In [1]:

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns
from IPython import get_ipython
import warnings
warnings.filterwarnings("ignore")
```

In [2]:

```
data = pd.read_csv('coursera.csv')
```

In [3]:

```
data.head()
```

Out[3]:

	Course Name	University	Difficulty Level	Course Rating	Course URL	Des
0	Write A Feature Length Screenplay For Film Or ...	Michigan State University	Beginner	4.8	https://www.coursera.org/learn/write-a-feature...	Wri Feat Scri
1	Business Strategy: Business Model Canvas Analy...	Coursera Project Network	Beginner	4.8	https://www.coursera.org/learn/canvas-analysis...	By th this pro
2	Silicon Thin Film Solar Cells	cole Polytechnique	Advanced	4.1	https://www.coursera.org/learn/silicon-thin-fi...	This cons prese
3	Finance for Managers	IESE Business School	Intermediate	4.8	https://www.coursera.org/learn/operational-fin...	c n
4	Retrieve Data using Single-Table SQL Queries	Coursera Project Network	Beginner	4.6	https://www.coursera.org/learn/single-table-sq...	In this you effe

In [4]:

```
1 data.tail()
```

Out[4]:

	Course Name	University	Difficulty Level	Course Rating	Course URL	De
3517	Capstone: Retrieving, Processing, and Visualiz...	University of Michigan	Beginner	4.6	https://www.coursera.org/learn/python-data-vis...	In the stu build a
3518	Patrick Henry: Forgotten Founder	University of Virginia	Intermediate	4.9	https://www.coursera.org/learn/henry	libe me Rem
3519	Business intelligence and data analytics: Gene...	Macquarie University	Advanced	4.6	https://www.coursera.org/learn/business-intell...	Mega
3520	Rigid Body Dynamics	Korea Advanced Institute of Science and Techno...	Beginner	4.6	https://www.coursera.org/learn/rigid-body-dyna...	Th dynar of th
3521	Architecting with Google Kubernetes Engine: Pr...	Google Cloud	Intermediate	4.7	https://www.coursera.org/learn/deploying-secur...	In thi "Ar wi

In [5]:

```
1 data.shape
```

Out[5]:

```
(3522, 7)
```

In [6]:

```
1 data.columns
```

Out[6]:

```
Index(['Course Name', 'University', 'Difficulty Level', 'Course Rating',
       'Course URL', 'Course Description', 'Skills'],
      dtype='object')
```

In [7]:

```
1 data.duplicated().sum()
```

Out[7]:

```
98
```

In [8]:

```
1 data = data.drop_duplicates()
```

In [9]:

```
1 data.isnull().sum()
```

Out[9]:

```
Course Name           0
University            0
Difficulty Level      0
Course Rating         0
Course URL            0
Course Description    0
Skills                0
dtype: int64
```

In [10]:

```
1 data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Int64Index: 3424 entries, 0 to 3521
Data columns (total 7 columns):
 #   Column              Non-Null Count  Dtype 
---  ------              --------------  ----- 
 0   Course Name         3424 non-null   object
 1   University          3424 non-null   object
 2   Difficulty Level    3424 non-null   object
 3   Course Rating       3424 non-null   object
 4   Course URL          3424 non-null   object
 5   Course Description  3424 non-null   object
 6   Skills              3424 non-null   object
dtypes: object(7)
memory usage: 214.0+ KB
```

In [11]:

```
1 data.nunique()
```

Out[11]:

```
Course Name           3416
University             184
Difficulty Level         5
Course Rating           31
Course URL            3424
Course Description    3397
Skills                3424
dtype: int64
```

In [12]:

```
1 data['Difficulty Level'].unique()
```

Out[12]:

```
array(['Beginner', 'Advanced', 'Intermediate', 'Not Calibrated',
       'Conversant'], dtype=object)
```

In [13]:

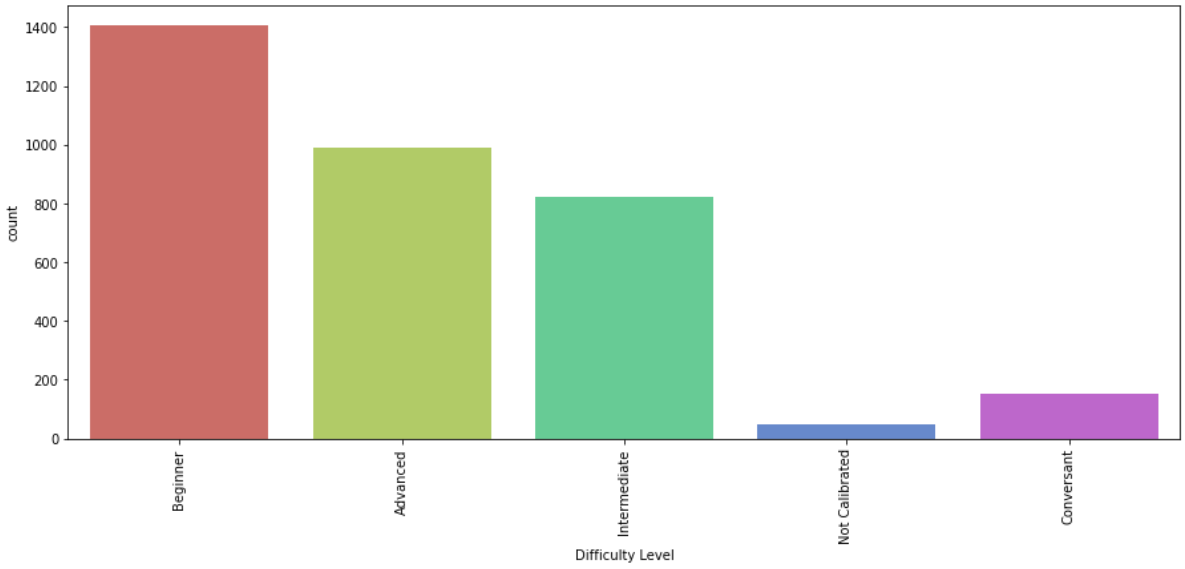
```
1 data['Difficulty Level'].value_counts()
```

Out[13]:

```
Beginner          1406
Advanced           991
Intermediate       823
Conversant         154
Not Calibrated      50
Name: Difficulty Level, dtype: int64
```

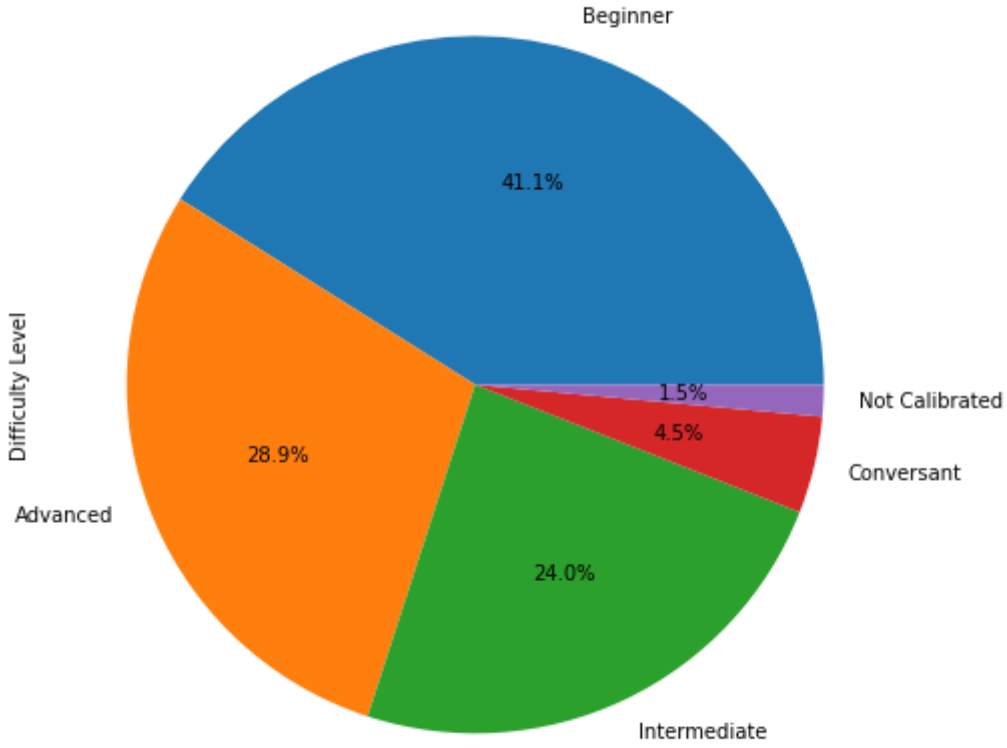
In [14]:

```
plt.figure(figsize=(15,6))
sns.countplot('Difficulty Level', data = data,
              palette = 'hls')
plt.xticks(rotation = 90)
plt.show()
```



In [15]:

```
data['Difficulty Level'].value_counts().plot(kind='pie',
                                             figsize=(8, 8),
                                             autopct='%1.1f%%')
plt.xticks(rotation = 90)
plt.show()
```



In [16]:

```
data['Course Rating'].unique()
```

Out[16]:

```
array(['4.8', '4.1', '4.6', '4.7', '3.3', '4.9', '4.3', 'Not Calibrated',
       '4', '4.4', '3.4', '4.5', '4.2', '5', '3.5', '3.7', '3', '3.6',
       '3.8', '3.9', '2.9', '2.6', '2.8', '2', '3.1', '3.2', '2.5', '2.4',
       '1', '1.9', '2.3'], dtype=object)
```

In [17]:

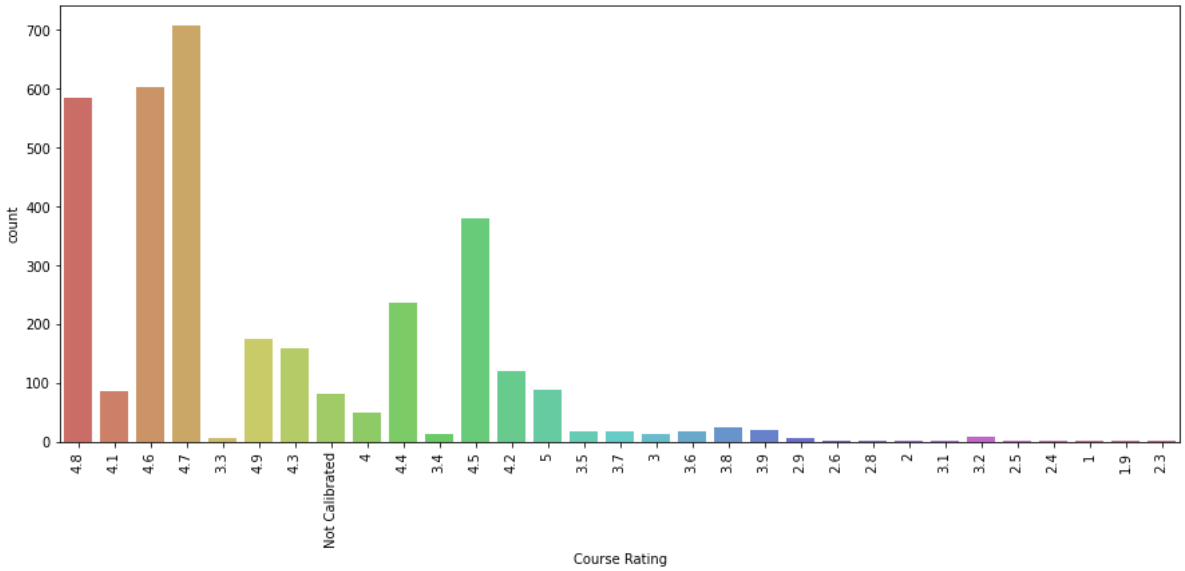
```
1 data['Course Rating'].value_counts()
```

Out[17]:

```
4.7               707
4.6               602
4.8               585
4.5               380
4.4               235
4.9               174
4.3               159
4.2               120
5                  89
4.1                85
Not Calibrated     82
4                  50
3.8                24
3.9                20
3.6                18
3.7                18
3.5                17
3.4                13
3                  12
3.2                 9
3.3                 6
2.9                 6
2.6                 2
2.8                 2
2.4                 2
1                   2
2                   1
3.1                 1
2.5                 1
1.9                 1
2.3                 1
Name: Course Rating, dtype: int64
```

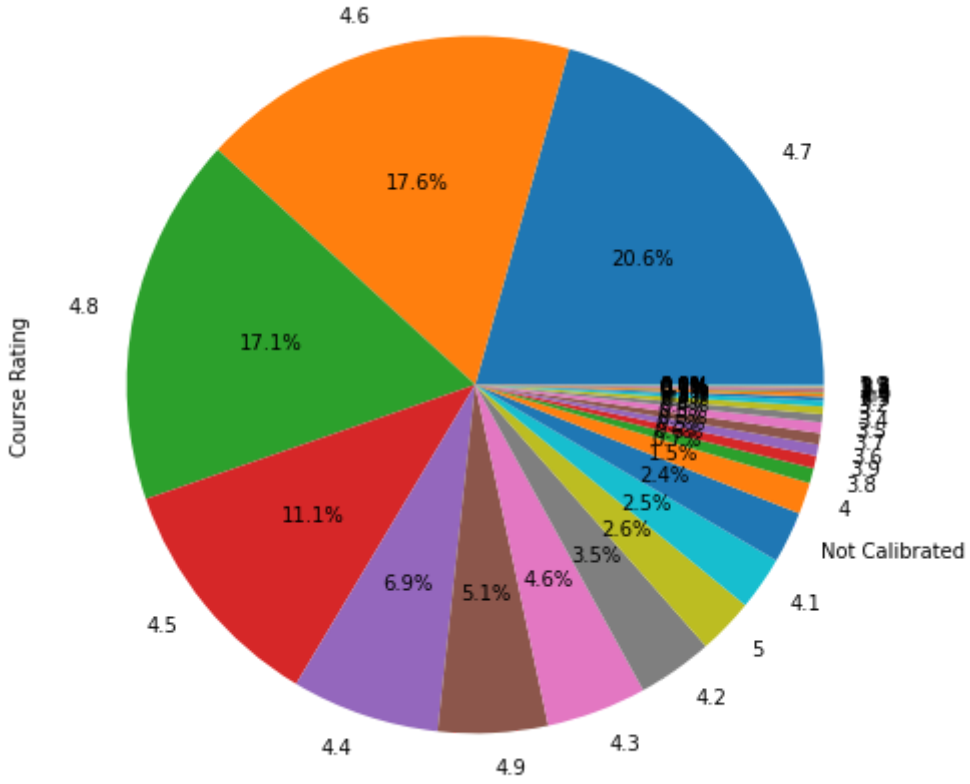
In [18]:

```
plt.figure(figsize=(15,6))
sns.countplot('Course Rating', data = data,
              palette = 'hls')
plt.xticks(rotation = 90)
plt.show()
```



In [19]:

```
data['Course Rating'].value_counts().plot(kind='pie',
                                          figsize=(8, 8),
                                          autopct='%1.1f%%')
plt.xticks(rotation = 90)
plt.show()
```



In [20]:

```
data['University'].unique()
```

Out[20]:

```
array(['Michigan State University', 'Coursera Project Network',
       '�cole Polytechnique', 'IESE Business School',
       'The Chinese University of Hong Kong', 'University of Washingto
n',
       'The State University of New York',
       'Saint Petersburg State University',
       'University of California, Irvine',
       'University of Colorado System', 'Rice University',
       'University of Geneva', 'University of California, Davis',
       '�cole Polytechnique F�d�rale de Lausanne', 'Google Cloud',
       'National Research Tomsk State University',
       'University of Florida', 'Johns Hopkins University',
       'Universiteit Leiden', 'The University of Edinburgh',
       'Columbia University',
       'Korea Advanced Institute of Science and Technology(KAIST)',
       'University of Maryland, College Park', 'IBM',
       'The Hong Kong University of Science and Technology',
```

In [21]:

```
1 data['University'].value_counts()
```

Out[21]:

```
Coursera Project Network                      562
University of Illinois at Urbana-Champaign    138
University of Michigan                        101
Johns Hopkins University                      101
University of Colorado Boulder                101
                                             ...
GitLab                                          1
Yeshiva University                              1
University of Glasgow                           1
Laureate Education                              1
The World Bank Group                            1
Name: University, Length: 184, dtype: int64
```

In [22]:

```
1 data_university = data['University'].value_counts()
```

In [23]:

```
1 data_university = pd.DataFrame(data_university)
```

In [24]:

```
1 data_university = data_university.reset_index()
```

In [25]:

```
1 data_university.head()
```

Out[25]:

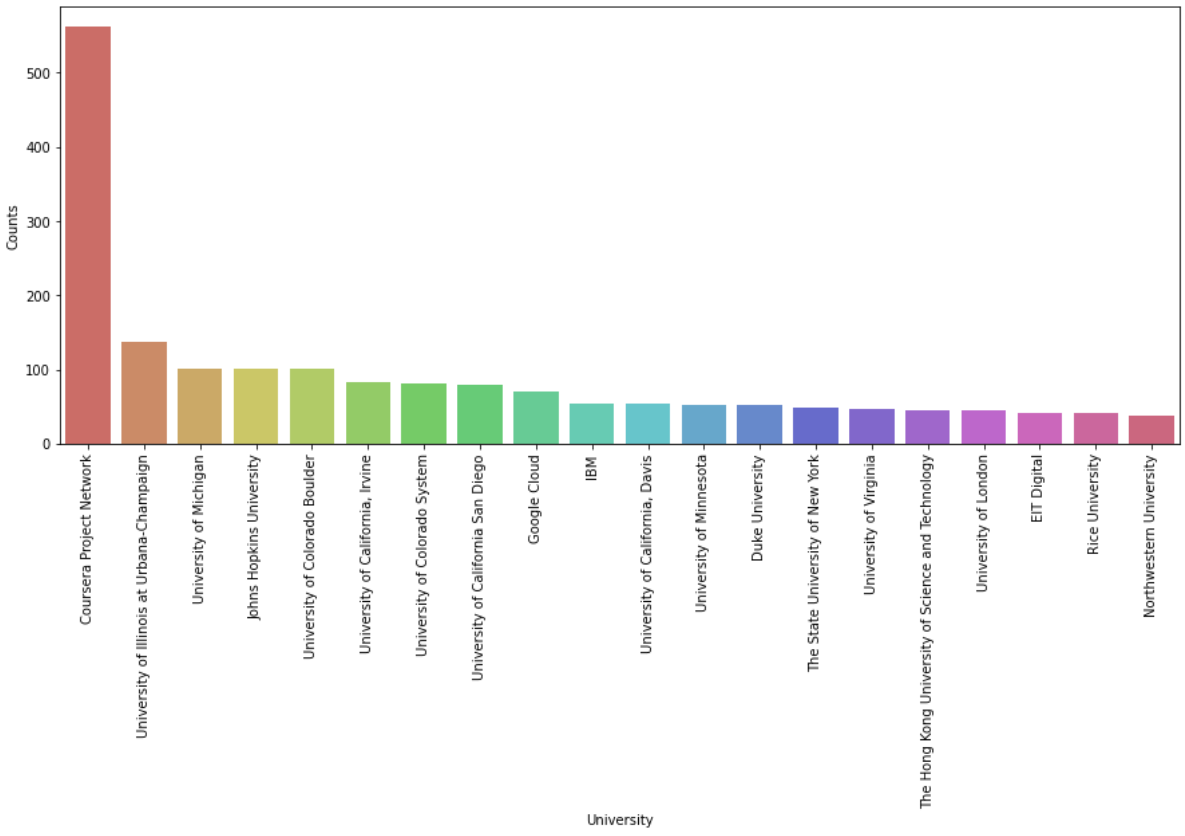
	index	University
0	Coursera Project Network	562
1	University of Illinois at Urbana-Champaign	138
2	University of Michigan	101
3	Johns Hopkins University	101
4	University of Colorado Boulder	101

In [26]:

```
1 data_university.rename(columns = {'index':'University', 'University':'Counts'}, inp
```

In [29]:

```
plt.figure(figsize=(15,6))
sns.barplot(x = 'University', y = 'Counts' ,
            data = data_university.head(20), palette = 'hls')
plt.xticks(rotation = 90)
plt.show()
```



In [30]:

```
data['Course Name'].unique()
```

Out[30]:

```
array(['Write A Feature Length Screenplay For Film Or Television',
       'Business Strategy: Business Model Canvas Analysis with Miro',
       'Silicon Thin Film Solar Cells', ...,
       'Business intelligence and data analytics: Generate insights',
       'Rigid Body Dynamics',
       'Architecting with Google Kubernetes Engine: Production'],
      dtype=object)
```

In [31]:

```
data['Course Name'].value_counts()
```

Out[31]:

```
Corporate Strategy                                    2
Introduction to Psychology                            2
Cryptography                                          2
Portfolio and Risk Management                         2
Python Data Structures                                2
                                                     ..
Symmetric Cryptography                                1
Using Descriptive Statistics to Analyze Data in R     1
BIM Fundamentals for Engineers                        1
Organizational Behavior: How to Manage People         1
Architecting with Google Kubernetes Engine: Production    1
Name: Course Name, Length: 3416, dtype: int64
```

In [32]:

```
data = data[['Course Name','Difficulty Level',
             'Course Description','Skills']]
```

In [33]:

```
data['Course Name'] = data['Course Name'].str.replace(' ',',')
data['Course Name'] = data['Course Name'].str.replace(',,',',')
data['Course Name'] = data['Course Name'].str.replace(':','')
data['Course Description'] = data['Course Description'].str.replace(' ',',')
data['Course Description'] = data['Course Description'].str.replace(',,',',')
data['Course Description'] = data['Course Description'].str.replace('_','')
data['Course Description'] = data['Course Description'].str.replace(':','')
data['Course Description'] = data['Course Description'].str.replace('(','')
data['Course Description'] = data['Course Description'].str.replace(')','')
data['Skills'] = data['Skills'].str.replace('(','')
data['Skills'] = data['Skills'].str.replace(')','')
```

In [34]:

```
data['Keywords'] = data['Course Name'] + data['Difficulty Level'] + data['Course De
```

In [35]:

```
1 data['Keywords'].iloc[1]
```

Out[35]:

```
'Business,Strategy,Business,Model,Canvas,Analysis,with,MiroBeginnerBy,the,
end,of,this,guided,project,you,will,be,fluent,in,identifying,and,creating,
Business,Model,Canvas,solutions,based,on,previous,high-level,analyses,and,
research,data.,This,will,enable,you,to,identify,and,map,the,elements,requi
red,for,new,products,and,services.,Furthermore,it,is,essential,for,generat
ing,positive,results,for,your,business,venture.,This,guided,project,is,des
igned,to,engage,and,harness,your,visionary,and,exploratory,abilities.,You,
will,use,proven,models,in,strategy,and,product,development,with,the,Miro,p
latform,to,explore,and,analyse,your,business,propositions.,,We,will,practi
ce,critically,examining,results,from,previous,analysis,and,research,result
s,in,deriving,the,values,for,each,of,the,business,model,sections.Finance
business plan  persona user experience  business model canvas  Planning  B
usiness  project  Product Development  presentation  Strategy business bus
iness-strategy'
```

In [36]:

```
1 data_new = data[['Course Name','Keywords']]
```

In [37]:

```
1 data_new['Keywords'] = data['Keywords'].str.replace(',',' ')
```

In [38]:

```
1 data_new['Course Name'] = data['Course Name'].str.replace(',',' ')
```

In [39]:

```
1 data_new['Keywords'] = data_new['Keywords'].apply(lambda x:x.lower())
```

In [40]:

```
1 data_new.shape
```

Out[40]:

```
(3424, 2)
```

In [41]:

```
1 from sklearn.feature_extraction.text import CountVectorizer
```

In [43]:

```
x = np.array(data_new["Keywords"])
y = np.array(data_new["Course Name"])
```

In [44]:

```
cv = CountVectorizer(max_features=5000,stop_words='english')
x = cv.fit_transform(x)
```

In [45]:

```
import nltk
```

In [46]:

```
from nltk.stem.porter import PorterStemmer
ps = PorterStemmer()
```

In [47]:

```
def stem(text):
    y=[]

    for i in text.split():
        y.append(ps.stem(i))

    return " ".join(y)
```

In [48]:

```
data_new['Keywords'] = data_new['Keywords'].apply(stem)
```

In [49]:

```
from sklearn.metrics.pairwise import cosine_similarity
```

In [50]:

```
similarity = cosine_similarity(x)
```

In [56]:

```
data_new.rename(columns = {'Course Name':'course_name'}, inplace = True)
```

In [57]:

```
def recommend(course):
    course_index = data_new[data_new['course_name'] == course].index[0]
    distances = similarity[course_index]
    course_list = sorted(list(enumerate(distances)),reverse=True, key=lambda x:x[1]
    for i in course_list:
        print(data_new.iloc[i[0]].course_name)
```

In [69]:

```
recommend('Finance for Managers')
```

```
Fundamentals of financial and management accounting
Introduction to Finance The Basics
The Language and Tools of Financial Analysis
Corporate finance Know your numbers 2
Finance for Non-Financial Professionals
Operations Management Analysis and Improvement Methods
```

In [70]:

```
recommend('Python Programming Essentials')
```

```
Python Data Representations
Python Data Analysis
Python Basics
Programming for Everybody (Getting Started with Python)
Python Functions Files and Dictionaries
Python Programming A Concise Introduction
```