In [11]:
```
!pip install pyPDF2
```

```
Requirement already satisfied: pyPDF2 in c:\users\nihar\anaconda3\lib\site-packages (3.0.1)
Requirement already satisfied: typing_extensions>=3.10.0.0 in c:\users\nihar\anaconda3\lib\si
te-packages (from pyPDF2) (4.8.0)
```

In [12]:
```
#!pip uninstall pyPDF2
```

In [13]:
```
#pip install pyPDF2=3.0.1
```

In [17]:
```
import PyPDF2
from PyPDF2 import PdfFileReader
```

In [18]:
```
PyPDF2.__version__
```

Out[18]: '3.0.1'

In [19]:
```
#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  Corporation
 of
Greater Mumbai for their invaluable  support, insights and contributions  towards ‘Working  P
aper 1
– Preparation  of Base Map’ for the preparation  of the Development  Plan for Greater Mumbai
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  Plan.
 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. Hiren
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 helping t
o
contact various MCGM departments  as well as in helping to establish contact with personnel
 from
other government  departments  and organizations.  Many thanks for the MCGM team, for
deploying  personnel,  particularly  Mr. Prasad Gharat, on extensive  field visits that hav
e helped in
understanding  actual ground conditions.

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

In [20]:
```
import PyPDF2, urllib , nltk
from io import BytesIO
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
```

In [21]:
```
1 nltk.download('stopwords')
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\nihar\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

Out[21]: True

In [22]:
```
1 wFile = urllib.request.urlopen('http://www.udri.org/pdf/02%20working%20paper%201.pdf')
2 pdfreader = PyPDF2.PdfReader(BytesIO(wFile.read()))
```

In [23]:
```
1 pageObj = pdfreader.pages[2]
2 page2 = pageObj.extract_text()
3 punctuations = ['(',')',';',':','[',']',',','...','.']
4 tokens = word_tokenize(page2)
5 stop_words = stopwords.words('english')
6 keywords = [word for word in tokens if not word in stop_words and not word in punctuations
```

In [24]:
```
1 keywords
```

Out[24]:
```
['Development',
 'Plan',
 'Greater',
 'Mumbai',
 '2014-2034',
 'Table',
 'Contents',
 'The',
 'Consultant',
 'wishes',
 'thank',
 'following',
 'individuals',
 'Municipal',
 'Corporation',
 'Greater',
 'Mumbai',
 'invaluable',
```

In [25]:
```
1 name_list = list()
2 check =  ['Mr.',  'Mrs.',  'Ms.']
3 for idx, token in enumerate(tokens):
4     if token.startswith(tuple(check)) and idx < len(tokens)-1:
5         name = token + tokens[idx+1] + ' ' + tokens[idx+2]
6         name_list.append(name)
7 print(name_list)
```

```
['Mr.Jagdish Talreja', 'Mr.Dinesh Naik', 'Mr.Hiren Daftardar', 'Ms.Anita Naik', 'Mr.Prasad Gh
arat']
```

In [26]:
```
1 wFile.close()
```

In [27]:
```
1 !pip install python_docx
```

```
Requirement already satisfied: python_docx in c:\users\nihar\anaconda3\lib\site-packages (1.
1.2)
Requirement already satisfied: lxml>=3.1.0 in c:\users\nihar\anaconda3\lib\site-packages (fro
m python_docx) (4.9.1)
Requirement already satisfied: typing-extensions>=4.9.0 in c:\users\nihar\anaconda3\lib\site-
packages (from python_docx) (4.12.2)
```

In [34]:
```
import docx
```

In [36]:
```
doc = open("Task-1-Answers.docx","rb")
document = docx.Document(doc)
```

In [37]:
```
docu=""
for para in document.paragraphs:
    docu+=para.text
print(docu)
```

In [38]:
```
!pip install bs4
```

```
Collecting bs4
  Downloading bs4-0.0.2-py2.py3-none-any.whl (1.2 kB)
Requirement already satisfied: beautifulsoup4 in c:\users\nihar\anaconda3\lib\site-packages
(from bs4) (4.11.1)
Requirement already satisfied: soupsieve>1.2 in c:\users\nihar\anaconda3\lib\site-packages (f
rom beautifulsoup4->bs4) (2.3.1)
Installing collected packages: bs4
Successfully installed bs4-0.0.2
```

In [39]:
```
import urllib.request as urllib2
from bs4 import BeautifulSoup
```

In [40]:
```
response = urllib2.urlopen('https://en.wikipedia.org/wiki/Natural_language_processing')
html_doc = response.read()
```

In [41]:
```
soup = BeautifulSoup(html_doc, 'html.parser')
strhtm = soup.prettify()
print (strhtm[:5000])
```

```
<!DOCTYPE html>
<html class="client-nojs vector-feature-language-in-header-enabled vector-feature-language-in
-main-page-header-disabled vector-feature-page-tools-pinned-disabled vector-feature-toc-pinne
d-clientpref-1 vector-feature-main-menu-pinned-disabled vector-feature-limited-width-clientpr
ef-1 vector-feature-limited-width-content-enabled vector-feature-custom-font-size-clientpref-
1 vector-feature-appearance-pinned-clientpref-1 vector-feature-night-mode-enabled skin-theme-
clientpref-day vector-sticky-header-enabled vector-toc-available" dir="ltr" lang="en">
 <head>
  <meta charset="utf-8"/>
  <title>
   Natural language processing - Wikipedia
  </title>
  <script>
   (function(){var className="client-js vector-feature-language-in-header-enabled vector-feat
ure-language-in-main-page-header-disabled vector-feature-page-tools-pinned-disabled vector-fe
ature-toc-pinned-clientpref-1 vector-feature-main-menu-pinned-disabled vector-feature-limited
-width-clientpref-1 vector-feature-limited-width-content-enabled vector-feature-custom-font-s
ize-clientpref-1 vector-feature-appearance-pinned-clientpref-1 vector-feature-night-mode-enab
led skin-theme-clientpref-day vector-sticky-header-enabled vector-toc-available";var cookie=d
ocument.cookie.match(/(?:^|; )enwikimwclientpreferences=([^;]+)/);if(cookie){cookie[1].split
('%2C').forEach(function(pref){className=className.replace(new RegExp('(^| )'+pref.replace(/-
clientpref-\w+$|[^\w-]+/g,'')+'-clientpref-\\w+( |$)'),'$1'+pref+'$2');});}document.documentE
lement.className=className;}());RLCONF={"wgBreakFrames":false,"wgSeparatorTransformTable":
["",""],"wgDigitTransformTable":["",""],"wgDefaultDateFormat":"dmy",
"wgMonthNames":["","January","February","March","April","May","June","July","August","Septemb
er","October","November","December"],"wgRequestId":"bb8cccbb-931f-4281-aaf8-3aff8c9effb8","wg
CanonicalNamespace":"","wgCanonicalSpecialPageName":false,"wgNamespaceNumber":0,"wgPageNam
e":"Natural_language_processing","wgTitle":"Natural language processing","wgCurRevisionId":12
74942014,"wgRevisionId":1274942014,"wgArticleId":21652,"wgIsArticle":true,"wgIsRedirect":fals
e,"wgAction":"view","wgUserName":null,"wgUserGroups":["*"],"wgCategories":["All accuracy disp
utes","Accuracy disputes from December 2013","Harv and Sfn no-target errors","CS1 errors: per
iodical ignored","CS1 maint: location","Articles with short description","Short description i
s different from Wikidata","Articles needing additional references from May 2024","All articl
es needing additional references","All articles with unsourced statements","Articles with uns
ourced statements from May 2024","Commons category link from Wikidata",
"Natural language processing","Computational fields of study","Computational linguistics","Sp
eech recognition"],"wgPageViewLanguage":"en","wgPageContentLanguage":"en","wgPageContentMode
l":"wikitext","wgRelevantPageName":"Natural_language_processing","wgRelevantArticleId":2165
2,"wgIsProbablyEditable":true,"wgRelevantPageIsProbablyEditable":true,"wgRestrictionEdit":
[],"wgRestrictionMove":[],"wgNoticeProject":"wikipedia","wgCiteReferencePreviewsActive":fals
e,"wgFlaggedRevsParams":{"tags":{"status":{"levels":1}}},"wgMediaViewerOnClick":true,"wgMedia
ViewerEnabledByDefault":true,"wgPopupsFlags":0,"wgVisualEditor":{"pageLanguageCode":"en","pag
eLanguageDir":"ltr","pageVariantFallbacks":"en"},"wgMFDisplayWikibaseDescriptions":{"search":
true,"watchlist":true,"tagline":false,"nearby":true},"wgWMESchemaEditAttemptStepOversample":f
alse,"wgWMEPageLength":60000,"wgEditSubmitButtonLabelPublish":true,"wgULSPosition":"interlang
uage","wgULSisCompactLinksEnabled":false,"wgVector2022LanguageInHeader":true,
"wgULSisLanguageSelectorEmpty":false,"wgWikibaseItemId":"Q30642","wgCheckUserClientHintsHeade
rsJsApi":["brands","architecture","bitness","fullVersionList","mobile","model","platform","pl
atformVersion"],"GEHomepageSuggestedEditsEnableTopics":true,"wgGETopicsMatchModeEnabled":fals
e,"wgGEStructuredTaskRejectionReasonTextInputEnabled":false,"wgGELevelingUpEnabledForUser":fa
lse};RLSTATE={"ext.globalCssJs.user.styles":"ready","site.styles":"ready","user.styles":"read
y","ext.globalCssJs.user":"ready","user":"ready","user.options":"loading","ext.cite.style
s":"ready","ext.math.styles":"ready","skins.vector.search.codex.styles":"ready","skins.vecto
r.styles":"ready","skins.vector.icons":"ready","jquery.makeCollapsible.styles":"ready","ext.w
ikimediamessages.styles":"ready","ext.visualEditor.desktopArticleTarget.noscript":"ready","ex
t.uls.interlanguage":"ready","wikibase.client.init":"ready","ext.wikimediaBadges":"ready"};RL
PAGEMODULES=["ext.cite.ux-enhancements","ext.scribunto.logs","site",
"mediawiki.page.ready","jquery.makeCollapsible","mediawiki.toc","skins.vector.js","ext.centra
lNotice.geoIP","ext.centralNotice.startUp","ext.gadget.ReferenceTooltips","ext.gadget.switche
r","ext.urlShortener.toolbar","ext.centralauth.centralautologin","mmv.bootstrap","ext.popup
s","ext.visualEditor.desktopArticleTarget.init","ext.visu
```

In [ ]:
```
```