In [1]:

```
import pandas as pd
```

In [2]:

```
df = pd.read_csv('EdX.csv')
```

In [3]:

```
df.head()
```

Out[3]:

	Name	University	Difficulty Level	Link	Ab
0	How to Learn Online	edX	Beginner	https://www.edx.org/course/how-to-learn-online	Le esser strategies succes or
1	Programming for Everybody (Getting Started wit...	The University of Michigan	Beginner	https://www.edx.org/course/programming-for-eve...	This cours a prerequis introduct
2	CS50's Introduction to Computer Science	Harvard University	Beginner	https://www.edx.org/course/cs50s-introduction-...	An introduc to intellec enterpris
3	The Analytics Edge	Massachusetts Institute of Technology	Intermediate	https://www.edx.org/course/the-analytics-edge	Thro inspi examples stor disc
4	Marketing Analytics: Marketing Measurement Str...	University of California, Berkeley	Beginner	https://www.edx.org/course/marketing-analytics...	This cours part MicroMaste Prog

In [4]:

```
df.tail()
```

Out[4]:

	Name	University	Difficulty Level	Link	A
715	Global China: From the Mongols to the Ming	Harvard University	Beginner	https://www.edx.org/course/global-china-from-t...	Explor impact c conc dynastie
716	Leaders in Citizen Security and Justice Manage...	Inter-American Development Bank	Intermediate	https://www.edx.org/course/leaders-in-citizen-...	Learn a the late prever polic
717	Computational Neuroscience: Neuronal Dynamics ...	École polytechnique fédérale de Lausanne	Advanced	https://www.edx.org/course/computational-neuro...	This cc explain mathem and co
718	Cities and the Challenge of Sustainable Develo...	SDG Academy	Beginner	https://www.edx.org/course/cities-and-the-chal...	Wha sustair city? L the b
719	MathTrackX: Special Functions	University of Adelaide	Beginner	https://www.edx.org/course/mathtrackx-special-...	Unders trigonom expone and lc

In [5]:

```
df.shape
```

Out[5]:

```
(720, 6)
```

In [6]:

```
df.columns
```

Out[6]:

```
Index(['Name', 'University', 'Difficulty Level', 'Link', 'About',
       'Course Description'],
      dtype='object')
```

In [7]:

```
df.duplicated().sum()
```

Out[7]:

```
1
```

In [8]:

```
df.isnull().sum()
```

Out[8]:

```
Name                  0
University            0
Difficulty Level      0
Link                  0
About                 0
Course Description    0
dtype: int64
```

In [9]:

```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 720 entries, 0 to 719
Data columns (total 6 columns):
 #   Column              Non-Null Count  Dtype 
---  ------              --------------  ----- 
 0   Name                720 non-null    object
 1   University          720 non-null    object
 2   Difficulty Level    720 non-null    object
 3   Link                720 non-null    object
 4   About               720 non-null    object
 5   Course Description  720 non-null    object
dtypes: object(6)
memory usage: 33.9+ KB
```

In [10]:

```
df.nunique()
```

Out[10]:

```
Name                  717
University            102
Difficulty Level        3
Link                  719
About                 698
Course Description    717
dtype: int64
```

In [11]:

```python
import matplotlib.pyplot as plt
import seaborn as sns
```

In [12]:

```python
import warnings
warnings.filterwarnings('ignore')
```

In [13]:

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

```
Out[13]:
```

```
array(['edX', 'The University of Michigan', 'Harvard University',
       'Massachusetts Institute of Technology',
       'University of California, Berkeley', 'University of Adelaide',
       'The University of Queensland', 'Curtin University',
       'Technische Universität München',
       'National Research Nuclear University', 'SDG Academy',
       'University of Hong Kong', 'IBM',
       'Wageningen University & Research', 'Boston University',
       'Educational Testing Service',
       'Universitat Politècnica de Valencia', 'The Linux Foundation',
       'Indian Institute of Management Bangalore',
       'The Georgia Institute of Technology', 'New York University',
       'The World Wide Web Consortium (W3C)', 'Columbia University',
       'Stanford University', 'Cornell University', 'Davidson College',
       'Berklee College of Music', 'The Hong Kong Polytechnic University',
       'University of Washington', 'Dartmouth_IMTx-Dartmouth College-IMT',
       'University of Pennsylvania', 'United Arab Emirates University',
       'The University of California, San Diego', 'MandarinX',
       'Tsinghua University', 'The International Monetary Fund',
       'University of British Columbia', 'IITBombay',
       'Delft University of Technology',
       'The Hong Kong University of Science and Technology',
       'The University of Maryland, College Park-University System of Mary
land',
       'IsraelX-Tel Aviv University', 'University of Oxford',
       'Doane University', 'University of Toronto',
       'University System of Maryland-The University of Maryland, College
Park',
       'Rochester Institute of Technology',
       'Imperial College Business School-Imperial College London',
       'Catalyst', 'Universidad Carlos III de Madrid',
       'Universidad Galileo', 'Université catholique de Louvain',
       'Kyoto University', 'The National University of Singapore',
       'Inter-American Development Bank', 'SchoolYourself',
       'Universidades Anáhuac',
       'The Wharton School of the University of Pennsylvania',
       'The University of Texas at Austin', 'Tecnológico de Monterrey',
       'Arizona State University', 'Rolls-Royce', 'Amnesty International',
       'Wits University', 'Waseda University',
       'The Smithsonian Institution',
       'École polytechnique fédérale de Lausanne',
       'Georgetown University', 'Babson College',
       'The University of Tokyo', 'Australian National University',
       'KU Leuven University', 'World Bank Group', 'Dartmouth College',
       'NYIF', 'Chalmers University of Technology',
       'Imperial College London',
       'Salisbury University-University System of Maryland',
       'Delft University of Technology-Wageningen University & Research-De
lft University & Wageningen University',
       'The University of Iceland', 'Western Governors University',
       'University System of Maryland-University of Maryland Global Campu
s',
       'Rice University', 'The University of Edinburgh',
       'Université de Montréal', 'University of Texas at Arlington',
       'ETH Zurich', 'Amazon Web Services',
       'Università degli Studi di Napoli Federico II',
       'Queen’s University',
       'DelftXRWTHx-Delft University of Technology-RWTH Aachen Universit
y',
       'The University of Texas of the Permian Basin',
       'Tokyo Institute of Technology', 'IMT', 'Davidson Next',
```

In [14]:

```
df['University'].value_counts()
```

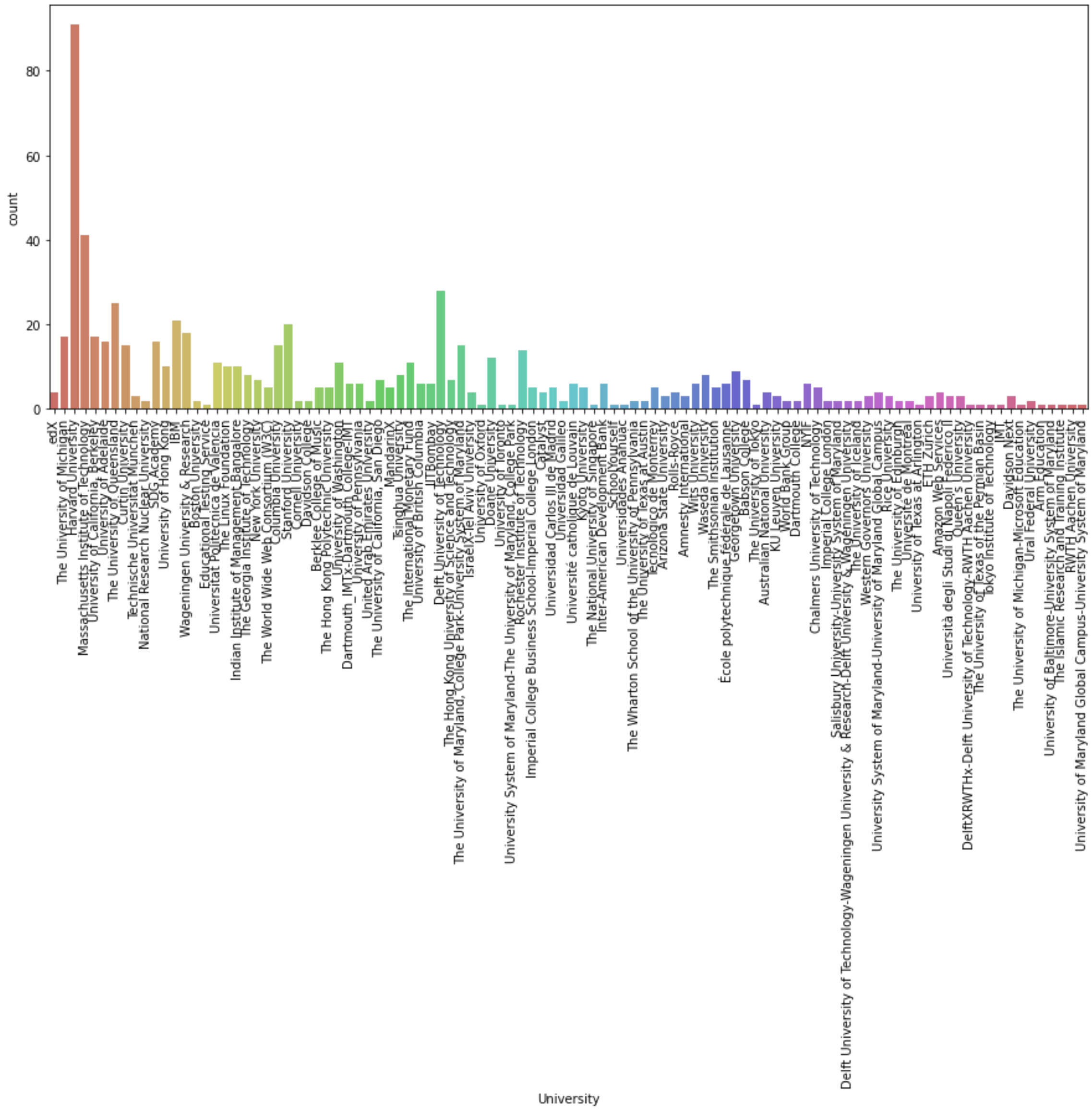
Out[14]:

```
Harvard University
91
Massachusetts Institute of Technology
41
Delft University of Technology
28
The University of Queensland
25
IBM
21
..
University of Oxford
1
University of Toronto
1
University System of Maryland-The University of Maryland, College Park
1
Universidades Anahuac
1
University of Maryland Global Campus-University System of Maryland
1
Name: University, Length: 102, dtype: int64
```

```
       'The University of Michigan-Microsoft Education',
       'Ural Federal University', 'Arm Education',
       'University of Baltimore-University System of Maryland',
       'The Islamic Research and Training Institute',
       'RWTH Aachen University',
       'University of Maryland Global Campus-University System of Marylan
d'],
      dtype=object)
```

In [15]:

```
plt.figure(figsize=(15,6))
sns.countplot(df['University'], data = df, palette ='hls')
plt.xticks(rotation = 90)
plt.show()
```



In [17]:

```
df['Difficulty Level'].unique()
```

Out[17]:

```
array(['Beginner', 'Intermediate', 'Advanced'], dtype=object)
```

In [18]:

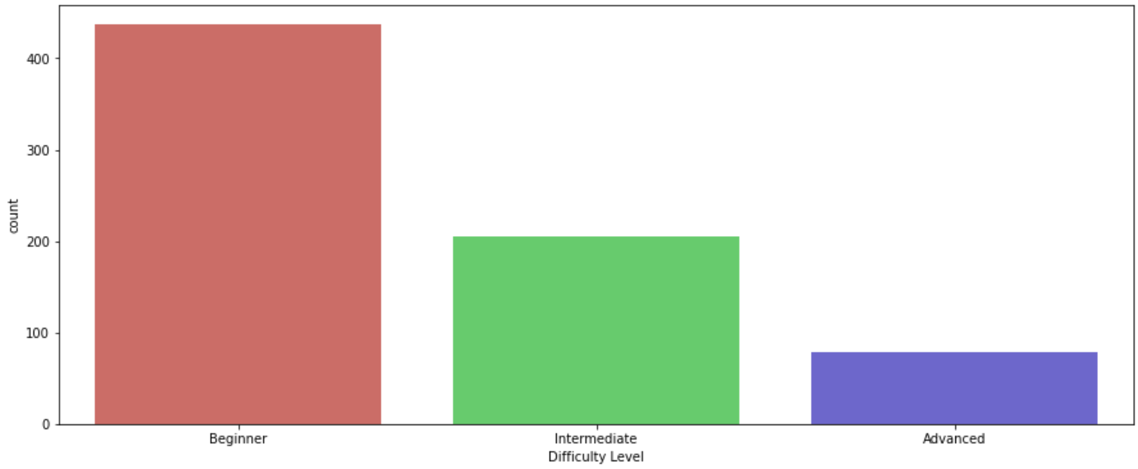
```
df['Difficulty Level'].value_counts()
```

Out[18]:

```
Beginner        437
Intermediate    205
Advanced         78
Name: Difficulty Level, dtype: int64
```

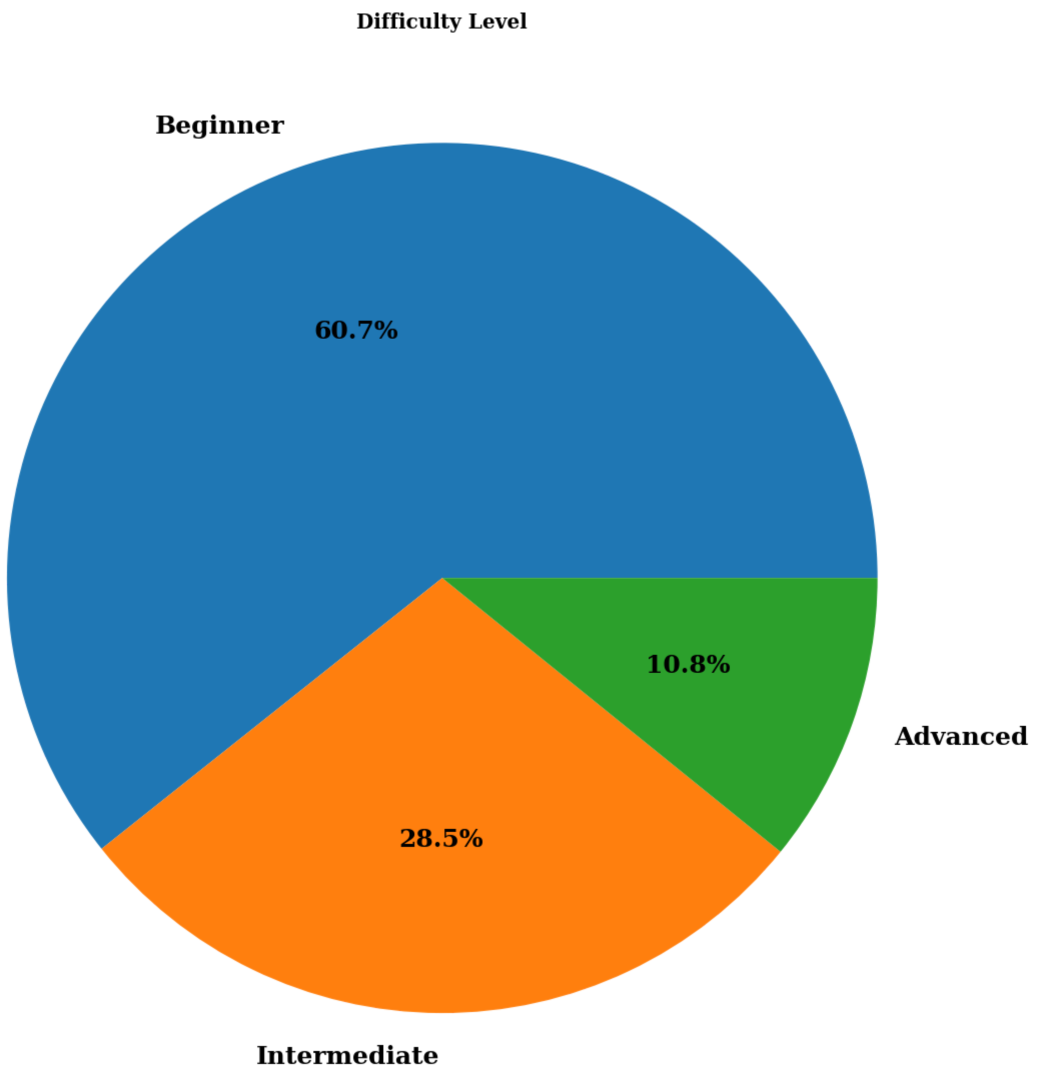
In [19]:

```
plt.figure(figsize=(15,6))
sns.countplot(df['Difficulty Level'], data = df, palette = 'hls')
plt.show()
```



In [20]:

```
plt.figure(figsize=(30,20))
plt.pie(df['Difficulty Level'].value_counts(), labels=df['Difficulty Level'].value_count
                                                    'color': 'black',
                                                    'weight': 'bold',
                                                    'family': 'serif' })
hfont = {'fontname':'serif', 'weight': 'bold'}
plt.title('Difficulty Level', size=20, **hfont)
plt.show()
```



In [21]:

```
df = df.drop(['Link'],axis=1)
```

In [27]:

```
import re
import string
```

In [23]:

```
def clean_text(text):
    '''Make text lowercase, remove text in square brackets,remove links,remove punctuati
    and remove words containing numbers.'''
    text = str(text).lower()
    text = re.sub('\[.*?\]', '', text)
    text = re.sub('https?://\S+|www\.\S+', '', text)
    text = re.sub('<.*?>+', '', text)
    text = re.sub('[%s]' % re.escape(string.punctuation), '', text)
    text = re.sub('\n', '', text)
    text = re.sub('\w*\d\w*', '', text)
    return text
```

In [24]:

```
df_new = df.copy()
```

In [28]:

```
df_new['About'] = df_new['About'].apply(clean_text)
df_new['Course Description'] = df_new['Course Description'].apply(clean_text)
```

In [29]:

```
from PIL import Image
from wordcloud import WordCloud, STOPWORDS, ImageColorGenerator
import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
```

In [30]:

```
stop_words = stopwords.words('english')
more_stopwords = ['u', 'im', 'c']
stop_words = stop_words + more_stopwords

def remove_stopwords(text):
    text = ' '.join(word for word in text.split(' ') if word not in stop_words)
    return text

df_new['About'] = df_new['About'].apply(remove_stopwords)
df_new['Course Description'] = df_new['Course Description'].apply(remove_stopwords)
```

Out[30]:

	Name	University	Difficulty Level	About	Course Description
0	How to Learn Online	edX	Beginner	Learn essential strategies for successful onli...	Designed for those who are new to elearning, t...
1	Programming for Everybody (Getting Started wit...	The University of Michigan	Beginner	This course is a "no prerequisite" introductio...	This course aims to teach everyone the basics ...
2	CS50's Introduction to Computer Science	Harvard University	Beginner	An introduction to the intellectual enterprise...	This is CS50x , Harvard University's introduct...
3	The Analytics Edge	Massachusetts Institute of Technology	Intermediate	Through inspiring examples and stories, discov...	In the last decade, the amount of data availab...
4	Marketing Analytics: Marketing Measurement Str...	University of California, Berkeley	Beginner	This course is part of a MicroMasters® Program	Begin your journey in a new career in marketin...

In [31]:

```
df_new.head()
```

Out[31]:

	Name	University	Difficulty Level	About	Course Description
0	How to Learn Online	edX	Beginner	learn essential strategies successful online l...	designed new elearning course prepare strategi...
1	Programming for Everybody (Getting Started wit...	The University of Michigan	Beginner	course prerequisite introduction python progra...	course aims teach everyone basics programming ...
2	CS50's Introduction to Computer Science	Harvard University	Beginner	introduction intellectual enterprises computer...	harvard universitys introduction intellectua...
3	The Analytics Edge	Massachusetts Institute of Technology	Intermediate	inspiring examples stories discover power data...	last decade amount data available organization...
4	Marketing Analytics: Marketing Measurement Str...	University of California, Berkeley	Beginner	course part micromasters® program	begin journey new career marketing analytics l...

In [32]:

```
stemmer = nltk.SnowballStemmer("english")

def stemm_text(text):
    text = ' '.join(stemmer.stem(word) for word in text.split(' '))
    return text
```

In [33]:

```
df_new['About'] = df_new['About'].apply(stemm_text)
df_new['Course Description'] = df_new['Course Description']
df_new.head()
```

Out[33]:

	Name	University	Difficulty Level	About	Course Description
0	How to Learn Online	edX	Beginner	learn essenti strategi success onlin learn	designed new elearning course prepare strategi...
1	Programming for Everybody (Getting Started wit...	The University of Michigan	Beginner	cours prerequisit introduct python program lea...	course aims teach everyone basics programming ...
2	CS50's Introduction to Computer Science	Harvard University	Beginner	introduct intellectu enterpris comput scienc a...	harvard universitys introduction intellectua...
3	The Analytics Edge	Massachusetts Institute of Technology	Intermediate	inspir exampl stori discov power data use anal...	last decade amount data available organization...
4	Marketing Analytics: Marketing Measurement Str...	University of California, Berkeley	Beginner	cours part micromasters® program	begin journey new career marketing analytics l...

In [36]:

```
def ngrams_func(i,j):
    count_vectoriser = CountVectorizer(ngram_range=(i,j))
    ngrams = count_vectoriser.fit_transform(df_new["Course Description"])
    count_values = ngrams.toarray().sum(axis=0)
    vocab=count_vectoriser.vocabulary_
    return count_values,vocab
```

In [35]:

```
from sklearn.feature_extraction.text import CountVectorizer
from nltk.tokenize import word_tokenize
```

In [37]:

```
count_values,vocab=ngrams_func(3,3)
df_trigrams = pd.DataFrame(sorted([(count_values[i],k)
                                   for k,i in vocab.items()], reverse=True)).rename(colum
df_trigrams.head(7)
```

Out[37]:

	Freq	Trigrams
0	52	data analysis statistics
1	40	you ll learn
2	40	professional certificate program
3	29	biology life sciences
4	24	course you ll
5	22	end course able
6	18	education teacher training

In [38]:

```
count_values,vocab= ngrams_func(2,2)
df_bigrams = pd.DataFrame(sorted([(count_values[i],k) for k,i in vocab.items()],reverse=
df_bigrams.head()
```

Out[38]:

	Freq	Bigrams
0	157	computer science
1	136	business management
2	122	you ll
3	112	course part
4	101	course learn

In [39]:

```
freq_of_words = pd.Series(' '.join(df_new["Course Description"]).split()).value_counts()
freq_of_words
```

Out[39]:

```
course           2058
learn             730
data              506
business          405
also              356
science           333
management        327
learning          296
skills            277
understanding     246
program           244
use               239
world             238
design            237
part              226
dtype: int64
```

In [40]:

```
from sklearn.preprocessing import LabelEncoder

le = LabelEncoder()
le.fit(df_new['Difficulty Level'])

df_new['Difficulty Level'] = le.transform(df_new['Difficulty Level'])
df_new.head()
```

Out[40]:

	Name	University	Difficulty Level	About	Course Description
0	How to Learn Online	edX	1	learn essenti strategi success onlin learn	designed new elearning course prepare strategi...
1	Programming for Everybody (Getting Started wit...	The University of Michigan	1	cours prerequisit introduct python program lea...	course aims teach everyone basics programming ...
2	CS50's Introduction to Computer Science	Harvard University	1	introduct intellectu enterpris comput scienc a...	harvard universitys introduction intellectua...
3	The Analytics Edge	Massachusetts Institute of Technology	2	inspir exampl stori discov power data use anal...	last decade amount data available organization...
4	Marketing Analytics: Marketing Measurement Str...	University of California, Berkeley	1	cours part micromasters® program	begin journey new career marketing analytics l...

In [41]:

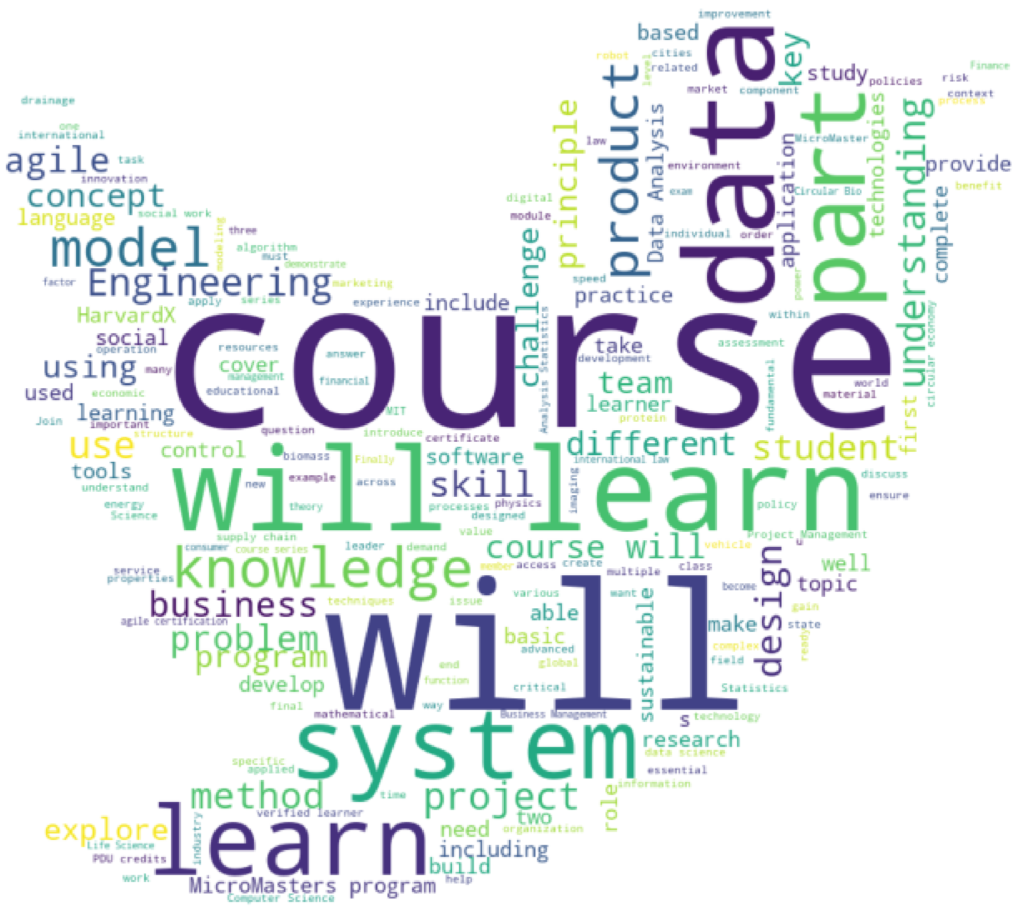
```
import numpy as np
```

In [42]:

```
twitter_mask = np.array(Image.open('twitter_mask.png'))

wc = WordCloud(
    background_color='white',
    max_words=200,
    mask=twitter_mask,
)
wc.generate(' '.join(text for text in df.loc[df_new['Difficulty Level'] == 0, 'Course De
plt.figure(figsize=(18,10))
plt.title('Top words for Course Description - Beginner',
          fontdict={'size': 22,  'verticalalignment': 'bottom'})
plt.imshow(wc)
plt.axis("off")
plt.show()
```

Top words for Course Description - Beginner

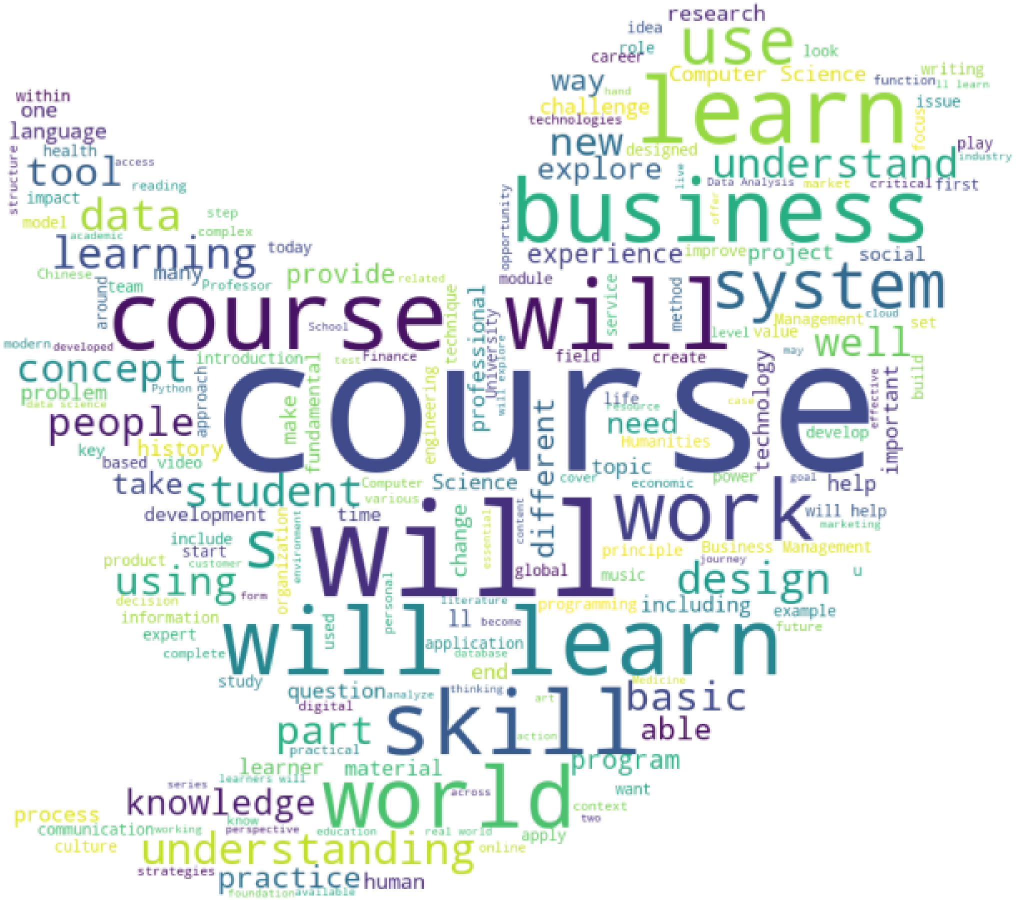


In [43]:

```
twitter_mask = np.array(Image.open('twitter_mask.png'))

wc = WordCloud(
    background_color='white',
    max_words=200,
    mask=twitter_mask,
)
wc.generate(' '.join(text for text in df.loc[df_new['Difficulty Level'] == 1, 'Course De
plt.figure(figsize=(18,10))
plt.title('Top words for Course Description - Intermediate',
          fontdict={'size': 22,  'verticalalignment': 'bottom'})
plt.imshow(wc)
plt.axis("off")
plt.show()
```

Top words for Course Description - Intermediate

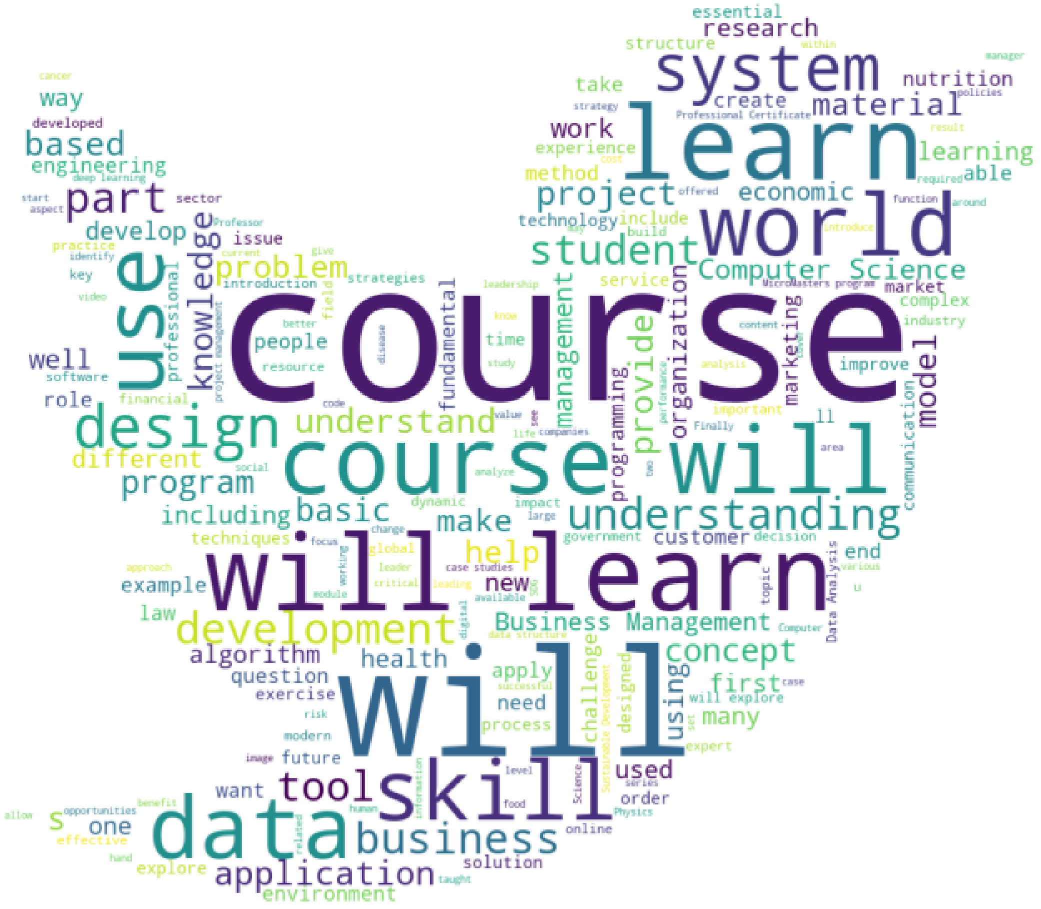


In [44]:

```
twitter_mask = np.array(Image.open('twitter_mask.png'))

wc = WordCloud(
    background_color='white',
    max_words=200,
    mask=twitter_mask,
)
wc.generate(' '.join(text for text in df.loc[df_new['Difficulty Level'] == 2, 'Course De
plt.figure(figsize=(18,10))
plt.title('Top words for Course Description - Advanced',
          fontdict={'size': 22,  'verticalalignment': 'bottom'})
plt.imshow(wc)
plt.axis("off")
plt.show()
```

Top words for Course Description - Advanced


course
will learn
course will
learn
will
world
system
data
skill
use
design
understand
understanding
development
student
knowledge
project
business
tool
application
problem
Computer Science
management
programming
provide
organization
marketing
model
part
based
research
material
nutrition
economic
learning
work
people
help
make
basic
program
including
different
Business Management
concept
algorithm
health
first
using
challenge
fundamental
professional
communication
essential
structure
take
create
able
method
experience
technology
include
develop
issue
engineering
way
well
role
improve
complex
industry
market
customer
example
law
question
exercise
need
apply
many
used
order
solution
environment
explore
one
future
want
modern
process
time
new
end
Data Analysis