Information Retrieval Systems

Lab Practical and date – Practical 1, Wednesday 22nd July 2020

Name and Roll Number- Het Shah, 17BIT103

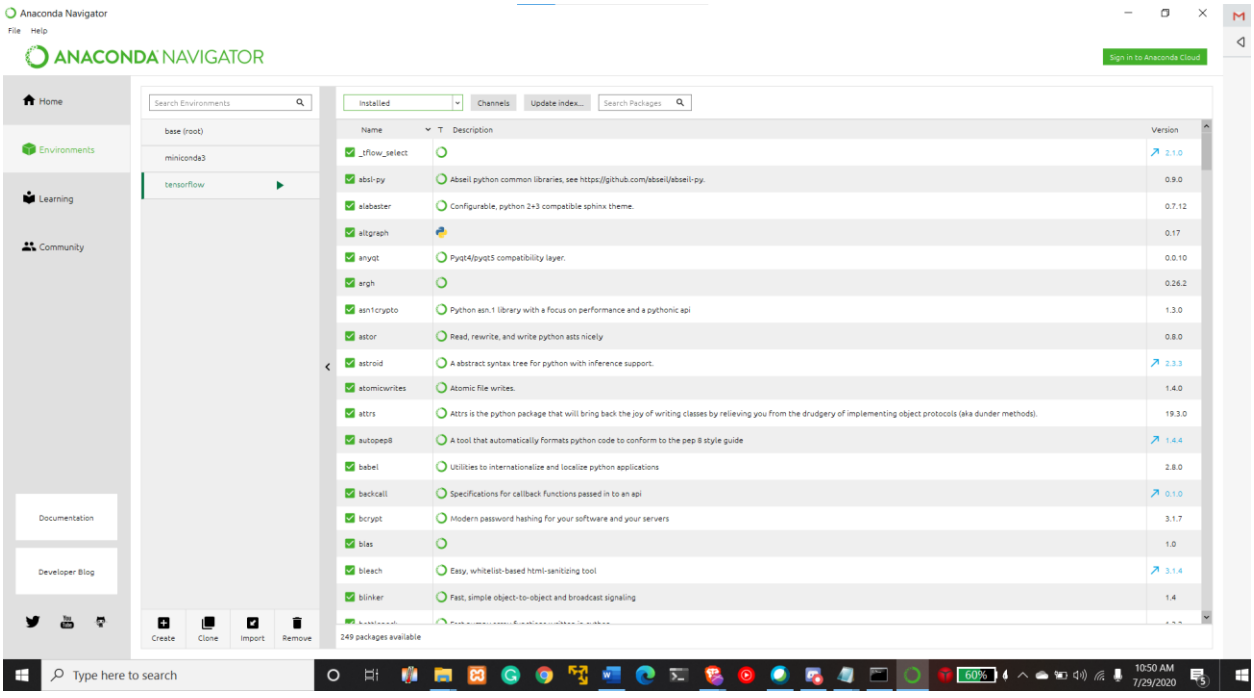
Practical Objective- Study and exploration of Python libraries for Information Retrieval and related tasks.

Background

1) Install the Anaconda Framework

Anaconda is a free and open-source distribution of the Python and R programming languages for scientific computing, that aims to simplify package management and deployment.

I installed anaconda and create an environment called tensorflow which keeps a track of all my libraries and packages.



As we can see, my tensorflow anaconda environment has 249 packages installed out of which many are default and I have installed some such as NumPy, matplotlib, etc.

2) Explore and study following libraries and packages:

a) NLTK

NLTK stands for Natural Language tookkit which is a collection of libraries and packages for NLP programming for the English language written in python. NLTK also consists of graphical demonstrations and sample data.

NLTK supports many IRS functionality such as classification, tokenization, stemming, tagging, parsing, and semantic reasoning functionalities.

It can also identify named entities, tokenize and tag some text as well as generate parse trees. As of July 2020, its latest version is NLTK 3.5

b) Sci-Kit Learn

Scikit-learn is a machine learning library written in the python language. It supports many machine learning classification, regression and clustering algorithms such as random forests, gradient boosting, k-means clustering, DBSCAN and support vector machines.

Sci-Kit learn was designed to operate with other computing libraries such as NumPy and SciPy. It also uses other python libraries such as matplotlib for plotting and visualizing the data, NumPy for high performance linear algebra and array operations and also pandas for its dataframes functionalities.

As of July 2020, its latest version is scikit learn 0.23.0

3) Write a small program that will accept a sentence from user, and provide number of words, and list of words as the output. Assume that the words are separated by blank space.

Package Used-

1) OS

OS package is used in python for the file management task such as creating and removing a directory, fetching contents of a file and changing the current directory.

Methods-

1) removePunctuations(inputString)

This custom-made method removes all the punctuations(including symbols) from the input string and returns the string without all the punctuations.

2) removeDigits(inputString)

This custom-made method removes all the numbers from the input string

3) strip()

This inbuild methods removes all the trailing and leading white spaces

Data Structure Used-

1) Set

All the words are stores in the set so they are displayed only once

I/P

hi i am het and i am 20 years old.

O/P

number of words in the sentence are :9
the word list of the sentence is {'years', 'i', 'and', 'het', 'am', 'old', 'hi'}

Since ‘I’ and ‘am’ is repeated twice it is not included in the word list however they are counted for the total word count

Conclusion

In this practical, we learned about the python anaconda environment and its setup. In addition we learned about 2 python libraries that will be useful for the IRS practical which are NLTK and Sci-kit Learn. In the end, we wrote a simple program to output the wordcount of the sentence after receiving it from the user.

In [1]:
```
import os
```

In [2]:
```
def removePunctuations(inputString):
    punctuationList = '''!()-[]{};:'"\,<>./?@#$%^&*_~'''
    resultString = ''
    for char in inputString:
        if char not in punctuationList:
            resultString = resultString + char
    return resultString
```

In [3]:
```
def removeDigits(inputString):
    resultString = ''.join([i for i in inputString if not i.isdigit()])
    return resultString
```

In [6]:
```
print("Enter content to search:")
searchString = input()
searchString= removePunctuations(searchString)

searchString= removeDigits(searchString)
print(searchString)
words= searchString.split(" ")
count = 0
wordset=set()
for i in words:
    if i is not ' ' and i is not '':
        wordset.add(i)
        count=count+1
```

```
Enter content to search:
hi i am het and i am 20 years old.
hi i am het and i am  years old
```

In [7]:
```
print('number of words in the sentence are :' + str(count))
print('the word list of the sentece is' + str(wordset))
```

```
number of words in the sentence are :9
the word list of the sentece is{'years', 'i', 'and', 'het', 'am', 'old', 'h
i'}
```

In [ ]:

In [ ]:

In [ ]: