In [1]:
```
!pip install PyPDF2
```

```
Collecting PyPDF2
  Obtaining dependency information for PyPDF2 from https://files.pythonhosted.org/packages/8e/5e/c86a5643653825d3c913719e788e41386bee415c2b87b4f955432f2de6b2/pypdf2-3.0.1-py3-none-any.whl.metadata (https://files.pythonhosted.org/packages/8e/5e/c86a5643653825d3c913719e788e41386bee415c2b87b4f955432f2de6b2/pypdf2-3.0.1-py3-none-any.whl.metadata)
  Downloading pypdf2-3.0.1-py3-none-any.whl.metadata (6.8 kB)
Downloading pypdf2-3.0.1-py3-none-any.whl (232 kB)
   ---------------------------------------- 0.0/232.6 kB ? eta -:--:--
   ---------------------------------------- 232.6/232.6 kB 4.7 MB/s eta 0:00:00
Installing collected packages: PyPDF2
Successfully installed PyPDF2-3.0.1
```

In [2]:
```
#!pip uninstall PyPDF2
```

In [3]:
```
#!pip install PyPDF2==3.0.1
```

In [4]:
```
import PyPDF2
from PyPDF2 import PdfFileReader
```

In [5]:
```
PyPDF2.__version__
```

Out[5]:
```
'3.0.1'
```

In [15]:
```
#Creating a pdf file object
pdf = open("file1pdf.pdf","rb")

#creating pdf reader object
pdf_reader = PyPDF2.PdfReader(pdf)

#checking number of pages in a pdf file
print("Number of pages:", len(pdf_reader.pages))

#creating a page object
page = pdf_reader.pages[1]

#finally extracting text from the page
print(page.extract_text())

#closing the pdf file
pdf.close()
```

```
Number of pages: 35
```

```
 Development  Plan for Greater Mumbai 2014-2034



Acknowledgements
The Consultant  wishes to thank the following  individuals  from the Municipal  Corporati
on  of
Greater Mumbai for their invaluable  support, insights and contributions  towards ‘Workin
g  Paper 1
– Preparation  of Base Map’ for the preparation  of the Development  Plan for Greater Mum
bai
2014-34.
 Mr. Subodh Kumar, IAS, Municipal  Commissioner;
 Mr. Rajeev Kuknoor, Chief Engineer Development  Plan;
 Mr. Sudhir Ghate, Deputy Chief Engineer Development  Plan;
 Mr. A.G. Marathe, Deputy Chief Engineer Development  Plan;
 Mr. R. Balachandran,  Executive  Engineer and Town Planning Officer, Development  Pla
n.
 Our gratitude  to the following  experts for their invaluable  insights and support:

Mr. V.K Phatak, Former Chief Town Planner (MMRDA);
 Mr. A.N Kale, Former Chief Engineer, (DP);
 Mr. A. S Jain Former Dy. Chief Engineer, (DP).
 We wish to especially  thank MCGM officers, Mr. Jagdish Talreja, Mr. Dinesh Naik, Mr. Hi
ren
Daftardar,  Ms. Anita Naik for their continual  support since the
 beginning  of the project and their
help towards familiarization  and data collection.  They have been instrumental  in helpi
ng to
contact various MCGM departments  as well as in helping to establish contact with personn
el  from
other government  departments  and organizations.  Many thanks for the MCGM team, for
deploying  personnel,  particularly  Mr. Prasad Gharat, on extensive  field visits that h
ave helped in
understanding  actual ground conditions.

We apologize  if we have inadvertently  omitted anyone to whom acknowledgement  is due. W
e hope
and anticipate  the work's usefulness  for the intended purpose.
```

In [21]:
```
import PyPDF2, urllib , nltk
from io import BytesIO
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
```

In [22]:
```
#nltk.download('stopwords')
```

In [23]:
```
wFile = urllib.request.urlopen('http://www.udri.org/pdf/02%20working%20paper%201.pdf')
pdfreader = PyPDF2.PdfReader(BytesIO(wFile.read()))
```

In [26]:
```
pageObj = pdfreader.pages[2]
page2 = pageObj.extract_text()
punctuations = ['(',')',';',':','[',']',',','...','.']
tokens = word_tokenize(page2)
stop_words = stopwords.words('english')
keywords = [word for word in tokens if not word in stop_words and not word in punctuations
```

In [27]:
```
keywords
```

```
 'insights',
 'contributions',
 'towards',
 '‘',
 'Working',
 'Paper',
 '1',
 '–',
 'Preparation',
 'Base',
 'Map',
 '’',
 'preparation',
 'Development',
 'Plan',
 'Greater',
 'Mumbai',
 '2014-34',
```

In [28]:
```
name_list = list()
check =  ['Mr.',  'Mrs.',  'Ms.']
for idx, token in enumerate(tokens):
    if token.startswith(tuple(check)) and idx < len(tokens)-1:
        name = token + tokens[idx+1] + ' ' + tokens[idx+2]
        name_list.append(name)
print(name_list)
```

```
['Mr.Jagdish Talreja', 'Mr.Dinesh Naik', 'Mr.Hiren Daftardar', 'Ms.Anita Naik', 'Mr.Prasa
d Gharat']
```

In [29]:
```
wFile.close()
```

In [30]:
```
!pip install python-docx
```

```
Collecting python-docx
  Obtaining dependency information for python-docx from https://files.pythonhosted.org/packages/3e/3d/330d9efbdb816d3f60bf2ad92f05e1708e4a1b9abe80461ac3444c83f749/python_docx-1.1.2-py3-none-any.whl.metadata (https://files.pythonhosted.org/packages/3e/3d/330d9efbdb816d3f60bf2ad92f05e1708e4a1b9abe80461ac3444c83f749/python_docx-1.1.2-py3-none-any.whl.metadata)
  Downloading python_docx-1.1.2-py3-none-any.whl.metadata (2.0 kB)
Requirement already satisfied: lxml>=3.1.0 in c:\users\pooja reddy\anaconda3\lib\site-packages (from python-docx) (4.9.2)
Requirement already satisfied: typing-extensions>=4.9.0 in c:\users\pooja reddy\anaconda3\lib\site-packages (from python-docx) (4.12.2)
Downloading python_docx-1.1.2-py3-none-any.whl (244 kB)
   ---------------------------------------- 0.0/244.3 kB ? eta -:--:--
   -------------------------------- -------- 194.6/244.3 kB 3.9 MB/s eta 0:00:01
   ---------------------------------------- 244.3/244.3 kB 3.8 MB/s eta 0:00:00
Installing collected packages: python-docx
Successfully installed python-docx-1.1.2
```

In [31]:
```
import docx
```

In [38]:
```
doc = open("Task-1-Answers.docx","rb")
document = docx.Document(doc)
```

In [ ]:
```

```