Assessment_1

January 8, 2020

WEB MINING ASSIGNMENT 1

NAME: ACHYUT TRIPATHI REGNO: 17BCE0954 Slot: L45+L46

Ques1: Write a program to extract the source content (excluding any tags) from the website (https://en.wikipedia.org/wiki/Web_mining). Display the number of terms and term frequency of each term present in them after applying stop word removal. Also, apply stemming and lemmatization to the same document and display the number of terms along with their corresponding stemmed as well as lemmatized words present in them. Count the total number of stemmed and lemmatized words.

[2]:
```python
from bs4 import BeautifulSoup
import requests
```

[3]:
```python
url = requests.get("https://en.wikipedia.org/wiki/Web_mining")
```

[4]:
```python
soup = BeautifulSoup(url.text)
```

[5]:
```python
for script in soup(["script", "style"]):
    script.decompose()
```

[6]:
```python
text = soup.get_text()
```

[7]:
```python
lines = (line.strip() for line in text.splitlines())
# break multi-headlines into a line each
chunks = (phrase.strip() for line in lines for phrase in line.split("  "))
# drop blank lines
text = '\n'.join(chunk for chunk in chunks if chunk)
```

[8]:
```python
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
```

[9]:
```python
stop_words = set(stopwords.words('english'))
word_tokens = word_tokenize(text)
```

[10]:
```python
filt_text = [w for w in word_tokens if not w in stop_words]
```

[11]:
```python
print("Number of Terms after removing Stop Words: ", len(filt_text))
```

```
Number of Terms after removing Stop Words:  2886
```

[12]:

```
term_count = {}
for i in filt_text:
    count = 0
    if i not in term_count.keys():
        for j in filt_text:
            if i==j:
                count+=1
        term_count[i] = count
```

[13]:

```
for i in term_count.keys():
    print("Count of ", i, ": ", term_count[i])
```

```
Count of  Web :  94
Count of  mining :  69
Count of  - :  3
Count of  Wikipedia :  7
Count of  From :  1
Count of  , :  232
Count of  free :  1
Count of  encyclopedia :  1
Count of  Jump :  2
Count of  navigation :  2
Count of  search :  3
Count of  This :  12
Count of  article :  4
Count of  may :  2
Count of  require :  2
Count of  cleanup :  4
Count of  meet :  1
Count of  's :  5
Count of  quality :  1
Count of  standards :  2
Count of  . :  146
Count of  No :  2
Count of  reason :  2
Count of  specified :  1
Count of  Please :  2
Count of  help :  3
Count of  improve :  3
Count of  ( :  33
Count of  June :  6
Count of  2009 :  2
Count of  ) :  33
Count of  Learn :  2
Count of  remove :  2
Count of  template :  2
Count of  message :  3
Count of  application :  6
Count of  data :  43
Count of  techniques :  8
Count of  discover :  4
Count of  patterns :  9
Count of  World :  6
Count of  Wide :  7
Count of  As :  3
Count of  name :  2
Count of  proposes :  1
Count of  information :  17
Count of  gathered :  1
Count of  web :  29
Count of  It :  4
Count of  makes :  3
Count of  utilization :  1
Count of  automated :  2
Count of  apparatuses :  1
Count of  reveal :  2
Count of  extricate :  1
Count of  servers :  2
Count of  web2 :  1
Count of  reports :  1
Count of  permits :  2
Count of  organizations :  1
Count of  get :  1
Count of  organized :  1
Count of  unstructured :  3
Count of  browser :  1
Count of  activities :  2
Count of  server :  5
Count of  logs :  5
Count of  website :  2
Count of  link :  2
Count of  structure :  27
Count of  page :  12
Count of  content :  15
Count of  different :  6
Count of  sources :  3
Count of  The :  21
Count of  goal :  1
Count of  generate :  2
Count of  structural :  4
Count of  summary :  1
Count of  site :  9
Count of  Technically :  1
Count of  mainly :  1
Count of  focuses :  1
Count of  inner-document :  1
Count of  tries :  2
Count of  hyperlinks :  4
Count of  inter-document :  1
Count of  level :  3
Count of  Based :  3
Count of  topology :  1
Count of  categorize :  3
Count of  pages :  5
Count of  similarity :  1
Count of  relationship :  3
Count of  sites :  2
Count of  also :  5
Count of  another :  1
Count of  direction :  1
Count of  - :  5
Count of  discovering :  1
Count of  document :  5
Count of  type :  3
Count of  used :  8
Count of  schema :  3
Count of  would :  1
Count of  good :  1
Count of  purpose :  2
Count of  make :  3
Count of  possible :  1
Count of  compare/integrate :  1
Count of  schemes :  1
Count of  facilitate :  1
Count of  introducing :  2
Count of  database :  4
Count of  accessing :  1
Count of  providing :  2
Count of  reference :  2
Count of  Contents :  2
Count of  1 :  5
Count of  types :  4
Count of  2 :  5
Count of  usage :  20
Count of  2.1 :  1
Count of  Pros :  2
Count of  2.2 :  1
Count of  Cons :  2
Count of  3 :  2
Count of  4 :  3
Count of  4.1 :  1
Count of  foreign :  2
Count of  languages :  2
Count of  4.1.1 :  1
Count of  Chinese :  4
Count of  5 :  2
Count of  See :  2
Count of  6 :  5
Count of  References :  2
Count of  7 :  3
Count of  Resources :  2
Count of  7.1 :  1
Count of  External :  2
Count of  links :  7
Count of  7.2 :  1
Count of  Books :  2
Count of  7.3 :  1
Count of  Bibliographic :  2
Count of  references :  3
Count of  [ :  26
Count of  edit :  14
Count of  ] :  26
Count of  divided :  2
Count of  three :  1
Count of  general :  2
Count of  categories :  4
Count of  objectives :  1
Count of  Comparison :  1
Count of  IR :  1
Count of  view :  6
Count of  DB :  2
Count of  View :  3
Count of  Unstructured :  1
Count of  Structured :  2
Count of  Semi-structured :  1
Count of  Link :  3
Count of  Interactivity :  1
Count of  Main :  2
Count of  Text :  2
Count of  documents :  11
Count of  Hypertext :  3
Count of  Server :  3
Count of  Browser :  1
Count of  Representation :  1
Count of  Bag :  1
Count of  words :  5
Count of  n-gram :  1
Count of  terms :  2
Count of  phrases :  1
Count of  concepts :  1
Count of  ontology :  1
Count of  Relational :  4
Count of  Edge :  1
Count of  labed :  1
Count of  graph :  5
Count of  Graph :  2
Count of  table :  1
Count of  Method :  2
Count of  Machine :  3
Count of  learning :  2
Count of  Statistical :  2
Count of  including :  2
Count of  NLP :  1
Count of  Proprietary :  2
Count of  algorithms :  5
Count of  Association :  4
Count of  rules :  3
Count of  Application :  3
Count of  Categorization :  2
Count of  Clustering :  3
Count of  Finding :  4
Count of  extract :  2
Count of  text :  4
Count of  frequent :  1
Count of  sub :  1
Count of  structures :  3
Count of  discovery :  2
Count of  Site :  2
Count of  construction :  1
Count of  Adaptation :  1
Count of  management :  3
Count of  interesting :  1
Count of  order :  2
Count of  understand :  1
Count of  better :  4
Count of  serve :  1
Count of  needs :  5
Count of  Web-based :  1
Count of  applications :  6
Count of  Usage :  8
Count of  captures :  1
Count of  identity :  1
Count of  origin :  1
Count of  users :  3
Count of  along :  1
Count of  browsing :  2
Count of  behavior :  2
Count of  classified :  1
Count of  depending :  1
Count of  kind :  1
Count of  considered :  3
Count of  : :  36
Count of  user :  10
Count of  collected :  2
Count of  Typical :  1
Count of  includes :  2
Count of  IP :  1
Count of  address :  2
Count of  access :  2
Count of  time :  2
Count of  Commercial :  1
Count of  significant :  1
Count of  features :  4
Count of  enable :  1
Count of  e-commerce :  2
Count of  built :  1
Count of  top :  1
Count of  little :  1
Count of  effort :  1
Count of  A :  3
Count of  key :  1
Count of  feature :  4
Count of  ability :  1
Count of  track :  1
Count of  various :  1
Count of  kinds :  3
Count of  business :  1
Count of  events :  3
Count of  log :  2
Count of  New :  1
Count of  defined :  3
Count of  logging :  1
Count of  turned :  1
Count of  thus :  2
Count of  generating :  1
Count of  histories :  1
Count of  specially :  1
Count of  Many :  1
Count of  end :  1
Count of  combination :  1
Count of  one :  4
Count of  applied :  3
Count of  Studies :  1
Count of  related :  1
Count of  work :  1
Count of  concerned :  1
Count of  two :  4
Count of  areas :  1
Count of  constraint-based :  1
Count of  developed :  1
Count of  software :  1
Count of  tools :  4
Count of  systems :  2
Count of  Costa :  2
Count of  Seco :  2
Count of  demonstrated :  1
Count of  semantic :  3
Count of  hyponymy :  1
Count of  relationships :  1
Count of  particular :  4
Count of  given :  2
Count of  community :  1
Count of  essentially :  2
Count of  many :  1
Count of  advantages :  1
Count of  technology :  6
Count of  attractive :  1
Count of  corporations :  1
Count of  government :  1
Count of  agencies :  2
Count of  enabled :  1
Count of  personalized :  1
Count of  marketing :  1
Count of  eventually :  1
Count of  results :  2
Count of  higher :  2
Count of  trade :  2
Count of  volumes :  1
Count of  Government :  1
Count of  using :  3
Count of  classify :  1
Count of  threats :  1
Count of  fight :  1
Count of  terrorism :  1
Count of  predicting :  1
Count of  capability :  2
Count of  benefit :  1
Count of  society :  1
Count of  identifying :  1
Count of  criminal :  1
Count of  Companies :  2
Count of  establish :  1
Count of  customer :  7
Count of  understanding :  1
Count of  reacting :  1
Count of  faster :  1
Count of  find :  3
Count of  attract :  1
Count of  retain :  2
Count of  customers :  3
Count of  ; :  17
Count of  save :  1
Count of  production :  1
Count of  costs :  1
Count of  utilizing :  1
Count of  acquired :  1
Count of  insight :  1
Count of  requirements :  1
Count of  They :  3
Count of  increase :  1
Count of  profitability :  1
Count of  target :  2
Count of  pricing :  1
Count of  based :  4
Count of  profiles :  3
Count of  created :  1
Count of  even :  1
Count of  might :  6
Count of  default :  1
Count of  competitor :  1
Count of  company :  2
Count of  try :  1
Count of  promotional :  1
Count of  offers :  1
Count of  specific :  4
Count of  reducing :  1
Count of  risk :  1
Count of  losing :  1
Count of  More :  2
Count of  benefits :  2
Count of  particularly :  1
Count of  area :  1
Count of  personalization :  3
Count of  outlined :  1
Count of  frameworks :  1
Count of  probabilistic :  1
Count of  latent :  1
Count of  analysis :  3
Count of  model :  2
Count of  offer :  1
Count of  additional :  3
Count of  pattern :  2
Count of  process :  2
Count of  provides :  1
Count of  relevant :  1
Count of  collaborative :  1
Count of  recommendation :  1
Count of  These :  3
Count of  models :  1
Count of  demonstrate :  1
Count of  problems :  1
Count of  associated :  1
Count of  traditional :  2
Count of  biases :  1
Count of  questions :  1
Count of  regarding :  1
Count of  validity :  2
Count of  since :  1
Count of  obtained :  6
Count of  subjective :  1
Count of  degrade :  1
Count of  There :  2
Count of  elements :  1
Count of  unique :  1
Count of  show :  1
Count of  include :  1
Count of  way :  1
Count of  knowledge :  4
Count of  interpreting :  1
Count of  analyzing :  1
Count of  reasoning :  1
Count of  phase :  1
Count of  create :  1
Count of  issues :  2
Count of  personal :  4
Count of  nature :  1
Count of  cause :  1
Count of  concerns :  1
Count of  criticized :  1
Count of  ethical :  2
Count of  issue :  1
Count of  involving :  1
Count of  invasion :  1
Count of  privacy :  3
Count of  Privacy :  5
Count of  lost :  1
Count of  concerning :  1
Count of  individual :  4
Count of  disseminated :  1
Count of  especially :  1
Count of  occurs :  1
Count of  without :  3
Count of  consent :  1
Count of  analyzed :  1
Count of  clustered :  1
Count of  form :  1
Count of  made :  2
Count of  anonymous :  3
Count of  clustering :  1
Count of  Thus :  1
Count of  de-individualize :  1
Count of  judging :  2
Count of  mouse :  1
Count of  clicks :  1
Count of  De-individualization :  1
Count of  tendency :  1
Count of  treating :  1
Count of  people :  1
Count of  basis :  1
Count of  group :  1
Count of  characteristics :  2
Count of  instead :  1
Count of  merits :  2
Count of  Another :  1
Count of  important :  2
Count of  concern :  1
Count of  companies :  3
Count of  collecting :  1
Count of  use :  5
Count of  totally :  1
Count of  purposes :  1
Count of  violates :  1
Count of  interests :  1
Count of  growing :  1
Count of  trend :  2
Count of  selling :  1
Count of  commodity :  1
Count of  encourages :  1
Count of  owners :  1
Count of  increased :  1
Count of  amount :  1
Count of  captured :  1
Count of  traded :  1
Count of  increasing :  1
Count of  likeliness :  1
Count of  invaded :  1
Count of  buy :  1
Count of  obliged :  1
Count of  authors :  1
Count of  release :  3
Count of  legally :  1
Count of  responsible :  1
Count of  contents :  1
Count of  inaccuracies :  1
Count of  result :  3
Count of  serious :  1
Count of  lawsuits :  1
Count of  law :  1
Count of  preventing :  1
Count of  trading :  1
Count of  Some :  1
Count of  controversial :  2
Count of  attributes :  3
Count of  like :  1
Count of  sex :  1
Count of  race :  2
Count of  religion :  2
Count of  sexual :  2
Count of  orientation :  2
Count of  individuals :  1
Count of  practices :  1
Count of  anti-discrimination :  1
Count of  legislation :  1
Count of  hard :  1
Count of  identify :  3
Count of  strong :  1
Count of  rule :  1
Count of  could :  1
Count of  denial :  1
Count of  service :  1
Count of  privilege :  1
Count of  situation :  1
Count of  avoided :  1
Count of  high :  1
Count of  maintained :  1
Count of  traced :  1
Count of  back :  1
Count of  look :  1
Count of  poses :  1
Count of  threat :  1
Count of  however :  1
Count of  inferred :  1
Count of  combining :  1
Count of  separate :  1
Count of  unscrupulous :  1
Count of  section :  2
Count of  expansion :  1
Count of  You :  1
Count of  adding :  1
Count of  2015 :  1
Count of  uses :  1
Count of  theory :  1
Count of  analyze :  1
Count of  node :  5
Count of  connection :  1
Count of  According :  1
Count of  Extracting :  1
Count of  hyperlink :  2
Count of  component :  2
Count of  connects :  1
Count of  location :  1
Count of  Mining :  20
Count of  tree-like :  1
Count of  describe :  1
Count of  HTML :  3
Count of  XML :  1
Count of  tag :  1
Count of  terminology :  1
Count of  directed :  1
Count of  representing :  1
Count of  edge :  1
Count of  degree :  2
Count of  number :  3
Count of  pointing :  2
Count of  generated :  1
Count of  Techniques :  2
Count of  PageRank :  1
Count of  algorithm :  3
Count of  Google :  1
Count of  rank :  2
Count of  Google-founder :  1
Count of  Larry :  1
Count of  Page :  1
Count of  decided :  1
Count of  extraction :  2
Count of  integration :  1
Count of  useful :  3
Count of  heterogeneity :  1
Count of  lack :  1
Count of  much :  1
Count of  ever-expanding :  1
Count of  hypertext :  1
Count of  organization :  3
Count of  indexing :  1
Count of  Internet :  2
Count of  Lycos :  1
Count of  Alta :  1
Count of  Vista :  1
Count of  WebCrawler :  1
Count of  Aliweb :  1
Count of  MetaCrawler :  1
Count of  others :  1
Count of  provide :  3
Count of  comfort :  1
Count of  generally :  1
Count of  filter :  1
Count of  interpret :  1
Count of  factors :  1
Count of  prompted :  1
Count of  researchers :  1
Count of  develop :  1
Count of  intelligent :  2
Count of  retrieval :  2
Count of  agents :  1
Count of  well :  1
Count of  extend :  1
Count of  semi-structured :  3
Count of  available :  2
Count of  agent-based :  1
Count of  approach :  1
Count of  involves :  1
Count of  development :  1
Count of  sophisticated :  1
Count of  AI :  1
Count of  act :  1
Count of  autonomously :  1
Count of  semi-autonomously :  1
Count of  behalf :  1
Count of  organize :  1
Count of  web-based :  1
Count of  differentiated :  1
Count of  points :  1
Count of  8 :  1
Count of  Information :  6
Count of  Retrieval :  1
Count of  Database :  1
Count of  9 :  1
Count of  summarized :  1
Count of  research :  1
Count of  works :  2
Count of  done :  1
Count of  shows :  1
Count of  researches :  1
Count of  bag :  1
Count of  statistics :  1
Count of  single :  2
Count of  isolation :  1
Count of  represent :  2
Count of  take :  1
Count of  word :  2
Count of  found :  1
Count of  training :  1
Count of  corpus :  1
Count of  For :  1
Count of  utilize :  1
Count of  inside :  1
Count of  utilized :  1
Count of  representation :  2
Count of  querying :  1
Count of  always :  1
Count of  infer :  1
Count of  transform :  2
Count of  become :  1
Count of  several :  1
Count of  ways :  1
Count of  vector :  2
Count of  space :  2
Count of  typically :  1
Count of  constitute :  1
Count of  whole :  1
Count of  realize :  1
Count of  importance :  1
Count of  To :  1
Count of  resolve :  1
Count of  tf-idf :  1
Count of  Term :  1
Count of  Frequency :  2
Count of  Times :  1
Count of  Inverse :  1
Count of  Document :  1
Count of  introduced :  1
Count of  By :  2
Count of  multi-scanning :  1
Count of  implement :  1
Count of  selection :  1
Count of  Under :  1
Count of  condition :  1
Count of  category :  1
Count of  rarely :  1
Count of  affected :  1
Count of  subset :  1
Count of  needed :  1
Count of  construct :  1
Count of  evaluating :  1
Count of  function :  1
Count of  evaluate :  1
Count of  set :  1
Count of  gain :  1
Count of  cross :  1
Count of  entropy :  1
Count of  mutual :  1
Count of  odds :  1
Count of  ratio :  1
Count of  usually :  1
Count of  classifier :  1
Count of  methods :  1
Count of  similar :  1
Count of  usual :  1
Count of  evaluative :  1
Count of  classification :  1
Count of  accuracy :  1
Count of  precision :  1
Count of  recall :  1
Count of  score :  1
Count of  pipeline :  1
Count of  portals :  1
Count of  confirmation :  1
Count of  verification :  1
Count of  integrity :  1
Count of  building :  1
Count of  taxonomies :  1
Count of  generation :  1
Count of  opinion :  1
Count of  10 :  1
Count of  language :  1
Count of  code :  4
Count of  complicated :  1
Count of  compared :  1
Count of  English :  1
Count of  GB :  1
Count of  Big5 :  1
Count of  HZ :  1
Count of  common :  1
Count of  codes :  1
Count of  Before :  1
Count of  standard :  1
Count of  inner :  1
Count of  intelligence :  1
Count of  analytics :  1
Count of  scraping :  2
Count of  Data :  16
Count of  ^ :  10
Count of  Galitsky :  2
Count of  B. :  5
Count of  Dobrocsi :  2
Count of  G. :  2
Count of  de :  3
Count of  la :  2
Count of  Rosa :  2
Count of  J. :  2
Count of  L. :  2
Count of  Kuznetsov :  2
Count of  S. :  1
Count of  O.. :  1
Count of  Using :  3
Count of  generalization :  2
Count of  syntactic :  2
Count of  parse :  2
Count of  trees :  2
Count of  taxonomy :  2
Count of  capture :  2
Count of  ICCS :  2
Count of  2011 :  2
Count of  8323 :  2
Count of  Weichbroth :  2
Count of  et :  1
Count of  al :  1
Count of  Ngu :  1
Count of  Anne :  1
Count of  Kitsuregawa :  1
Count of  Masaru :  1
Count of  Chung :  1
Count of  Jen-Yao :  1
Count of  Neuhold :  1
Count of  Erich :  1
Count of  Sheng :  1
Count of  Quan :  1
Count of  2005 :  6
Count of  Systems :  2
Count of  Engineering :  1
Count of  WISE :  1
Count of  Berlin :  2
Count of  Springer :  5
Count of  p. :  3
Count of  15 :  1
Count of  ISBN :  3
Count of  9783540300175 :  1
Count of  Bauknecht :  1
Count of  Kurt :  1
Count of  Madria :  1
Count of  Sanjay :  1
Count of  Pernul :  1
Count of  Gunther :  1
Count of  2000 :  5
Count of  Electronic :  1
Count of  Commerce :  1
Count of  Technologies :  1
Count of  First :  1
Count of  International :  3
Count of  Conference :  3
Count of  EC-Web :  1
Count of  London :  1
Count of  UK :  1
Count of  September :  3
Count of  4-6 :  1
Count of  Proceedings :  4
Count of  165 :  1
Count of  978-3540679813 :  1
Count of  Scime :  1
Count of  Anthony :  1
Count of  Applications :  2
Count of  Hershey :  1
Count of  PA :  1
Count of  Idea :  2
Count of  Group :  2
Count of  Publishing :  1
Count of  pp :  6
Count of  282 :  1
Count of  978-1591404149 :  1
Count of  b :  1
Count of  c :  1
Count of  Lita :  1
Count of  van :  1
Count of  Wel :  1
Count of  & :  2
Count of  Lambèr :  1
Count of  Royakkers :  1
Count of  2004 :  4
Count of  `` :  15
Count of  Ethical :  2
Count of  '' :  19
Count of  PDF :  2
Count of  Issues :  3
Count of  Mining.. :  2
Count of  Kirsten :  1
Count of  Maelstrom :  1
Count of  John :  1
Count of  F. :  2
Count of  Rodrick :  1
Count of  Vladimir :  1
Count of  Estivill-Castro :  1
Count of  Denise :  1
Count of  Vries :  1
Count of  2007 :  4
Count of  Legal :  2
Count of  Technical :  2
Count of  Preservation :  2
Count of  Wang :  2
Count of  Yan :  1
Count of  Knowledge :  4
Count of  Discovery :  5
Count of  Patterns :  5
Count of  Kosala :  1
Count of  Raymond :  1
Count of  Hendrik :  1
Count of  Blockeel :  1
Count of  July :  1
Count of  Research :  2
Count of  Survey :  1
Count of  SIGKDD :  1
Count of  Explorations :  1
Count of  arXiv :  1
Count of  cs.LG/0011033 :  1
Count of  B :  1
Count of  G :  1
Count of  JL :  1
Count of  SO :  1
Count of  list :  1
Count of  remain :  1
Count of  unclear :  1
Count of  insufficient :  1
Count of  inline :  1
Count of  citations :  4
Count of  precise :  1
Count of  Future :  1
Count of  Sites :  1
Count of  = :  1
Count of  Services :  1
Count of  Zdravko :  1
Count of  Markov :  1
Count of  Daniel :  4
Count of  T. :  2
Count of  Larose :  1
Count of  Uncovering :  1
Count of  Content :  1
Count of  Structure :  1
Count of  Wiley :  1
Count of  Jesus :  1
Count of  Mena :  1
Count of  Your :  1
Count of  Website :  2
Count of  Digital :  1
Count of  Press :  2
Count of  1999 :  3
Count of  Soumen :  1
Count of  Chakrabarti :  1
Count of  Analysis :  5
Count of  Semi :  1
Count of  Morgan :  1
Count of  Kaufmann :  1
Count of  2002 :  1
Count of  Bing :  2
Count of  Liu :  2
Count of  Exploring :  1
Count of  Hyperlinks :  1
Count of  Advances :  1
Count of  revised :  2
Count of  papers :  2
Count of  th :  2
Count of  workshop :  2
Count of  Olfa :  2
Count of  Nasraoui :  5
Count of  Osmar :  1
Count of  Zaiane :  1
Count of  Myra :  1
Count of  Spiliopoulou :  1
Count of  Bamshad :  2
Count of  Mobasher :  6
Count of  Philip :  1
Count of  Yu :  1
Count of  Brij :  2
Count of  Masand :  2
Count of  Eds. :  2
Count of  Lecture :  2
Count of  Notes :  2
Count of  Artificial :  4
Count of  Intelligence :  4
Count of  LNAI :  1
Count of  4198 :  1
Count of  2006 :  5
Count of  Mike :  1
Count of  Thelwall :  1
Count of  An :  1
Count of  Science :  1
Count of  Approach :  1
Count of  Academic :  1
Count of  Baraglia :  1
Count of  R. :  6
Count of  Silvestri :  1
Count of  Dynamic :  1
Count of  intervention :  1
Count of  In :  3
Count of  Communications :  2
Count of  ACM :  3
Count of  50 :  1
Count of  63-67 :  1
Count of  Cooley :  3
Count of  Srivastave :  1
Count of  J :  3
Count of  1997 :  1
Count of  " :  10
Count of  Pattern :  2
Count of  " :  10
Count of  9th :  1
Count of  IEEE :  1
Count of  Tool :  1
Count of  Srivastava :  2
Count of  Preparation :  1
Count of  Browsing :  2
Count of  Journal :  2
Count of  System :  1
Count of  Vol.1 :  1
Count of  Issue :  2
Count of  5-32 :  1
Count of  RP :  1
Count of  N. :  1
Count of  Hyponymy :  1
Count of  Extraction :  1
Count of  Search :  2
Count of  Behavior :  1
Count of  On :  1
Count of  Query :  1
Count of  Reformulation :  1
Count of  11th :  1
Count of  Ibero-American :  1
Count of  2008 :  1
Count of  October :  1
Count of  Kohavi :  1
Count of  Mason :  1
Count of  Zheng :  1
Count of  Z :  1
Count of  Lessons :  1
Count of  Challenges :  1
Count of  Retail :  1
Count of  E-commerce :  1
Count of  Learning :  1
Count of  Vol :  3
Count of  57 :  1
Count of  83-113 :  1
Count of  Lillian :  1
Count of  Clark :  1
Count of  I-Hsien :  3
Count of  Ting :  3
Count of  Chris :  3
Count of  Kimble :  3
Count of  Peter :  1
Count of  Wright :  1
Count of  Kudenko :  3
Count of  Combining :  2
Count of  ethnographic :  1
Count of  clickstream :  1
Count of  strategies :  1
Count of  11 :  1
Count of  January :  2
Count of  Eirinaki :  1
Count of  M. :  5
Count of  Vazirgiannis :  1
Count of  2003 :  5
Count of  Personalization :  5
Count of  Transactions :  1
Count of  Technology :  1
Count of  Vol.3 :  1
Count of  No.1 :  1
Count of  February :  1
Count of  Automatic :  1
Count of  43 :  1
Count of  No.8 :  1
Count of  142-151 :  1
Count of  Dai :  1
Count of  H. :  2
Count of  Luo :  1
Count of  Nakagawa :  1
Count of  2001 :  2
Count of  Effective :  2
Count of  Rule :  1
Count of  Discover :  1
Count of  WIDM :  1
Count of  Atlanta :  1
Count of  GA :  1
Count of  USA :  1
Count of  9-15 :  1
Count of  O. :  3
Count of  Petenes :  1
Count of  C. :  3
Count of  Fuzzy :  3
Count of  Inference :  1
Count of  Proc :  1
Count of  WebKDD :  1
Count of  KDD :  1
Count of  Workshop :  1
Count of  Premise :  1
Count of  Intelligent :  1
Count of  Washington :  1
Count of  DC :  1
Count of  August :  2
Count of  37 :  1
Count of  Frigui :  1
Count of  Joshi :  1
Count of  A. :  1
Count of  Krishnapuram :  1
Count of  Access :  1
Count of  Logs :  1
Count of  Competitive :  1
Count of  Eighth :  1
Count of  Congress :  1
Count of  Hsinchu :  1
Count of  Taiwan :  1
Count of  Invited :  1
Count of  chapter :  1
Count of  Encyclopedia :  1
Count of  Warehousing :  1
Count of  Ed :  1
Count of  Pierrakos :  1
Count of  D. :  2
Count of  Paliouras :  1
Count of  Papatheodorou :  1
Count of  Spyropoulos :  1
Count of  tool :  1
Count of  survey :  1
Count of  User :  2
Count of  modelling :  1
Count of  adapted :  1
Count of  interaction :  1
Count of  journal :  1
Count of  Vol.13 :  1
Count of  311-372 :  1
Count of  Restore :  1
Count of  Restoring :  1
Count of  Missing :  1
Count of  Side :  1
Count of  Clickstream :  2
Count of  UBB :  1
Count of  Unexpected :  1
Count of  Behaviour :  1
Count of  ' :  1
Count of  Design :  1
Count of  P. :  1
Count of  Owoc :  1
Count of  Pleszkun :  1
Count of  2012 :  1
Count of  Navigation :  3
Count of  WWW :  1
Count of  Log :  1
Count of  Files :  1
Count of  Retrieved :  1
Count of  https :  1
Count of  //en.wikipedia.org/w/index.php :  1
Count of  ? :  1
Count of  title=Web_mining :  1
Count of  oldid=933573148 :  1
Count of  Categories :  1
Count of  analyticsData :  1
Count of  miningWorld :  1
Count of  WebHidden :  1
Count of  Articles :  1
Count of  needing :  3
Count of  2009All :  2
Count of  cleanupCleanup :  1
Count of  tagged :  1
Count of  articles :  3
Count of  field :  1
Count of  2009Wikipedia :  1
Count of  2009Articles :  1
Count of  expanded :  1
Count of  2015All :  1
Count of  expandedArticles :  1
Count of  small :  1
Count of  boxesArticles :  1
Count of  lacking :  2
Count of  in-text :  2
Count of  menu :  1
Count of  Personal :  1
Count of  Not :  1
Count of  logged :  1
Count of  inTalkContributionsCreate :  1
Count of  accountLog :  1
Count of  Namespaces :  1
Count of  ArticleTalk :  1
Count of  Variants :  1
Count of  Views :  1
Count of  ReadEditView :  1
Count of  history :  1
Count of  pageContentsFeatured :  1
Count of  contentCurrent :  1
Count of  eventsRandom :  1
Count of  articleDonate :  1
Count of  WikipediaWikipedia :  1
Count of  store :  1
Count of  Interaction :  1
Count of  HelpAbout :  1
Count of  WikipediaCommunity :  1
Count of  portalRecent :  1
Count of  changesContact :  1
Count of  Tools :  1
Count of  What :  1
Count of  hereRelated :  1
Count of  changesUpload :  1
Count of  fileSpecial :  1
Count of  pagesPermanent :  1
Count of  linkPage :  1
Count of  informationWikidata :  1
Count of  itemCite :  1
Count of  Print/export :  1
Count of  Create :  1
Count of  bookDownload :  1
Count of  PDFPrintable :  1
Count of  version :  1
Count of  Languages :  1
Count of  Français    HrvatskiMagyar  Por   DeutschEspañolEuskara
tuguês    Slovenčina   :  1
Count of  Edit :  1
Count of  last :  1
Count of  edited :  1
Count of  2020 :  1
Count of  20:51 :  1
Count of  UTC :  1
Count of  Creative :  1
Count of  Commons :  1
Count of  Attribution-ShareAlike :  1
Count of  License :  1
Count of  apply :  1
Count of  agree :  1
Count of  Terms :  1
Count of  Use :  1
Count of  Policy :  1
Count of  Wikipedia® :  1
Count of  registered :  1
Count of  trademark :  1
Count of  Wikimedia :  1
Count of  Foundation :  1
Count of  Inc. :  1
Count of  non-profit :  1
Count of  policy :  1
Count of  About :  1
Count of  Disclaimers :  1
Count of  Contact :  1
Count of  Developers :  1
Count of  Statistics :  1
Count of  Cookie :  1
Count of  statement :  1
Count of  Mobile :  1
```

[14]:
```
print("Number of Unique Terms after removing Stop Words: ", len(term_count.
 ↪keys()))
```

```
Number of Unique Terms after removing Stop Words:  1158
```

[15]:
```
from nltk.stem import PorterStemmer
from nltk.tokenize import sent_tokenize, word_tokenize
from nltk.stem import WordNetLemmatizer
```

[16]:
```
ps = PorterStemmer()
lemmatizer = WordNetLemmatizer()
```

[17]:
```
stem_groups = []
lem_groups = []
groups = []
for w in list(term_count.keys()):
    stem_groups.append(ps.stem(w))
    lem_groups.append(lemmatizer.lemmatize(w))
    groups.append(w)
```

[18]:
```
print("Number of Stemmed Words: ", len(list(dict.fromkeys(stem_groups))))
```

```
Number of Stemmed Words:  913
```

[19]:
```
print("Number of Lemmatized Words: ", len(list(dict.fromkeys(lem_groups))))
```

```
Number of Lemmatized Words:  1110
```

[20]:
```
import pandas as pd
```

[21]:
```
data = {"Word": groups, "Stemmed Word": stem_groups, "Lemmatized Word":␣
 ↪lem_groups}
df = pd.DataFrame(data, columns = ['Word', 'Stemmed Word', 'Lemmatized Word'])
```

[22]:
```
df.head()
```

[22]:
```
        Word Stemmed Word Lemmatized Word
0        Web          web             Web
1     mining         mine          mining
2          -            -               -
3  Wikipedia    wikipedia       Wikipedia
4       From         from            From
```

Ques2: Add one new word to NLTK stopword list and filter the content extracted from the website given in Q. No. 1 in order to display the number of terms present in them after excluding newly added stopwords and their term frequency count. Display the POS tag for all the stopwords, which are removed from the content.

[23]:
```
stop_words = (stopwords.words('english'))
stop_words.append('patterns')
stop_words = set(stop_words)
word_tokens = word_tokenize(text)
```

[24]:
```
filt_text = [w for w in word_tokens if not w in stop_words]
```

[25]:
```
print("Number of Terms after removing Stop Words: ", len(filt_text))
```

```
Number of Terms after removing Stop Words:  2877
```

[26]:
```
term_count = {}
for i in filt_text:
    count = 0
    if i not in term_count.keys():
        for j in filt_text:
            if i==j:
                count+=1
        term_count[i] = count
```

[27]:
```
for i in term_count.keys():
    print("Count of ", i, ": ", term_count[i])
```

```
Count of  Web :  94
Count of  mining :  69
Count of  - :  3
Count of  Wikipedia :  7
Count of  From :  1
Count of  , :  232
Count of  free :  1
Count of  encyclopedia :  1
Count of  Jump :  2
Count of  navigation :  2
Count of  search :  3
Count of  This :  12
Count of  article :  4
Count of  may :  2
Count of  require :  2
Count of  cleanup :  4
Count of  meet :  1
Count of  's :  5
Count of  quality :  1
Count of  standards :  2
Count of  . :  146
Count of  No :  2
Count of  reason :  2
Count of  specified :  1
Count of  Please :  2
Count of  help :  3
Count of  improve :  3
Count of  ( :  33
Count of  June :  6
Count of  2009 :  2
Count of  ) :  33
Count of  Learn :  2
Count of  remove :  2
Count of  template :  2
Count of  message :  3
Count of  application :  6
Count of  data :  43
Count of  techniques :  8
Count of  discover :  4
Count of  World :  6
Count of  Wide :  7
Count of  As :  3
Count of  name :  2
Count of  proposes :  1
Count of  information :  17
Count of  gathered :  1
Count of  web :  29
Count of  It :  4
Count of  makes :  3
Count of  utilization :  1
Count of  automated :  2
Count of  apparatuses :  1
Count of  reveal :  2
Count of  extricate :  1
Count of  servers :  2
Count of  web2 :  1
Count of  reports :  1
Count of  permits :  2
Count of  organizations :  1
Count of  get :  1
Count of  organized :  1
Count of  unstructured :  3
Count of  browser :  1
Count of  activities :  2
Count of  server :  5
Count of  logs :  5
Count of  website :  2
Count of  link :  2
Count of  structure :  27
Count of  page :  12
Count of  content :  15
Count of  different :  6
Count of  sources :  3
Count of  The :  21
Count of  goal :  1
Count of  generate :  2
Count of  structural :  4
Count of  summary :  1
Count of  site :  9
Count of  Technically :  1
Count of  mainly :  1
Count of  focuses :  1
Count of  inner-document :  1
Count of  tries :  2
Count of  hyperlinks :  4
Count of  inter-document :  1
Count of  level :  3
Count of  Based :  3
Count of  topology :  1
Count of  categorize :  3
Count of  pages :  5
Count of  similarity :  1
Count of  relationship :  3
Count of  sites :  2
Count of  also :  5
Count of  another :  1
Count of  direction :  1
Count of  - :  5
Count of  discovering :  1
Count of  document :  5
Count of  type :  3
Count of  used :  8
Count of  schema :  3
Count of  would :  1
Count of  good :  1
Count of  purpose :  2
Count of  make :  3
Count of  possible :  1
Count of  compare/integrate :  1
Count of  schemes :  1
Count of  facilitate :  1
Count of  introducing :  2
Count of  database :  4
Count of  accessing :  1
Count of  providing :  2
Count of  reference :  2
Count of  Contents :  2
Count of  1 :  5
Count of  types :  4
Count of  2 :  5
Count of  usage :  20
Count of  2.1 :  1
Count of  Pros :  2
Count of  2.2 :  1
Count of  Cons :  2
Count of  3 :  2
Count of  4 :  3
Count of  4.1 :  1
Count of  foreign :  2
Count of  languages :  2
Count of  4.1.1 :  1
Count of  Chinese :  4
Count of  5 :  2
Count of  See :  2
Count of  6 :  5
Count of  References :  2
Count of  7 :  3
Count of  Resources :  2
Count of  7.1 :  1
Count of  External :  2
Count of  links :  7
Count of  7.2 :  1
Count of  Books :  2
Count of  7.3 :  1
Count of  Bibliographic :  2
Count of  references :  3
Count of  [ :  26
Count of  edit :  14
Count of  ] :  26
Count of  divided :  2
Count of  three :  1
Count of  general :  2
Count of  categories :  4
Count of  objectives :  1
Count of  Comparison :  1
Count of  IR :  1
Count of  view :  6
Count of  DB :  2
Count of  View :  3
Count of  Unstructured :  1
Count of  Structured :  2
Count of  Semi-structured :  1
Count of  Link :  3
Count of  Interactivity :  1
Count of  Main :  2
Count of  Text :  2
Count of  documents :  11
Count of  Hypertext :  3
Count of  Server :  3
Count of  Browser :  1
Count of  Representation :  1
Count of  Bag :  1
Count of  words :  5
Count of  n-gram :  1
Count of  terms :  2
Count of  phrases :  1
Count of  concepts :  1
Count of  ontology :  1
Count of  Relational :  4
Count of  Edge :  1
Count of  labed :  1
Count of  graph :  5
Count of  Graph :  2
Count of  table :  1
Count of  Method :  2
Count of  Machine :  3
Count of  learning :  2
Count of  Statistical :  2
Count of  including :  2
Count of  NLP :  1
Count of  Proprietary :  2
Count of  algorithms :  5
Count of  Association :  4
Count of  rules :  3
Count of  Application :  3
Count of  Categorization :  2
Count of  Clustering :  3
Count of  Finding :  4
Count of  extract :  2
Count of  text :  4
Count of  frequent :  1
Count of  sub :  1
Count of  structures :  3
Count of  discovery :  2
Count of  Site :  2
Count of  construction :  1
Count of  Adaptation :  1
Count of  management :  3
Count of  interesting :  1
Count of  order :  2
Count of  understand :  1
Count of  better :  4
Count of  serve :  1
Count of  needs :  5
Count of  Web-based :  1
Count of  applications :  6
Count of  Usage :  8
Count of  captures :  1
Count of  identity :  1
Count of  origin :  1
Count of  users :  3
Count of  along :  1
Count of  browsing :  2
Count of  behavior :  2
Count of  classified :  1
Count of  depending :  1
Count of  kind :  1
Count of  considered :  3
Count of  : :  36
Count of  user :  10
Count of  collected :  2
Count of  Typical :  1
Count of  includes :  2
Count of  IP :  1
Count of  address :  2
Count of  access :  2
Count of  time :  2
Count of  Commercial :  1
Count of  significant :  1
Count of  features :  4
Count of  enable :  1
Count of  e-commerce :  2
Count of  built :  1
Count of  top :  1
Count of  little :  1
Count of  effort :  1
Count of  A :  3
Count of  key :  1
Count of  feature :  4
Count of  ability :  1
Count of  track :  1
Count of  various :  1
Count of  kinds :  3
Count of  business :  1
Count of  events :  3
Count of  log :  2
Count of  New :  1
Count of  defined :  3
Count of  logging :  1
Count of  turned :  1
Count of  thus :  2
Count of  generating :  1
Count of  histories :  1
Count of  specially :  1
Count of  Many :  1
Count of  end :  1
Count of  combination :  1
Count of  one :  4
Count of  applied :  3
Count of  Studies :  1
Count of  related :  1
Count of  work :  1
Count of  concerned :  1
Count of  two :  4
Count of  areas :  1
Count of  constraint-based :  1
Count of  developed :  1
Count of  software :  1
Count of  tools :  4
Count of  systems :  2
Count of  Costa :  2
Count of  Seco :  2
Count of  demonstrated :  1
Count of  semantic :  3
Count of  hyponymy :  1
Count of  relationships :  1
Count of  particular :  4
Count of  given :  2
Count of  community :  1
Count of  essentially :  2
Count of  many :  1
Count of  advantages :  1
Count of  technology :  6
Count of  attractive :  1
Count of  corporations :  1
Count of  government :  1
Count of  agencies :  2
Count of  enabled :  1
Count of  personalized :  1
Count of  marketing :  1
Count of  eventually :  1
Count of  results :  2
Count of  higher :  2
Count of  trade :  2
Count of  volumes :  1
Count of  Government :  1
Count of  using :  3
Count of  classify :  1
Count of  threats :  1
Count of  fight :  1
Count of  terrorism :  1
Count of  predicting :  1
Count of  capability :  2
Count of  benefit :  1
Count of  society :  1
Count of  identifying :  1
Count of  criminal :  1
Count of  Companies :  2
Count of  establish :  1
Count of  customer :  7
Count of  understanding :  1
Count of  reacting :  1
Count of  faster :  1
Count of  find :  3
Count of  attract :  1
Count of  retain :  2
Count of  customers :  3
Count of  ; :  17
Count of  save :  1
Count of  production :  1
Count of  costs :  1
Count of  utilizing :  1
Count of  acquired :  1
Count of  insight :  1
Count of  requirements :  1
Count of  They :  3
Count of  increase :  1
Count of  profitability :  1
Count of  target :  2
Count of  pricing :  1
Count of  based :  4
Count of  profiles :  3
Count of  created :  1
Count of  even :  1
Count of  might :  6
Count of  default :  1
Count of  competitor :  1
Count of  company :  2
Count of  try :  1
Count of  promotional :  1
Count of  offers :  1
Count of  specific :  4
Count of  reducing :  1
Count of  risk :  1
Count of  losing :  1
Count of  More :  2
Count of  benefits :  2
Count of  particularly :  1
Count of  area :  1
Count of  personalization :  3
Count of  outlined :  1
Count of  frameworks :  1
Count of  probabilistic :  1
Count of  latent :  1
Count of  analysis :  3
Count of  model :  2
Count of  offer :  1
Count of  additional :  3
Count of  pattern :  2
Count of  process :  2
Count of  provides :  1
Count of  relevant :  1
Count of  collaborative :  1
Count of  recommendation :  1
Count of  These :  3
Count of  models :  1
Count of  demonstrate :  1
Count of  problems :  1
Count of  associated :  1
Count of  traditional :  2
Count of  biases :  1
Count of  questions :  1
Count of  regarding :  1
Count of  validity :  2
Count of  since :  1
Count of  obtained :  6
Count of  subjective :  1
Count of  degrade :  1
Count of  There :  2
Count of  elements :  1
Count of  unique :  1
Count of  show :  1
Count of  include :  1
Count of  way :  1
Count of  knowledge :  4
Count of  interpreting :  1
Count of  analyzing :  1
Count of  reasoning :  1
Count of  phase :  1
Count of  create :  1
Count of  issues :  2
Count of  personal :  4
Count of  nature :  1
Count of  cause :  1
Count of  concerns :  1
Count of  criticized :  1
Count of  ethical :  2
Count of  issue :  1
Count of  involving :  1
Count of  invasion :  1
Count of  privacy :  3
Count of  Privacy :  5
Count of  lost :  1
Count of  concerning :  1
Count of  individual :  4
Count of  disseminated :  1
Count of  especially :  1
Count of  occurs :  1
Count of  without :  3
Count of  consent :  1
Count of  analyzed :  1
Count of  clustered :  1
Count of  form :  1
Count of  made :  2
Count of  anonymous :  3
Count of  clustering :  1
Count of  Thus :  1
Count of  de-individualize :  1
Count of  judging :  2
Count of  mouse :  1
Count of  clicks :  1
Count of  De-individualization :  1
Count of  tendency :  1
Count of  treating :  1
Count of  people :  1
Count of  basis :  1
Count of  group :  1
Count of  characteristics :  2
Count of  instead :  1
Count of  merits :  2
Count of  Another :  1
Count of  important :  2
Count of  concern :  1
Count of  companies :  3
Count of  collecting :  1
Count of  use :  5
Count of  totally :  1
Count of  purposes :  1
Count of  violates :  1
Count of  interests :  1
Count of  growing :  1
Count of  trend :  2
Count of  selling :  1
Count of  commodity :  1
Count of  encourages :  1
Count of  owners :  1
Count of  increased :  1
Count of  amount :  1
Count of  captured :  1
Count of  traded :  1
Count of  increasing :  1
Count of  likeliness :  1
Count of  invaded :  1
Count of  buy :  1
Count of  obliged :  1
Count of  authors :  1
Count of  release :  3
Count of  legally :  1
Count of  responsible :  1
Count of  contents :  1
Count of  inaccuracies :  1
Count of  result :  3
Count of  serious :  1
Count of  lawsuits :  1
Count of  law :  1
Count of  preventing :  1
Count of  trading :  1
Count of  Some :  1
Count of  controversial :  2
Count of  attributes :  3
Count of  like :  1
Count of  sex :  1
Count of  race :  2
Count of  religion :  2
Count of  sexual :  2
Count of  orientation :  2
Count of  individuals :  1
Count of  practices :  1
Count of  anti-discrimination :  1
Count of  legislation :  1
Count of  hard :  1
Count of  identify :  3
Count of  strong :  1
Count of  rule :  1
Count of  could :  1
Count of  denial :  1
Count of  service :  1
Count of  privilege :  1
Count of  situation :  1
Count of  avoided :  1
Count of  high :  1
Count of  maintained :  1
Count of  traced :  1
Count of  back :  1
Count of  look :  1
Count of  poses :  1
Count of  threat :  1
Count of  however :  1
Count of  inferred :  1
Count of  combining :  1
Count of  separate :  1
Count of  unscrupulous :  1
Count of  section :  2
Count of  expansion :  1
Count of  You :  1
Count of  adding :  1
Count of  2015 :  1
Count of  uses :  1
Count of  theory :  1
Count of  analyze :  1
Count of  node :  5
Count of  connection :  1
Count of  According :  1
Count of  Extracting :  1
Count of  hyperlink :  2
Count of  component :  2
Count of  connects :  1
Count of  location :  1
Count of  Mining :  20
Count of  tree-like :  1
Count of  describe :  1
Count of  HTML :  3
Count of  XML :  1
Count of  tag :  1
Count of  terminology :  1
Count of  directed :  1
Count of  representing :  1
Count of  edge :  1
Count of  degree :  2
Count of  number :  3
Count of  pointing :  2
Count of  generated :  1
Count of  Techniques :  2
Count of  PageRank :  1
Count of  algorithm :  3
Count of  Google :  1
Count of  rank :  2
Count of  Google-founder :  1
Count of  Larry :  1
Count of  Page :  1
Count of  decided :  1
Count of  extraction :  2
Count of  integration :  1
Count of  useful :  3
Count of  heterogeneity :  1
Count of  lack :  1
Count of  much :  1
Count of  ever-expanding :  1
Count of  hypertext :  1
Count of  organization :  3
Count of  indexing :  1
Count of  Internet :  2
Count of  Lycos :  1
Count of  Alta :  1
Count of  Vista :  1
Count of  WebCrawler :  1
Count of  Aliweb :  1
Count of  MetaCrawler :  1
Count of  others :  1
Count of  provide :  3
Count of  comfort :  1
Count of  generally :  1
Count of  filter :  1
Count of  interpret :  1
Count of  factors :  1
Count of  prompted :  1
Count of  researchers :  1
Count of  develop :  1
Count of  intelligent :  2
Count of  retrieval :  2
Count of  agents :  1
Count of  well :  1
Count of  extend :  1
Count of  semi-structured :  3
Count of  available :  2
Count of  agent-based :  1
Count of  approach :  1
Count of  involves :  1
Count of  development :  1
Count of  sophisticated :  1
Count of  AI :  1
Count of  act :  1
Count of  autonomously :  1
Count of  semi-autonomously :  1
Count of  behalf :  1
Count of  organize :  1
Count of  web-based :  1
Count of  differentiated :  1
Count of  points :  1
Count of  8 :  1
Count of  Information :  6
Count of  Retrieval :  1
Count of  Database :  1
Count of  9 :  1
Count of  summarized :  1
Count of  research :  1
Count of  works :  2
Count of  done :  1
Count of  shows :  1
Count of  researches :  1
Count of  bag :  1
Count of  statistics :  1
Count of  single :  2
Count of  isolation :  1
Count of  represent :  2
Count of  take :  1
Count of  word :  2
Count of  found :  1
Count of  training :  1
Count of  corpus :  1
Count of  For :  1
Count of  utilize :  1
Count of  inside :  1
Count of  utilized :  1
Count of  representation :  2
Count of  querying :  1
Count of  always :  1
Count of  infer :  1
Count of  transform :  2
Count of  become :  1
Count of  several :  1
Count of  ways :  1
Count of  vector :  2
Count of  space :  2
Count of  typically :  1
Count of  constitute :  1
Count of  whole :  1
Count of  realize :  1
Count of  importance :  1
Count of  To :  1
Count of  resolve :  1
Count of  tf-idf :  1
Count of  Term :  1
Count of  Frequency :  2
Count of  Times :  1
Count of  Inverse :  1
Count of  Document :  1
Count of  introduced :  1
Count of  By :  2
Count of  multi-scanning :  1
Count of  implement :  1
Count of  selection :  1
Count of  Under :  1
Count of  condition :  1
Count of  category :  1
Count of  rarely :  1
Count of  affected :  1
Count of  subset :  1
Count of  needed :  1
Count of  construct :  1
Count of  evaluating :  1
Count of  function :  1
Count of  evaluate :  1
Count of  set :  1
Count of  gain :  1
Count of  cross :  1
Count of  entropy :  1
Count of  mutual :  1
Count of  odds :  1
Count of  ratio :  1
Count of  usually :  1
Count of  classifier :  1
Count of  methods :  1
Count of  similar :  1
Count of  usual :  1
Count of  evaluative :  1
Count of  classification :  1
Count of  accuracy :  1
Count of  precision :  1
Count of  recall :  1
Count of  score :  1
Count of  pipeline :  1
Count of  portals :  1
Count of  confirmation :  1
Count of  verification :  1
Count of  integrity :  1
Count of  building :  1
Count of  taxonomies :  1
Count of  generation :  1
Count of  opinion :  1
Count of  10 :  1
Count of  language :  1
Count of  code :  4
Count of  complicated :  1
Count of  compared :  1
Count of  English :  1
Count of  GB :  1
Count of  Big5 :  1
Count of  HZ :  1
Count of  common :  1
Count of  codes :  1
Count of  Before :  1
Count of  standard :  1
Count of  inner :  1
Count of  intelligence :  1
Count of  analytics :  1
Count of  scraping :  2
Count of  Data :  16
Count of  ^ :  10
Count of  Galitsky :  2
Count of  B. :  5
Count of  Dobrocsi :  2
Count of  G. :  2
Count of  de :  3
Count of  la :  2
Count of  Rosa :  2
Count of  J. :  2
Count of  L. :  2
Count of  Kuznetsov :  2
Count of  S. :  1
Count of  O.. :  1
Count of  Using :  3
Count of  generalization :  2
Count of  syntactic :  2
Count of  parse :  2
Count of  trees :  2
Count of  taxonomy :  2
Count of  capture :  2
Count of  ICCS :  2
Count of  2011 :  2
Count of  8323 :  2
Count of  Weichbroth :  2
Count of  et :  1
Count of  al :  1
Count of  Ngu :  1
Count of  Anne :  1
Count of  Kitsuregawa :  1
Count of  Masaru :  1
Count of  Chung :  1
Count of  Jen-Yao :  1
Count of  Neuhold :  1
Count of  Erich :  1
Count of  Sheng :  1
Count of  Quan :  1
Count of  2005 :  6
Count of  Systems :  2
Count of  Engineering :  1
Count of  WISE :  1
Count of  Berlin :  2
Count of  Springer :  5
Count of  p. :  3
Count of  15 :  1
Count of  ISBN :  3
Count of  9783540300175 :  1
Count of  Bauknecht :  1
Count of  Kurt :  1
Count of  Madria :  1
Count of  Sanjay :  1
Count of  Pernul :  1
Count of  Gunther :  1
Count of  2000 :  5
Count of  Electronic :  1
Count of  Commerce :  1
Count of  Technologies :  1
Count of  First :  1
Count of  International :  3
Count of  Conference :  3
Count of  EC-Web :  1
Count of  London :  1
Count of  UK :  1
Count of  September :  3
Count of  4-6 :  1
Count of  Proceedings :  4
Count of  165 :  1
Count of  978-3540679813 :  1
Count of  Scime :  1
Count of  Anthony :  1
Count of  Applications :  2
Count of  Hershey :  1
Count of  PA :  1
Count of  Idea :  2
Count of  Group :  2
Count of  Publishing :  1
Count of  pp :  6
Count of  282 :  1
Count of  978-1591404149 :  1
Count of  b :  1
Count of  c :  1
Count of  Lita :  1
Count of  van :  1
Count of  Wel :  1
Count of  & :  2
Count of  Lambèr :  1
Count of  Royakkers :  1
Count of  2004 :  4
Count of  `` :  15
Count of  Ethical :  2
Count of  '' :  19
Count of  PDF :  2
Count of  Issues :  3
Count of  Mining.. :  2
Count of  Kirsten :  1
Count of  Maelstrom :  1
Count of  John :  1
Count of  F. :  2
Count of  Rodrick :  1
Count of  Vladimir :  1
Count of  Estivill-Castro :  1
Count of  Denise :  1
Count of  Vries :  1
Count of  2007 :  4
Count of  Legal :  2
Count of  Technical :  2
Count of  Preservation :  2
Count of  Wang :  2
Count of  Yan :  1
Count of  Knowledge :  4
Count of  Discovery :  5
Count of  Patterns :  5
Count of  Kosala :  1
Count of  Raymond :  1
Count of  Hendrik :  1
Count of  Blockeel :  1
Count of  July :  1
Count of  Research :  2
Count of  Survey :  1
Count of  SIGKDD :  1
Count of  Explorations :  1
Count of  arXiv :  1
Count of  cs.LG/0011033 :  1
Count of  B :  1
Count of  G :  1
Count of  JL :  1
Count of  SO :  1
Count of  list :  1
Count of  remain :  1
Count of  unclear :  1
Count of  insufficient :  1
Count of  inline :  1
Count of  citations :  4
Count of  precise :  1
Count of  Future :  1
Count of  Sites :  1
Count of  = :  1
Count of  Services :  1
Count of  Zdravko :  1
Count of  Markov :  1
Count of  Daniel :  4
Count of  T. :  2
Count of  Larose :  1
Count of  Uncovering :  1
Count of  Content :  1
Count of  Structure :  1
Count of  Wiley :  1
Count of  Jesus :  1
Count of  Mena :  1
Count of  Your :  1
Count of  Website :  2
Count of  Digital :  1
Count of  Press :  2
Count of  1999 :  3
Count of  Soumen :  1
Count of  Chakrabarti :  1
Count of  Analysis :  5
Count of  Semi :  1
Count of  Morgan :  1
Count of  Kaufmann :  1
Count of  2002 :  1
Count of  Bing :  2
Count of  Liu :  2
Count of  Exploring :  1
Count of  Hyperlinks :  1
Count of  Advances :  1
Count of  revised :  2
Count of  papers :  2
Count of  th :  2
Count of  workshop :  2
Count of  Olfa :  2
Count of  Nasraoui :  5
Count of  Osmar :  1
Count of  Zaiane :  1
Count of  Myra :  1
Count of  Spiliopoulou :  1
Count of  Bamshad :  2
Count of  Mobasher :  6
Count of  Philip :  1
Count of  Yu :  1
Count of  Brij :  2
Count of  Masand :  2
Count of  Eds. :  2
Count of  Lecture :  2
Count of  Notes :  2
Count of  Artificial :  4
Count of  Intelligence :  4
Count of  LNAI :  1
Count of  4198 :  1
Count of  2006 :  5
Count of  Mike :  1
Count of  Thelwall :  1
Count of  An :  1
Count of  Science :  1
Count of  Approach :  1
Count of  Academic :  1
Count of  Baraglia :  1
Count of  R. :  6
Count of  Silvestri :  1
Count of  Dynamic :  1
Count of  intervention :  1
Count of  In :  3
Count of  Communications :  2
Count of  ACM :  3
Count of  50 :  1
Count of  63-67 :  1
Count of  Cooley :  3
Count of  Srivastave :  1
Count of  J :  3
Count of  1997 :  1
Count of  " :  10
Count of  Pattern :  2
Count of  " :  10
Count of  9th :  1
Count of  IEEE :  1
Count of  Tool :  1
Count of  Srivastava :  2
Count of  Preparation :  1
Count of  Browsing :  2
Count of  Journal :  2
Count of  System :  1
Count of  Vol.1 :  1
Count of  Issue :  2
Count of  5-32 :  1
Count of  RP :  1
Count of  N. :  1
Count of  Hyponymy :  1
Count of  Extraction :  1
Count of  Search :  2
Count of  Behavior :  1
Count of  On :  1
Count of  Query :  1
Count of  Reformulation :  1
Count of  11th :  1
Count of  Ibero-American :  1
Count of  2008 :  1
Count of  October :  1
Count of  Kohavi :  1
Count of  Mason :  1
Count of  Zheng :  1
Count of  Z :  1
Count of  Lessons :  1
Count of  Challenges :  1
Count of  Retail :  1
Count of  E-commerce :  1
Count of  Learning :  1
Count of  Vol :  3
Count of  57 :  1
Count of  83-113 :  1
Count of  Lillian :  1
Count of  Clark :  1
Count of  I-Hsien :  3
Count of  Ting :  3
Count of  Chris :  3
Count of  Kimble :  3
Count of  Peter :  1
Count of  Wright :  1
Count of  Kudenko :  3
Count of  Combining :  2
Count of  ethnographic :  1
Count of  clickstream :  1
Count of  strategies :  1
Count of  11 :  1
Count of  January :  2
Count of  Eirinaki :  1
Count of  M. :  5
Count of  Vazirgiannis :  1
Count of  2003 :  5
Count of  Personalization :  5
Count of  Transactions :  1
Count of  Technology :  1
Count of  Vol.3 :  1
Count of  No.1 :  1
Count of  February :  1
Count of  Automatic :  1
Count of  43 :  1
Count of  No.8 :  1
Count of  142-151 :  1
Count of  Dai :  1
Count of  H. :  2
Count of  Luo :  1
Count of  Nakagawa :  1
Count of  2001 :  2
Count of  Effective :  2
Count of  Rule :  1
Count of  Discover :  1
Count of  WIDM :  1
Count of  Atlanta :  1
Count of  GA :  1
Count of  USA :  1
Count of  9-15 :  1
Count of  O. :  3
Count of  Petenes :  1
Count of  C. :  3
Count of  Fuzzy :  3
Count of  Inference :  1
Count of  Proc :  1
Count of  WebKDD :  1
Count of  KDD :  1
Count of  Workshop :  1
Count of  Premise :  1
Count of  Intelligent :  1
Count of  Washington :  1
Count of  DC :  1
Count of  August :  2
Count of  37 :  1
Count of  Frigui :  1
Count of  Joshi :  1
Count of  A. :  1
Count of  Krishnapuram :  1
Count of  Access :  1
Count of  Logs :  1
Count of  Competitive :  1
Count of  Eighth :  1
Count of  Congress :  1
Count of  Hsinchu :  1
Count of  Taiwan :  1
Count of  Invited :  1
Count of  chapter :  1
Count of  Encyclopedia :  1
Count of  Warehousing :  1
Count of  Ed :  1
Count of  Pierrakos :  1
Count of  D. :  2
Count of  Paliouras :  1
Count of  Papatheodorou :  1
Count of  Spyropoulos :  1
Count of  tool :  1
Count of  survey :  1
Count of  User :  2
Count of  modelling :  1
Count of  adapted :  1
Count of  interaction :  1
Count of  journal :  1
Count of  Vol.13 :  1
Count of  311-372 :  1
Count of  Restore :  1
Count of  Restoring :  1
Count of  Missing :  1
Count of  Side :  1
Count of  Clickstream :  2
Count of  UBB :  1
Count of  Unexpected :  1
Count of  Behaviour :  1
Count of  ’ :  1
Count of  Design :  1
Count of  P. :  1
Count of  Owoc :  1
Count of  Pleszkun :  1
Count of  2012 :  1
Count of  Navigation :  3
Count of  WWW :  1
Count of  Log :  1
Count of  Files :  1
Count of  Retrieved :  1
Count of  https :  1
Count of  //en.wikipedia.org/w/index.php :  1
Count of  ? :  1
Count of  title=Web_mining :  1
Count of  oldid=933573148 :  1
Count of  Categories :  1
Count of  analyticsData :  1
Count of  miningWorld :  1
Count of  WebHidden :  1
Count of  Articles :  1
Count of  needing :  3
Count of  2009All :  2
Count of  cleanupCleanup :  1
Count of  tagged :  1
Count of  articles :  3
Count of  field :  1
Count of  2009Wikipedia :  1
Count of  2009Articles :  1
Count of  expanded :  1
Count of  2015All :  1
Count of  expandedArticles :  1
Count of  small :  1
Count of  boxesArticles :  1
Count of  lacking :  2
Count of  in-text :  2
Count of  menu :  1
Count of  Personal :  1
Count of  Not :  1
Count of  logged :  1
Count of  inTalkContributionsCreate :  1
Count of  accountLog :  1
Count of  Namespaces :  1
Count of  ArticleTalk :  1
Count of  Variants :  1
Count of  Views :  1
Count of  ReadEditView :  1
Count of  history :  1
Count of  pageContentsFeatured :  1
Count of  contentCurrent :  1
Count of  eventsRandom :  1
Count of  articleDonate :  1
Count of  WikipediaWikipedia :  1
Count of  store :  1
Count of  Interaction :  1
Count of  HelpAbout :  1
Count of  WikipediaCommunity :  1
Count of  portalRecent :  1
Count of  changesContact :  1
Count of  Tools :  1
Count of  What :  1
Count of  hereRelated :  1
Count of  changesUpload :  1
Count of  fileSpecial :  1
Count of  pagesPermanent :  1
Count of  linkPage :  1
Count of  informationWikidata :  1
Count of  itemCite :  1
Count of  Print/export :  1
Count of  Create :  1
Count of  bookDownload :  1
Count of  PDFPrintable :  1
Count of  version :  1
Count of  Languages :  1
Count of  Français    HrvatskiMagyar  Por   DeutschEspañolEuskara
tuguês    Slovenčina   :  1
Count of  Edit :  1
Count of  last :  1
Count of  edited :  1
Count of  2020 :  1
Count of  20:51 :  1
Count of  UTC :  1
Count of  Creative :  1
Count of  Commons :  1
Count of  Attribution-ShareAlike :  1
Count of  License :  1
Count of  apply :  1
Count of  agree :  1
Count of  Terms :  1
Count of  Use :  1
Count of  Policy :  1
Count of  Wikipedia® :  1
Count of  registered :  1
Count of  trademark :  1
Count of  Wikimedia :  1
Count of  Foundation :  1
Count of  Inc. :  1
Count of  non-profit :  1
Count of  policy :  1
Count of  About :  1
Count of  Disclaimers :  1
Count of  Contact :  1
Count of  Developers :  1
Count of  Statistics :  1
Count of  Cookie :  1
Count of  statement :  1
Count of  Mobile :  1
```

[28]:

```
import nltk
```

[29]:

```
url = requests.get("https://en.wikipedia.org/wiki/Web_mining")
soup = BeautifulSoup(url.text)
for script in soup(["script", "style"]):
    script.decompose()
text = soup.get_text()
lines = (line.strip() for line in text.splitlines())
# break multi-headlines into a line each
chunks = (phrase.strip() for line in lines for phrase in line.split("  "))
# drop blank lines
text = '\n'.join(chunk for chunk in chunks if chunk)
```

[30]:

```
text = text.split("\n")
```

[31]:

```
for line in text:
    line = line.split(".")
    for sub_line in line:
        wordsList = nltk.word_tokenize(sub_line)
        for i in wordsList:
            if i in stop_words:
                wordsList2 = nltk.word_tokenize(i)
                tagged = nltk.pos_tag(wordsList2)
                print(tagged)
```

```
[('the', 'DT')]
[('to', 'TO')]
[('to', 'TO')]
[('to', 'TO')]
[('has', 'VBZ')]
[('been', 'VBN')]
[('this', 'DT')]
[('if', 'IN')]
[('you', 'PRP')]
[('can', 'MD')]
[('how', 'WRB')]
[('and', 'CC')]
[('when', 'WRB')]
[('to', 'TO')]
[('this', 'DT')]
[('is', 'VBZ')]
[('the', 'DT')]
[('of', 'IN')]
[('to', 'TO')]
```

```
[('patterns', 'NNS')]
[('from', 'IN')]
[('the', 'DT')]
[('the', 'DT')]
[('this', 'DT')]
[('is', 'VBZ')]
[('by', 'IN')]
[('the', 'DT')]
[('of', 'IN')]
[('to', 'TO')]
[('and', 'CC')]
[('from', 'IN')]
[('and', 'CC')]
[('and', 'CC')]
[('it', 'PRP')]
[('to', 'TO')]
[('to', 'TO')]
[('both', 'DT')]
[('and', 'CC')]
[('from', 'IN')]
[('and', 'CC')]
[('and', 'CC')]
[('of', 'IN')]
[('is', 'VBZ')]
[('to', 'TO')]
[('about', 'IN')]
[('the', 'DT')]
[('and', 'CC')]
[('on', 'IN')]
[('the', 'DT')]
[('of', 'IN')]
[('while', 'IN')]
[('to', 'TO')]
[('the', 'DT')]
[('of', 'IN')]
[('the', 'DT')]
[('at', 'IN')]
[('the', 'DT')]
[('on', 'IN')]
[('the', 'DT')]
[('of', 'IN')]
[('the', 'DT')]
[('will', 'MD')]
[('the', 'DT')]
[('and', 'CC')]
[('the', 'DT')]
[('such', 'JJ')]
[('as', 'IN')]
```

```
[('the', 'DT')]
[('and', 'CC')]
[('between', 'IN')]
[('can', 'MD')]
[('have', 'VB')]
[('the', 'DT')]
[('of', 'IN')]
[('itself', 'PRP')]
[('of', 'IN')]
[('can', 'MD')]
[('be', 'VB')]
[('to', 'TO')]
[('the', 'DT')]
[('of', 'IN')]
[('this', 'DT')]
[('be', 'VB')]
[('for', 'IN')]
[('and', 'CC')]
[('it', 'PRP')]
[('to', 'TO')]
[('of', 'IN')]
[('will', 'MD')]
[('for', 'IN')]
[('in', 'IN')]
[('by', 'IN')]
[('a', 'DT')]
[('in', 'IN')]
[('can', 'MD')]
[('be', 'VB')]
[('into', 'IN')]
[('and', 'CC')]
[('between', 'IN')]
[('the', 'DT')]
[('of', 'IN')]
[('and', 'CC')]
[('of', 'IN')]
[('of', 'IN')]
[('of', 'IN')]
[('as', 'IN')]
[('of', 'IN')]
[('or', 'CC')]
[('patterns', 'NNS')]
[('in', 'IN')]
[('and', 'CC')]
[('is', 'VBZ')]
[('the', 'DT')]
[('of', 'IN')]
[('to', 'TO')]
```

```
[('patterns', 'NNS')]
[('from', 'IN')]
[('in', 'IN')]
[('to', 'TO')]
[('and', 'CC')]
[('the', 'DT')]
[('of', 'IN')]
[('the', 'DT')]
[('or', 'CC')]
[('of', 'IN')]
[('with', 'IN')]
[('their', 'PRP$')]
[('at', 'IN')]
[('a', 'DT')]
[('itself', 'PRP')]
[('can', 'MD')]
[('be', 'VB')]
[('further', 'RB')]
[('on', 'IN')]
[('the', 'DT')]
[('of', 'IN')]
[('are', 'VBP')]
[('by', 'IN')]
[('the', 'DT')]
[('and', 'CC')]
[('have', 'VB')]
[('to', 'TO')]
[('to', 'TO')]
[('be', 'VB')]
[('on', 'IN')]
[('of', 'IN')]
[('them', 'PRP')]
[('with', 'IN')]
[('is', 'VBZ')]
[('the', 'DT')]
[('to', 'TO')]
[('of', 'IN')]
[('and', 'CC')]
[('them', 'PRP')]
[('in', 'IN')]
[('of', 'IN')]
[('can', 'MD')]
[('be', 'VB')]
[('in', 'IN')]
[('an', 'DT')]
[('and', 'CC')]
[('can', 'MD')]
[('be', 'VB')]
```

```
[('on', 'IN')]
[('for', 'IN')]
[('them', 'PRP')]
[('of', 'IN')]
[('these', 'DT')]
[('a', 'DT')]
[('of', 'IN')]
[('or', 'CC')]
[('more', 'RBR')]
[('of', 'IN')]
[('the', 'DT')]
[('in', 'IN')]
[('the', 'DT')]
[('above', 'IN')]
[('to', 'TO')]
[('are', 'VBP')]
[('with', 'IN')]
[('in', 'IN')]
[('and', 'CC')]
[('and', 'CC')]
[('that', 'IN')]
[('can', 'MD')]
[('be', 'VB')]
[('to', 'TO')]
[('in', 'IN')]
[('about', 'IN')]
[('the', 'DT')]
[('and', 'CC')]
[('a', 'DT')]
[('has', 'VBZ')]
[('which', 'WDT')]
[('this', 'DT')]
[('to', 'TO')]
[('has', 'VBZ')]
[('to', 'TO')]
[('do', 'VB')]
[('which', 'WDT')]
[('in', 'IN')]
[('are', 'VBP')]
[('this', 'DT')]
[('to', 'TO')]
[('and', 'CC')]
[('against', 'IN')]
[('of', 'IN')]
[('can', 'MD')]
[('by', 'IN')]
[('can', 'MD')]
[('by', 'IN')]
```

```
[('the', 'DT')]
[('of', 'IN')]
[('the', 'DT')]
[('and', 'CC')]
[('to', 'TO')]
[('can', 'MD')]
[('and', 'CC')]
[('they', 'PRP')]
[('can', 'MD')]
[('on', 'IN')]
[('by', 'IN')]
[('the', 'DT')]
[('of', 'IN')]
[('can', 'MD')]
[('by', 'IN')]
[('on', 'IN')]
[('the', 'DT')]
[('can', 'MD')]
[('who', 'WP')]
[('to', 'TO')]
[('a', 'DT')]
[('the', 'DT')]
[('will', 'MD')]
[('to', 'TO')]
[('the', 'DT')]
[('by', 'IN')]
[('to', 'TO')]
[('the', 'DT')]
[('the', 'DT')]
[('of', 'IN')]
[('a', 'DT')]
[('or', 'CC')]
[('of', 'IN')]
[('in', 'IN')]
[('the', 'DT')]
[('of', 'IN')]
[('are', 'VBP')]
[('in', 'IN')]
[('such', 'JJ')]
[('as', 'IN')]
[('the', 'DT')]
[('which', 'WDT')]
[('to', 'TO')]
[('the', 'DT')]
[('and', 'CC')]
[('is', 'VBZ')]
[('because', 'IN')]
[('the', 'DT')]
```

```
[('the', 'DT')]
[('with', 'IN')]
[('more', 'RBR')]
[('through', 'IN')]
[('a', 'DT')]
[('in', 'IN')]
[('to', 'TO')]
[('with', 'IN')]
[('such', 'JJ')]
[('as', 'IN')]
[('and', 'CC')]
[('the', 'DT')]
[('and', 'CC')]
[('patterns', 'NNS')]
[('are', 'VBP')]
[('not', 'RB')]
[('and', 'CC')]
[('do', 'VB')]
[('not', 'RB')]
[('over', 'IN')]
[('are', 'VBP')]
[('to', 'TO')]
[('that', 'IN')]
[('can', 'MD')]
[('the', 'DT')]
[('and', 'CC')]
[('these', 'DT')]
[('the', 'DT')]
[('is', 'VBZ')]
[('when', 'WRB')]
[('and', 'CC')]
[('about', 'IN')]
[('patterns', 'NNS')]
[('during', 'IN')]
[('the', 'DT')]
[('by', 'IN')]
[('itself', 'PRP')]
[('does', 'VBZ')]
[('not', 'RB')]
[('but', 'CC')]
[('this', 'DT')]
[('when', 'WRB')]
[('on', 'IN')]
[('of', 'IN')]
[('most', 'JJS')]
[('is', 'VBZ')]
[('the', 'DT')]
[('of', 'IN')]
```

```
[('is', 'VBZ')]
[('when', 'WRB')]
[('an', 'DT')]
[('is', 'VBZ')]
[('or', 'CC')]
[('if', 'IN')]
[('this', 'DT')]
[('their', 'PRP$')]
[('or', 'CC')]
[('will', 'MD')]
[('be', 'VB')]
[('and', 'CC')]
[('to', 'TO')]
[('the', 'DT')]
[('will', 'MD')]
[('be', 'VB')]
[('before', 'IN')]
[('so', 'RB')]
[('that', 'IN')]
[('there', 'RB')]
[('are', 'VBP')]
[('no', 'DT')]
[('these', 'DT')]
[('the', 'DT')]
[('by', 'IN')]
[('them', 'PRP')]
[('by', 'IN')]
[('their', 'PRP$')]
[('can', 'MD')]
[('be', 'VB')]
[('as', 'IN')]
[('a', 'DT')]
[('of', 'IN')]
[('and', 'CC')]
[('on', 'IN')]
[('the', 'DT')]
[('of', 'IN')]
[('of', 'IN')]
[('on', 'IN')]
[('their', 'PRP$')]
[('own', 'JJ')]
[('and', 'CC')]
[('is', 'VBZ')]
[('that', 'IN')]
[('the', 'DT')]
[('the', 'DT')]
[('for', 'IN')]
[('a', 'DT')]
```

```
[('the', 'DT')]
[('for', 'IN')]
[('and', 'CC')]
[('this', 'DT')]
[('the', 'DT')]
[('of', 'IN')]
[('as', 'IN')]
[('a', 'DT')]
[('to', 'TO')]
[('from', 'IN')]
[('their', 'PRP$')]
[('has', 'VBZ')]
[('the', 'DT')]
[('of', 'IN')]
[('being', 'VBG')]
[('and', 'CC')]
[('the', 'DT')]
[('of', 'IN')]
[('being', 'VBG')]
[('which', 'WDT')]
[('the', 'DT')]
[('are', 'VBP')]
[('it', 'PRP')]
[('and', 'CC')]
[('these', 'DT')]
[('are', 'VBP')]
[('of', 'IN')]
[('any', 'DT')]
[('of', 'IN')]
[('patterns', 'NNS')]
[('are', 'VBP')]
[('for', 'IN')]
[('the', 'DT')]
[('of', 'IN')]
[('the', 'DT')]
[('any', 'DT')]
[('in', 'IN')]
[('the', 'DT')]
[('will', 'MD')]
[('in', 'IN')]
[('but', 'CC')]
[('there', 'RB')]
[('is', 'VBZ')]
[('no', 'DT')]
[('them', 'PRP')]
[('from', 'IN')]
[('the', 'DT')]
[('or', 'CC')]
```

```
[('to', 'TO')]
[('be', 'VB')]
[('against', 'IN')]
[('the', 'DT')]
[('it', 'PRP')]
[('to', 'TO')]
[('the', 'DT')]
[('of', 'IN')]
[('such', 'JJ')]
[('and', 'CC')]
[('there', 'RB')]
[('is', 'VBZ')]
[('no', 'DT')]
[('against', 'IN')]
[('the', 'DT')]
[('of', 'IN')]
[('such', 'JJ')]
[('with', 'IN')]
[('such', 'JJ')]
[('in', 'IN')]
[('of', 'IN')]
[('or', 'CC')]
[('a', 'DT')]
[('to', 'TO')]
[('an', 'DT')]
[('on', 'IN')]
[('his', 'PRP$')]
[('or', 'CC')]
[('can', 'MD')]
[('be', 'VB')]
[('by', 'IN')]
[('the', 'DT')]
[('by', 'IN')]
[('the', 'DT')]
[('is', 'VBZ')]
[('being', 'VBG')]
[('so', 'RB')]
[('that', 'IN')]
[('the', 'DT')]
[('and', 'CC')]
[('the', 'DT')]
[('patterns', 'NNS')]
[('can', 'MD')]
[('not', 'RB')]
[('be', 'VB')]
[('to', 'TO')]
[('an', 'DT')]
[('as', 'IN')]
```

```
[('if', 'IN')]
[('this', 'DT')]
[('no', 'DT')]
[('to', 'TO')]
[('can', 'MD')]
[('be', 'VB')]
[('by', 'IN')]
[('the', 'DT')]
[('by', 'IN')]
[('from', 'IN')]
[('the', 'DT')]
[('can', 'MD')]
[('by', 'IN')]
[('to', 'TO')]
[('it', 'PRP')]
[('to', 'TO')]
[('the', 'DT')]
[('and', 'CC')]
[('of', 'IN')]
[('a', 'DT')]
[('to', 'TO')]
[('the', 'DT')]
[('of', 'IN')]
[('can', 'MD')]
[('be', 'VB')]
[('into', 'IN')]
[('patterns', 'NNS')]
[('from', 'IN')]
[('in', 'IN')]
[('the', 'DT')]
[('a', 'DT')]
[('is', 'VBZ')]
[('a', 'DT')]
[('that', 'IN')]
[('the', 'DT')]
[('to', 'TO')]
[('a', 'DT')]
[('the', 'DT')]
[('of', 'IN')]
[('the', 'DT')]
[('of', 'IN')]
[('to', 'TO')]
[('or', 'CC')]
[('in', 'IN')]
[('in', 'IN')]
[('of', 'IN')]
[('to', 'TO')]
[('out', 'IN')]
```

```
[('of', 'IN')]
[('from', 'IN')]
[('of', 'IN')]
[('this', 'DT')]
[('is', 'VBZ')]
[('by', 'IN')]
[('to', 'TO')]
[('of', 'IN')]
[('this', 'DT')]
[('is', 'VBZ')]
[('by', 'IN')]
[('of', 'IN')]
[('a', 'DT')]
[('is', 'VBZ')]
[('by', 'IN')]
[('the', 'DT')]
[('of', 'IN')]
[('to', 'TO')]
[('the', 'DT')]
[('is', 'VBZ')]
[('the', 'DT')]
[('and', 'CC')]
[('of', 'IN')]
[('and', 'CC')]
[('from', 'IN')]
[('and', 'CC')]
[('the', 'DT')]
[('of', 'IN')]
[('that', 'IN')]
[('of', 'IN')]
[('the', 'DT')]
[('on', 'IN')]
[('the', 'DT')]
[('such', 'JJ')]
[('as', 'IN')]
[('and', 'CC')]
[('and', 'CC')]
[('of', 'IN')]
[('the', 'DT')]
[('and', 'CC')]
[('the', 'DT')]
[('such', 'JJ')]
[('as', 'IN')]
[('and', 'CC')]
[('some', 'DT')]
[('to', 'TO')]
[('but', 'CC')]
[('they', 'PRP')]
```

```
[('do', 'VB')]
[('not', 'RB')]
[('nor', 'CC')]
[('or', 'CC')]
[('have', 'VB')]
[('to', 'TO')]
[('more', 'RBR')]
[('for', 'IN')]
[('such', 'JJ')]
[('as', 'IN')]
[('as', 'IN')]
[('as', 'IN')]
[('to', 'TO')]
[('and', 'CC')]
[('to', 'TO')]
[('a', 'DT')]
[('of', 'IN')]
[('for', 'IN')]
[('on', 'IN')]
[('the', 'DT')]
[('to', 'TO')]
[('the', 'DT')]
[('of', 'IN')]
[('that', 'IN')]
[('can', 'MD')]
[('or', 'CC')]
[('on', 'IN')]
[('of', 'IN')]
[('a', 'DT')]
[('to', 'TO')]
[('and', 'CC')]
[('is', 'VBZ')]
[('from', 'IN')]
[('of', 'IN')]
[('and', 'CC')]
[('the', 'DT')]
[('for', 'IN')]
[('and', 'CC')]
[('from', 'IN')]
[('that', 'IN')]
[('most', 'JJS')]
[('of', 'IN')]
[('the', 'DT')]
[('of', 'IN')]
[('which', 'WDT')]
[('is', 'VBZ')]
[('on', 'IN')]
[('the', 'DT')]
```

```
[('about', 'IN')]
[('in', 'IN')]
[('to', 'TO')]
[('and', 'CC')]
[('in', 'IN')]
[('the', 'DT')]
[('as', 'IN')]
[('the', 'DT')]
[('all', 'DT')]
[('the', 'DT')]
[('the', 'DT')]
[('the', 'DT')]
[('and', 'CC')]
[('some', 'DT')]
[('the', 'DT')]
[('between', 'IN')]
[('the', 'DT')]
[('for', 'IN')]
[('for', 'IN')]
[('the', 'DT')]
[('in', 'IN')]
[('to', 'TO')]
[('have', 'VB')]
[('the', 'DT')]
[('and', 'CC')]
[('on', 'IN')]
[('the', 'DT')]
[('the', 'DT')]
[('to', 'TO')]
[('the', 'DT')]
[('of', 'IN')]
[('the', 'DT')]
[('to', 'TO')]
[('a', 'DT')]
[('to', 'TO')]
[('a', 'DT')]
[('are', 'VBP')]
[('to', 'TO')]
[('is', 'VBZ')]
[('the', 'DT')]
[('does', 'VBZ')]
[('not', 'RB')]
[('the', 'DT')]
[('of', 'IN')]
[('in', 'IN')]
[('a', 'DT')]
[('this', 'DT')]
[('is', 'VBZ')]
```

```
[('the', 'DT')]
[('we', 'PRP')]
[('can', 'MD')]
[('the', 'DT')]
[('that', 'IN')]
[('the', 'DT')]
[('is', 'VBZ')]
[('the', 'DT')]
[('of', 'IN')]
[('is', 'VBZ')]
[('is', 'VBZ')]
[('to', 'TO')]
[('an', 'DT')]
[('to', 'TO')]
[('the', 'DT')]
[('and', 'CC')]
[('are', 'VBP')]
[('and', 'CC')]
[('of', 'IN')]
[('are', 'VBP')]
[('very', 'RB')]
[('to', 'TO')]
[('are', 'VBP')]
[('and', 'CC')]
[('and', 'CC')]
[('is', 'VBZ')]
[('an', 'DT')]
[('of', 'IN')]
[('for', 'IN')]
[('is', 'VBZ')]
[('in', 'IN')]
[('and', 'CC')]
[('and', 'CC')]
[('and', 'CC')]
[('in', 'IN')]
[('of', 'IN')]
[('is', 'VBZ')]
[('very', 'RB')]
[('to', 'TO')]
[('that', 'IN')]
[('of', 'IN')]
[('and', 'CC')]
[('are', 'VBP')]
[('in', 'IN')]
[('to', 'TO')]
[('the', 'DT')]
[('of', 'IN')]
[('the', 'DT')]
```

```
[('and', 'CC')]
[('it', 'PRP')]
[('into', 'IN')]
[('then', 'RB')]
[('other', 'JJ')]
[('to', 'TO')]
[('and', 'CC')]
[('patterns', 'NNS')]
[('of', 'IN')]
[('for', 'IN')]
[('on', 'IN')]
[('the', 'DT')]
[('and', 'CC')]
[('and', 'CC')]
[('a', 'DT')]
[('in', 'IN')]
[('in', 'IN')]
[('and', 'CC')]
[('of', 'IN')]
[('in', 'IN')]
[('and', 'CC')]
[('of', 'IN')]
[('in', 'IN')]
[('and', 'CC')]
[('of', 'IN')]
[('of', 'IN')]
[('for', 'IN')]
[('on', 'IN')]
[('the', 'DT')]
[('a', 'DT')]
[('of', 'IN')]
[('but', 'CC')]
[('its', 'PRP$')]
[('because', 'IN')]
[('it', 'PRP')]
[('has', 'VBZ')]
[('to', 'TO')]
[('this', 'DT')]
[('by', 'IN')]
[('more', 'RBR')]
[('how', 'WRB')]
[('and', 'CC')]
[('when', 'WRB')]
[('to', 'TO')]
[('this', 'DT')]
[('of', 'IN')]
[('with', 'IN')]
[('a', 'DT')]
```

```
[('on', 'IN')]
[('the', 'DT')]
[('in', 'IN')]
[('and', 'CC')]
[('the', 'DT')]
[('of', 'IN')]
[('and', 'CC')]
[('and', 'CC')]
[('in', 'IN')]
[('and', 'CC')]
[('from', 'IN')]
[('on', 'IN')]
[('on', 'IN')]
[('the', 'DT')]
[('in', 'IN')]
[('and', 'CC')]
[('from', 'IN')]
[('on', 'IN')]
[('on', 'IN')]
[('the', 'DT')]
[('in', 'IN')]
[('of', 'IN')]
[('of', 'IN')]
[('the', 'DT')]
[('and', 'CC')]
[('and', 'CC')]
[('on', 'IN')]
[('the', 'DT')]
[('of', 'IN')]
[('the', 'DT')]
[('on', 'IN')]
[('with', 'IN')]
[('and', 'CC')]
[('for', 'IN')]
[('of', 'IN')]
[('and', 'CC')]
[('and', 'CC')]
[('and', 'CC')]
[('on', 'IN')]
[('and', 'CC')]
[('and', 'CC')]
[('from', 'IN')]
[('and', 'CC')]
[('to', 'TO')]
[('of', 'IN')]
[('for', 'IN')]
[('on', 'IN')]
[('and', 'CC')]
```

```
[('on', 'IN')]
[('of', 'IN')]
[('the', 'DT')]
[('and', 'CC')]
[('on', 'IN')]
[('from', 'IN')]
[('of', 'IN')]
[('and', 'CC')]
[('for', 'IN')]
[('in', 'IN')]
[('of', 'IN')]
[('on', 'IN')]
[('as', 'IN')]
[('a', 'DT')]
[('to', 'TO')]
[('and', 'CC')]
[('and', 'CC')]
[('of', 'IN')]
[('the', 'DT')]
[('in', 'IN')]
[('of', 'IN')]
[('and', 'CC')]
[('as', 'IN')]
[('a', 'DT')]
[('for', 'IN')]
[('a', 'DT')]
[('and', 'CC')]
[('for', 'IN')]
[('in', 'IN')]
[('in', 'IN')]
[('to', 'TO')]
[('a', 'DT')]
[('s', 'NN')]
[('from', 'IN')]
[('from', 'IN')]
[('from', 'IN')]
[('a', 'DT')]
[('from', 'IN')]
[('from', 'IN')]
[('to', 'TO')]
[('be', 'VB')]
[('from', 'IN')]
[('to', 'TO')]
[('be', 'VB')]
[('from', 'IN')]
[('in', 'IN')]
[('to', 'TO')]
[('this', 'DT')]
```

```
[('a', 'DT')]
[('as', 'IN')]
[('was', 'VBD')]
[('on', 'IN')]
[('at', 'IN')]
[('is', 'VBZ')]
[('under', 'IN')]
[('the', 'DT')]
[('this', 'DT')]
[('you', 'PRP')]
[('to', 'TO')]
[('the', 'DT')]
[('of', 'IN')]
[('and', 'CC')]
[('is', 'VBZ')]
[('a', 'DT')]
[('of', 'IN')]
[('the', 'DT')]
[('a', 'DT')]
```

Ques3: Write a program to extract the contents (excluding any tags) from two websites (https://en.wikipedia.org/wiki/Web_mining & https://en.wikipedia.org/wiki/Data_mining) and save the content in two separate .doc file. Remove stopwords from the content and represent the documents using Boolean, Bag-of-words and Complete representation. Process a search a query and compare the contents of the both pages with the processed query, display the similarity result based on highest matching count (bag-of-words).

[32]:

```python
url1 = requests.get("https://en.wikipedia.org/wiki/Web_mining")
url2 = requests.get("https://en.wikipedia.org/wiki/Data_mining")
```

[33]:

```python
soup1 = BeautifulSoup(url1.text)
soup2 = BeautifulSoup(url2.text)
for script in soup1(["script", "style"]):
    script.decompose()
for script in soup2(["script", "style"]):
    script.decompose()
text1 = soup1.get_text()
text2 = soup2.get_text()

lines1 = (line.strip() for line in text1.splitlines())
chunks1 = (phrase.strip() for line in lines1 for phrase in line.split("  "))
text1 = '\n'.join(chunk for chunk in chunks1 if chunk)
text1 = text1.lower()

lines2 = (line.strip() for line in text2.splitlines())
chunks2 = (phrase.strip() for line in lines2 for phrase in line.split("  "))
text2 = '\n'.join(chunk for chunk in chunks2 if chunk)
text2 = text2.lower()
```

[34]:
```
doc1 = open("web_mining.doc", "w", encoding='utf-8')
doc1.write(text1)
doc1.close()
```

[35]:
```
doc2 = open("data_mining.doc", "w", encoding='utf-8')
doc2.write(text2)
doc2.close()
```

[36]:
```
stop_words = set(stopwords.words('english'))
word_tokens1 = word_tokenize(text1)
word_tokens2 = word_tokenize(text2)
```

[37]:
```
filt_text1 = [w for w in word_tokens1 if not w in stop_words]
filt_text2 = [w for w in word_tokens2 if not w in stop_words]
```

[38]:
```
all_words = filt_text1 + filt_text2
```

[39]:
```
all_words = list(dict.fromkeys(all_words))
```

[40]:
```
word_in_doc1 = [None]*len(all_words)
word_in_doc2 = [None]*len(all_words)
```

[41]:
```
k = 0
for i in all_words:
    for j in filt_text1:
        if i==j:
            word_in_doc1[k] = 1
            break
    if word_in_doc1[k] != 1:
        word_in_doc1[k] = 0

    for j in filt_text2:
        if i==j:
            word_in_doc2[k] = 1
            break
    if word_in_doc2[k] != 1:
        word_in_doc2[k] = 0
    k+=1
```

[42]:
```
import pandas as pd
```

[43]:
```
boolean_rep = pd.DataFrame({"Words": all_words, "Web Mining Doc": word_in_doc1,␣
 ↪"Data Mining Doc": word_in_doc2})
```

[44]:
```
boolean_rep.head()
```

[44]:

```
       Words  Web Mining Doc  Data Mining Doc
0        web               1                1
1     mining               1                1
2          -               1                1
3  wikipedia               1                1
4          ,               1                1
```

[45]:

```
doc1 = open("web_mining.doc", "r", encoding='utf-8')
doc2 = open("data_mining.doc", "r", encoding='utf-8')
```

[46]:

```
groups1 = {}
groups2 = {}
row = 0
col = 0
#for line in doc1:
#    temp = line.split()
#    row+=1
for word1 in all_words:
    for line in doc1:
        row+=1
        temp = line.split()
        for word2 in temp:
            if(word1 == word2):
                if word1 not in groups1.keys():
                    groups1[word1] = list()
                groups1[word1].append((row, col+1))
            col+=len(word2)+1
        col=0
    if word1 not in groups1.keys():
        groups1[word1] = list()
    doc1.seek(0, 0)

    row=0

    for line in doc2:
        row+=1
        temp = line.split()
        for word2 in temp:
            if(word1 == word2):
                if word1 not in groups2.keys():
                    groups2[word1] = list()
                groups2[word1].append((row, col+1))
            col+=len(word2)+1
        col=0
    if word1 not in groups2.keys():
        groups2[word1] = list()
    doc2.seek(0, 0)
    row=0
```

[47]:
```
complete_repr = pd.DataFrame({'Word': all_words, 'Position in Web Mining Doc':␣
 ↪list(groups1.values()), 'Position in Data Mining Doc': list(groups2.
 ↪values())})
```

[48]:
```
complete_repr.head()
```

[48]:
```
        Word                           Position in Web Mining Doc  \
0        web  [(1, 1), (2, 1), (7, 1), (8, 13), (8, 78), (8,…
1     mining  [(1, 5), (2, 5), (7, 5), (7, 39), (7, 157), (8…
2          -                    [(1, 12), (145, 52), (146, 40)]
3  wikipedia           [(1, 14), (122, 23), (195, 7), (197, 9)]
4          ,                                                 []

                        Position in Data Mining Doc
0  [(203, 90), (213, 49), (280, 1), (302, 1), (35…
1  [(1, 6), (2, 6), (6, 26), (99, 6), (99, 172), …
2  [(1, 13), (207, 575), (218, 6), (327, 37), (33…
3                    [(1, 15), (647, 7), (649, 9)]
4                                               []
```

[49]:
```
count1 = [None]*complete_repr.shape[0]
count2 = [None]*complete_repr.shape[0]
for i in range(complete_repr.shape[0]):
    count1[i] = len(complete_repr.loc[i, "Position in Web Mining Doc"])
    count2[i] = len(complete_repr.loc[i, "Position in Data Mining Doc"])
```

[50]:
```
bag_of_words = pd.DataFrame({'Word': all_words, 'Count in Web Mining Doc':␣
 ↪count1, 'Count in Data Mining Doc': count2})
```

[52]:
```
bag_of_words.head()
```

[52]:
```
        Word  Count in Web Mining Doc  Count in Data Mining Doc
0        web                      102                         6
1     mining                       70                       115
2          -                        3                         7
3  wikipedia                        4                         3
4          ,                        0                         0
```

[53]:
```
search_word = str(input("Enter a sentence to search: "))
```

```
Enter a sentence to search: web mining wikipedia
```

[54]:
```
search_word = search_word.split()
search_word
```

[54]:
```
['web', 'mining', 'wikipedia']
```

[55]:
```
count1 = 0
count2 = 0
for i in search_word:
    indx = bag_of_words.loc[bag_of_words['Word'] == i]
    count1+=indx["Count in Web Mining Doc"][indx.index.tolist()[0]]
    count2+=indx["Count in Data Mining Doc"][indx.index.tolist()[0]]
print("Frequency of search query in Web Mining Doc is", count1)
print("Frequency of search query in Data Mining Doc is", count2)
```

```
Frequency of search query in Web Mining Doc is 176
Frequency of search query in Data Mining Doc is 124
```

Quest4: Write a program to show the implementation of sentence paraphrasing through synonyms (retaining semantic meaning) for the following four sentences. Display at least three other paraphrased sentences for each sentence mentioned below. a. The quick brown fox jumps over the lazy dog b. Obama and Putin met the previous week c. At least 12 people were killed in the battle last week d. I will go home and come back tomorrow.

[56]:
```
from nltk.tokenize import word_tokenize
from nltk.tag import pos_tag
from nltk.corpus import wordnet as wn
import random

def tag(sentence):
    words = word_tokenize(sentence)
    words = pos_tag(words)
    return words

def paraphraseable(tag):
    return tag.startswith('NN') or tag == 'VB' or tag.startswith('JJ')

def pos(tag):
    if tag.startswith('NN'):
        return wn.NOUN
    elif tag.startswith('V'):
        return wn.VERB

def synonyms(word, tag):
    lemma_lists = [ss.lemmas() for ss in wn.synsets(word, pos(tag))]
    lemmas = [lemma.name() for lemma in sum(lemma_lists, [])]
    return set(lemmas)


def question(sentence):
    directory = {}
    for (word, t) in tag(sentence):
        if paraphraseable(t):
            syns = synonyms(word, t)
            if syns:
                if len(syns) > 1:
                    directory[word] = list(syns)
                    continue
        directory[word] = []
    new_sentence = ["", "", ""]
    for i in range(3):
        for word in sentence.split():
            if len(directory[word]) == 0:
                new_sentence[i] = new_sentence[i]+word+" "
            else:
                new_sentence[i] = new_sentence[i]+random.
↪choice(directory[word])+" "
    print("Paraphrase for", "\"",sentence, "\"", "----->")
    for i in new_sentence:
        print(i)
```

[57]:

```
question("The quick brown fox jumps over the lazy dog")
question("Obama and Putin met the previous week")
question("At least 12 people were killed in the battle last week")
question(" I will go home and come back tomorrow")
```

```
Paraphrase for " The quick brown fox jumps over the lazy dog " ----->
The prompt Brown_University slyboots jumps over the lazy dog
The promptly John_Brown George_Fox jumps over the work-shy dog-iron
The immediate brownness dodger jumps over the slothful cad
Paraphrase for " Obama and Putin met the previous week " ----->
Obama and Putin met the late week
Obama and Vladimir_Vladimirovich_Putin met the old workweek
Obama and Vladimir_Putin met the late hebdomad
Paraphrase for " At least 12 people were killed in the battle last week " ----->
At least 12 people were killed in the conflict endure workweek
At least 12 mass were killed in the fight final week
At least 12 hoi_polloi were killed in the battle in_conclusion week
Paraphrase for "  I will go home and come back tomorrow " ----->
I will endure home_plate and add_up back tomorrow
I will proceed dwelling and amount back tomorrow
I will hold_up base and come_up back tomorrow
```

[ ]: