(entry)

Tokenization: segmenting texts into words

Named entity recognition: labelling the named real world objects.

Stop words: list of common words that are often useful to filter out.

POS tagging: assigning word types of tokens so that spaCy can parse and tag a text.

Lemmatizers: used to assign base forms of words. eg: ‘compute’ for ‘computer’.

Stemmatizers: idea that relates to linguistic normalization. Eg; ‘wait’ for ‘waiting’.

Models: hybrid model is the one which combines rule based model with linguistic rules and statistical model with probabilities.

Wordnet: lexical database of English.

WORD VECTORIZATION, REGRESSION AND CLASSIFICATION.

- Word vectorization: process of turning a collection of text documents into numeric feature vectors.
- Bag of words model: focuses only on occurrence of word in a document and not on the order of information in the words.
- CountVectorizer: tokenizes ,builds vocabulary and encodes new documents using that vocabulary.
- N-gram model: can be constructed by finding pairs of words that occur next to each other.
- Skip gram model: predicts the surrounding words given a centre word. Reverse of how bag of words model works.
- TF-IDF method: term frequency-tells how often a given word appears within a document.
- inverse document frequency-downscales the words that appear a lot across the document.
- Uses TfidfVectorizer instead of CountVectorizer.
- Word embedding: vector representation of a particular word.
- Linear regression: models the relationship between 2 variables by fitting a linear equation to observed data. This is commonly done by method of least squares.
- Logistic regression: useful when dependent variable is binary.
- Logistic function: also called sigmoid function , takes any real values number, maps into a value between 0 an 1but never exactly to those limits.
- SVM: support vector machines, analyses data used for classification and regression analysis.
- Naïve baye's theorem: it is a classification technique which assumes independence among predictors.

NEURAL NETWORKS

- Neural networks: modelled after human brain, to recognize patterns that are contained in vectors. They help us cluster and classify when they have labelled dataset to train on.
- Has layers which are made of nodes, where the computation happens. Nodes combines input from the data with weights, this sum is passed through activation function and further through the network to affect the ultimate outcome.
- Gradient descent: (slope) optimization function: adjusts weights according to the error they caused. Their relationship is a derivative.
- Stochastic gradient: helps to avoid the problem of finding local extremities.
- Activation function: determines the output a node will generate based on its input.
- Recurrent neural networks: recursively applies computation to every instance of an input sequence conditioned on previous computed results.
- Long shot term memory: help preserve the error that can be backpropagated through time and layers.
- Encoder decoder networks: has 3 units: encoder, encoder vector, decoder.
- Each recurrent unit accepts a hidden state from the previous unit and produces output as well as its own hidden state.

CONCLUSION:

In this course I have learnt about classification algorithms, working in python using NLTK , about regression, vectorization and neural network and I have understood that natural language processing has many real world applications in major scenarios like speech recognition, sentiment analysis, and recommendation systems. This helped me to take a step towards working on a dataset, analyse it and understand it.

THANK YOU