# 1. Getting the Data

In [1]:

```
%pip install requests

import requests

url = "https://en.wikipedia.org/wiki/Data_science"
text = requests.get(url).content.decode("utf-8")
print(text[:1000])
```

```
Requirement already satisfied: requests in d:\projects\mlprojects\applied-python-training
\.venv\lib\site-packages (2.32.3)
Requirement already satisfied: charset-normalizer<4,>=2 in d:\projects\mlprojects\applied
-python-training\.venv\lib\site-packages (from requests) (3.3.2)
Requirement already satisfied: idna<4,>=2.5 in d:\projects\mlprojects\applied-python-trai
ning\.venv\lib\site-packages (from requests) (3.7)
Requirement already satisfied: urllib3<3,>=1.21.1 in d:\projects\mlprojects\applied-pytho
n-training\.venv\lib\site-packages (from requests) (2.2.2)
Requirement already satisfied: certifi>=2017.4.17 in d:\projects\mlprojects\applied-pytho
n-training\.venv\lib\site-packages (from requests) (2024.7.4)
Note: you may need to restart the kernel to use updated packages.
<!DOCTYPE html>
<html class="client-nojs vector-feature-language-in-header-enabled vector-feature-languag
e-in-main-page-header-disabled vector-feature-sticky-header-disabled vector-feature-page-
tools-pinned-disabled vector-feature-toc-pinned-clientpref-1 vector-feature-main-menu-pin
ned-disabled vector-feature-limited-width-clientpref-1 vector-feature-limited-width-conte
nt-enabled vector-feature-custom-font-size-clientpref-1 vector-feature-appearance-enabled
vector-feature-appearance-pinned-clientpref-1 vector-feature-night-mode-enabled skin-them
e-clientpref-day vector-toc-available" lang="en" dir="ltr">
<head>
<meta charset="UTF-8">
<title>Data science - Wikipedia</title>
<script>(function(){var className="client-js vector-feature-language-in-header-enabled ve
ctor-feature-language-in-main-page-header-disabled vector-feature-sticky-header-disabled
vector-feature-page-tools-pinned-disabled vector-feature-toc-pinned-clientpref-1 vector-f
eature-main-menu-pinned-disabled vector-feature-limi
```

# 2. Transforming the Data

In [3]:

```
from html.parser import HTMLParser

class MyHTMLParser(HTMLParser):
    script = False
    res = ""

    def handle_starttag(self, tag, attrs):
        if tag.lower() in ["script", "style"]:
            self.script = True

    def handle_endtag(self, tag):
        if tag.lower() in ["script", "style"]:
            self.script = False

    def handle_data(self, data):
        if str.strip(data) == "" or self.script:
            return
        self.res += " " + data.replace("[ edit ]", "")

parser = MyHTMLParser()
parser.feed(text)
```

```
text = parser.res
print(text[:1000])
```

```
 Data science - Wikipedia Jump to content Main menu Main menu move to sidebar hide
  Navigation
  Main page Contents Current events Random article About Wikipedia Contact us Donate
  Contribute
  Help Learn to edit Community portal Recent changes Upload file Search Search Appearance
Create account Log in Personal tools Create account Log in
  Pages for logged out editors  learn more Contributions Talk Contents move to sidebar hi
de (Top) 1 Foundations Toggle Foundations subsection 1.1 Relationship to statistics 2 Ety
mology Toggle Etymology subsection 2.1 Early usage 2.2 Modern usage 3 Data science and da
ta analysis 4 Cloud computing for data science 5 Ethical consideration in data science 6 
See also 7 References Toggle the table of contents Data science 48 languages العربية Azər
baycanca বাংলা Български Català Čeština Deutsch Eesti Ελληνικά Español Esperanto Euskara ف
ارسی Français Galego 한국어 Հայերեն हिन्दी Bahasa Indonesia IsiZulu Italiano עברית ಕನ್ನಡ Қазақ
ша Latviešu Македонски Bahas
```

# 3. Extracting Keywords

In [4]:

```
%pip install nlp_rake

import nlp_rake

extractor = nlp_rake.Rake(max_words=2, min_freq=3, min_chars=5)
res = extractor.apply(text)
print(res)
```

```
Requirement already satisfied: nlp_rake in d:\projects\mlprojects\applied-python-training
\.venv\lib\site-packages (0.0.2)
Requirement already satisfied: langdetect>=1.0.8 in d:\projects\mlprojects\applied-python
-training\.venv\lib\site-packages (from nlp_rake) (1.0.9)
Requirement already satisfied: numpy>=1.14.4 in d:\projects\mlprojects\applied-python-tra
ining\.venv\lib\site-packages (from nlp_rake) (2.1.0)
Requirement already satisfied: pyrsistent>=0.14.2 in d:\projects\mlprojects\applied-pytho
n-training\.venv\lib\site-packages (from nlp_rake) (0.20.0)
Requirement already satisfied: regex>=2018.6.6 in d:\projects\mlprojects\applied-python-t
raining\.venv\lib\site-packages (from nlp_rake) (2024.7.24)
Requirement already satisfied: six in d:\projects\mlprojects\applied-python-training\.ven
v\lib\site-packages (from langdetect>=1.0.8->nlp_rake) (1.16.0)
Note: you may need to restart the kernel to use updated packages.
[('data scientist', 4.0), ('data visualization', 4.0), ('machine learning', 4.0), ('data 
mining', 4.0), ('sexiest job', 4.0), ('21st century', 4.0), ('big data', 4.0), ('data sci
entists', 4.0), ('data science', 3.901408450704225), ('computer science', 3.9014084507042
25), ('statistical learning', 3.9), ('information science', 3.8244853737811484), ('^ dave
nport', 3.8), ('cloud computing', 3.75), ('data analysis', 3.7058823529411766), ('extract
 insights', 3.5277777777777777), ('science', 1.9014084507042253), ('analysis', 1.705882352
9411764), ('field', 1.4285714285714286), ('computational', 1.4), ('process', 1.25), ('sta
tistics', 1.2173913043478262), ('thomas', 1.2), ('mathematics', 1.0), ('education', 1.0),
('communications', 1.0), ('archived', 1.0), ('original', 1.0), ('chikio', 1.0), ('forbes'
, 1.0)]
```

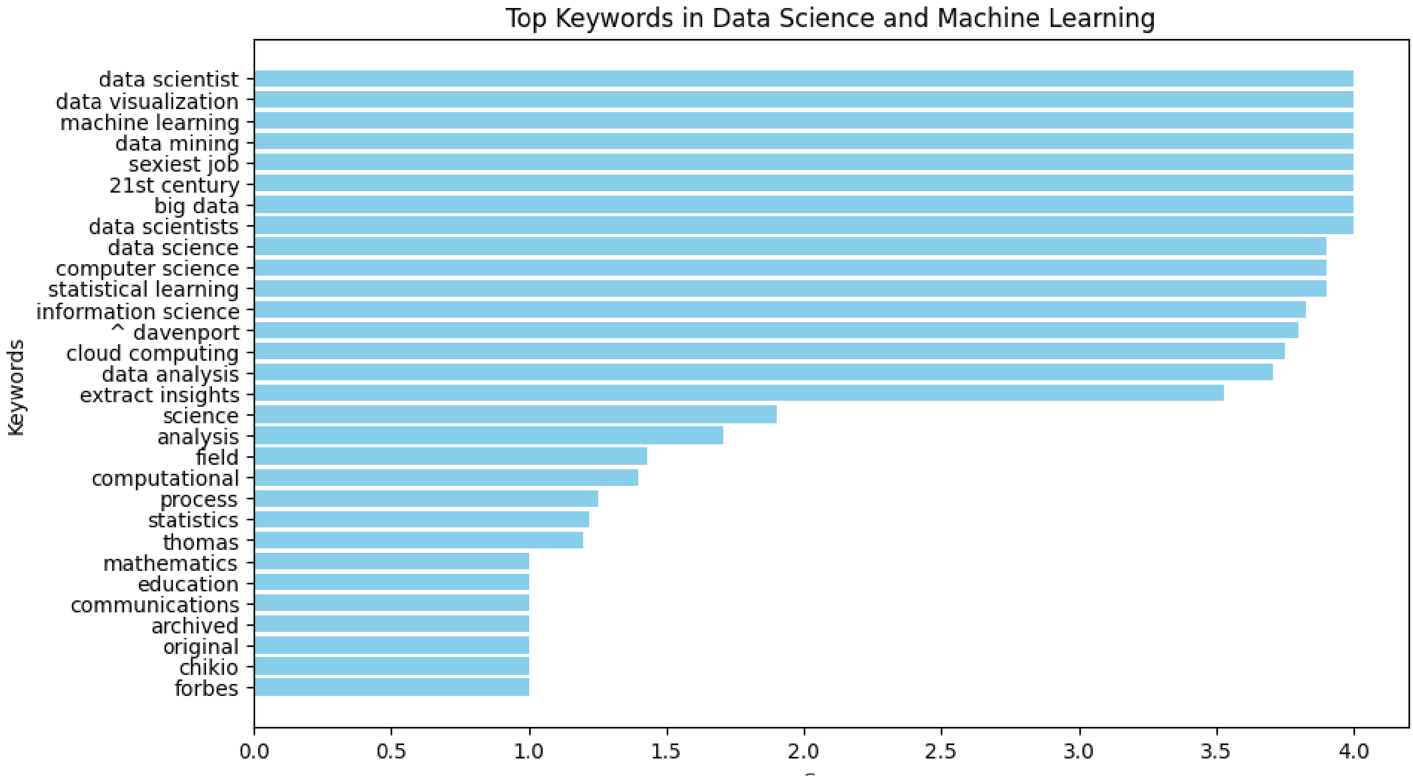
# 4. Visualizing

In [9]:

```
%pip install matplotlib wordcloud

import matplotlib.pyplot as plt
from wordcloud import WordCloud

# Extracting keywords and their scores
keywords, scores = zip(*res[:100])

# Plotting the results
plt.figure(figsize=(10, 6))
```

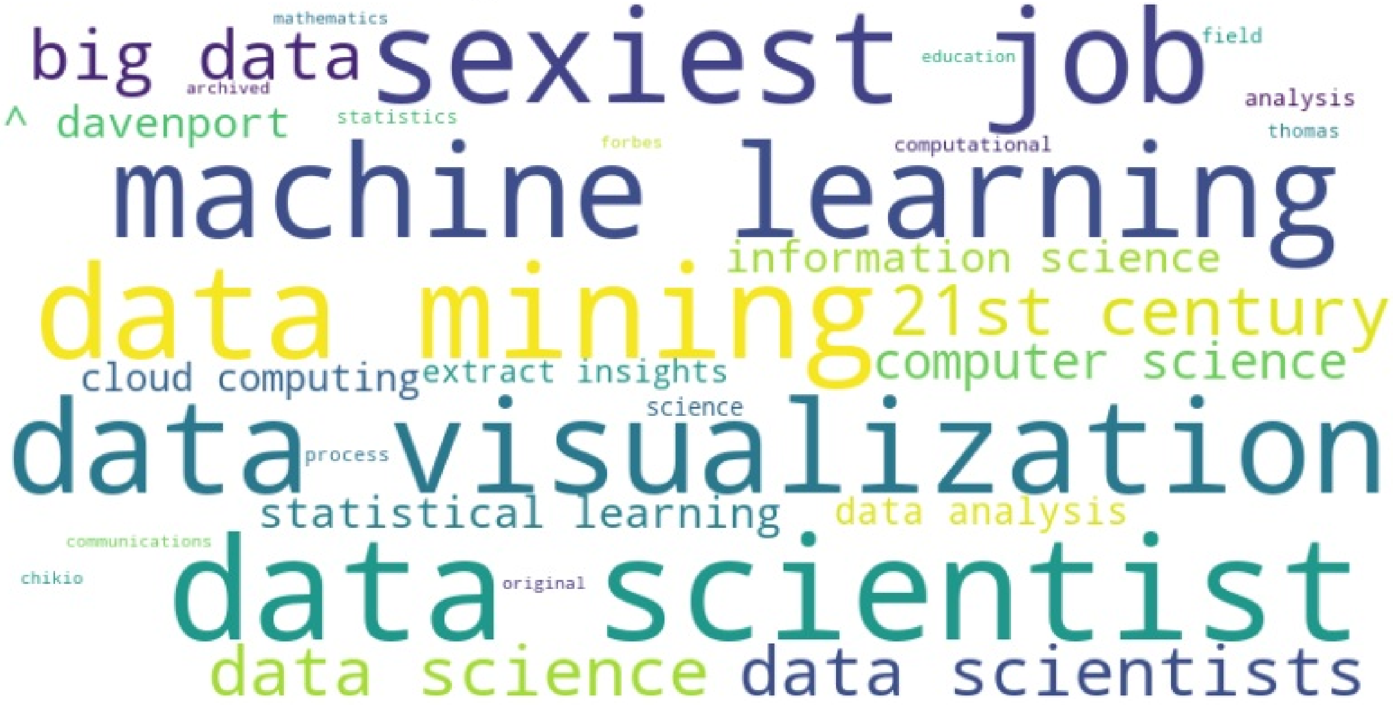
```python
plt.barh(keywords, scores, color='skyblue')
plt.xlabel('Score')
plt.ylabel('Keywords')
plt.title('Top Keywords in Data Science and Machine Learning')
plt.gca().invert_yaxis()
plt.show()

# Creating a dictionary for the WordCloud
wordcloud_dict = dict(res)

# Generating the WordCloud
wordcloud = WordCloud(width=800, height=400, background_color='white').generate_from_frequencies(wordcloud_dict)

# Plotting the WordCloud
plt.figure(figsize=(10, 6))
plt.imshow(wordcloud, interpolation='bilinear')
plt.axis('off')
plt.title('WordCloud of Keywords in Data Science and Machine Learning')
plt.show()
```

```
Requirement already satisfied: matplotlib in d:\projects\mlprojects\applied-python-training\.venv\lib\site-packages (3.9.2)
Requirement already satisfied: wordcloud in d:\projects\mlprojects\applied-python-training\.venv\lib\site-packages (1.9.3)
Requirement already satisfied: contourpy>=1.0.1 in d:\projects\mlprojects\applied-python-training\.venv\lib\site-packages (from matplotlib) (1.2.1)
Requirement already satisfied: cycler>=0.10 in d:\projects\mlprojects\applied-python-training\.venv\lib\site-packages (from matplotlib) (0.12.1)
Requirement already satisfied: fonttools>=4.22.0 in d:\projects\mlprojects\applied-python-training\.venv\lib\site-packages (from matplotlib) (4.53.1)
Requirement already satisfied: kiwisolver>=1.3.1 in d:\projects\mlprojects\applied-python-training\.venv\lib\site-packages (from matplotlib) (1.4.5)
Requirement already satisfied: numpy>=1.23 in d:\projects\mlprojects\applied-python-training\.venv\lib\site-packages (from matplotlib) (2.1.0)
Requirement already satisfied: packaging>=20.0 in d:\projects\mlprojects\applied-python-training\.venv\lib\site-packages (from matplotlib) (24.1)
Requirement already satisfied: pillow>=8 in d:\projects\mlprojects\applied-python-training\.venv\lib\site-packages (from matplotlib) (10.4.0)
Requirement already satisfied: pyparsing>=2.3.1 in d:\projects\mlprojects\applied-python-training\.venv\lib\site-packages (from matplotlib) (3.1.2)
Requirement already satisfied: python-dateutil>=2.7 in d:\projects\mlprojects\applied-python-training\.venv\lib\site-packages (from matplotlib) (2.9.0.post0)
Requirement already satisfied: six>=1.5 in d:\projects\mlprojects\applied-python-training\.venv\lib\site-packages (from python-dateutil>=2.7->matplotlib) (1.16.0)
Note: you may need to restart the kernel to use updated packages.
```



Score

WordCloud of Keywords in Data Science and Machine Learning


mathematics
sexiest job
big data
field
education
archived
analysis
^ davenport
statistics
forbes
computational
thomas
machine learning
information science
data mining
21st century
computer science
cloud computing
extract insights
science
data visualization
process
statistical learning
data analysis
communications
chikio
original
data scientist
data science
data scientists