Module 4: Text Classification 1

Case Study

edureka!

Case Study

1 Write a program to input three sentences from user and creates the corpus

Example:

Let’s say these 3 are your strings:

S1=” India won the match”
S2=” England won the cricket match”
S3=” Australia won the final match”

Then Corpus (list of union of all words from all strings) is:

[India, England, Australia, won, the, match, cricket, final]

Create a function named “MakeCorpus” which will take list of string as an input and will return a list having union of all words. Save this function in a python file named “Corpus”. This can be used for future applications

2.Write a program to input three sentences from user and convert them into vectors. Use presence and absence of words to build the vectors.

Example:

Let’s say these 3 are your strings:

S1=” India won the match”
S2=” England won the cricket match”
S3=” Australia won the final match”

Then Corpus (list of union of all words from all strings) is:

[India, England, Australia, won, the, match, cricket, final]

So, S1 will be [1,0,0,1,1,1,0,0]
S2 will be [0,1,0,1,1,1,1,0]
S3 will be [0,0,1,1,1,1,0,1]

Create a function named “PresenceAbsenceVectorization” which will take list of string as an input and will return a list of vectors. Save this function in a python file named “Vectorization”. This can be used for future applications

3. Write a program to enter 3 strings from a user and vectorise them on basis of their counts.

Example:

Let's say these 3 are your strings:

S1=" A lives with B. A plays with C. "
S2=" B lives with C. B plays with D"
S3=" C lives with D. C plays with A"

Then Corpus (list of union of all words from all strings) is:

[A, B, C, D, lives, with, plays]

So, S1 will be [2,1,1,0,1,2,1]
S2 will be [0,2,1,1,1,2,1]
S3 will be [1,0,2,1,1,2,1]

Create a function named "CountVectorization" which will take list of string as an input and will return a list of vectors. Save this function in a python file named "Vectorization". This can be used for future applications

4. Write a program to input 3 strings but vectorise them using TF-IDF (Term Frequency and Inverse Document Frequency) and print the strings along with the vectors.

You can use already available python TF-IDF Vectorizer
(Refer : http://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html)

Create a function named "TFIDFVectorization" which will take list of string as an input and will return a list of vectors. Save this function in a python file named "Vectorization". This can be used for future applications