Importing Modules

In [1]:

```
pip install goose3
```

```
Requirement already satisfied: goose3 in /Users/jasjitsingh/opt/anaconda3/l
ib/python3.9/site-packages (3.1.12)
Requirement already satisfied: lxml in /Users/jasjitsingh/opt/anaconda3/lib
/python3.9/site-packages (from goose3) (4.6.3)
Requirement already satisfied: requests in /Users/jasjitsingh/opt/anaconda3
/lib/python3.9/site-packages (from goose3) (2.26.0)
Requirement already satisfied: pyahocorasick in /Users/jasjitsingh/opt/anac
onda3/lib/python3.9/site-packages (from goose3) (1.4.4)
Requirement already satisfied: langdetect in /Users/jasjitsingh/opt/anacond
a3/lib/python3.9/site-packages (from goose3) (1.0.9)
Requirement already satisfied: python-dateutil in /Users/jasjitsingh/opt/an
aconda3/lib/python3.9/site-packages (from goose3) (2.8.2)
Requirement already satisfied: beautifulsoup4 in /Users/jasjitsingh/opt/ana
conda3/lib/python3.9/site-packages (from goose3) (4.10.0)
Requirement already satisfied: Pillow in /Users/jasjitsingh/opt/anaconda3/l
ib/python3.9/site-packages (from goose3) (8.4.0)
Requirement already satisfied: cssselect in /Users/jasjitsingh/opt/anaconda
3/lib/python3.9/site-packages (from goose3) (1.2.0)
Requirement already satisfied: soupsieve>1.2 in /Users/jasjitsingh/opt/anac
onda3/lib/python3.9/site-packages (from beautifulsoup4->goose3) (2.2.1)
Requirement already satisfied: six in /Users/jasjitsingh/opt/anaconda3/lib/
python3.9/site-packages (from langdetect->goose3) (1.16.0)
Requirement already satisfied: certifi>=2017.4.17 in /Users/jasjitsingh/opt
/anaconda3/lib/python3.9/site-packages (from requests->goose3) (2021.10.8)
Requirement already satisfied: urllib3<1.27,>=1.21.1 in /Users/jasjitsingh/
opt/anaconda3/lib/python3.9/site-packages (from requests->goose3) (1.26.7)
Requirement already satisfied: idna<4,>=2.5 in /Users/jasjitsingh/opt/anaco
nda3/lib/python3.9/site-packages (from requests->goose3) (3.2)
Requirement already satisfied: charset-normalizer~=2.0.0 in /Users/jasjitsi
ngh/opt/anaconda3/lib/python3.9/site-packages (from requests->goose3) (2.0.
4)
Note: you may need to restart the kernel to use updated packages.
```

In [2]:

```
import nltk
nltk.download('punkt')

import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import string
import sklearn
#for fetching data from website, Using goose.
from goose3 import Goose
```

```
[nltk_data] Downloading package punkt to
[nltk_data]     /Users/jasjitsingh/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
```

In [3]:

```
url = "https://en.wikipedia.org/wiki/Machine_learning"
```

In [4]:

```
article = Goose().extract(url)
```

In [5]:

```
article
```

Out[5]:

```
<goose3.article.Article at 0x12c6e48b0>
```

In [ ]:

```

```

In [ ]:

```

```

Data set ('Machine Learning')

In [6]:

```
article.title
```

Out[6]:

```
'Machine learning - Wikipedia'
```

In [7]:

```
Mydataset = article.cleaned_text
```

In [8]:

```
Mydataset
```

Out[8]:

```
'Study of algorithms that improve automatically through experience\n\nMachi
ne learning (ML) is a field of inquiry devoted to understanding and buildin
g methods that \'learn\', that is, methods that leverage data to improve pe
rformance on some set of tasks.[1] It is seen as a part of artificial intel
ligence. Machine learning algorithms build a model based on sample data, kn
own as training data, in order to make predictions or decisions without bei
ng explicitly programmed to do so.[2] Machine learning algorithms are used
in a wide variety of applications, such as in medicine, email filtering, sp
eech recognition, agriculture, and computer vision, where it is difficult o
r unfeasible to develop conventional algorithms to perform the needed tasks
.[3][4] A subset of machine learning is closely related to computational st
atistics, which focuses on making predictions using computers, but not all
machine learning is statistical learning. The study of mathematical optimiz
ation delivers methods, theory and application domains to the field of mach
ine learning. Data mining is a related field of study, focusing on explorat
ory data analysis through unsupervised learning.[6][7] Some implementations
of machine learning use data and neural networks in a way that mimics the w
orking of a biological brain.[8][9] In its application across business prob
lems, machine learning is also referred to as predictive analytics.\n\nLear
ning algorithms work on the basis that strategies, algorithms, and inferenc
es that worked well in the past are likely to continue working well in the
future. These inferences can be obvious, such as "since the sun rose every
morning for the last 10,000 days, it will probably rise tomorrow morning as
well". They can be nuanced, such as "X% of families have geographically sep
arate species with color variants, so there is a Y% chance that undiscovere
d black swans exist".\n\nMachine learning programs can perform tasks withou
t being explicitly programmed to do so. It involves computers learning from
data provided so that they carry out certain tasks. For simple tasks assign
ed to computers, it is possible to program algorithms telling the machine h

ow to execute all steps required to solve the problem at hand; on the compu
ter\'s part, no learning is needed. For more advanced tasks, it can be chal
lenging for a human to manually create the needed algorithms. In practice,
it can turn out to be more effective to help the machine develop its own al
gorithm, rather than having human programmers specify every needed step.[11
]\n\nThe discipline of machine learning employs various approaches to teach
computers to accomplish tasks where no fully satisfactory algorithm is avai
lable. In cases where vast numbers of potential answers exist, one approach
is to label some of the correct answers as valid. This can then be used as
training data for the computer to improve the algorithm(s) it uses to deter
mine correct answers. For example, to train a system for the task of digita
l character recognition, the MNIST dataset of handwritten digits has often
been used.[11]\n\nHistory and relationships to other fields\n\nThe term mac
hine learning was coined in 1959 by Arthur Samuel, an IBM employee and pion
eer in the field of computer gaming and artificial intelligence.[12][13] Al
so the synonym self-teaching computers were used in this time period.[14][1
5]\n\nBy the early 1960s an experimental "learning machine" with punched ta
pe memory, called Cybertron, had been developed by Raytheon Company to anal
```

yze sonar signals, electrocardiograms, and speech patterns using rudimentar
y reinforcement learning. It was repetitively "trained" by a human operator
/teacher to recognize patterns and equipped with a "goof" button to cause i
t to re-evaluate incorrect decisions.[16] A representative book on research
into machine learning during the 1960s was Nilsson\'s book on Learning Mach
ines, dealing mostly with machine learning for pattern classification.[17]
Interest related to pattern recognition continued into the 1970s, as descri
bed by Duda and Hart in 1973.[18] In 1981 a report was given on using teach
ing strategies so that a neural network learns to recognize 40 characters (
26 letters, 10 digits, and 4 special symbols) from a computer terminal.[19]
\n\nTom M. Mitchell provided a widely quoted, more formal definition of the
algorithms studied in the machine learning field: "A computer program is sa
id to learn from experience E with respect to some class of tasks T and per
formance measure P if its performance at tasks in T, as measured by P, impr
oves with experience E."[20] This definition of the tasks in which machine
learning is concerned offers a fundamentally operational definition rather
than defining the field in cognitive terms. This follows Alan Turing\'s pro
posal in his paper "Computing Machinery and Intelligence", in which the que
stion "Can machines think?" is replaced with the question "Can machines do
what we (as thinking entities) can do?".[21]\n\nModern-day machine learning
has two objectives, one is to classify data based on models which have been
developed, the other purpose is to make predictions for future outcomes bas
ed on these models. A hypothetical algorithm specific to classifying data m
ay use computer vision of moles coupled with supervised learning in order t
o train it to classify the cancerous moles. A machine learning algorithm fo
r stock trading may inform the trader of future potential predictions.[22]\
n\nAs a scientific endeavor, machine learning grew out of the quest for art
ificial intelligence. In the early days of AI as an academic discipline, so
me researchers were interested in having machines learn from data. They att
empted to approach the problem with various symbolic methods, as well as wh
at was then termed "neural networks"; these were mostly perceptrons and oth
er models that were later found to be reinventions of the generalized linea
r models of statistics.[25] Probabilistic reasoning was also employed, espe
cially in automated medical diagnosis.[26]: 488\n\nHowever, an increasing e
mphasis on the logical, knowledge-based approach caused a rift between AI a
nd machine learning. Probabilistic systems were plagued by theoretical and
practical problems of data acquisition and representation.[26]: 488 By 1980
, expert systems had come to dominate AI, and statistics was out of favor.[
27] Work on symbolic/knowledge-based learning did continue within AI, leadi
ng to inductive logic programming, but the more statistical line of researc
h was now outside the field of AI proper, in pattern recognition and inform
ation retrieval.[26]: 708–710, 755 Neural networks research had been abando
ned by AI and computer science around the same time. This line, too, was co
ntinued outside the AI/CS field, as "connectionism", by researchers from ot
her disciplines including Hopfield, Rumelhart, and Hinton. Their main succe
ss came in the mid-1980s with the reinvention of backpropagation.[26]: 25\n
\nMachine learning (ML), reorganized as a separate field, started to flouri
sh in the 1990s. The field changed its goal from achieving artificial intel
ligence to tackling solvable problems of a practical nature. It shifted foc
us away from the symbolic approaches it had inherited from AI, and toward m
ethods and models borrowed from statistics, fuzzy logic, and probability th
eory.[27]\n\nThe difference between ML and AI is frequently misunderstood.
ML learns and predicts based on passive observations, whereas AI implies an
agent interacting with the environment to learn and take actions that maxim
ize its chance of successfully achieving its goals.[28]\n\nAs of 2020, many
sources continue to assert that ML remains a subfield of AI.[29][30][27] Ot
hers have the view that not all ML is part of AI, but only an \'intelligent
subset\' of ML should be considered AI.[5][31][32]\n\nMachine learning and
data mining often employ the same methods and overlap significantly, but wh
ile machine learning focuses on prediction, based on known properties learn
ed from the training data, data mining focuses on the discovery of (previou
sly) unknown properties in the data (this is the analysis step of knowledge
discovery in databases). Data mining uses many machine learning methods, bu
t with different goals; on the other hand, machine learning also employs da
ta mining methods as "unsupervised learning" or as a preprocessing step to
improve learner accuracy. Much of the confusion between these two research
communities (which do often have separate conferences and separate journals
, ECML PKDD being a major exception) comes from the basic assumptions they
work with: in machine learning, performance is usually evaluated with respe
ct to the ability to reproduce known knowledge, while in knowledge discover

y and data mining (KDD) the key task is the discovery of previously unknown
knowledge. Evaluated with respect to known knowledge, an uninformed (unsupe
rvised) method will easily be outperformed by other supervised methods, whi
le in a typical KDD task, supervised methods cannot be used due to the unav
ailability of training data.\n\nMachine learning also has intimate ties to
optimization: many learning problems are formulated as minimization of some
loss function on a training set of examples. Loss functions express the dis
crepancy between the predictions of the model being trained and the actual
problem instances (for example, in classification, one wants to assign a la
bel to instances, and models are trained to correctly predict the pre-assig
ned labels of a set of examples).[33]\n\nThe difference between optimizatio
n and machine learning arises from the goal of generalization: while optimi
zation algorithms can minimize the loss on a training set, machine learning
is concerned with minimizing the loss on unseen samples. Characterizing the
generalization of various learning algorithms is an active topic of current
research, especially for deep learning algorithms.\n\nMachine learning and
statistics are closely related fields in terms of methods, but distinct in
their principal goal: statistics draws population inferences from a sample,
while machine learning finds generalizable predictive patterns.[34] Accordi
ng to Michael I. Jordan, the ideas of machine learning, from methodological
principles to theoretical tools, have had a long pre-history in statistics.
[35] He also suggested the term data science as a placeholder to call the o
verall field.[35]\n\nLeo Breiman distinguished two statistical modeling par
adigms: data model and algorithmic model,[29] wherein "algorithmic model" m
eans more or less the machine learning algorithms like Random forest.\n\nSo
me statisticians have adopted methods from machine learning, leading to a c
ombined field that they call statistical learning.[30]\n\nA core objective
of a learner is to generalize from its experience.[5][31] Generalization in
this context is the ability of a learning machine to perform accurately on
new, unseen examples/tasks after having experienced a learning data set. Th
e training examples come from some generally unknown probability distributi
on (considered representative of the space of occurrences) and the learner
has to build a general model about this space that enables it to produce su
fficiently accurate predictions in new cases.\n\nThe computational analysis
of machine learning algorithms and their performance is a branch of theoret
ical computer science known as computational learning theory via the Probab
ly Approximately Correct Learning (PAC) model. Because training sets are fi
nite and the future is uncertain, learning theory usually does not yield gu
arantees of the performance of algorithms. Instead, probabilistic bounds on
the performance are quite common. The bias–variance decomposition is one wa
y to quantify generalization error.\n\nFor the best performance in the cont
ext of generalization, the complexity of the hypothesis should match the co
mplexity of the function underlying the data. If the hypothesis is less com
plex than the function, then the model has under fitted the data. If the co
mplexity of the model is increased in response, then the training error dec
reases. But if the hypothesis is too complex, then the model is subject to
overfitting and generalization will be poorer.[32]\n\nIn addition to perfor
mance bounds, learning theorists study the time complexity and feasibility
of learning. In computational learning theory, a computation is considered
feasible if it can be done in polynomial time. There are two kinds of time
complexity results: Positive results show that a certain class of functions
can be learned in polynomial time. Negative results show that certain class
es cannot be learned in polynomial time.\n\nMachine learning approaches are
traditionally divided into three broad categories, which correspond to lear
ning paradigms, depending on the nature of the "signal" or "feedback" avail
able to the learning system:\n• Supervised learning: The computer is presen
ted with example inputs and their desired outputs, given by a "teacher", an
d the goal is to learn a general rule that maps inputs to outputs.\n• Unsup
ervised learning: No labels are given to the learning algorithm, leaving it
on its own to find structure in its input. Unsupervised learning can be a g
oal in itself (discovering hidden patterns in data) or a means towards an e
nd (feature learning).\n• Reinforcement learning: A computer program intera
cts with a dynamic environment in which it must perform a certain goal (suc
h as driving a vehicle or playing a game against an opponent). As it naviga
tes its problem space, the program is provided feedback that\'s analogous t
o rewards, which it tries to maximize.[5]\n\nSupervised learning algorithms
build a mathematical model of a set of data that contains both the inputs a
nd the desired outputs.[36] The data is known as training data, and consist
s of a set of training examples. Each training example has one or more inpu
ts and the desired output, also known as a supervisory signal. In the mathe

matical model, each training example is represented by an array or vector,
sometimes called a feature vector, and the training data is represented by
a matrix. Through iterative optimization of an objective function, supervis
ed learning algorithms learn a function that can be used to predict the out
put associated with new inputs.[37] An optimal function will allow the algo
rithm to correctly determine the output for inputs that were not a part of
the training data. An algorithm that improves the accuracy of its outputs o
r predictions over time is said to have learned to perform that task.[20]\n
\nTypes of supervised-learning algorithms include active learning, classifi
cation and regression.[28] Classification algorithms are used when the outp
uts are restricted to a limited set of values, and regression algorithms ar
e used when the outputs may have any numerical value within a range. As an
example, for a classification algorithm that filters emails, the input woul
d be an incoming email, and the output would be the name of the folder in w
hich to file the email.\n\nSimilarity learning is an area of supervised mac
hine learning closely related to regression and classification, but the goa
l is to learn from examples using a similarity function that measures how s
imilar or related two objects are. It has applications in ranking, recommen
dation systems, visual identity tracking, face verification, and speaker ve
rification.\n\nUnsupervised learning algorithms take a set of data that con
tains only inputs, and find structure in the data, like grouping or cluster
ing of data points. The algorithms, therefore, learn from test data that ha
s not been labeled, classified or categorized. Instead of responding to fee
dback, unsupervised learning algorithms identify commonalities in the data
and react based on the presence or absence of such commonalities in each ne
w piece of data. A central application of unsupervised learning is in the f
ield of density estimation in statistics, such as finding the probability d
ensity function.[38] Though unsupervised learning encompasses other domains
involving summarizing and explaining data features.\n\nCluster analysis is
the assignment of a set of observations into subsets (called clusters) so t
hat observations within the same cluster are similar according to one or mo
re predesignated criteria, while observations drawn from different clusters
are dissimilar. Different clustering techniques make different assumptions
on the structure of the data, often defined by some similarity metric and e
valuated, for example, by internal compactness, or the similarity between m
embers of the same cluster, and separation, the difference between clusters
. Other methods are based on estimated density and graph connectivity.\n\nS
emi-supervised learning falls between unsupervised learning (without any la
beled training data) and supervised learning (with completely labeled train
ing data). Some of the training examples are missing training labels, yet m
any machine-learning researchers have found that unlabeled data, when used
in conjunction with a small amount of labeled data, can produce a considera
ble improvement in learning accuracy.\n\nIn weakly supervised learning, the
training labels are noisy, limited, or imprecise; however, these labels are
often cheaper to obtain, resulting in larger effective training sets.[39]\n
\nReinforcement learning is an area of machine learning concerned with how
software agents ought to take actions in an environment so as to maximize s
ome notion of cumulative reward. Due to its generality, the field is studie
d in many other disciplines, such as game theory, control theory, operation
s research, information theory, simulation-based optimization, multi-agent
systems, swarm intelligence, statistics and genetic algorithms. In machine
learning, the environment is typically represented as a Markov decision pro
cess (MDP). Many reinforcement learning algorithms use dynamic programming
techniques.[40] Reinforcement learning algorithms do not assume knowledge o
f an exact mathematical model of the MDP, and are used when exact models ar
e infeasible. Reinforcement learning algorithms are used in autonomous vehi
cles or in learning to play a game against a human opponent.\n\nDimensional
ity reduction is a process of reducing the number of random variables under
consideration by obtaining a set of principal variables.[41] In other words
, it is a process of reducing the dimension of the feature set, also called
the "number of features". Most of the dimensionality reduction techniques c
an be considered as either feature elimination or extraction. One of the po
pular methods of dimensionality reduction is principal component analysis (
PCA). PCA involves changing higher-dimensional data (e.g., 3D) to a smaller
space (e.g., 2D). This results in a smaller dimension of data (2D instead o
f 3D), while keeping all original variables in the model without changing t
he data.[42] The manifold hypothesis proposes that high-dimensional data se
ts lie along low-dimensional manifolds, and many dimensionality reduction t
echniques make this assumption, leading to the area of manifold learning an
d manifold regularization.\n\nOther approaches have been developed which do

a manifold regularization.\n\nOther approaches have been developed which don\'t fit neatly into this three-fold categorization, and sometimes more than one is used by the same machine learning system. For example topic modeling, meta-learning.[43]\n\nAs of 2022, deep learning is the dominant approach for much ongoing work in the field of machine learning.[11]\n\nSelf-learning, as a machine learning paradigm was introduced in 1982 along with a neural network capable of self-learning, named crossbar adaptive array (CAA).[44] It is learning with no external rewards and no external teacher advice. The CAA self-learning algorithm computes, in a crossbar fashion, both decisions about actions and emotions (feelings) about consequence situations. The system is driven by the interaction between cognition and emotion.[45] The self-learning algorithm updates a memory matrix W =||w(a,s)|| such that in each iteration executes the following machine learning routine:\n• compute emotion of being in consequence situation v(s\')\n• update crossbar memory w\'(a,s) = w(a,s) + v(s\')\n\nIt is a system with only one input, situation, and only one output, action (or behavior) a. There is neither a separate reinforcement input nor an advice input from the environment. The backpropagated value (secondary reinforcement) is the emotion toward the consequence situation. The CAA exists in two environments, one is the behavioral environment where it behaves, and the other is the genetic environment, wherefrom it initially and only once receives initial emotions about situations to be encountered in the behavioral environment. After receiving the genome (species) vector from the genetic environment, the CAA learns a goal-seeking behavior, in an environment that contains both desirable and undesirable situations.[46]\n\nSeveral learning algorithms aim at discovering better representations of the inputs provided during training.[47] Classic examples include principal components analysis and cluster analysis. Feature learning algorithms, also called representation learning algorithms, often attempt to preserve the information in their input but also transform it in a way that makes it useful, often as a pre-processing step before performing classification or predictions. This technique allows reconstruction of the inputs coming from the unknown data-generating distribution, while not being necessarily faithful to configurations that are implausible under that distribution. This replaces manual feature engineering, and allows a machine to both learn the features and use them to perform a specific task.\n\nFeature learning can be either supervised or unsupervised. In supervised feature learning, features are learned using labeled input data. Examples include artificial neural networks, multilayer perceptrons, and supervised dictionary learning. In unsupervised feature learning, features are learned with unlabeled input data. Examples include dictionary learning, independent component analysis, autoencoders, matrix factorization[48] and various forms of clustering.[49][50][51]\n\nManifold learning algorithms attempt to do so under the constraint that the learned representation is low-dimensional. Sparse coding algorithms attempt to do so under the constraint that the learned representation is sparse, meaning that the mathematical model has many zeros. Multilinear subspace learning algorithms aim to learn low-dimensional representations directly from tensor representations for multidimensional data, without reshaping them into higher-dimensional vectors.[52] Deep learning algorithms discover multiple levels of representation, or a hierarchy of features, with higher-level, more abstract features defined in terms of (or generating) lower-level features. It has been argued that an intelligent machine is one that learns a representation that disentangles the underlying factors of variation that explain the observed data.[53]\n\nFeature learning is motivated by the fact that machine learning tasks such as classification often require input that is mathematically and computationally convenient to process. However, real-world data such as images, video, and sensory data has not yielded attempts to algorithmically define specific features. An alternative is to discover such features or representations through examination, without relying on explicit algorithms.\n\nSparse dictionary learning is a feature learning method where a training example is represented as a linear combination of basis functions, and is assumed to be a sparse matrix. The method is strongly NP-hard and difficult to solve approximately.[54] A popular heuristic method for sparse dictionary learning is the K-SVD algorithm. Sparse dictionary learning has been applied in several contexts. In classification, the problem is to determine the class to which a previously unseen training example belongs. For a dictionary where each class has already been built, a new training example is associated with the class that is best sparsely represented by the corresponding dictionary. Sparse dictionary learning has also been applied in image de-noising. The key idea is that a clean image patch can be sparsely represented by an image dictionary, but the noise cannot.[55]\n\nIn data mining, anomaly detection, also known as o

the noise cannot.[55]\n\nIn data mining, anomaly detection, also known as o
utlier detection, is the identification of rare items, events or observatio
ns which raise suspicions by differing significantly from the majority of t
he data.[56] Typically, the anomalous items represent an issue such as bank
fraud, a structural defect, medical problems or errors in a text. Anomalies
are referred to as outliers, novelties, noise, deviations and exceptions.[5
7]\n\nIn particular, in the context of abuse and network intrusion detectio
n, the interesting objects are often not rare objects, but unexpected burst
s of inactivity. This pattern does not adhere to the common statistical def
inition of an outlier as a rare object. Many outlier detection methods (in
particular, unsupervised algorithms) will fail on such data unless aggregat
ed appropriately. Instead, a cluster analysis algorithm may be able to dete
ct the micro-clusters formed by these patterns.[58]\n\nThree broad categori
es of anomaly detection techniques exist.[59] Unsupervised anomaly detectio
n techniques detect anomalies in an unlabeled test data set under the assum
ption that the majority of the instances in the data set are normal, by loo
king for instances that seem to fit the least to the remainder of the data
set. Supervised anomaly detection techniques require a data set that has be
en labeled as "normal" and "abnormal" and involves training a classifier (t
he key difference to many other statistical classification problems is the
inherently unbalanced nature of outlier detection). Semi-supervised anomaly
detection techniques construct a model representing normal behavior from a
given normal training data set and then test the likelihood of a test insta
nce to be generated by the model.\n\nRobot learning is inspired by a multit
ude of machine learning methods, starting from supervised learning, reinfor
cement learning,[60][61] and finally meta-learning (e.g. MAML).\n\nAssociat
ion rule learning is a rule-based machine learning method for discovering r
elationships between variables in large databases. It is intended to identi
fy strong rules discovered in databases using some measure of "interestingn
ess".[62]\n\nRule-based machine learning is a general term for any machine
learning method that identifies, learns, or evolves "rules" to store, manip
ulate or apply knowledge. The defining characteristic of a rule-based machi
ne learning algorithm is the identification and utilization of a set of rel
ational rules that collectively represent the knowledge captured by the sys
tem. This is in contrast to other machine learning algorithms that commonly
identify a singular model that can be universally applied to any instance i
n order to make a prediction.[63] Rule-based machine learning approaches in
clude learning classifier systems, association rule learning, and artificia
l immune systems.\n\nBased on the concept of strong rules, Rakesh Agrawal,
Tomasz Imieliński and Arun Swami introduced association rules for discoveri
ng regularities between products in large-scale transaction data recorded b
y point-of-sale (POS) systems in supermarkets.[64] For example, the rule fo
und in the sales data of a supermarket would indicate that if a customer bu
ys onions and potatoes together, they are likely to also buy hamburger meat
. Such information can be used as the basis for decisions about marketing a
ctivities such as promotional pricing or product placements. In addition to
market basket analysis, association rules are employed today in application
areas including Web usage mining, intrusion detection, continuous productio
n, and bioinformatics. In contrast with sequence mining, association rule l
earning typically does not consider the order of items either within a tran
saction or across transactions.\n\nLearning classifier systems (LCS) are a
family of rule-based machine learning algorithms that combine a discovery c
omponent, typically a genetic algorithm, with a learning component, perform
ing either supervised learning, reinforcement learning, or unsupervised lea
rning. They seek to identify a set of context-dependent rules that collecti
vely store and apply knowledge in a piecewise manner in order to make predi
ctions.[65]\n\nInductive logic programming (ILP) is an approach to rule lea
rning using logic programming as a uniform representation for input example
s, background knowledge, and hypotheses. Given an encoding of the known bac
kground knowledge and a set of examples represented as a logical database o
f facts, an ILP system will derive a hypothesized logic program that entail
s all positive and no negative examples. Inductive programming is a related
field that considers any kind of programming language for representing hypo
theses (and not only logic programming), such as functional programs.\n\nIn
ductive logic programming is particularly useful in bioinformatics and natu
ral language processing. Gordon Plotkin and Ehud Shapiro laid the initial t
heoretical foundation for inductive machine learning in a logical setting.[
66][67][68] Shapiro built their first implementation (Model Inference Syste
m) in 1981: a Prolog program that inductively inferred logic programs from
positive and negative examples.[69] The term inductive here refers to philo

sophical induction, suggesting a theory to explain observed facts, rather than mathematical induction, proving a property for all members of a well-ordered set.\n\nPerforming machine learning involves creating a model, which is trained on some training data and then can process additional data to make predictions. Various types of models have been used and researched for machine learning systems.\n\nArtificial neural networks (ANNs), or connectionist systems, are computing systems vaguely inspired by the biological neural networks that constitute animal brains. Such systems "learn" to perform tasks by considering examples, generally without being programmed with any task-specific rules.\n\nAn ANN is a model based on a collection of connected units or nodes called "artificial neurons", which loosely model the neurons in a biological brain. Each connection, like the synapses in a biological brain, can transmit information, a "signal", from one artificial neuron to another. An artificial neuron that receives a signal can process it and then signal additional artificial neurons connected to it. In common ANN implementations, the signal at a connection between artificial neurons is a real number, and the output of each artificial neuron is computed by some non-linear function of the sum of its inputs. The connections between artificial neurons are called "edges". Artificial neurons and edges typically have a weight that adjusts as learning proceeds. The weight increases or decreases the strength of the signal at a connection. Artificial neurons may have a threshold such that the signal is only sent if the aggregate signal crosses that threshold. Typically, artificial neurons are aggregated into layers. Different layers may perform different kinds of transformations on their inputs. Signals travel from the first layer (the input layer) to the last layer (the output layer), possibly after traversing the layers multiple times.\n\nThe original goal of the ANN approach was to solve problems in the same way that a human brain would. However, over time, attention moved to performing specific tasks, leading to deviations from biology. Artificial neural networks have been used on a variety of tasks, including computer vision, speech recognition, machine translation, social network filtering, playing board and video games and medical diagnosis.\n\nDeep learning consists of multiple hidden layers in an artificial neural network. This approach tries to model the way the human brain processes light and sound into vision and hearing. Some successful applications of deep learning are computer vision and speech recognition.[70]\n\nDecision tree learning uses a decision tree as a predictive model to go from observations about an item (represented in the branches) to conclusions about the item\'s target value (represented in the leaves). It is one of the predictive modeling approaches used in statistics, data mining, and machine learning. Tree models where the target variable can take a discrete set of values are called classification trees; in these tree structures, leaves represent class labels, and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees. In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. In data mining, a decision tree describes data, but the resulting classification tree can be an input for decision-making.\n\nSupport-vector machines (SVMs), also known as support-vector networks, are a set of related supervised learning methods used for classification and regression. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that predicts whether a new example falls into one category.[71] An SVM training algorithm is a non-probabilistic, binary, linear classifier, although methods such as Platt scaling exist to use SVM in a probabilistic classification setting. In addition to performing linear classification, SVMs can efficiently perform a non-linear classification using what is called the kernel trick, implicitly mapping their inputs into high-dimensional feature spaces.\n\nRegression analysis encompasses a large variety of statistical methods to estimate the relationship between input variables and their associated features. Its most common form is linear regression, where a single line is drawn to best fit the given data according to a mathematical criterion such as ordinary least squares. The latter is often extended by regularization methods to mitigate overfitting and bias, as in ridge regression. When dealing with non-linear problems, go-to models include polynomial regression (for example, used for trendline fitting in Microsoft Excel[72]), logistic regression (often used in statistical classification) or even kernel regression, which introduces non-linearity by taking advantage of the kernel trick to implicitly map input variables to higher-dimensional space.\n\nA Bayesian network, belief network, or directed acyclic graphical model is a probabilistic graphical model that represents a set of random variables and their conditional independence with a directed acyclic graph (DAG). For example, a Bayesian network could represent the probabilistic relationships between diseases and symptoms. Given symptoms, the network can be used to compute the probabilities of the presence of various diseases. Efficient algorithms exist that perform inference and learning. Bayesian networks that model sequences of variables, like speech signals or protein sequences, are called dynamic Bayesian networks. Generalizations of Bayesian networks that can represent and solve decision problems under uncertainty are called influence diagrams.\n\nA Gaussian process is a stochastic process in which every finite collection of the random variables in the process has a multivariate normal distribution, and it relies on a pre-defined covariance function, or kernel, that models how pairs of points relate to each other depending on their locations.\n\nGiven a set of observed points, or input–output examples, the distribution of the (unobserved) output of a new point as function of its input data, can be directly computed by looking as the observed points and the covariances between those points and the new, unobserved point.\n\nGaussian processes are popular surrogate models in Bayesian optimization used to do hyperparameter optimization.\n\nA genetic algorithm (GA) is a search algorithm and heuristic technique that mimics the process of natural selection, using methods such as mutation and crossover to generate new genotypes in the hope of finding good solutions to a given problem. In machine learning, genetic algorithms were used in the 1980s and 1990s.[74][75] Conversely, machine learning techniques have been used to improve the performance of genetic and evolutionary algorithms.[76]\n\nTypically, machine learning models require a high quantity of reliable data in order for the models to perform accurate predictions. When training a machine learning model, machine learning engineers need to target and collect a large and representative sample of data. Data from the training set can be as varied as a corpus of text, a collection of images, sensor data, and data collected from individual users of a service. Overfitting is something to watch out for when training a machine learning model. Trained models derived from biased or non-evaluated data can result in skewed or undesired predictions. Bias models may result in detrimental outcomes thereby furthering the negative impacts on society or objectives. Algorithmic bias is a potential result of data not being fully prepared for training. Machine learning ethics is becoming a field of study and notably be integrated within machine learning engineering teams.\n\nFederated learning is an adapted form of distributed artificial intelligence to training machine learning models that decentralizes the training process, allowing for users\' privacy to be maintained by not needing to send their data to a centralized server. This also increases efficiency by decentralizing the training process to many devices. For example, Gboard uses federated machine learning to train search query prediction models on users\' mobile phones without having to send individual searches back to Google.[77]\n\nThere are many applications for machine learning, including:\n\nIn 2006, the media-services provider Netflix held the first "Netflix Prize" competition to find a program to better predict user preferences and improve the accuracy of its existing Cinematch movie recommendation algorithm by at least 10%. A joint team made up of researchers from AT&T Labs-Research in collaboration with the teams Big Chaos and Pragmatic Theory built an ensemble model to win the Grand Prize in 2009 for $1 million.[79] Shortly after the prize was awarded, Netflix realized that viewers\' ratings were not the best indicators of their viewing patterns ("everything is a recommendation") and they changed their recommendation engine accordingly.[80] In 2010 The Wall Street Journal wrote about the firm Rebellion Research and their use of machine learning to predict the financial crisis.[81] In 2012, co-founder of Sun Microsystems, Vinod Khosla, predicted that 80% of medical doctors jobs would be lost in the next two decades to automated machine learning medical diagnostic software.[82] In 2014, it was reported that a machine learning algorithm had been applied in the field of art history to study fine art paintings and that it may have revealed previously unrecognized influences among artists.[83] In 2019 Springer Nature published the first research book created using machine learning.[84] In 2020, machine learning technology was used to help make diagnoses and aid researchers in developing a cure for COVID-19.[85] Machine learning is recently applied to predict the green behavior of human-being.[86] Recently, machine learning technology is also applied to optimise smartphone\'s performance and thermal behaviour based on the user\'s interaction with the phone.[87][88][89]\n\nAlthough machine learning has been transformative in some fields, machine-learning programs often fail to deliver expected results.[90][91][92] Reasons for this are numerous: lack of (suitable) data, lack of access t o the data, data bias, privacy problems, badly chosen tasks and algorithms, wrong tools and people, lack of resources, and evaluation problems.[93]\n\nIn 2018, a self-driving car from Uber failed to detect a pedestrian, who was killed after a collision.[94] Attempts to use machine learning in healthcare with the IBM Watson system failed to deliver even after years of time and billions of dollars invested.[95][96]\n\nMachine learning has been used as a strategy to update the evidence related to a systematic review and increased reviewer burden related to the growth of biomedical literature. While it has improved with training sets, it has not yet developed sufficiently to reduce the workload burden without limiting the necessary sensitivity for the findings research themselves.[97]\n\nMachine learning approaches in particular can suffer from different data biases. A machine learning system trained specifically on current customers may not be able to predict the needs of new customer groups that are not represented in the training data. When trained on man-made data, machine learning is likely to pick up the constitutional and unconscious biases already present in society.[98] Language models learned from data have been shown to contain human-like biases.[99][100] Machine learning systems used for criminal risk assessment have been found to be biased against black people.[101][102] In 2015, Google photos would often tag black people as gorillas,[103] and in 2018 this still was not well resolved, but Google reportedly was still using the workaround to remove all gorillas from the training data, and thus was not able to recognize real gorillas at all.[104] Similar issues with recognizing non-white people have been found in many other systems.[105] In 2016, Microsoft tested a chatbot that learned from Twitter, and it quickly picked up racist and sexist language.[106] Because of such challenges, the effective use of machine learning may take longer to be adopted in other domains.[107] Concern for fairness in machine learning, that is, reducing bias in machine learning and propelling its use for human good is increasingly expressed by artificial intelligence scientists, including Fei-Fei Li, who reminds engineers that "There\'s nothing artificial about AI...It\'s inspired by people, it\'s created by people, and—most importantly—it impacts people. It is a powerful tool we are only just beginning to understand, and that is a profound responsibility."[108]\n\nExplainable AI (XAI), or Interpretable AI, or Explainable Machine Learning (XML), is artificial intelligence (AI) in which humans can understand the decisions or predictions made by the AI. It contrasts with the "black box" concept in machine learning where even its designers cannot explain why an AI arrived at a specific decision. By refining the mental models of users of AI-powered systems and dismantling their misconceptions, XAI promises to help users perform more effectively. XAI may be an implementation of the social right to explanation.\n\nSettling on a bad, overly complex theory gerrymandered to fit all the past training data is known as overfitting. Many systems attempt to reduce overfitting by rewarding a theory in accordance with how well it fits the data, but penalizing the theory in accordance with how complex the theory is.\n\nLearners can also disappoint by "learning the wrong lesson". A toy example is that an image classifier trained only on pictures of brown horses and black cats might conclude that all brown patches are likely to be horses. A real-world example is that, unlike humans, current image classifiers often don\'t primarily make judgments from the spatial relationship between components of the picture, and they learn relationships between pixels that humans are oblivious to, but that still correlate with images of certain types of real objects. Modifying these patterns on a legitimate image can result in "adversarial" images that the system misclassifies.[110][111]\n\nAdversarial vulnerabilities can also result in nonlinear systems, or from non-pattern perturbations. Some systems are so brittle that changing a single adversarial pixel predictably induces misclassification.[citation needed] Machine learning models are often vulnerable to manipulation and/or evasion via adversarial machine learning.[112]\n\nResearchers have demonstrated how backdoors can be placed undetectably into classifying (e.g. for categories "spam" and well-visible "not spam" of posts) machine learning models which are often developed and/or trained by third parties. Parties can change the classification of any input, including in cases for which a type of data/software transparency is provided, possibly including white-box access.[113][114][115]\n\nClassification of machine learning models can be validated by accuracy estimation techniques like the holdout method, which splits the data in a training and test set (conventionally 2/3 training set and 1/3 test set designation) and evaluates the performance of the training model on the test set. In comparison, the K-fold-cross-validation method randomly partitions the data into K subsets and then K experiments are performed each respectively considering 1 subset for evalua tion and the remaining K-1 subsets for training the model. In addition to the holdout and cross-validation methods, bootstrap, which samples n instances with replacement from the dataset, can be used to assess model accuracy.[116]\n\nIn addition to overall accuracy, investigators frequently report sensitivity and specificity meaning True Positive Rate (TPR) and True Negative Rate (TNR) respectively. Similarly, investigators sometimes report the false positive rate (FPR) as well as the false negative rate (FNR). However, these rates are ratios that fail to reveal their numerators and denominators. The total operating characteristic (TOC) is an effective method to express a model\'s diagnostic ability. TOC shows the numerators and denominators of the previously mentioned rates, thus TOC provides more information than the commonly used receiver operating characteristic (ROC) and ROC\'s associated area under the curve (AUC).[117]\n\nMachine learning poses a host of ethical questions. Systems that are trained on datasets collected with biases may exhibit these biases upon use (algorithmic bias), thus digitizing cultural prejudices.[118] For example, in 1988, the UK\'s Commission for Racial Equality found that St. George\'s Medical School had been using a computer program trained from data of previous admissions staff and this program had denied nearly 60 candidates who were found to be either women or had non-European sounding names.[98] Using job hiring data from a firm with racist hiring policies may lead to a machine learning system duplicating the bias by scoring job applicants by similarity to previous successful applicants.[119][120] Responsible collection of data and documentation of algorithmic rules used by a system thus is a critical part of machine learning.\n\nAI can be well-equipped to make decisions in technical fields, which rely heavily on data and historical information. These decisions rely on the objectivity and logical reasoning.[121] Because human languages contain biases, machines trained on language corpora will necessarily also learn these biases.[122][123]\n\nOther forms of ethical challenges, not related to personal biases, are seen in health care. There are concerns among health care professionals that these systems might not be designed in the public\'s interest but as income-generating machines.[124] This is especially true in the United States where there is a long-standing ethical dilemma of improving health care, but also increase profits. For example, the algorithms could be designed to provide patients with unnecessary tests or medication in which the algorithm\'s proprietary owners hold stakes. There is potential for machine learning in health care to provide professionals an additional tool to diagnose, medicate, and plan recovery paths for patients, but this requires these biases to be mitigated.[125]\n\nSince the 2010s, advances in both machine learning algorithms and computer hardware have led to more efficient methods for training deep neural networks (a particular narrow subdomain of machine learning) that contain many layers of non-linear hidden units.[126] By 2019, graphic processing units (GPUs), often with AI-specific enhancements, had displaced CPUs as the dominant method of training large-scale commercial cloud AI.[127] OpenAI estimated the hardware computing used in the largest deep learning projects from AlexNet (2012) to AlphaZero (2017), and found a 300,000-fold increase in the amount of compute required, with a doubling-time trendline of 3.4 months.[128][129]\n\nA physical neural network or Neuromorphic computer is a type of artificial neural network in which an electrically adjustable material is used to emulate the function of a neural synapse. "Physical" neural network is used to emphasize the reliance on physical hardware used to emulate neurons as opposed to software-based approaches. More generally the term is applicable to other artificial neural networks in which a memristor or other electrically adjustable resistance material is used to emulate a neural synapse.[130][131]\n\nEmbedded Machine Learning is a sub-field of machine learning, where the machine learning model is run on embedded systems with limited computing resources such as wearable computers, edge devices and microcontrollers.[132][133][134] Running machine learning model in embedded devices removes the need for transferring and storing data on cloud servers for further processing, henceforth, reducing data breaches and privacy leaks happening because of transferring data, and also minimizes theft of intellectual properties, personal data and business secrets. Embedded Machine Learning could be applied through several techniques including hardware acceleration,[135][136] using approximate computing,[137] optimization of machine learning models and many more.[138][139]\n\nSoftware suites containing a variety of machine learning algorithms include the following:\n• European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD)\n• International Conference on Computational Intelligence Methods for Bioinformatics and Biostatistics (CIBB)\n• Automated machine learning – Process of automating the ap

```
plication of machine learning\n• None Domingos, Pedro (September 22, 2015).
The Master Algorithm: How the Quest for the Ultimate Learning Machine Will
Remake Our World. Basic Books. ISBN .\n• None Nilsson, Nils (1998). . Morga
n Kaufmann. ISBN . Archived from the original on 26 July 2020 .\n• None Poo
le, David; Mackworth, Alan; Goebel, Randy (1998). Computational Intelligenc
e: A Logical Approach. New York: Oxford University Press. ISBN . Archived f
rom the original on 26 July 2020 .\n• mloss is an academic database of open
-source machine learning software.'
```

Tokenization

In [9]:

```python
tokens = nltk.word_tokenize(Mydataset)
len(tokens) #No. of tokens
```

Out[9]:

```
9026
```

In [10]:

```python
tokens[:100]
```

Out[10]:

```
['Study',
 'of',
 'algorithms',
 'that',
 'improve',
 'automatically',
 'through',
 'experience',
 'Machine',
 'learning',
 '(',
 'ML',
 ')',
 'is',
 'a',
 'field',
 'of',
 'inquiry',
 'devoted',
 'to',
 'understanding',
 'and',
 'building',
 'methods',
 'that',
 "'learn",
 "'",
 ',',
 'that',
 'is',
 ',',
 'methods',
 'that',
 'leverage',
 'data',
 'to',
 'improve',
 'performance',
 'on',
 'some',
 'set',
 'of',
 'tasks',
 '.',
 '[',
 '1',
 ']',
 'It',
```

```
 'is',
 'seen',
 'as',
 'a',
 'part',
 'of',
 'artificial',
 'intelligence',
 '.',
 'Machine',
 'learning',
 'algorithms',
 'build',
 'a',
 'model',
 'based',
 'on',
 'sample',
 'data',
 ',',
 'known',
 'as',
 'training',
 'data',
 ',',
 'in',
 'order',
 'to',
 'make',
 'predictions',
 'or',
 'decisions',
 'without',
 'being',
 'explicitly',
 'programmed',
 'to',
 'do',
 'so',
 '.',
 '[',
 '2',
 ']',
 'Machine',
 'learning',
 'algorithms',
 'are',
 'used',
 'in',
 'a',
 'wide',
 'variety']
```

In [11]:

```
from nltk.probability import FreqDist
fdist = FreqDist()
```

In [12]:

```
for word in tokens:
  fdist[word.lower()]+=1
fdist #dictrionary of all the distinct words with how many times they appea
```

Out[12]:

```
FreqDist({',': 408, 'the': 370, '.': 337, 'of': 240, 'a': 236, 'to': 219, '
learning': 198, 'and': 182, 'in': 181, '[': 151, ...})
```

In [13]:

```
len(fdist) #no. of distinct words (No of Unigrams)
```

Out[13]:

```
2076
```

In [14]:
```
#Top ten words
top10_repeated = fdist.most_common(10)
top10_repeated
```

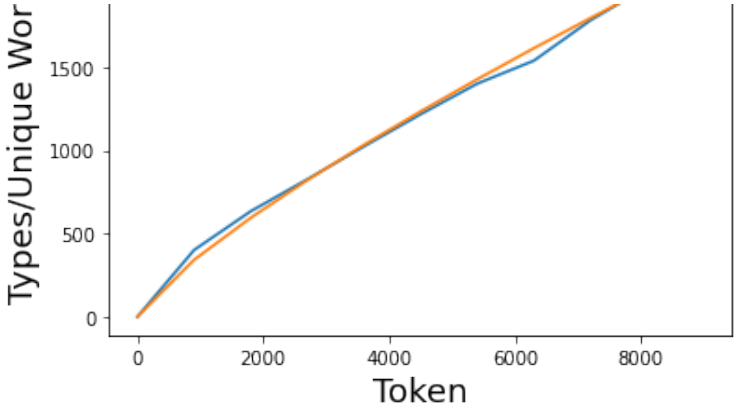
Out[14]:
```
[(',', 408),
 ('the', 370),
 ('.', 337),
 ('of', 240),
 ('a', 236),
 ('to', 219),
 ('learning', 198),
 ('and', 182),
 ('in', 181),
 ('[', 151)]
```

Herdan's Law

In [15]:
```
import re
import string
import nltk
from operator import itemgetter
from random import shuffle
from scipy.optimize import curve_fit
```

In [16]:
```
def function (n,k,bt):
    return k*(n**bt)
def heap_law(tokens_):
    #pass the list of all the tokens
    #this law tells how the overall vocabulary(types) grows with the size c
    #x is the diffent number of tokens and y is the diffent number of cor
    #then we use the scilearn curve_fit to get the esitmated paramters
    token_sample = [tokens_[0:i] for i in range(0,len(tokens_),900)]
    type_samples = [len(set(j)) for j in token_sample]

    #keeps the lenght of the list type
    x = [len(tokens_[0:i]) for i in range(0,len(tokens_),900)]
    y = type_samples

    plt.plot(x,y,label = "according to data")
    plt.legend()
    #curve fitting to estimate paramters
    parameters = curve_fit(function, x, y)#paramter order - function that i
    k,bt = parameters[0]

    plt.plot(x,k*(x**bt), label = "curve fit")
    plt.legend()
    plt.xlabel('Token', fontsize=18)
    plt.ylabel('Types/Unique Words', fontsize=18)
    plt.show()
    #print estimated value of k and beta
    print ("The estimated value of k is :",k)
    print ("The estimated value of beta is:",bt)
```

In [17]:
```
heap_law(tokens)
```




```
The estimated value of k is : 1.5085934988281775
The estimated value of beta is: 0.7973305268784159
```

Zipf's Law

In [18]:

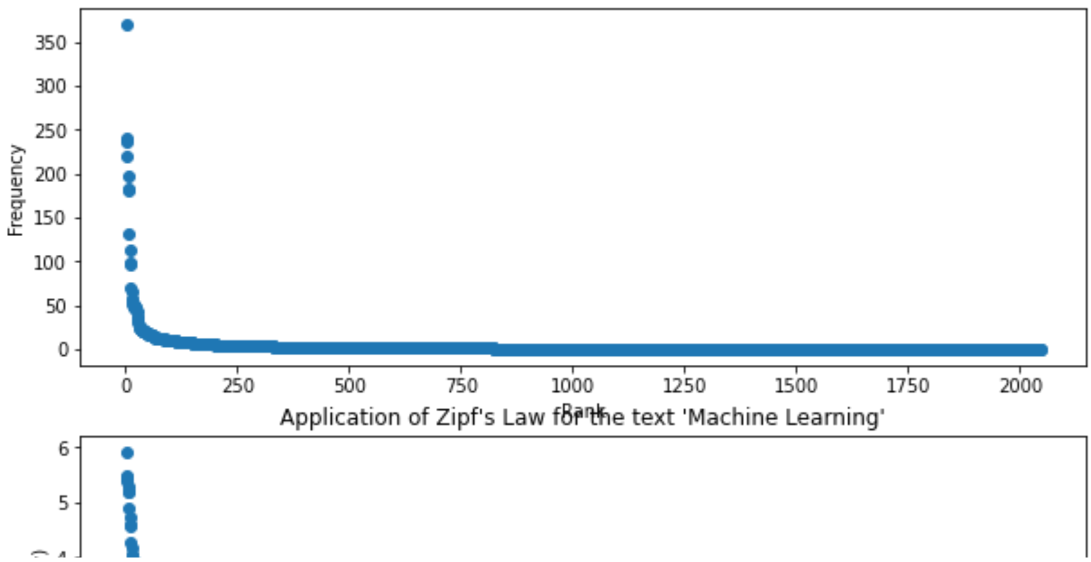
```
l2 = []
for word in tokens:
  word = word.lower()
  l2.append(word)
```

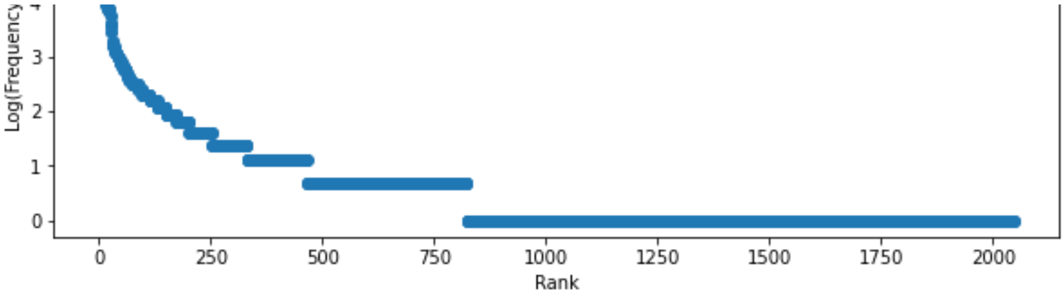
In [19]:

```
from collections import Counter
import math
def punc(text):
    exclude = set(string.punctuation)
    text = [''.join(x for x in y if x not in exclude) for y in text]
    text = [''.join(x.split()) for x in text]
    text = [x for x in text if x != '']
    return text

l2 = punc(l2)
```

```
#Term frequency
tf = Counter(l2)
values = [k for k,l in sorted([(j,i) for i,j in tf.items()], reverse=True)]
log_values = [math.log(k) for k,l in sorted([(j,i) for i,j in tf.items()],
x = [i+1 for i in range(len(values))]
#plotting,
fig = plt.figure(figsize= (10,8))

plt.subplot(2,1,1)
plt.plot(x,values, 'o')
plt.xlabel('Rank')
plt.ylabel('Frequency')

plt.subplot(2,1,2)
plt.plot(x, log_values, 'o' )
plt.xlabel('Rank')
plt.ylabel('Log(Frequency)')

plt.title("Application of Zipf's Law for the text 'Machine Learning'")
```

Out[19]: Text(0.5, 1.0, "Application of Zipf's Law for the text 'Machine Learning'")





N-Grams

In [20]:

```
from nltk.util import bigrams, trigrams, ngrams
```

In [ ]:

```

```

In [21]:

```
bigrams = list(nltk.bigrams(tokens))
len(bigrams) #no. of bigrams
```

Out[21]:

```
9025
```

In [22]:

```
bigrams[:100]
```

Out[22]:

```
[('Study', 'of'),
 ('of', 'algorithms'),
 ('algorithms', 'that'),
 ('that', 'improve'),
 ('improve', 'automatically'),
 ('automatically', 'through'),
 ('through', 'experience'),
 ('experience', 'Machine'),
 ('Machine', 'learning'),
 ('learning', '('),
 ('(', 'ML'),
 ('ML', ')'),
 (')', 'is'),
 ('is', 'a'),
 ('a', 'field'),
 ('field', 'of'),
 ('of', 'inquiry'),
 ('inquiry', 'devoted'),
 ('devoted', 'to'),
 ('to', 'understanding'),
 ('understanding', 'and'),
 ('and', 'building'),
 ('building', 'methods'),
 ('methods', 'that'),
 ('that', "'learn"),
 ("'learn", "'"),
 ("'", ','),
 (',', 'that'),
 ('that', 'is'),
 ('is', ','),
 (',', 'methods'),
 ('methods', 'that'),
 ('that', 'leverage'),
 ('leverage', 'data'),
 ('data', 'to'),
 ('to', 'improve'),
 ('improve', 'performance'),
 ('performance', 'on'),
 ('on', 'some')
```

```
 ('some', 'set'),
 ('set', 'of'),
 ('of', 'tasks'),
 ('tasks', '.'),
 ('.', '['),
 ('[', '1'),
 ('1', ']'),
 (']', 'It'),
 ('It', 'is'),
 ('is', 'seen'),
 ('seen', 'as'),
 ('as', 'a'),
 ('a', 'part'),
 ('part', 'of'),
 ('of', 'artificial'),
 ('artificial', 'intelligence'),
 ('intelligence', '.'),
 ('.', 'Machine'),
 ('Machine', 'learning'),
 ('learning', 'algorithms'),
 ('algorithms', 'build'),
 ('build', 'a'),
 ('a', 'model'),
 ('model', 'based'),
 ('based', 'on'),
 ('on', 'sample'),
 ('sample', 'data'),
 ('data', ','),
 (',', 'known'),
 ('known', 'as'),
 ('as', 'training'),
 ('training', 'data'),
 ('data', ','),
 (',', 'in'),
 ('in', 'order'),
 ('order', 'to'),
 ('to', 'make'),
 ('make', 'predictions'),
 ('predictions', 'or'),
 ('or', 'decisions'),
 ('decisions', 'without'),
 ('without', 'being'),
 ('being', 'explicitly'),
 ('explicitly', 'programmed'),
 ('programmed', 'to'),
 ('to', 'do'),
 ('do', 'so'),
 ('so', '.'),
 ('.', '['),
 ('[', '2'),
 ('2', ']'),
 (']', 'Machine'),
 ('Machine', 'learning'),
 ('learning', 'algorithms'),
 ('algorithms', 'are'),
 ('are', 'used'),
 ('used', 'in'),
 ('in', 'a'),
 ('a', 'wide'),
 ('wide', 'variety'),
 ('variety', 'of')]
```

In [23]:

```python
temp = set(bigrams)
unq_bigrams = list(temp)
unq_bigrams[:10]
```

Out[23]:

```
[('computing', 'used'),
 ('dynamic', 'programming'),
 ('vaguely', 'inspired'),
```

```
 ('for', 'much'),
 ('properties', 'learned'),
 ('goals', ';'),
 ('neurons', "'"),
 (')', 'can'),
 ('135', ']'),
 ('example', 'has')]
```

In [24]:
```
#Number of unique bigrams
len(unq_bigrams)
```

Out[24]:
```
6477
```

In [25]:
```
###########################
trigrams = list(nltk.trigrams(tokens))
len(trigrams)
```

Out[25]:
```
9024
```

In [26]:
```
trigrams[:100]
```

Out[26]:
```
[('Study', 'of', 'algorithms'),
 ('of', 'algorithms', 'that'),
 ('algorithms', 'that', 'improve'),
 ('that', 'improve', 'automatically'),
 ('improve', 'automatically', 'through'),
 ('automatically', 'through', 'experience'),
 ('through', 'experience', 'Machine'),
 ('experience', 'Machine', 'learning'),
 ('Machine', 'learning', '('),
 ('learning', '(', 'ML'),
 ('(', 'ML', ')'),
 ('ML', ')', 'is'),
 (')', 'is', 'a'),
 ('is', 'a', 'field'),
 ('a', 'field', 'of'),
 ('field', 'of', 'inquiry'),
 ('of', 'inquiry', 'devoted'),
 ('inquiry', 'devoted', 'to'),
 ('devoted', 'to', 'understanding'),
 ('to', 'understanding', 'and'),
 ('understanding', 'and', 'building'),
 ('and', 'building', 'methods'),
 ('building', 'methods', 'that'),
 ('methods', 'that', "'learn"),
 ('that', "'learn", "'"),
 ("'learn", "'", ','),
 ("'", ',', 'that'),
 (',', 'that', 'is'),
 ('that', 'is', ','),
 ('is', ',', 'methods'),
 (',', 'methods', 'that'),
 ('methods', 'that', 'leverage'),
 ('that', 'leverage', 'data'),
 ('leverage', 'data', 'to'),
 ('data', 'to', 'improve'),
 ('to', 'improve', 'performance'),
 ('improve', 'performance', 'on'),
 ('performance', 'on', 'some'),
 ('on', 'some', 'set'),
 ('some', 'set', 'of'),
 ('set', 'of', 'tasks'),
 ('of', 'tasks', '.'),
 ('tasks', '.', '['),
 ('.', '[', '1'),
```

```
 ('[', '1', ']'),
 ('1', ']', 'It'),
 (']', 'It', 'is'),
 ('It', 'is', 'seen'),
 ('is', 'seen', 'as'),
 ('seen', 'as', 'a'),
 ('as', 'a', 'part'),
 ('a', 'part', 'of'),
 ('part', 'of', 'artificial'),
 ('of', 'artificial', 'intelligence'),
 ('artificial', 'intelligence', '.'),
 ('intelligence', '.', 'Machine'),
 ('.', 'Machine', 'learning'),

 ('Machine', 'learning', 'algorithms'),
 ('learning', 'algorithms', 'build'),
 ('algorithms', 'build', 'a'),
 ('build', 'a', 'model'),
 ('a', 'model', 'based'),
 ('model', 'based', 'on'),
 ('based', 'on', 'sample'),
 ('on', 'sample', 'data'),
 ('sample', 'data', ','),
 ('data', ',', 'known'),
 (',', 'known', 'as'),
 ('known', 'as', 'training'),
 ('as', 'training', 'data'),
 ('training', 'data', ','),
 ('data', ',', 'in'),
 (',', 'in', 'order'),
 ('in', 'order', 'to'),
 ('order', 'to', 'make'),
 ('to', 'make', 'predictions'),
 ('make', 'predictions', 'or'),
 ('predictions', 'or', 'decisions'),
 ('or', 'decisions', 'without'),
 ('decisions', 'without', 'being'),
 ('without', 'being', 'explicitly'),
 ('being', 'explicitly', 'programmed'),
 ('explicitly', 'programmed', 'to'),
 ('programmed', 'to', 'do'),
 ('to', 'do', 'so'),
 ('do', 'so', '.'),
 ('so', '.', '['),
 ('.', '[', '2'),
 ('[', '2', ']'),
 ('2', ']', 'Machine'),
 (']', 'Machine', 'learning'),
 ('Machine', 'learning', 'algorithms'),
 ('learning', 'algorithms', 'are'),
 ('algorithms', 'are', 'used'),
 ('are', 'used', 'in'),
 ('used', 'in', 'a'),
 ('in', 'a', 'wide'),
 ('a', 'wide', 'variety'),
 ('wide', 'variety', 'of'),
 ('variety', 'of', 'applications')]
```

In [27]:

```
temp1 = set(trigrams)
unq_trigrams = list(temp1)
unq_trigrams[:10]
```

Out[27]:

```
[('and', 'the', 'actual'),
 (',', 'a', 'cluster'),
 ('to', 'model', 'the'),
 ('accuracy', 'estimation', 'techniques'),
 ('verification', '.', 'Unsupervised'),
 ('in', 'which', 'it'),
 (')', '+', 'v'),
 ('by', 'an', 'array'),
```

```
('.', '[', '21'),
('|', '[', '91')]
```

In [28]:
```python
len(unq_trigrams)#no. of unique trigrams
```

Out[28]:
```
8397
```

In [29]:
```python
######################################
four_grams = list(nltk.ngrams(tokens, 4))
len(four_grams)
```

Out[29]:
```
9023
```

In [30]:
```python
four_grams[1:100]
```

Out[30]:
```
[('of', 'algorithms', 'that', 'improve'),
 ('algorithms', 'that', 'improve', 'automatically'),
 ('that', 'improve', 'automatically', 'through'),
 ('improve', 'automatically', 'through', 'experience'),
 ('automatically', 'through', 'experience', 'Machine'),
 ('through', 'experience', 'Machine', 'learning'),
 ('experience', 'Machine', 'learning', '('),
 ('Machine', 'learning', '(', 'ML'),
 ('learning', '(', 'ML', ')'),
 ('(', 'ML', ')', 'is'),
 ('ML', ')', 'is', 'a'),
 (')', 'is', 'a', 'field'),
 ('is', 'a', 'field', 'of'),
 ('a', 'field', 'of', 'inquiry'),
 ('field', 'of', 'inquiry', 'devoted'),
 ('of', 'inquiry', 'devoted', 'to'),
 ('inquiry', 'devoted', 'to', 'understanding'),
 ('devoted', 'to', 'understanding', 'and'),
 ('to', 'understanding', 'and', 'building'),
 ('understanding', 'and', 'building', 'methods'),
 ('and', 'building', 'methods', 'that'),
 ('building', 'methods', 'that', "'learn"),
 ('methods', 'that', "'learn", "'"),
 ('that', "'learn", "'", ','),
 ("'learn", "'", ',', 'that'),
 ("'", ',', 'that', 'is'),
 (',', 'that', 'is', ','),
 ('that', 'is', ',', 'methods'),
 ('is', ',', 'methods', 'that'),
 (',', 'methods', 'that', 'leverage'),
 ('methods', 'that', 'leverage', 'data'),
 ('that', 'leverage', 'data', 'to'),
 ('leverage', 'data', 'to', 'improve'),
 ('data', 'to', 'improve', 'performance'),
 ('to', 'improve', 'performance', 'on'),
 ('improve', 'performance', 'on', 'some'),
 ('performance', 'on', 'some', 'set'),
 ('on', 'some', 'set', 'of'),
 ('some', 'set', 'of', 'tasks'),
 ('set', 'of', 'tasks', '.'),
 ('of', 'tasks', '.', '['),
 ('tasks', '.', '[', '1'),
 ('.', '[', '1', ']'),
 ('[', '1', ']', 'It'),

 ('1', ']', 'It', 'is'),
 (']', 'It', 'is', 'seen'),
 ('It', 'is', 'seen', 'as'),
 ('is', 'seen', 'as', 'a'),
```

```
 ('seen', 'as', 'a', 'part'),
 ('as', 'a', 'part', 'of'),
 ('a', 'part', 'of', 'artificial'),
 ('part', 'of', 'artificial', 'intelligence'),
 ('of', 'artificial', 'intelligence', '.'),
 ('artificial', 'intelligence', '.', 'Machine'),
 ('intelligence', '.', 'Machine', 'learning'),
 ('.', 'Machine', 'learning', 'algorithms'),
 ('Machine', 'learning', 'algorithms', 'build'),
 ('learning', 'algorithms', 'build', 'a'),
 ('algorithms', 'build', 'a', 'model'),
 ('build', 'a', 'model', 'based'),
 ('a', 'model', 'based', 'on'),
 ('model', 'based', 'on', 'sample'),
 ('based', 'on', 'sample', 'data'),
 ('on', 'sample', 'data', ','),
 ('sample', 'data', ',', 'known'),
 ('data', ',', 'known', 'as'),
 (',', 'known', 'as', 'training'),
 ('known', 'as', 'training', 'data'),
 ('as', 'training', 'data', ','),
 ('training', 'data', ',', 'in'),
 ('data', ',', 'in', 'order'),
 (',', 'in', 'order', 'to'),
 ('in', 'order', 'to', 'make'),
 ('order', 'to', 'make', 'predictions'),
 ('to', 'make', 'predictions', 'or'),
 ('make', 'predictions', 'or', 'decisions'),
 ('predictions', 'or', 'decisions', 'without'),
 ('or', 'decisions', 'without', 'being'),
 ('decisions', 'without', 'being', 'explicitly'),
 ('without', 'being', 'explicitly', 'programmed'),
 ('being', 'explicitly', 'programmed', 'to'),
 ('explicitly', 'programmed', 'to', 'do'),
 ('programmed', 'to', 'do', 'so'),
 ('to', 'do', 'so', '.'),
 ('do', 'so', '.', '['),
 ('so', '.', '[', '2'),
 ('.', '[', '2', ']'),
 ('[', '2', ']', 'Machine'),
 ('2', ']', 'Machine', 'learning'),
 (']', 'Machine', 'learning', 'algorithms'),
 ('Machine', 'learning', 'algorithms', 'are'),
 ('learning', 'algorithms', 'are', 'used'),
 ('algorithms', 'are', 'used', 'in'),
 ('are', 'used', 'in', 'a'),
 ('used', 'in', 'a', 'wide'),
 ('in', 'a', 'wide', 'variety'),
 ('a', 'wide', 'variety', 'of'),
 ('wide', 'variety', 'of', 'applications'),
 ('variety', 'of', 'applications', ',')]
```

In [31]:

```
temp2 = set(four_grams)
unq_four_grams = list(temp2)
unq_four_grams
```

Out[31]:

```
[('fail', 'to', 'reveal', 'their'),
 ('underlying', 'the', 'data', '.'),
 ('a', 'potential', 'result', 'of'),
 ('The', 'bias–variance', 'decomposition', 'is'),
 ('methods', 'to', 'estimate', 'the'),
 ('frequently', 'misunderstood', '.', 'ML'),
 ('Feature', 'learning', 'can', 'be'),
 ('a', 'field', 'of', 'study'),
 ('heavily', 'on', 'data', 'and'),
 ('(', 'PAC', ')', 'model'),
 ('medicine', ',', 'email', 'filtering'),
 ('require', 'input', 'that', 'is'),
 ('learning', ',', 'and', 'artificial'),
 ('known', 'as', 'overfitting', '.')
```

```
('known', 'as', 'overfitting', '.'),
('perform', 'a', 'certain', 'goal'),
('by', 'people', ',', 'it'),
(',', 'and', 'it', 'quickly'),
('or', 'had', 'non-European', 'sounding'),
('not', 'rare', 'objects', ','),
('from', 'the', 'training', 'set'),
(':', 'Oxford', 'University', 'Press'),
('Intelligence', ':', 'A', 'Logical'),
('care', '.', 'There', 'are'),
('signal', 'crosses', 'that', 'threshold'),
('2016', ',', 'Microsoft', 'tested'),
('amount', 'of', 'labeled', 'data'),
('being', 'a', 'major', 'exception'),
('neuron', 'to', 'another', '.'),
('scientific', 'endeavor', ',', 'machine'),
('research', 'was', 'now', 'outside'),
('Machine', 'learning', 'also', 'has'),
('Machine', 'learning', 'and', 'data'),
('and', 'graph', 'connectivity', '.'),
('.', 'New', 'York', ':'),
('predict', 'user', 'preferences', 'and'),
('learning', 'algorithm', ',', 'leaving'),
('techniques', '.', '[', '40'),
('modeling', ',', 'meta-learning', '.'),
('expressed', 'by', 'artificial', 'intelligence'),
('classification', 'or', 'predictions', '.'),
('AI', 'and', 'computer', 'science'),
('data', 'to', 'a', 'centralized'),
('structure', 'of', 'the', 'data'),
('in', 'which', 'every', 'finite'),
('Watson', 'system', 'failed', 'to'),
('Grand', 'Prize', 'in', '2009'),
('of', 'facts', ',', 'an'),
('applicants', 'by', 'similarity', 'to'),
('This', 'replaces', 'manual', 'feature'),
('FPR', ')', 'as', 'well'),
('items', 'either', 'within', 'a'),
('set', 'of', 'examples', 'represented'),
(',', 's', ')', '='),
('signal', 'is', 'only', 'sent'),
('reliable', 'data', 'in', 'order'),
('analysis', ',', 'autoencoders', ','),
('only', 'an', "'intelligent", 'subset'),
('gorillas', 'at', 'all', '.'),
('probabilistic', 'classification', 'setting', '.'),
('implies', 'an', 'agent', 'interacting'),
('data', 'provided', 'so', 'that'),

('a', '``', 'signal', "''"),
('genetic', 'algorithms', 'were', 'used'),
('the', 'teams', 'Big', 'Chaos'),
('is', 'associated', 'with', 'the'),
('predictions', 'or', 'decisions', 'without'),
('•', 'Unsupervised', 'learning', ':'),
('algorithms', 'and', 'computer', 'hardware'),
('transformative', 'in', 'some', 'fields'),
('that', 'commonly', 'identify', 'a'),
('autoencoders', ',', 'matrix', 'factorization'),
('dictionary', 'learning', 'is', 'a'),
('or', 'from', 'non-pattern', 'perturbations'),
('regression', ',', 'where', 'a'),
('(', '2017', ')', ','),
('an', 'intelligent', 'machine', 'is'),
('examples', '.', '[', '69'),
('can', 'transmit', 'information', ','),
('analysis', 'step', 'of', 'knowledge'),
('order', 'for', 'the', 'models'),
('tag', 'black', 'people', 'as'),
('AI', 'and', 'machine', 'learning'),
('of', 'the', 'function', 'underlying'),
```

```
('1998', ')', '.', '.'),
('notably', 'be', 'integrated', 'within'),
('.', 'For', 'more', 'advanced'),
('genetic', 'and', 'evolutionary', 'algorithms'),
('.', '[', '70', ']'),
('When', 'trained', 'on', 'man-made'),
('and', 'finally', 'meta-learning', '('),
('effective', 'use', 'of', 'machine'),
('engineering', 'teams', '.', 'Federated'),
('attention', 'moved', 'to', 'performing'),
('The', 'data', 'is', 'known'),
('represent', 'decisions', 'and', 'decision'),
('led', 'to', 'more', 'efficient'),
('for', 'the', 'task', 'of'),
('associated', 'area', 'under', 'the'),
('supervised', 'learning', 'in', 'order'),
('``', 'goof', "''", 'button'),
('typically', 'have', 'a', 'weight'),
('it', 'to', 'produce', 'sufficiently'),
('learning', 'technology', 'is', 'also'),
('input', '.', 'Unsupervised', 'learning'),
('components', 'analysis', 'and', 'cluster'),
('detection', ',', 'the', 'interesting'),
('test', 'set', '(', 'conventionally'),
(',', 'badly', 'chosen', 'tasks'),
('characteristic', 'of', 'a', 'rule-based'),
('the', 'objectivity', 'and', 'logical'),
('human', 'languages', 'contain', 'biases'),
('separate', 'journals', ',', 'ECML'),
('by', '``', 'learning', 'the'),
('that', 'all', 'brown', 'patches'),
('In', '2018', ',', 'a'),
('learning', '•', 'None', 'Domingos'),
('that', 'contain', 'many', 'layers'),
(',', 'David', ';', 'Mackworth'),
('the', 'class', 'that', 'is'),
('the', 'bias', 'by', 'scoring'),
('and', 'propelling', 'its', 'use'),
('within', 'machine', 'learning', 'engineering'),
(')', ',', 'or', 'Interpretable'),
('and', 'equipped', 'with', 'a'),
('How', 'the', 'Quest', 'for'),
('answers', 'as', 'valid', '.'),
('the', 'field', 'of', 'machine'),
('.', 'Different', 'clustering', 'techniques'),
('while', 'in', 'knowledge', 'discovery'),
('on', 'using', 'teaching', 'strategies'),
('association', 'rules', 'for', 'discovering'),
('such', 'as', 'promotional', 'pricing'),
('mentioned', 'rates', ',', 'thus'),
('Ehud', 'Shapiro', 'laid', 'the'),
('.', '[', '82', ']'),
('in', '``', 'adversarial', "''"),
('rule', 'learning', 'using', 'logic'),
('of', 'random', 'variables', 'under'),
('XAI', ')', ',', 'or'),
('Can', 'machines', 'do', 'what'),
('not', 'yet', 'developed', 'sufficiently'),
('in', 'each', 'iteration', 'executes'),
(']', 'OpenAI', 'estimated', 'the'),
('the', 'application', 'of', 'machine'),
('input', 'but', 'also', 'transform'),
('trees', 'where', 'the', 'target'),
('a', 'set', 'of', 'related'),
('Reasons', 'for', 'this', 'are'),
('.', 'Data', 'mining', 'is'),
('without', 'being', 'explicitly', 'programmed'),
('to', 'philosophical', 'induction', ','),
(']', 'Several', 'learning', 'algorithms'),
('86', ']', 'Recently', ','),
```

```
('and', 'regression', '.', 'Given'),
('and', 'find', 'structure', 'in'),
('COVID-19', '.', '[', '85'),
('.', 'As', 'it', 'navigates'),
('[', '105', ']', 'In'),
('The', 'system', 'is', 'driven'),
('and', 'their', 'associated', 'features'),
('significantly', 'from', 'the', 'majority'),
('For', 'the', 'best', 'performance'),
('challenges', ',', 'the', 'effective'),
('wrote', 'about', 'the', 'firm'),
('dismantling', 'their', 'misconceptions', ','),
('learning', 'system', 'trained', 'specifically'),
('Interpretable', 'AI', ',', 'or'),
('trained', 'only', 'on', 'pictures'),
('more', 'effective', 'to', 'help'),
('variables', 'in', 'large', 'databases'),
('method', 'randomly', 'partitions', 'the'),
('for', 'representing', 'hypotheses', '('),
('principal', 'goal', ':', 'statistics'),
('brain', '.', 'Each', 'connection'),
('from', 'statistics', ',', 'fuzzy'),
('patterns', 'using', 'rudimentary', 'reinforcement'),
('(', 'algorithmic', 'bias', ')'),
('AlphaZero', '(', '2017', ')'),
('overall', 'accuracy', ',', 'investigators'),
('gorillas', ',', '[', '103'),
('(', 'September', '22', ','),
('knowledge-based', 'approach', 'caused', 'a'),
('statistical', 'definition', 'of', 'an'),
('attempt', 'to', 'do', 'so'),
('Alan', ';', 'Goebel', ','),
('learns', 'a', 'representation', 'that'),
('exploratory', 'data', 'analysis', 'through'),
('worked', 'well', 'in', 'the'),
('brain', 'processes', 'light', 'and'),
('firm', 'Rebellion', 'Research', 'and'),
('machine', 'learning', 'algorithms', 'include'),
('sparsely', 'represented', 'by', 'the'),
('multitude', 'of', 'machine', 'learning'),
('known', 'background', 'knowledge', 'and'),
('of', 'two', 'categories', ','),
('journals', ',', 'ECML', 'PKDD'),
('correlate', 'with', 'images', 'of'),
('Types', 'of', 'supervised-learning', 'algorithms'),
('background', 'knowledge', ',', 'and'),
('improve', 'learner', 'accuracy', '.'),
('also', 'learn', 'these', 'biases'),
('other', 'disciplines', ',', 'such'),
('The', 'backpropagated', 'value', '('),
('this', 'requires', 'these', 'biases'),
('classification', 'setting', '.', 'In'),
('that', 'makes', 'it', 'useful'),
(')', '.', '[', '117'),
('training', 'data', 'and', 'then'),
('multivariate', 'normal', 'distribution', ','),
('statistics', 'was', 'out', 'of'),
('pattern', 'classification', '.', '['),
('system', 'will', 'derive', 'a'),
('to', 'fit', 'the', 'least'),
('had', 'been', 'applied', 'in'),
('.', 'Feature', 'learning', 'algorithms'),
('a', 'single', 'line', 'is'),
('AI', ',', 'and', 'toward'),
('the', 'spatial', 'relationship', 'between'),
('of', 'backpropagation', '.', '['),
('their', 'misconceptions', ',', 'XAI'),
('detrimental', 'outcomes', 'thereby', 'furthering'),
('and', 'the', 'covariances', 'between'),
('and', 'consists', 'of', 'a'),
```

```
('by', 'Arthur', 'Samuel', ','),
('larger', 'effective', 'training', 'sets'),
('Modern-day', 'machine', 'learning', 'has'),
('of', 'the', 'training', 'model'),
('trained', 'on', 'language', 'corpora'),
('1990s', '.', 'The', 'field'),
(',', 'implicitly', 'mapping', 'their'),
(',', 'rather', 'than', 'having'),
('of', 'favor', '.', '['),
('learning', '.', '[', '30'),
('(', 's', "'", ')'),
('%', 'of', 'families', 'have'),
(',', 'visual', 'identity', 'tracking'),
('[', '125', ']', 'Since'),
('widely', 'quoted', ',', 'more'),
('competition', 'to', 'find', 'a'),
('.', '[', '26', ']'),
('of', 'families', 'have', 'geographically'),
('the', 'field', 'in', 'cognitive'),
('reinforcement', 'learning', 'algorithms', 'use'),
('specific', 'task', '.', 'Feature'),
('his', 'paper', '``', 'Computing'),
('in', 'other', 'domains', '.'),
('in', 'order', 'for', 'the'),
('63', ']', 'Rule-based', 'machine'),
('understanding', 'and', 'building', 'methods'),
(',', 'machine', 'translation', ','),
('patches', 'are', 'likely', 'to'),
('autonomous', 'vehicles', 'or', 'in'),
('health', 'care', 'professionals', 'that'),
('rather', 'than', 'mathematical', 'induction'),
('research', 'communities', '(', 'which'),
(',', 'and', 'many', 'dimensionality'),
('image', 'classifiers', 'often', 'do'),
('9', ']', 'In', 'its'),
('distribution', '.', 'This', 'replaces'),
('find', 'a', 'program', 'to'),
('a', 'learning', 'machine', 'to'),
('meaning', 'that', 'the', 'mathematical'),
('(', 'PCA', ')', '.'),
('training', 'examples', 'come', 'from'),
('the', 'training', 'model', 'on'),
('model', 'based', 'on', 'a'),
('gerrymandered', 'to', 'fit', 'all'),
('to', 'the', 'data', ','),
(',', 'while', 'in', 'knowledge'),
('translation', ',', 'social', 'network'),
('are', 'many', 'applications', 'for'),
('on', 'the', 'structure', 'of'),
('undesirable', 'situations', '.', '['),
('opponent', ')', '.', 'As'),
('algorithms', ',', 'often', 'attempt'),
('inputs', 'coming', 'from', 'the'),
('used', '.', '[', '11'),
('reinforcement', 'input', 'nor', 'an'),
('as', 'thinking', 'entities', ')'),
('to', 'predict', 'the', 'green'),
('activities', 'such', 'as', 'promotional'),
('has', 'many', 'zeros', '.'),
('density', 'function', '.', '['),
('concerned', 'with', 'minimizing', 'the'),
('only', 'one', 'input', ','),
('rules', 'for', 'discovering', 'regularities'),
('.', 'It', 'has', 'applications'),
('.', 'The', 'computational', 'analysis'),
('order', 'to', 'make', 'predictions'),
('weakly', 'supervised', 'learning', ','),
('are', 'given', 'to', 'the'),
('learning', 'in', 'health', 'care'),
('image', 'dictionary', ',', 'but'),
```

```
('have', 'adopted', 'methods', 'from'),
('as', 'a', 'logical', 'database'),
('so', 'there', 'is', 'a'),
('input', 'for', 'decision-making', '.'),
('provided', 'so', 'that', 'they'),
('more', 'than', 'one', 'is'),
('(', 'ILP', ')', 'is'),
('not', 'consider', 'the', 'order'),
('on', 'Learning', 'Machines', ','),
('training', 'a', 'machine', 'learning'),
('a', 'way', 'that', 'mimics'),
('[', '14', ']', '['),
('allowing', 'for', 'users', "'"),
('to', 'Google', '.', '['),
('[', '31', ']', '['),
(',', 'and', 'toward', 'methods'),
('finding', 'the', 'probability', 'density'),
('the', 'Grand', 'Prize', 'in'),
('well', 'it', 'fits', 'the'),
('known', 'as', 'a', 'supervisory'),
('or', 'protein', 'sequences', ','),
('graphical', 'model', 'is', 'a'),
('network', 'in', 'which', 'an'),
('from', 'the', 'original', 'on'),
('a', 'learner', 'is', 'to'),
('of', 'Bayesian', 'networks', 'that'),
('these', 'tree', 'structures', ','),
('replaces', 'manual', 'feature', 'engineering'),
('people', ',', 'it', "'s"),
('94', ']', 'Attempts', 'to'),
('(', 'e.g', '.', 'for'),
('new', 'genotypes', 'in', 'the'),
('specificity', 'meaning', 'True', 'Positive'),
('supervised', 'methods', 'can', 'not'),
('the', 'function', 'underlying', 'the'),
('58', ']', 'Three', 'broad'),
('of', 'each', 'artificial', 'neuron'),
('machine', 'learning', 'field', ':'),
('results', ':', 'Positive', 'results'),
('of', 'an', 'exact', 'mathematical'),
('evidence', 'related', 'to', 'a'),
('MAML', ')', '.', 'Association'),
('to', 'maximize', 'some', 'notion'),
('models', 'where', 'the', 'target'),
('learning', 'field', ':', '``'),
('these', 'were', 'mostly', 'perceptrons'),
('to', 'both', 'learn', 'the'),
('transform', 'it', 'in', 'a'),
('anomalous', 'items', 'represent', 'an'),
('quantity', 'of', 'reliable', 'data'),
('had', 'displaced', 'CPUs', 'as'),
(',', 'or', 'the', 'similarity'),
(',', 'therefore', ',', 'learn'),
('from', 'the', 'spatial', 'relationship'),
('of', 'its', 'inputs', '.'),
('may', 'not', 'be', 'able'),
('responsibility', '.', '``', '['),
('applied', 'in', 'the', 'field'),
('outputs', 'or', 'predictions', 'over'),
('either', 'feature', 'elimination', 'or'),
('data', 'bias', ',', 'privacy'),
('In', 'unsupervised', 'feature', 'learning'),
('After', 'receiving', 'the', 'genome'),
('should', 'match', 'the', 'complexity'),
('from', 'machine', 'learning', ','),
('all', 'steps', 'required', 'to'),
('1982', 'along', 'with', 'a'),
('True', 'Positive', 'Rate', '('),
(',', 'starting', 'from', 'supervised'),
('the', 'view', 'that', 'not'),
```

```
('.', '[', '83', ']'),
(',', 'and', 'that', 'is'),
('a', 'non-linear', 'classification', 'using'),
('chance', 'of', 'successfully', 'achieving'),
('data', ',', 'lack', 'of'),
('a', 'limited', 'set', 'of'),
('[', '139', ']', 'Software'),
('training', 'examples', '.', 'Each'),
('resolved', ',', 'but', 'Google'),
('variables', 'and', 'their', 'associated'),
('varied', 'as', 'a', 'corpus'),
(')', ',', 'is', 'artificial'),
(',', 'in', 'which', 'the'),
('health', 'care', 'to', 'provide'),
('a', 'new', 'point', 'as'),
('animal', 'brains', '.', 'Such'),
('internal', 'compactness', ',', 'or'),
('subset', 'for', 'evaluation', 'and'),
('algorithms', 'discover', 'multiple', 'levels'),
('for', 'Bioinformatics', 'and', 'Biostatistics'),
('is', 'a', 'field', 'of'),
('process', 'it', 'and', 'then'),
('electrocardiograms', ',', 'and', 'speech'),
('necessarily', 'also', 'learn', 'these'),
('been', 'developed', ',', 'the'),
('formulated', 'as', 'minimization', 'of'),
('Loss', 'functions', 'express', 'the'),
('the', 'ability', 'of', 'a'),
('order', 'of', 'items', 'either'),
('machine', 'learning', 'systems', '.'),
(']', '[', '9', ']'),
('(', 'feelings', ')', 'about'),
('By', '1980', ',', 'expert'),
('to', 'determine', 'correct', 'answers'),
('outside', 'the', 'AI/CS', 'field'),
(',', 'that', 'is', ','),
('consists', 'of', 'multiple', 'hidden'),
('and', 'then', 'can', 'process'),
('is', 'increasingly', 'expressed', 'by'),
('disciplines', ',', 'such', 'as'),
('certain', 'goal', '(', 'such'),
('draws', 'population', 'inferences', 'from'),
('it', 'uses', 'to', 'determine'),
('classifier', 'systems', ',', 'association'),
('correspond', 'to', 'learning', 'paradigms'),
('which', 'every', 'finite', 'collection'),
('imprecise', ';', 'however', ','),
('.', 'Machine', 'learning', 'algorithms'),
('that', 'filters', 'emails', ','),
('used', 'to', 'do', 'hyperparameter'),
('related', 'to', 'a', 'systematic'),
('a', 'piecewise', 'manner', 'in'),
('in', 'large', 'databases', '.'),
('collision', '.', '[', '94'),
('interacts', 'with', 'a', 'dynamic'),
('to', 'emulate', 'a', 'neural'),
('Randy', '(', '1998', ')'),
('[', '43', ']', 'As'),
(',', 'one', 'approach', 'is'),
('intrusion', 'detection', ',', 'continuous'),
('The', 'manifold', 'hypothesis', 'proposes'),
('Uber', 'failed', 'to', 'detect'),
('speech', 'recognition', ',', 'agriculture'),
('in', 'having', 'machines', 'learn'),
(';', 'on', 'the', 'other'),
('rudimentary', 'reinforcement', 'learning', '.'),
('identify', 'strong', 'rules', 'discovered'),
('of', 'dimensionality', 'reduction', 'is'),
('are', 'used', 'when', 'the'),
('field', 'that', 'considers', 'any'),
```

```
('tracking', ',', 'face', 'verification'),
('develop', 'its', 'own', 'algorithm'),
(',', 'which', 'it', 'tries'),
('a', 'part', 'of', 'artificial'),
('abstract', 'features', 'defined', 'in'),
('set', '.', 'Performing', 'machine'),
('output', ',', 'also', 'known'),
('emails', ',', 'the', 'input'),
('that', 'an', 'image', 'classifier'),
('that', 'are', 'trained', 'on'),
('algorithms', ',', 'therefore', ','),
('as', 'well', 'as', 'what'),
('data', 'features', '.', 'Cluster'),
('the', 'overall', 'field', '.'),
('in', 'databases', ')', '.'),
('did', 'continue', 'within', 'AI'),
('[', '41', ']', 'In'),
(',', 'rather', 'than', 'mathematical'),
('these', 'two', 'research', 'communities'),
('data', ')', 'and', 'supervised'),
('self-teaching', 'computers', 'were', 'used'),
('.', 'There', 'are', 'two'),
('then', 'signal', 'additional', 'artificial'),
('new', ',', 'unobserved', 'point'),
('ethical', 'questions', '.', 'Systems'),
('knowledge', 'and', 'a', 'set'),
('situation', ',', 'and', 'only'),
('for', 'decision-making', '.', 'Support-vector'),
(',', 'and', 'evaluation', 'problems'),
('answers', '.', 'For', 'example'),
('uses', 'to', 'determine', 'correct'),
('false', 'negative', 'rate', '('),
('25', ']', 'Probabilistic', 'reasoning'),
("''", 'means', 'more', 'or'),
('machine', 'learning', 'models', 'that'),
('one', 'approach', 'is', 'to'),
('time', 'is', 'said', 'to'),
('members', 'of', 'the', 'same'),
('cheaper', 'to', 'obtain', ','),
('not', 'spam', "''", 'of'),
('normal', 'training', 'data', 'set'),
('external', 'rewards', 'and', 'no'),
('is', 'the', 'identification', 'of'),
('of', 'training', 'examples', ','),
('a', 'threshold', 'such', 'that'),
('estimated', 'density', 'and', 'graph'),
('an', 'agent', 'interacting', 'with'),
('model', 'is', 'subject', 'to'),
('(', 'for', 'example', ','),
('employs', 'various', 'approaches', 'to'),
('less', 'the', 'machine', 'learning'),
('within', 'AI', ',', 'leading'),
('and', 'black', 'cats', 'might'),
('on', 'estimated', 'density', 'and'),
('generated', 'by', 'the', 'model'),
('learning', 'algorithms', 'take', 'a'),
('Model', 'Inference', 'System', ')'),
('or', 'Explainable', 'Machine', 'Learning'),
('27', ']', 'The', 'difference'),
(',', 'overly', 'complex', 'theory'),
('routine', ':', '•', 'compute'),
('data', 'according', 'to', 'a'),
('in', 'terms', 'of', 'methods'),
('example', ',', 'by', 'internal'),
('provided', ',', 'possibly', 'including'),
('called', 'the', '``', 'number'),
('bias–variance', 'decomposition', 'is', 'one'),
('Cybertron', ',', 'had', 'been'),
('the', 'input', 'would', 'be'),
('employed', 'today', 'in', 'application')
```

```
('employed', 'today', 'in', 'application'),
('and', 'billions', 'of', 'dollars'),
('.', 'Many', 'reinforcement', 'learning'),
('Kaufmann', '.', 'ISBN', '.'),
('examples', 'come', 'from', 'some'),
('[', '101', ']', '['),
('be', 'generated', 'by', 'the'),
('model', 'without', 'changing', 'the'),
('linear', 'combination', 'of', 'basis'),
('algorithm', 'computes', ',', 'in'),
('When', 'dealing', 'with', 'non-linear'),
('firm', 'with', 'racist', 'hiring'),
('finally', 'meta-learning', '(', 'e.g'),
('the', 'fact', 'that', 'machine'),
('the', 'AI', '.', 'It'),
("''", '.', 'Artificial', 'neurons'),
('Characterizing', 'the', 'generalization', 'of'),
('the', '``', 'signal', "''"),
('we', '(', 'as', 'thinking'),
('that', 'decentralizes', 'the', 'training'),
('with', ':', 'in', 'machine'),
('.', '[', '112', ']'),
('and', 'neural', 'networks', 'in'),
('longer', 'to', 'be', 'adopted'),
('machine', 'learning', '.', 'AI'),
('likely', 'to', 'continue', 'working'),
('environment', 'to', 'learn', 'and'),
('neurons', 'are', 'called', '``'),
('data', 'of', 'previous', 'admissions'),
('use', 'them', 'to', 'perform'),
('numerous', ':', 'lack', 'of'),
('Data', 'from', 'the', 'training'),
('computers', 'learning', 'from', 'data'),
('including', 'hardware', 'acceleration', ','),
('.', 'For', 'a', 'dictionary'),
('of', 'outlier', 'detection', ')'),
('of', 'intellectual', 'properties', ','),
('synapses', 'in', 'a', 'biological'),
('on', 'cloud', 'servers', 'for'),
('and', 'relationships', 'to', 'other'),
('algorithm', 'is', 'a', 'non-probabilistic'),
('of', 'the', 'folder', 'in'),
('variable', 'can', 'take', 'continuous'),
('.', 'Sparse', 'coding', 'algorithms'),
('of', 'machine', 'learning', ')'),
(',', 'noise', ',', 'deviations'),
('rates', 'are', 'ratios', 'that'),
('for', 'users', "'", 'privacy'),
('turn', 'out', 'to', 'be'),
('and', 'genetic', 'algorithms', '.'),
('and', 'decision', 'making', '.'),
('bias', ',', 'as', 'in'),
('information', '.', 'These', 'decisions'),
('media-services', 'provider', 'Netflix', 'held'),
('approaches', 'to', 'teach', 'computers'),
('set', '.', 'Supervised', 'anomaly'),
('.', 'In', 'supervised', 'feature'),
('and', 'that', 'it', 'may'),
('optimization', 'and', 'machine', 'learning'),
('hypothesis', 'is', 'less', 'complex'),
('many', 'dimensionality', 'reduction', 'techniques'),
('Pragmatic', 'Theory', 'built', 'an'),
(',', 'Nils', '(', '1998'),
('were', 'plagued', 'by', 'theoretical'),
('Reinforcement', 'learning', ':', 'A'),
('referred', 'to', 'as', 'predictive'),
('datasets', 'collected', 'with', 'biases'),
('the', 'interaction', 'between', 'cognition'),
('interest', 'but', 'as', 'income-generating'),
(')', '.', 'Semi-supervised', 'anomaly'),
('are', 'aggregated', 'into', 'layers'),
```

('are', 'aggregated', 'into', 'layers'),
('to', 'train', 'a', 'system'),
('more', 'inputs', 'and', 'the'),
('it', 'can', 'be', 'done'),
('evolves', '``', 'rules', "''"),
('Computing', 'Machinery', 'and', 'Intelligence'),
('form', 'is', 'linear', 'regression'),
('of', '(', 'or', 'generating'),
(']', 'In', 'data', 'mining'),
('optimization', 'used', 'to', 'do'),
(']', 'In', 'addition', 'to'),
('receives', 'a', 'signal', 'can'),
('artists', '.', '[', '83'),
('algorithms', 'exist', 'that', 'perform'),
('and', 'feasibility', 'of', 'learning'),
(',', 'without', 'relying', 'on'),
('Different', 'layers', 'may', 'perform'),
('the', 'output', 'would', 'be'),
('in', 'the', 'behavioral', 'environment'),
('intended', 'to', 'identify', 'strong'),
('and', 'microcontrollers', '.', '['),
('aggregated', 'into', 'layers', '.'),
('systems', ',', 'are', 'computing'),
('a', 'combined', 'field', 'that'),
('normal', 'behavior', 'from', 'a'),
('machine', 'learning', 'has', 'two'),
('increases', 'or', 'decreases', 'the'),
('labeled', 'training', 'data', ')'),
('chatbot', 'that', 'learned', 'from'),
('regression', '.', 'Given', 'a'),
('that', 'distribution', '.', 'This'),
('learning', ',', 'including', ':'),
(',', 'but', 'only', 'an'),
('heuristic', 'method', 'for', 'sparse'),
('dictionary', 'where', 'each', 'class'),
('algorithms', '.', 'Machine', 'learning'),
('fraud', ',', 'a', 'structural'),
(')', 'vector', 'from', 'the'),
('reducing', 'the', 'number', 'of'),
('.', 'Many', 'systems', 'attempt'),
('that', 'ML', 'remains', 'a'),
('45', ']', 'The', 'self-learning'),
('that', 'if', 'a', 'customer'),
('``', 'signal', "''", ','),
('closely', 'related', 'to', 'regression'),
('in', 'the', 'amount', 'of'),
('how', 'complex', 'the', 'theory'),
('a', 'data', 'set', 'that'),
('algorithm', 'builds', 'a', 'model'),
(',', 'linear', 'classifier', ','),
('to', 'help', 'users', 'perform'),
('[', '38', ']', 'Though'),
('its', 'inputs', '.', 'The'),
('related', 'to', 'the', 'growth'),
('.', 'Other', 'approaches', 'have'),
('provide', 'professionals', 'an', 'additional'),
('termed', '``', 'neural', 'networks'),
('applied', 'through', 'several', 'techniques'),
('can', 'then', 'be', 'used'),
('symptoms', '.', 'Given', 'symptoms'),
('line', 'of', 'research', 'was'),
('networks', "''", ';', 'these'),
('exist', "''", '.', 'Machine'),
('cloud', 'AI', '.', '['),
('travel', 'from', 'the', 'first'),
('a', 'set', 'of', 'observations'),
('identifies', ',', 'learns', ','),
('in', 'the', 'model', 'without'),
('some', 'measure', 'of', '``'),
('include', 'dictionary', 'learning', ','),
('detection', 'techniques', 'detect', 'anomalies')

```
('detection', 'techniques', 'detect', 'anomalies'),
('remaining', 'K-1', 'subsets', 'for'),
('density', 'estimation', 'in', 'statistics'),
('the', 'predictive', 'modeling', 'approaches'),
('is', 'a', 'general', 'term'),
('even', 'its', 'designers', 'can'),
('lower-level', 'features', '.', 'It'),
('a', 'widely', 'quoted', ','),
('would', '.', 'However', ','),
('implementations', ',', 'the', 'signal'),
(']', '[', '123', ']'),
('one', 'that', 'learns', 'a'),
(',', 'association', 'rule', 'learning'),
('inferred', 'logic', 'programs', 'from'),
('ought', 'to', 'take', 'actions'),
('reinforcement', 'learning', ',', 'or'),
('vision', 'and', 'speech', 'recognition'),
('to', 'detect', 'the', 'micro-clusters'),
('the', 'knowledge', 'captured', 'by'),
('used', 'to', 'predict', 'the'),
('unsupervised', 'learning', 'is', 'in'),
(',', 'unsupervised', 'algorithms', ')'),
('exceptions', '.', '[', '57'),
('learning', 'problems', 'are', 'formulated'),
('problems', 'is', 'the', 'inherently'),
('noisy', ',', 'limited', ','),
('was', 'to', 'solve', 'problems'),
('training', '.', '[', '47'),
('higher-dimensional', 'data', '(', 'e.g.'),
('to', 'solve', 'problems', 'in'),
('82', ']', 'In', '2014'),
('are', 'based', 'on', 'estimated'),
('unconscious', 'biases', 'already', 'present'),
('random', 'variables', 'in', 'the'),
('to', 'algorithmically', 'define', 'specific'),
('examples', ',', 'each', 'marked'),
('vision', 'and', 'hearing', '.'),
('model', 'and', 'algorithmic', 'model'),
('fuzzy', 'logic', ',', 'and'),
("''", 'by', 'a', 'human'),
('53', ']', 'Feature', 'learning'),
('neurons', 'are', 'aggregated', 'into'),
('goal', '(', 'such', 'as'),
('acquisition', 'and', 'representation', '.'),
('vast', 'numbers', 'of', 'potential'),
('non-pattern', 'perturbations', '.', 'Some'),
('any', 'numerical', 'value', 'within'),
('two', 'objects', 'are', '.'),
('systems', '``', 'learn', "''"),
(']', 'Machine', 'learning', 'is'),
('variables', 'to', 'higher-dimensional', 'space'),
('95', ']', '[', '96'),
('such', 'commonalities', 'in', 'each'),
('on', 'a', 'legitimate', 'image'),
('contrasts', 'with', 'the', '``'),
('a', 'goal-seeking', 'behavior', ','),
('black', 'cats', 'might', 'conclude'),
('is', 'artificial', 'intelligence', '('),
('the', 'Ultimate', 'Learning', 'Machine'),
('of', 'generalization', ':', 'while'),
('fact', 'that', 'machine', 'learning'),
('by', 'the', 'model', '.'),
('data', ',', 'and', 'also'),
('of', 'the', 'same', 'cluster'),
('the', 'environment', 'is', 'typically'),
('the', 'remaining', 'K-1', 'subsets'),
('crisis', '.', '[', '81'),
('.', '[', '138', ']'),
('items', 'represent', 'an', 'issue'),
('in', 'large-scale', 'transaction', 'data'),
('these', 'patterns', '.', '[')
```

```
('these', 'patterns', '.', '['),
('neurons', 'and', 'edges', 'typically'),
('language', '.', '[', '106'),
('.', 'This', 'approach', 'tries'),
('users', 'perform', 'more', 'effectively'),
('classifying', '(', 'e.g', '.'),
('True', 'Negative', 'Rate', '('),
('predictably', 'induces', 'misclassification', '.'),
('something', 'to', 'watch', 'out'),
('units', '.', '[', '126'),
('refining', 'the', 'mental', 'models'),
('by', 'a', '``', 'teacher'),
('I.', 'Jordan', ',', 'the'),
(',', 'a', 'self-driving', 'car'),
('had', 'denied', 'nearly', '60'),
('along', 'with', 'a', 'neural'),
('suffer', 'from', 'different', 'data'),
('machine', 'learning', 'may', 'take'),
('that', 'is', ',', 'methods'),
('depending', 'on', 'the', 'nature'),
('(', 'ECML', 'PKDD', ')'),
('for', 'discovering', 'regularities', 'between'),
('features', ',', 'with', 'higher-level'),
('financial', 'crisis', '.', '['),
('optimization', 'algorithms', 'can', 'minimize'),
('do', 'what', 'we', '('),
('from', 'data', 'have', 'been'),
('learning', ',', 'where', 'the'),
('Rate', '(', 'TPR', ')'),
('to', 'build', 'a', 'general'),
('unsupervised', 'learning', 'algorithms', 'identify'),
('.', '[', '64', ']'),
('fit', 'the', 'least', 'to'),
('in', 'which', 'an', 'electrically'),
('related', 'fields', 'in', 'terms'),
('in', 'which', 'to', 'file'),
('learning', 'engineers', 'need', 'to'),
(',', 'then', 'the', 'model'),
('of', 'ethical', 'challenges', ','),
('common', '.', 'The', 'bias–variance'),
('low-dimensional', 'manifolds', ',', 'and'),
('input', 'variables', 'to', 'higher-dimensional'),
('their', 'data', 'to', 'a'),
(',', 'while', 'not', 'being'),
('different', 'kinds', 'of', 'transformations'),
('customer', 'buys', 'onions', 'and'),
('from', 'AT', '&', 'T'),
('None', 'Nilsson', ',', 'Nils'),
(',', 'sometimes', 'called', 'a'),
('the', 'context', 'of', 'generalization'),
(']', 'Inductive', 'logic', 'programming'),
('between', 'artificial', 'neurons', 'are'),
(']', 'Machine', 'learning', 'systems'),
('For', 'example', ',', 'to'),
('image', 'patch', 'can', 'be'),
('Master', 'Algorithm', ':', 'How'),
('react', 'based', 'on', 'the'),
(':', '25', 'Machine', 'learning'),
('called', 'the', 'kernel', 'trick'),
('is', 'a', 'recommendation', "''"),
('a', 'game', 'against', 'an'),
('is', 'an', 'effective', 'method'),
('behavior', 'from', 'a', 'given'),
('and', 'manifold', 'regularization', '.'),
('quantify', 'generalization', 'error', '.'),
('a', 'supervisory', 'signal', '.'),
('yet', 'developed', 'sufficiently', 'to'),
('.', 'The', 'bias–variance', 'decomposition'),
('or', 'unsupervised', 'learning', '.'),
('vector', ',', 'sometimes', 'called'),
```

```
('assumed', 'to', 'be', 'a'),
('called', '``', 'edges', "''"),
('•', 'None', 'Nilsson', ','),
('via', 'the', 'Probably', 'Approximately'),
('This', 'line', ',', 'too'),
('with', 'punched', 'tape', 'memory'),
('The', 'defining', 'characteristic', 'of'),
('on', 'a', 'pre-defined', 'covariance'),
('come', 'from', 'some', 'generally'),
('social', 'network', 'filtering', ','),
('neural', 'network', 'or', 'Neuromorphic'),
('[', '135', ']', '['),
('data', 'recorded', 'by', 'point-of-sale'),
(']', '[', '32', ']'),
('emulate', 'neurons', 'as', 'opposed'),
(';', 'on', 'the', 'computer'),
('supervised', 'learning', 'algorithms', 'learn'),
(',', 'while', 'observations', 'drawn'),
('it', 'useful', ',', 'often'),
('.', 'Machine', 'learning', 'and'),
('of', 'occurrences', ')', 'and'),
('process', 'of', 'reducing', 'the'),
('help', 'users', 'perform', 'more'),
('system', 'for', 'the', 'task'),
('or', 'medication', 'in', 'which'),
('of', '3D', ')', ','),
('mining', 'methods', 'as', '``'),
('started', 'to', 'flourish', 'in'),
('typically', 'does', 'not', 'consider'),
('19', ']', 'Tom', 'M.'),
('suites', 'containing', 'a', 'variety'),
(']', 'In', 'particular', ','),
('the', 'dimensionality', 'reduction', 'techniques'),
('is', 'one', 'way', 'to'),
('of', 'data/software', 'transparency', 'is'),
('that', 'receives', 'a', 'signal'),
("'s", 'interaction', 'with', 'the'),
('of', 'ML', 'should', 'be'),
('not', 'only', 'logic', 'programming'),
('ordinary', 'least', 'squares', '.'),
('of', 'improving', 'health', 'care'),
('hardware', 'computing', 'used', 'in'),
('51', ']', 'Manifold', 'learning'),
('moles', 'coupled', 'with', 'supervised'),
('it', 'must', 'perform', 'a'),
('been', 'developed', 'by', 'Raytheon'),
('learning', '.', 'They', 'seek'),
('bounds', 'on', 'the', 'performance'),
(')', '.', 'Some', 'of'),
('win', 'the', 'Grand', 'Prize'),
('of', 'strong', 'rules', ','),
(',', 'such', 'as', '``'),
('data', 'and', 'neural', 'networks'),
('.', 'There', 'is', 'potential'),
('is', 'concerned', 'with', 'minimizing'),
('a', 'set', 'of', 'relational'),
('.', 'For', 'example', 'topic'),
(':', 'data', 'model', 'and'),
('models', 'that', 'decentralizes', 'the'),
('that', 'they', 'call', 'statistical'),
('million', '.', '[', '79'),
('and', 'utilization', 'of', 'a'),
('this', 'time', 'period', '.'),
(',', 'machine', 'learning', 'grew'),
('network', 'is', 'used', 'to'),
('CAA', ')', '.', '['),
('dataset', ',', 'can', 'be'),
('are', 'called', 'regression', 'trees'),
('2012', ',', 'co-founder', 'of'),
('experience', 'Machine', 'learning', '('),
```

```
(';', 'Mackworth', ',', 'Alan'),
('.', 'The', 'CAA', 'self-learning'),
('are', 'performed', 'each', 'respectively'),
('biases', 'to', 'be', 'mitigated'),
('specific', 'features', '.', 'An'),
('actions', 'that', 'maximize', 'its'),
('co-founder', 'of', 'Sun', 'Microsystems'),
('.', '[', '116', ']'),
('curve', '(', 'AUC', ')'),
('models', 'are', 'trained', 'to'),
('remainder', 'of', 'the', 'data'),
('feature', 'engineering', ',', 'and'),
('class', 'has', 'already', 'been'),
('how', 'software', 'agents', 'ought'),
('does', 'not', 'yield', 'guarantees'),
('reveal', 'their', 'numerators', 'and'),
('[', '53', ']', 'Feature'),
('are', 'traditionally', 'divided', 'into'),
(']', 'In', '2014', ','),
('16', ']', 'A', 'representative'),
('best', 'performance', 'in', 'the'),
('plan', 'recovery', 'paths', 'for'),
('data', ',', 'can', 'produce'),
('other', 'artificial', 'neural', 'networks'),
('a', 'Y', '%', 'chance'),
('quoted', ',', 'more', 'formal'),
('is', 'applicable', 'to', 'other'),
('the', 'future', 'is', 'uncertain'),
(')', 'are', 'a', 'family'),
('in', 'particular', 'can', 'suffer'),
('a', 'major', 'exception', ')'),
('diagnosis', '.', '[', '26'),
('language', 'for', 'representing', 'hypotheses'),
('operating', 'characteristic', '(', 'ROC'),
('text', '.', 'Anomalies', 'are'),
('partitions', 'the', 'data', 'into'),
('35', ']', 'Leo', 'Breiman'),
('which', 'have', 'been', 'developed'),
('it', 'can', 'be', 'challenging'),
('that', 'seem', 'to', 'fit'),
('way', 'that', 'a', 'human'),
('the', 'synonym', 'self-teaching', 'computers'),
('in', '2018', 'this', 'still'),
('is', 'potential', 'for', 'machine'),
('were', 'later', 'found', 'to'),
('It', 'is', 'one', 'of'),
('is', 'computed', 'by', 'some'),
('[', '58', ']', 'Three'),
('KDD', 'task', ',', 'supervised'),
('exists', 'in', 'two', 'environments'),
('learning', 'approaches', 'in', 'particular'),
('including', 'Fei-Fei', 'Li', ','),
('ensemble', 'model', 'to', 'win'),
('kinds', 'of', 'time', 'complexity'),
('machine', 'learning', 'is', 'likely'),
("''", '.', '[', '21'),
('to', 'improve', 'performance', 'on'),
('influences', 'among', 'artists', '.'),
('Netflix', 'realized', 'that', 'viewers'),
('in', 'the', 'machine', 'learning'),
('a', 'subfield', 'of', 'AI'),
('into', 'layers', '.', 'Different'),
('is', 'neither', 'a', 'separate'),
('physical', 'hardware', 'used', 'to'),
('The', 'original', 'goal', 'of'),
('learning', 'models', 'are', 'often'),
('for', 'future', 'outcomes', 'based'),
('59', ']', 'Unsupervised', 'anomaly'),
('[', '69', ']', 'The'),
('machine', "''", 'with', 'punched'),
```

```
('features', '.', 'It', 'has'),
(']', 'Reinforcement', 'learning', 'is'),
('unsupervised', 'learning', 'encompasses', 'other'),
('functions', ',', 'and', 'is'),
('including', 'Web', 'usage', 'mining'),
('that', 'a', 'human', 'brain'),
('e.g.', ',', '2D', ')'),
('users', "'", 'mobile', 'phones'),
('objects', 'are', 'often', 'not'),
('of', 'supervised-learning', 'algorithms', 'include'),
('learning', ',', 'the', 'training'),
('having', 'human', 'programmers', 'specify'),
('134', ']', 'Running', 'machine'),
('individual', 'users', 'of', 'a'),
('is', 'represented', 'by', 'an'),
(')', 'it', 'uses', 'to'),
('PKDD', 'being', 'a', 'major'),
('categorization', ',', 'and', 'sometimes'),
('outcomes', 'thereby', 'furthering', 'the'),
('clean', 'image', 'patch', 'can'),
('this', 'are', 'numerous', ':'),
('on', 'the', 'concept', 'of'),
('called', 'representation', 'learning', 'algorithms'),
('main', 'success', 'came', 'in'),
('machine', 'learning', 'medical', 'diagnostic'),
('in', 'statistics', ',', 'data'),
('that', 'threshold', '.', 'Typically'),
('(', 'discovering', 'hidden', 'patterns'),
('learning', '.', 'It', 'was'),
('polynomial', 'time', '.', 'Negative'),
(',', 'novelties', ',', 'noise'),
('software-based', 'approaches', '.', 'More'),
('Because', 'human', 'languages', 'contain'),
('.', 'Feature', 'learning', 'can'),
('several', 'contexts', '.', 'In'),
('In', 'addition', 'to', 'performing'),
('[', '42', ']', 'The'),
('been', 'applied', 'in', 'the'),
('data', '.', '[', '56'),
(']', 'Other', 'forms', 'of'),
('Probabilistic', 'systems', 'were', 'plagued'),
('and', 'aid', 'researchers', 'in'),
('of', 'finding', 'good', 'solutions'),
('an', 'additional', 'tool', 'to'),
('emotion', 'of', 'being', 'in'),
('of', 'AI-powered', 'systems', 'and'),
('or', 'as', 'a', 'preprocessing'),
('113', ']', '[', '114'),
('factors', 'of', 'variation', 'that'),
('learning', '.', 'Data', 'mining'),
('to', 'solve', 'approximately', '.'),
('Instead', ',', 'a', 'cluster'),
('machine', 'to', 'both', 'learn'),
('data', 'for', 'the', 'computer'),
('.', 'Learning', 'classifier', 'systems'),
('transparency', 'is', 'provided', ','),
('decision', '.', 'By', 'refining'),
('and', 'denominators', 'of', 'the'),
('divided', 'into', 'three', 'broad'),
('using', 'methods', 'such', 'as'),
('emulate', 'a', 'neural', 'synapse'),
('will', 'probably', 'rise', 'tomorrow'),
('Responsible', 'collection', 'of', 'data'),
('by', 'looking', 'for', 'instances'),
(')', 'and', 'ROC', "'s"),
('Other', 'methods', 'are', 'based'),
('statistics', '.', '[', '35'),
(',', 'the', 'training', 'labels'),
('language', 'processing', '.', 'Gordon'),
('non-linear', 'function', 'of', 'the'),
```

```
 ('models', '.', 'A', 'hypothetical'),
 ('approach', 'tries', 'to', 'model'),
 ('.', 'Learners', 'can', 'also'),
 ('is', 'the', 'assignment', 'of'),
 ('higher-dimensional', 'space', '.', 'A'),
 ('such', 'as', 'game', 'theory'),
 ('assumption', 'that', 'the', 'majority'),
 ('e.g', '.', 'MAML', ')'),
 ...]
```

In [32]:
```
len(unq_four_grams)
```

Out[32]:
```
8835
```

In [33]:
```
five_gram = list(nltk.ngrams(tokens, 5))
len(five_gram) #No. of unique four grams
```

Out[33]:
```
9022
```

In [34]:
```
five_gram[1:100]
```

Out[34]:
```
[('of', 'algorithms', 'that', 'improve', 'automatically'),
 ('algorithms', 'that', 'improve', 'automatically', 'through'),
 ('that', 'improve', 'automatically', 'through', 'experience'),
 ('improve', 'automatically', 'through', 'experience', 'Machine'),
 ('automatically', 'through', 'experience', 'Machine', 'learning'),
 ('through', 'experience', 'Machine', 'learning', '('),
 ('experience', 'Machine', 'learning', '(', 'ML'),
 ('Machine', 'learning', '(', 'ML', ')'),
 ('learning', '(', 'ML', ')', 'is'),
 ('(', 'ML', ')', 'is', 'a'),
 ('ML', ')', 'is', 'a', 'field'),
 (')', 'is', 'a', 'field', 'of'),
 ('is', 'a', 'field', 'of', 'inquiry'),
 ('a', 'field', 'of', 'inquiry', 'devoted'),
 ('field', 'of', 'inquiry', 'devoted', 'to'),
 ('of', 'inquiry', 'devoted', 'to', 'understanding'),
 ('inquiry', 'devoted', 'to', 'understanding', 'and'),
 ('devoted', 'to', 'understanding', 'and', 'building'),
 ('to', 'understanding', 'and', 'building', 'methods'),
 ('understanding', 'and', 'building', 'methods', 'that'),
 ('and', 'building', 'methods', 'that', "'learn"),
 ('building', 'methods', 'that', "'learn", "'"),
 ('methods', 'that', "'learn", "'", ','),
 ('that', "'learn", "'", ',', 'that'),
 ("'learn", "'", ',', 'that', 'is'),
 ("'", ',', 'that', 'is', ','),
 (',', 'that', 'is', ',', 'methods'),
 ('that', 'is', ',', 'methods', 'that'),
 ('is', ',', 'methods', 'that', 'leverage'),
 (',', 'methods', 'that', 'leverage', 'data'),
 ('methods', 'that', 'leverage', 'data', 'to'),
 ('that', 'leverage', 'data', 'to', 'improve'),
 ('leverage', 'data', 'to', 'improve', 'performance'),
 ('data', 'to', 'improve', 'performance', 'on'),
 ('to', 'improve', 'performance', 'on', 'some'),
 ('improve', 'performance', 'on', 'some', 'set'),
 ('performance', 'on', 'some', 'set', 'of'),
 ('on', 'some', 'set', 'of', 'tasks'),
 ('some', 'set', 'of', 'tasks', '.'),
 ('set', 'of', 'tasks', '.', '['),
 ('of', 'tasks', '.', '[', '1'),
 ('tasks', '.', '[', '1', ']'),
 ('.', '[', '1', ']', 'It'),
 ('[', '1', ']', 'It', 'is'),
 ('1', ']', 'It', 'is', 'seen'),
```

```
(']', 'It', 'is', 'seen', 'as'),
('It', 'is', 'seen', 'as', 'a'),
('is', 'seen', 'as', 'a', 'part'),
('seen', 'as', 'a', 'part', 'of'),
('as', 'a', 'part', 'of', 'artificial'),
('a', 'part', 'of', 'artificial', 'intelligence'),
('part', 'of', 'artificial', 'intelligence', '.'),
('of', 'artificial', 'intelligence', '.', 'Machine'),
('artificial', 'intelligence', '.', 'Machine', 'learning'),
('intelligence', '.', 'Machine', 'learning', 'algorithms'),
('.', 'Machine', 'learning', 'algorithms', 'build'),
('Machine', 'learning', 'algorithms', 'build', 'a'),
('learning', 'algorithms', 'build', 'a', 'model'),
('algorithms', 'build', 'a', 'model', 'based'),
('build', 'a', 'model', 'based', 'on'),
('a', 'model', 'based', 'on', 'sample'),
('model', 'based', 'on', 'sample', 'data'),
('based', 'on', 'sample', 'data', ','),
('on', 'sample', 'data', ',', 'known'),
('sample', 'data', ',', 'known', 'as'),
('data', ',', 'known', 'as', 'training'),
(',', 'known', 'as', 'training', 'data'),
('known', 'as', 'training', 'data', ','),
('as', 'training', 'data', ',', 'in'),
('training', 'data', ',', 'in', 'order'),
('data', ',', 'in', 'order', 'to'),
(',', 'in', 'order', 'to', 'make'),
('in', 'order', 'to', 'make', 'predictions'),
('order', 'to', 'make', 'predictions', 'or'),
('to', 'make', 'predictions', 'or', 'decisions'),
('make', 'predictions', 'or', 'decisions', 'without'),
('predictions', 'or', 'decisions', 'without', 'being'),
('or', 'decisions', 'without', 'being', 'explicitly'),
('decisions', 'without', 'being', 'explicitly', 'programmed'),
('without', 'being', 'explicitly', 'programmed', 'to'),
('being', 'explicitly', 'programmed', 'to', 'do'),
('explicitly', 'programmed', 'to', 'do', 'so'),
('programmed', 'to', 'do', 'so', '.'),
('to', 'do', 'so', '.', '['),
('do', 'so', '.', '[', '2'),
('so', '.', '[', '2', ']'),
('.', '[', '2', ']', 'Machine'),
('[', '2', ']', 'Machine', 'learning'),
('2', ']', 'Machine', 'learning', 'algorithms'),
(']', 'Machine', 'learning', 'algorithms', 'are'),
('Machine', 'learning', 'algorithms', 'are', 'used'),
('learning', 'algorithms', 'are', 'used', 'in'),
('algorithms', 'are', 'used', 'in', 'a'),
('are', 'used', 'in', 'a', 'wide'),
('used', 'in', 'a', 'wide', 'variety'),
('in', 'a', 'wide', 'variety', 'of'),
('a', 'wide', 'variety', 'of', 'applications'),
('wide', 'variety', 'of', 'applications', ','),
('variety', 'of', 'applications', ',', 'such')]
```

In [35]:

```
temp3 = set(five_gram)
unq_five_gram = list(temp3)
unq_five_gram[:100]
```

Out[35]:

```
[('smaller', 'dimension', 'of', 'data', '('),
 ('the', 'holdout', 'and', 'cross-validation', 'methods'),
 ('of', 'a', 'new', 'point', 'as'),
 ('ML', 'and', 'AI', 'is', 'frequently'),
 ('a', 'signal', 'can', 'process', 'it'),
 (',', 'machines', 'trained', 'on', 'language'),
 ('AI/CS', 'field', ',', 'as', '``'),
 ('to', 'be', 'encountered', 'in', 'the'),
 ('necessary', 'sensitivity', 'for', 'the', 'findings'),
 ('algorithms', 'and', 'computer', 'hardware', 'have'),
```

```
('the', 'algorithms', 'studied', 'in', 'the'),
('in', 'the', 'data', 'and', 'react'),
('learning', 'has', 'been', 'applied', 'in'),
('.', '[', '110', ']', '['),
('learning', 'theory', ',', 'a', 'computation'),
(',', 'and', 'computer', 'vision', ','),
('a', 'system', 'with', 'only', 'one'),
('a', 'uniform', 'representation', 'for', 'input'),
('[', '87', ']', '[', '88'),
('biomedical', 'literature', '.', 'While', 'it'),
('related', 'two', 'objects', 'are', '.'),
(',', 'or', 'imprecise', ';', 'however'),
(',', 'in', 'pattern', 'recognition', 'and'),
('signal', 'at', 'a', 'connection', 'between'),
('a', 'means', 'towards', 'an', 'end'),
('programming', ')', ',', 'such', 'as'),
('.', 'Multilinear', 'subspace', 'learning', 'algorithms'),
('remains', 'a', 'subfield', 'of', 'AI'),
('97', ']', 'Machine', 'learning', 'approaches'),
('Discovery', 'in', 'Databases', '(', 'ECML'),
('specific', 'to', 'classifying', 'data', 'may'),
('analysis', 'algorithm', 'may', 'be', 'able'),
(')', 'is', 'an', 'approach', 'to'),
('Cinematch', 'movie', 'recommendation', 'algorithm', 'by'),
('constitutional', 'and', 'unconscious', 'biases', 'already'),
('[', '17', ']', 'Interest', 'related'),
('dynamic', 'programming', 'techniques', '.', '['),
('reproduce', 'known', 'knowledge', ',', 'while'),
('of', 'theoretical', 'computer', 'science', 'known'),
('Rule-based', 'machine', 'learning', 'approaches', 'include'),
('machine', 'learning', 'technology', 'was', 'used'),
('the', 'Grand', 'Prize', 'in', '2009'),
('are', 'ratios', 'that', 'fail', 'to'),
('technology', 'was', 'used', 'to', 'help'),
('performance', 'measure', 'P', 'if', 'its'),
('.', 'The', 'bias–variance', 'decomposition', 'is'),
('learning', 'algorithms', 'that', 'combine', 'a'),
('behaves', ',', 'and', 'the', 'other'),
('process', '(', 'MDP', ')', '.'),
(')', 'are', 'called', 'regression', 'trees'),
('task', 'is', 'the', 'discovery', 'of'),
('often', 'tag', 'black', 'people', 'as'),
('sales', 'data', 'of', 'a', 'supermarket'),
('absence', 'of', 'such', 'commonalities', 'in'),
('Learning', 'Machines', ',', 'dealing', 'mostly'),
('by', 'artificial', 'intelligence', 'scientists', ','),
('Leo', 'Breiman', 'distinguished', 'two', 'statistical'),
('so', 'there', 'is', 'a', 'Y'),
('changing', 'the', 'data', '.', '['),
('data', 'of', 'previous', 'admissions', 'staff'),
('``', 'Can', 'machines', 'think', '?'),
('dynamic', 'Bayesian', 'networks', '.', 'Generalizations'),
(',', 'sensor', 'data', ',', 'and'),
('Using', 'job', 'hiring', 'data', 'from'),
('allows', 'a', 'machine', 'to', 'both'),
('one', 'of', 'the', 'predictive', 'modeling'),
('examples/tasks', 'after', 'having', 'experienced', 'a'),
('In', 'the', 'early', 'days', 'of'),
('in', 'the', 'field', 'of', 'machine'),
(')', '.', '•', 'Reinforcement', 'learning'),
('range', '.', 'As', 'an', 'example'),
('IBM', 'Watson', 'system', 'failed', 'to'),
('point-of-sale', '(', 'POS', ')', 'systems'),
('logic', ',', 'and', 'probability', 'theory'),
('the', 'common', 'statistical', 'definition', 'of'),
(')', 'and', 'True', 'Negative', 'Rate'),
('sources', 'continue', 'to', 'assert', 'that'),
('such', 'as', 'promotional', 'pricing', 'or'),
('learning', '.', '[', '84', ']'),
('than', 'the', 'function', ',', 'then'),
```

```
('realized', 'that', 'viewers', "'", 'ratings'),
('Labs-Research', 'in', 'collaboration', 'with', 'the'),
(',', 'each', 'training', 'example', 'is'),
('performing', 'specific', 'tasks', ',', 'leading'),
('and', 'models', 'borrowed', 'from', 'statistics'),
('and', 'supervised', 'dictionary', 'learning', '.'),
('conditional', 'independence', 'with', 'a', 'directed'),
('learning', 'techniques', 'have', 'been', 'used'),
('Sparse', 'dictionary', 'learning', 'has', 'also'),
('pioneer', 'in', 'the', 'field', 'of'),
('and', 'Pragmatic', 'Theory', 'built', 'an'),
('using', 'labeled', 'input', 'data', '.'),
('Goebel', ',', 'Randy', '(', '1998'),
('a', 'rule-based', 'machine', 'learning', 'method'),
('method', 'to', 'express', 'a', 'model'),
('did', 'continue', 'within', 'AI', ','),
('anomaly', 'detection', 'techniques', 'require', 'a'),
('used', 'to', 'predict', 'the', 'output'),
('network', 'learns', 'to', 'recognize', '40'),
('learning', 'machine', 'to', 'perform', 'accurately')]
```

In [36]:

```
len(unq_five_gram) #No of unique Five grams
```

Out[36]:

```
8956
```

Training Model

In [37]:

```
listOfBigrams = []
bigramCounts = {}
unigramCounts = {}
def createBigram(data):
   for i in range(len(data)-1):
      if i < len(data) - 1 and data[i+1].islower():

         listOfBigrams.append((data[i], data[i + 1]))

         if (data[i], data[i+1]) in bigramCounts:
            bigramCounts[(data[i], data[i + 1])] += 1
         else:
            bigramCounts[(data[i], data[i + 1])] = 1

      if data[i] in unigramCounts:
         unigramCounts[data[i]] += 1
      else:
         unigramCounts[data[i]] = 1
   return listOfBigrams, unigramCounts, bigramCounts


def calcBigramProb(listOfBigrams, unigramCounts, bigramCounts):
    listOfProb = {}
    for bigram in listOfBigrams:
        word1 = bigram[0]
        word2 = bigram[1]
        listOfProb[bigram] = (bigramCounts.get(bigram))/(unigramCounts.get(
    return listOfProb
```

In [38]:

```
createBigram(tokens)
```

Out[38]:

```
([('Study', 'of'),
  ('of', 'algorithms'),
  ('algorithms', 'that'),
  ('that', 'improve'),
  ('improve', 'automatically'),
  ('automatically', 'through'),
  ('through', 'experience'),
```

```
('through', 'experience'),
('Machine', 'learning'),
(')', 'is'),
('is', 'a'),
('a', 'field'),
('field', 'of'),
('of', 'inquiry'),
('inquiry', 'devoted'),
('devoted', 'to'),

('to', 'understanding'),
('understanding', 'and'),
('and', 'building'),
('building', 'methods'),
('methods', 'that'),
('that', "'learn"),
(',', 'that'),
('that', 'is'),
(',', 'methods'),
('methods', 'that'),
('that', 'leverage'),
('leverage', 'data'),
('data', 'to'),
('to', 'improve'),
('improve', 'performance'),
('performance', 'on'),
('on', 'some'),
('some', 'set'),
('set', 'of'),
('of', 'tasks'),
('It', 'is'),
('is', 'seen'),
('seen', 'as'),
('as', 'a'),
('a', 'part'),
('part', 'of'),
('of', 'artificial'),
('artificial', 'intelligence'),
('Machine', 'learning'),
('learning', 'algorithms'),
('algorithms', 'build'),
('build', 'a'),
('a', 'model'),
('model', 'based'),
('based', 'on'),
('on', 'sample'),
('sample', 'data'),
(',', 'known'),
('known', 'as'),
('as', 'training'),
('training', 'data'),
(',', 'in'),
('in', 'order'),
('order', 'to'),
('to', 'make'),
('make', 'predictions'),
('predictions', 'or'),
('or', 'decisions'),
('decisions', 'without'),
('without', 'being'),
('being', 'explicitly'),
('explicitly', 'programmed'),
('programmed', 'to'),
('to', 'do'),
('do', 'so'),
('Machine', 'learning'),
('learning', 'algorithms'),
('algorithms', 'are'),
('are', 'used'),
('used', 'in'),
('in', 'a'),
```

```
('a', 'wide'),
('wide', 'variety'),
('variety', 'of'),
('of', 'applications'),
(',', 'such'),
('such', 'as'),
('as', 'in'),
('in', 'medicine'),
(',', 'email'),
('email', 'filtering'),
(',', 'speech'),
('speech', 'recognition'),
(',', 'agriculture'),
(',', 'and'),
('and', 'computer'),
('computer', 'vision'),
(',', 'where'),
('where', 'it'),
('it', 'is'),
('is', 'difficult'),
('difficult', 'or'),
('or', 'unfeasible'),
('unfeasible', 'to'),
('to', 'develop'),
('develop', 'conventional'),
('conventional', 'algorithms'),
('algorithms', 'to'),
('to', 'perform'),
('perform', 'the'),
('the', 'needed'),
('needed', 'tasks'),
('A', 'subset'),
('subset', 'of'),
('of', 'machine'),
('machine', 'learning'),
('learning', 'is'),
('is', 'closely'),
('closely', 'related'),
('related', 'to'),
('to', 'computational'),
('computational', 'statistics'),
(',', 'which'),
('which', 'focuses'),
('focuses', 'on'),
('on', 'making'),
('making', 'predictions'),
('predictions', 'using'),
('using', 'computers'),
(',', 'but'),
('but', 'not'),
('not', 'all'),
('all', 'machine'),
('machine', 'learning'),
('learning', 'is'),
('is', 'statistical'),
('statistical', 'learning'),
('The', 'study'),
('study', 'of'),
('of', 'mathematical'),
('mathematical', 'optimization'),
('optimization', 'delivers'),
('delivers', 'methods'),
(',', 'theory'),
('theory', 'and'),
('and', 'application'),
('application', 'domains'),
('domains', 'to'),
('to', 'the'),
('the', 'field'),
('field', 'of')
```

```
('field', 'of'),
('of', 'machine'),
('machine', 'learning'),
('Data', 'mining'),
('mining', 'is'),
('is', 'a'),
('a', 'related'),
('related', 'field'),
('field', 'of'),
('of', 'study'),
(',', 'focusing'),
('focusing', 'on'),
('on', 'exploratory'),
('exploratory', 'data'),
('data', 'analysis'),
('analysis', 'through'),
('through', 'unsupervised'),
('unsupervised', 'learning'),
('Some', 'implementations'),
('implementations', 'of'),
('of', 'machine'),
('machine', 'learning'),
('learning', 'use'),
('use', 'data'),
('data', 'and'),
('and', 'neural'),
('neural', 'networks'),
('networks', 'in'),
('in', 'a'),
('a', 'way'),
('way', 'that'),
('that', 'mimics'),
('mimics', 'the'),
('the', 'working'),
('working', 'of'),
('of', 'a'),
('a', 'biological'),
('biological', 'brain'),
('In', 'its'),
('its', 'application'),
('application', 'across'),
('across', 'business'),
('business', 'problems'),
(',', 'machine'),
('machine', 'learning'),
('learning', 'is'),
('is', 'also'),
('also', 'referred'),
('referred', 'to'),
('to', 'as'),
('as', 'predictive'),
('predictive', 'analytics'),
('Learning', 'algorithms'),
('algorithms', 'work'),
('work', 'on'),
('on', 'the'),
('the', 'basis'),
('basis', 'that'),
('that', 'strategies'),
(',', 'algorithms'),
(',', 'and'),
('and', 'inferences'),
('inferences', 'that'),
('that', 'worked'),
('worked', 'well'),
('well', 'in'),
('in', 'the'),
('the', 'past'),
('past', 'are'),
('are', 'likely'),
('likely', 'to')
```

```
('likely', 'to'),
('to', 'continue'),
('continue', 'working'),
('working', 'well'),
('well', 'in'),
('in', 'the'),
('the', 'future'),
('These', 'inferences'),
('inferences', 'can'),
('can', 'be'),
('be', 'obvious'),
(',', 'such'),
('such', 'as'),
('``', 'since'),
('since', 'the'),
('the', 'sun'),
('sun', 'rose'),
('rose', 'every'),
('every', 'morning'),
('morning', 'for'),
('for', 'the'),
('the', 'last'),
('10,000', 'days'),
(',', 'it'),
('it', 'will'),
('will', 'probably'),
('probably', 'rise'),
('rise', 'tomorrow'),
('tomorrow', 'morning'),
('morning', 'as'),
('as', 'well'),
('They', 'can'),
('can', 'be'),
('be', 'nuanced'),
(',', 'such'),
('such', 'as'),
('%', 'of'),
('of', 'families'),
('families', 'have'),
('have', 'geographically'),
('geographically', 'separate'),
('separate', 'species'),
('species', 'with'),
('with', 'color'),
('color', 'variants'),
(',', 'so'),
('so', 'there'),
('there', 'is'),
('is', 'a'),
('%', 'chance'),
('chance', 'that'),
('that', 'undiscovered'),
('undiscovered', 'black'),
('black', 'swans'),
('swans', 'exist'),
('Machine', 'learning'),
('learning', 'programs'),
('programs', 'can'),
('can', 'perform'),
('perform', 'tasks'),
('tasks', 'without'),
('without', 'being'),
('being', 'explicitly'),
('explicitly', 'programmed'),
('programmed', 'to'),
('to', 'do'),
('do', 'so'),
('It', 'involves'),
('involves', 'computers'),
('computers', 'learning'),
('learning', 'from')
```

```
('learning', 'from'),
('from', 'data'),
('data', 'provided'),
('provided', 'so'),
('so', 'that'),
('that', 'they'),
('they', 'carry'),
('carry', 'out'),
('out', 'certain'),
('certain', 'tasks'),
('For', 'simple'),
('simple', 'tasks'),
('tasks', 'assigned'),
('assigned', 'to'),
('to', 'computers'),
(',', 'it'),
('it', 'is'),
('is', 'possible'),
('possible', 'to'),
('to', 'program'),
('program', 'algorithms'),
('algorithms', 'telling'),
('telling', 'the'),
('the', 'machine'),
('machine', 'how'),
('how', 'to'),
('to', 'execute'),
('execute', 'all'),
('all', 'steps'),
('steps', 'required'),
('required', 'to'),
('to', 'solve'),
('solve', 'the'),
('the', 'problem'),
('problem', 'at'),
('at', 'hand'),
(';', 'on'),
('on', 'the'),
('the', 'computer'),
('computer', "'s"),
("'s", 'part'),
(',', 'no'),
('no', 'learning'),
('learning', 'is'),
('is', 'needed'),
('For', 'more'),
('more', 'advanced'),
('advanced', 'tasks'),
(',', 'it'),
('it', 'can'),
('can', 'be'),
('be', 'challenging'),
('challenging', 'for'),
('for', 'a'),
('a', 'human'),
('human', 'to'),
('to', 'manually'),
('manually', 'create'),
('create', 'the'),
('the', 'needed'),
('needed', 'algorithms'),
('In', 'practice'),
(',', 'it'),
('it', 'can'),
('can', 'turn'),
('turn', 'out'),
('out', 'to'),
('to', 'be'),
('be', 'more'),
('more', 'effective'),
```

```
('effective', 'to'),
('to', 'help'),
('help', 'the'),
('the', 'machine'),
('machine', 'develop'),
('develop', 'its'),
('its', 'own'),
('own', 'algorithm'),
(',', 'rather'),
('rather', 'than'),
('than', 'having'),
('having', 'human'),
('human', 'programmers'),
('programmers', 'specify'),
('specify', 'every'),
('every', 'needed'),
('needed', 'step'),
('The', 'discipline'),
('discipline', 'of'),
('of', 'machine'),
('machine', 'learning'),
('learning', 'employs'),
('employs', 'various'),
('various', 'approaches'),
('approaches', 'to'),
('to', 'teach'),
('teach', 'computers'),
('computers', 'to'),
('to', 'accomplish'),
('accomplish', 'tasks'),
('tasks', 'where'),
('where', 'no'),
('no', 'fully'),
('fully', 'satisfactory'),
('satisfactory', 'algorithm'),
('algorithm', 'is'),
('is', 'available'),
('In', 'cases'),
('cases', 'where'),
('where', 'vast'),
('vast', 'numbers'),
('numbers', 'of'),
('of', 'potential'),
('potential', 'answers'),
('answers', 'exist'),
(',', 'one'),
('one', 'approach'),
('approach', 'is'),
('is', 'to'),
('to', 'label'),
('label', 'some'),
('some', 'of'),
('of', 'the'),
('the', 'correct'),
('correct', 'answers'),
('answers', 'as'),
('as', 'valid'),
('This', 'can'),
('can', 'then'),
('then', 'be'),
('be', 'used'),
('used', 'as'),
('as', 'training'),
('training', 'data'),
('data', 'for'),
('for', 'the'),
('the', 'computer'),
('computer', 'to'),
('to', 'improve'),
('improve', 'the'),
```

```
('the', 'algorithm'),
('(', 's'),
(')', 'it'),
('it', 'uses'),
('uses', 'to'),
('to', 'determine'),
('determine', 'correct'),
('correct', 'answers'),
('For', 'example'),
(',', 'to'),
('to', 'train'),
('train', 'a'),
('a', 'system'),
('system', 'for'),
('for', 'the'),
('the', 'task'),
('task', 'of'),
('of', 'digital'),
('digital', 'character'),
('character', 'recognition'),
(',', 'the'),
('MNIST', 'dataset'),
('dataset', 'of'),
('of', 'handwritten'),
('handwritten', 'digits'),
('digits', 'has'),
('has', 'often'),
('often', 'been'),
('been', 'used'),
('History', 'and'),
('and', 'relationships'),
('relationships', 'to'),
('to', 'other'),
('other', 'fields'),
('The', 'term'),
('term', 'machine'),
('machine', 'learning'),
('learning', 'was'),
('was', 'coined'),
('coined', 'in'),
('1959', 'by'),
(',', 'an'),
('IBM', 'employee'),
('employee', 'and'),
('and', 'pioneer'),
('pioneer', 'in'),
('in', 'the'),
('the', 'field'),
('field', 'of'),
('of', 'computer'),
('computer', 'gaming'),
('gaming', 'and'),
('and', 'artificial'),
('artificial', 'intelligence'),
('Also', 'the'),
('the', 'synonym'),
('synonym', 'self-teaching'),
('self-teaching', 'computers'),
('computers', 'were'),
('were', 'used'),
('used', 'in'),
('in', 'this'),
('this', 'time'),
('time', 'period'),
('By', 'the'),
('the', 'early'),
('early', '1960s'),
('1960s', 'an'),
('an', 'experimental'),
('``', 'learning'),
```

```
('learning', 'machine'),
("''", 'with'),
('with', 'punched'),
('punched', 'tape'),
('tape', 'memory'),
(',', 'called'),
(',', 'had'),
('had', 'been'),
('been', 'developed'),
('developed', 'by'),
('Company', 'to'),
('to', 'analyze'),
('analyze', 'sonar'),
('sonar', 'signals'),
(',', 'electrocardiograms'),
(',', 'and'),
('and', 'speech'),
('speech', 'patterns'),
('patterns', 'using'),
('using', 'rudimentary'),
('rudimentary', 'reinforcement'),
('reinforcement', 'learning'),
('It', 'was'),
('was', 'repetitively'),
('``', 'trained'),
("''", 'by'),
('by', 'a'),
('a', 'human'),
('human', 'operator/teacher'),
('operator/teacher', 'to'),
('to', 'recognize'),
('recognize', 'patterns'),
('patterns', 'and'),
('and', 'equipped'),
('equipped', 'with'),
('with', 'a'),
('``', 'goof'),
("''", 'button'),
('button', 'to'),
('to', 'cause'),
('cause', 'it'),
('it', 'to'),
('to', 're-evaluate'),
('re-evaluate', 'incorrect'),
('incorrect', 'decisions'),
('A', 'representative'),
('representative', 'book'),
('book', 'on'),
('on', 'research'),
('research', 'into'),
('into', 'machine'),
('machine', 'learning'),
('learning', 'during'),
('during', 'the'),
('the', '1960s'),
('1960s', 'was'),
('Nilsson', "'s"),
("'s", 'book'),
('book', 'on'),
(',', 'dealing'),
('dealing', 'mostly'),
('mostly', 'with'),
('with', 'machine'),
('machine', 'learning'),
('learning', 'for'),
('for', 'pattern'),
('pattern', 'classification'),
('Interest', 'related'),
('related', 'to'),
('to', 'pattern'),
```

```
('pattern', 'recognition'),
('recognition', 'continued'),
('continued', 'into'),
('into', 'the'),
('the', '1970s'),
(',', 'as'),
('as', 'described'),
('described', 'by'),
('Duda', 'and'),
('Hart', 'in'),
('1981', 'a'),
('a', 'report'),
('report', 'was'),
('was', 'given'),
('given', 'on'),
('on', 'using'),
('using', 'teaching'),
('teaching', 'strategies'),
('strategies', 'so'),
('so', 'that'),
('that', 'a'),
('a', 'neural'),
('neural', 'network'),
('network', 'learns'),
('learns', 'to'),
('to', 'recognize'),
('40', 'characters'),
('26', 'letters'),
('10', 'digits'),
(',', 'and'),
('4', 'special'),
('special', 'symbols'),
(')', 'from'),
('from', 'a'),
('a', 'computer'),
('computer', 'terminal'),
('Mitchell', 'provided'),
('provided', 'a'),
('a', 'widely'),
('widely', 'quoted'),
(',', 'more'),
('more', 'formal'),
('formal', 'definition'),
('definition', 'of'),
('of', 'the'),
('the', 'algorithms'),
('algorithms', 'studied'),
('studied', 'in'),
('in', 'the'),
('the', 'machine'),
('machine', 'learning'),
('learning', 'field'),
('A', 'computer'),
('computer', 'program'),
('program', 'is'),
('is', 'said'),
('said', 'to'),
('to', 'learn'),
('learn', 'from'),
('from', 'experience'),
('E', 'with'),
('with', 'respect'),
('respect', 'to'),
('to', 'some'),
('some', 'class'),
('class', 'of'),
('of', 'tasks'),
('T', 'and'),
('and', 'performance'),
('performance', 'measure'),
```

```
('P', 'if'),
('if', 'its'),
('its', 'performance'),
('performance', 'at'),
('at', 'tasks'),
('tasks', 'in'),
(',', 'as'),
('as', 'measured'),
('measured', 'by'),
(',', 'improves'),
('improves', 'with'),
('with', 'experience'),
('This', 'definition'),
('definition', 'of'),
('of', 'the'),
('the', 'tasks'),
('tasks', 'in'),
('in', 'which'),
('which', 'machine'),
('machine', 'learning'),
('learning', 'is'),
('is', 'concerned'),
('concerned', 'offers'),
('offers', 'a'),
('a', 'fundamentally'),
('fundamentally', 'operational'),
('operational', 'definition'),
('definition', 'rather'),
('rather', 'than'),
('than', 'defining'),
('defining', 'the'),
('the', 'field'),
('field', 'in'),
('in', 'cognitive'),
('cognitive', 'terms'),
('This', 'follows'),
('Turing', "'s"),
("'s", 'proposal'),
('proposal', 'in'),
('in', 'his'),
('his', 'paper'),
('Machinery', 'and'),
(',', 'in'),
('in', 'which'),
('which', 'the'),
('the', 'question'),
('Can', 'machines'),
('machines', 'think'),
("''", 'is'),
('is', 'replaced'),
('replaced', 'with'),
('with', 'the'),
('the', 'question'),
('Can', 'machines'),
('machines', 'do'),
('do', 'what'),
('what', 'we'),
('(', 'as'),
('as', 'thinking'),
('thinking', 'entities'),
(')', 'can'),
('can', 'do'),
('Modern-day', 'machine'),
('machine', 'learning'),
('learning', 'has'),
('has', 'two'),
('two', 'objectives'),
(',', 'one'),
('one', 'is'),
('is', 'to'),
```

```
('to', 'classify'),
('classify', 'data'),
('data', 'based'),
('based', 'on'),
('on', 'models'),
('models', 'which'),
('which', 'have'),
('have', 'been'),
('been', 'developed'),
(',', 'the'),
('the', 'other'),
('other', 'purpose'),
('purpose', 'is'),
('is', 'to'),
('to', 'make'),
('make', 'predictions'),
('predictions', 'for'),
('for', 'future'),
('future', 'outcomes'),
('outcomes', 'based'),
('based', 'on'),
('on', 'these'),
('these', 'models'),
('A', 'hypothetical'),
('hypothetical', 'algorithm'),
('algorithm', 'specific'),
('specific', 'to'),
('to', 'classifying'),
('classifying', 'data'),
('data', 'may'),
('may', 'use'),
('use', 'computer'),
('computer', 'vision'),
('vision', 'of'),
('of', 'moles'),
('moles', 'coupled'),
('coupled', 'with'),
('with', 'supervised'),
('supervised', 'learning'),
('learning', 'in'),
('in', 'order'),
('order', 'to'),
('to', 'train'),
('train', 'it'),
('it', 'to'),
('to', 'classify'),
('classify', 'the'),
('the', 'cancerous'),
('cancerous', 'moles'),
('A', 'machine'),
('machine', 'learning'),
('learning', 'algorithm'),
('algorithm', 'for'),
('for', 'stock'),
('stock', 'trading'),
('trading', 'may'),
('may', 'inform'),
('inform', 'the'),
('the', 'trader'),
('trader', 'of'),
('of', 'future'),
('future', 'potential'),
('potential', 'predictions'),
('As', 'a'),
('a', 'scientific'),
('scientific', 'endeavor'),
(',', 'machine'),
('machine', 'learning'),
('learning', 'grew'),
('grew', 'out'),
```

```
('out', 'of'),
('of', 'the'),
('the', 'quest'),
('quest', 'for'),
('for', 'artificial'),
('artificial', 'intelligence'),
('In', 'the'),
('the', 'early'),
('early', 'days'),
('days', 'of'),
('AI', 'as'),
('as', 'an'),
('an', 'academic'),
('academic', 'discipline'),
(',', 'some'),
('some', 'researchers'),
('researchers', 'were'),
('were', 'interested'),
('interested', 'in'),
('in', 'having'),
('having', 'machines'),
('machines', 'learn'),
('learn', 'from'),
('from', 'data'),
('They', 'attempted'),
('attempted', 'to'),
('to', 'approach'),
('approach', 'the'),
('the', 'problem'),
('problem', 'with'),
('with', 'various'),
('various', 'symbolic'),
('symbolic', 'methods'),
(',', 'as'),
('as', 'well'),
('well', 'as'),
('as', 'what'),
('what', 'was'),
('was', 'then'),
('then', 'termed'),
('``', 'neural'),
('neural', 'networks'),
(';', 'these'),
('these', 'were'),
('were', 'mostly'),
('mostly', 'perceptrons'),
('perceptrons', 'and'),
('and', 'other'),
('other', 'models'),
('models', 'that'),
('that', 'were'),
('were', 'later'),
('later', 'found'),
('found', 'to'),
('to', 'be'),
('be', 'reinventions'),
('reinventions', 'of'),
('of', 'the'),
('the', 'generalized'),
('generalized', 'linear'),
('linear', 'models'),
('models', 'of'),
('of', 'statistics'),
('Probabilistic', 'reasoning'),
('reasoning', 'was'),
('was', 'also'),
('also', 'employed'),
(',', 'especially'),
('especially', 'in'),
('in', 'automated'),
```

```
('automated', 'medical'),
('medical', 'diagnosis'),
(',', 'an'),
('an', 'increasing'),
('increasing', 'emphasis'),
('emphasis', 'on'),
('on', 'the'),
('the', 'logical'),
(',', 'knowledge-based'),
('knowledge-based', 'approach'),
('approach', 'caused'),
('caused', 'a'),
('a', 'rift'),
('rift', 'between'),
('AI', 'and'),
('and', 'machine'),
('machine', 'learning'),
('Probabilistic', 'systems'),
('systems', 'were'),
('were', 'plagued'),
('plagued', 'by'),
('by', 'theoretical'),
('theoretical', 'and'),
('and', 'practical'),
('practical', 'problems'),
('problems', 'of'),
('of', 'data'),
('data', 'acquisition'),
('acquisition', 'and'),
('and', 'representation'),
(',', 'expert'),
('expert', 'systems'),
('systems', 'had'),
('had', 'come'),
('come', 'to'),
('to', 'dominate'),
(',', 'and'),
('and', 'statistics'),
('statistics', 'was'),
('was', 'out'),
('out', 'of'),
('of', 'favor'),
('Work', 'on'),
('on', 'symbolic/knowledge-based'),
('symbolic/knowledge-based', 'learning'),
('learning', 'did'),
('did', 'continue'),
('continue', 'within'),
(',', 'leading'),
('leading', 'to'),
('to', 'inductive'),
('inductive', 'logic'),
('logic', 'programming'),
(',', 'but'),
('but', 'the'),
('the', 'more'),
('more', 'statistical'),
('statistical', 'line'),
('line', 'of'),
('of', 'research'),
('research', 'was'),
('was', 'now'),
('now', 'outside'),
('outside', 'the'),
('the', 'field'),
('field', 'of'),
('AI', 'proper'),
(',', 'in'),
('in', 'pattern'),
('pattern', 'recognition'),
```

```
('recognition', 'and'),
('and', 'information'),
('information', 'retrieval'),
('Neural', 'networks'),
('networks', 'research'),
('research', 'had'),
('had', 'been'),
('been', 'abandoned'),
('abandoned', 'by'),
('AI', 'and'),
('and', 'computer'),
('computer', 'science'),
('science', 'around'),
('around', 'the'),
('the', 'same'),
('same', 'time'),
('This', 'line'),
(',', 'too'),
(',', 'was'),
('was', 'continued'),
('continued', 'outside'),
('outside', 'the'),
('AI/CS', 'field'),
(',', 'as'),
('``', 'connectionism'),
(',', 'by'),
('by', 'researchers'),
('researchers', 'from'),
('from', 'other'),
('other', 'disciplines'),
('disciplines', 'including'),
(',', 'and'),
('Their', 'main'),
('main', 'success'),
('success', 'came'),
('came', 'in'),
('in', 'the'),
('the', 'mid-1980s'),
('mid-1980s', 'with'),
('with', 'the'),
('the', 'reinvention'),
('reinvention', 'of'),
('of', 'backpropagation'),
('Machine', 'learning'),
(',', 'reorganized'),
('reorganized', 'as'),
('as', 'a'),
('a', 'separate'),
('separate', 'field'),
(',', 'started'),
('started', 'to'),
('to', 'flourish'),
('flourish', 'in'),
('in', 'the'),
('the', '1990s'),
('The', 'field'),
('field', 'changed'),
('changed', 'its'),
('its', 'goal'),
('goal', 'from'),
('from', 'achieving'),
('achieving', 'artificial'),
('artificial', 'intelligence'),
('intelligence', 'to'),
('to', 'tackling'),
('tackling', 'solvable'),
('solvable', 'problems'),
('problems', 'of'),
('of', 'a'),
('a', 'practical'),
```

```
 ('practical', 'nature'),
 ('It', 'shifted'),
 ('shifted', 'focus'),
 ('focus', 'away'),
 ('away', 'from'),
 ('from', 'the'),
 ('the', 'symbolic'),
 ('symbolic', 'approaches'),
 ('approaches', 'it'),
 ('it', 'had'),
 ('had', 'inherited'),
 ('inherited', 'from'),
 (',', 'and'),
 ('and', 'toward'),
 ('toward', 'methods'),
 ...],
{'Study': 1,
 'of': 240,
 'algorithms': 45,
 'that': 97,
 'improve': 6,
 'automatically': 1,
 'through': 4,
 'experience': 4,
 'Machine': 21,
 'learning': 189,
 '(': 79,
 'ML': 7,
 ')': 79,
 'is': 132,
 'a': 218,
 'field': 18,
 'inquiry': 1,
 'devoted': 1,
 'to': 219,
 'understanding': 1,
 'and': 182,
 'building': 1,
 'methods': 23,
 "'learn": 1,
 "'": 8,
 ',': 408,
 'leverage': 1,
 'data': 95,
 'performance': 12,
 'on': 57,
 'some': 12,
 'set': 31,
 'tasks': 15,
 '.': 336,
 '[': 151,
 '1': 3,
 ']': 151,
 'It': 13,
 'seen': 2,
 'as': 66,
 'part': 5,
 'artificial': 21,
 'intelligence': 8,
 'build': 3,
 'model': 36,
 'based': 9,
 'sample': 3,
 'known': 11,
 'training': 48,
 'in': 146,
 'order': 6,
 'make': 10,
 'predictions': 13,
```

'decisions': 8,
'without': 9,
'being': 8,
'explicitly': 3,
'programmed': 3,
'do': 11,
'so': 10,
'2': 1,
'are': 56,
'used': 35,
'wide': 1,
'variety': 4,
'applications': 4,
'such': 21,
'medicine': 1,
'email': 3,
'filtering': 2,
'speech': 5,
'recognition': 6,
'agriculture': 1,
'computer': 16,
'vision': 5,
'where': 12,
'it': 30,
'difficult': 2,
'unfeasible': 1,
'develop': 2,
'conventional': 1,
'perform': 12,
'the': 341,
'needed': 5,
'3': 1,
'4': 2,
'A': 18,
'subset': 3,
'machine': 93,
'closely': 3,
'related': 11,
'computational': 4,
'statistics': 10,
'which': 25,
'focuses': 3,
'making': 2,
'using': 13,
'computers': 6,
'but': 17,
'not': 27,
'all': 10,
'statistical': 8,
'The': 29,
'study': 5,
'mathematical': 7,
'optimization': 9,
'delivers': 1,
'theory': 13,
'application': 5,
'domains': 3,
'Data': 3,
'mining': 11,
'focusing': 1,
'exploratory': 1,
'analysis': 12,
'unsupervised': 11,
'6': 1,
'7': 1,
'Some': 5,
'implementations': 2,
'use': 10,
'neural': 16,

```
'networks': 13,
'way': 5,
'mimics': 2,
'working': 2,
'biological': 4,
'brain': 5,
'8': 1,
'9': 1,
'In': 35,
'its': 17,
'across': 2,
'business': 2,
'problems': 11,
'also': 20,
'referred': 2,
'predictive': 4,
'analytics': 1,
'Learning': 9,
'work': 3,
'basis': 3,
'strategies': 2,
'inferences': 3,
'worked': 1,
'well': 7,
'past': 2,
'likely': 4,
'continue': 3,
'future': 4,
'These': 2,
'can': 45,
'be': 48,
'obvious': 1,
'``': 35,
'since': 1,
'sun': 1,
'rose': 1,
'every': 3,
'morning': 2,
'for': 54,
'last': 2,
'10,000': 1,
'days': 2,
'will': 7,
'probably': 1,
'rise': 1,
'tomorrow': 1,
"''": 33,
'They': 3,
'nuanced': 1,
'X': 1,
'%': 4,
'families': 1,
'have': 21,
'geographically': 1,
'separate': 5,
'species': 2,
'with': 49,
'color': 1,
'variants': 1,
'there': 2,
'Y': 1,
'chance': 2,
'undiscovered': 1,
'black': 5,
'swans': 1,
'exist': 5,
'programs': 4,
'involves': 4,
'from': 49,
'provided': 5,
```

```
'they': 6,
'carry': 1,
'out': 5,
'certain': 5,
'For': 11,
'simple': 1,
'assigned': 1,
'possible': 1,
'program': 9,
'telling': 1,
'how': 7,
'execute': 1,
'steps': 1,
'required': 2,
'solve': 4,
'problem': 6,
'at': 8,
'hand': 2,
';': 7,
"'s": 16,
'no': 5,
'more': 13,
'advanced': 1,
'challenging': 1,
'human': 8,
'manually': 1,
'create': 1,
'practice': 1,
'turn': 1,
'effective': 4,
'help': 3,
'own': 2,
'algorithm': 22,
'rather': 3,
'than': 6,
'having': 4,
'programmers': 1,
'specify': 1,
'step': 4,
'11': 3,
'discipline': 2,
'employs': 2,
'various': 5,
'approaches': 8,
'teach': 1,
'accomplish': 1,
'fully': 2,
'satisfactory': 1,
'available': 2,
'cases': 3,
'vast': 1,
'numbers': 2,
'potential': 4,
'answers': 3,
'one': 15,
'approach': 7,
'label': 2,
'correct': 2,
'valid': 1,
'This': 12,
'then': 9,
's': 6,
'uses': 4,
'determine': 3,
'example': 20,
'train': 3,
'system': 12,
'task': 5,
'digital': 1,
'character': 1,
```

'MNIST': 1,
'dataset': 2,
'handwritten': 1,
'digits': 2,
'has': 20,
'often': 17,
'been': 21,
'History': 1,
'relationships': 4,
'other': 17,
'fields': 4,
'term': 5,
'was': 18,
'coined': 1,
'1959': 1,
'by': 47,
'Arthur': 1,
'Samuel': 1,
'an': 41,
'IBM': 2,
'employee': 1,
'pioneer': 1,
'gaming': 1,
'12': 1,
'13': 1,
'Also': 1,
'synonym': 1,
'self-teaching': 1,
'were': 9,
'this': 10,
'time': 10,
'period': 1,
'14': 1,
'15': 1,
'By': 4,
'early': 2,
'1960s': 2,
'experimental': 1,
'punched': 1,
'tape': 1,
'memory': 3,
'called': 12,
'Cybertron': 1,
'had': 10,
'developed': 5,
'Raytheon': 1,
'Company': 1,
'analyze': 1,
'sonar': 1,
'signals': 2,
'electrocardiograms': 1,
'patterns': 7,
'rudimentary': 1,
'reinforcement': 6,
'repetitively': 1,
'trained': 11,
'operator/teacher': 1,
'recognize': 3,
'equipped': 1,
'goof': 1,
'button': 1,
'cause': 1,
're-evaluate': 1,
'incorrect': 1,
'16': 1,
'representative': 3,
'book': 3,
'research': 8,
'into': 12,
'during': 2,

'Nilsson': 2,
'Machines': 1,
'dealing': 2,
'mostly': 2,
'pattern': 4,
'classification': 17,
'17': 1,
'Interest': 1,
'continued': 2,
'1970s': 1,
'described': 1,
'Duda': 1,
'Hart': 1,
'1973': 1,
'18': 1,
'1981': 2,
'report': 3,
'given': 6,
'teaching': 1,
'network': 12,
'learns': 5,
'40': 2,
'characters': 1,
'26': 7,
'letters': 1,
'10': 2,
'special': 1,
'symbols': 1,
'terminal': 1,
'19': 1,
'Tom': 1,
'M.': 1,
'Mitchell': 1,
'widely': 1,
'quoted': 1,
'formal': 1,
'definition': 4,
'studied': 2,
':': 23,
'said': 2,
'learn': 12,
'E': 1,
'respect': 3,
'class': 7,
'T': 3,
'measure': 2,
'P': 2,
'if': 5,
'measured': 1,
'improves': 2,
'E.': 1,
'20': 2,
'concerned': 3,
'offers': 1,
'fundamentally': 1,
'operational': 1,
'defining': 2,
'cognitive': 1,
'terms': 3,
'follows': 1,
'Alan': 2,
'Turing': 1,
'proposal': 1,
'his': 1,
'paper': 1,
'Computing': 1,
'Machinery': 1,
'Intelligence': 3,
'question': 2,
'Can': 2,

```
'machines': 6,
'think': 1,
'?': 2,
'replaced': 1,
'what': 3,
'we': 2,
'thinking': 1,
'entities': 1,
'21': 1,
'Modern-day': 1,
'two': 8,
'objectives': 2,
'classify': 2,
'models': 24,
'purpose': 1,
'outcomes': 2,
'these': 12,
'hypothetical': 1,
'specific': 5,
'classifying': 2,
'may': 13,
'moles': 2,
'coupled': 1,
'supervised': 13,
'cancerous': 1,
'stock': 1,
'trading': 1,
'inform': 1,
'trader': 1,
'22': 2,
'As': 5,
'scientific': 1,
'endeavor': 1,
'grew': 1,
'quest': 1,
'AI': 20,
'academic': 2,
'researchers': 5,
'interested': 1,
'attempted': 1,
'symbolic': 2,
'termed': 1,
'perceptrons': 2,
'later': 1,
'found': 8,
'reinventions': 1,
'generalized': 1,
'linear': 5,
'25': 2,
'Probabilistic': 2,
'reasoning': 2,
'employed': 2,
'especially': 3,
'automated': 2,
'medical': 5,
'diagnosis': 2,
'488': 2,
'However': 4,
'increasing': 1,
'emphasis': 1,
'logical': 4,
'knowledge-based': 1,
'caused': 1,
'rift': 1,
'between': 18,
'systems': 20,
'plagued': 1,
'theoretical': 4,
'practical': 2,
'acquisition': 1,
```

'representation': 7,
'1980': 1,
'expert': 1,
'come': 2,
'dominate': 1,
'favor': 1,
'27': 3,
'Work': 1,
'symbolic/knowledge-based': 1,
'did': 1,
'within': 5,
'leading': 4,
'inductive': 3,
'logic': 8,
'programming': 8,
'line': 3,
'now': 1,
'outside': 2,
'proper': 1,
'information': 7,
'retrieval': 1,
'708–710': 1,
'755': 1,
'Neural': 1,
'abandoned': 1,
'science': 3,
'around': 1,
'same': 6,
'too': 2,
'AI/CS': 1,
'connectionism': 1,
'disciplines': 2,
'including': 8,
'Hopfield': 1,
'Rumelhart': 1,
'Hinton': 1,
'Their': 1,
'main': 1,
'success': 1,
'came': 1,
'mid-1980s': 1,
'reinvention': 1,
'backpropagation': 1,
'reorganized': 1,
'started': 1,
'flourish': 1,
'1990s': 2,
'changed': 2,
'goal': 8,
'achieving': 2,
'tackling': 1,
'solvable': 1,
'nature': 3,
'shifted': 1,
'focus': 1,
'away': 1,
'inherited': 1,
'toward': 2,
'borrowed': 1,
'fuzzy': 1,
'probability': 3,
'difference': 4,
'frequently': 2,
'misunderstood': 1,
'predicts': 2,
'passive': 1,
'observations': 6,
'whereas': 1,
'implies': 1,
'agent': 1,

```
'interacting': 1,
'environment': 10,
'take': 6,
'actions': 3,
'maximize': 3,
'successfully': 1,
'goals': 2,
'28': 2,
'2020': 4,
'many': 13,
'sources': 1,
'assert': 1,
'remains': 1,
'subfield': 1,
'29': 2,
'30': 2,
'Others': 1,
'view': 1,
'only': 9,
"'intelligent": 1,
'should': 2,
'considered': 4,
'5': 3,
'31': 2,
'32': 2,
'employ': 1,
'overlap': 1,
'significantly': 2,
'while': 8,
'prediction': 3,
'properties': 3,
'learned': 10,
'discovery': 5,
'previously': 5,
'unknown': 4,
'knowledge': 11,
'databases': 3,
'different': 5,
'preprocessing': 1,
'learner': 3,
'accuracy': 7,
'Much': 1,
'confusion': 1,
'communities': 1,
'conferences': 1,
'journals': 1,
'ECML': 2,
'PKDD': 2,
'major': 1,
'exception': 1,
'comes': 1,
'basic': 1,
'assumptions': 2,
'usually': 2,
'evaluated': 2,
'ability': 3,
'reproduce': 1,
'KDD': 2,
'key': 3,
'Evaluated': 1,
'uninformed': 1,
'method': 10,
'easily': 1,
'outperformed': 1,
'typical': 1,
'due': 1,
'unavailability': 1,
'intimate': 1,
'ties': 1,
'formulated': 1,
```

'minimization': 1,
'loss': 3,
'function': 12,
'examples': 14,
'Loss': 1,
'functions': 3,
'express': 2,
'discrepancy': 1,
'actual': 1,
'instances': 5,
'wants': 1,
'assign': 1,
'correctly': 2,
'predict': 6,
'pre-assigned': 1,
'labels': 7,
'33': 1,
'arises': 1,
'generalization': 5,
'minimize': 1,
'minimizing': 1,
'unseen': 3,
'samples': 2,
'Characterizing': 1,
'active': 2,
'topic': 2,
'current': 3,
'deep': 5,
'distinct': 1,
'their': 16,
'principal': 4,
'draws': 1,
'population': 1,
'finds': 1,
'generalizable': 1,
'34': 1,
'According': 1,
'Michael': 1,
'I.': 1,
'Jordan': 1,
'ideas': 1,
'methodological': 1,
'principles': 1,
'tools': 2,
'long': 1,
'pre-history': 1,
'35': 2,
'He': 1,
'suggested': 1,
'placeholder': 1,
'call': 2,
'overall': 2,
'Leo': 1,
'Breiman': 1,
'distinguished': 1,
'modeling': 3,
'paradigms': 2,
'algorithmic': 4,
'wherein': 1,
'means': 2,
'less': 2,
'like': 5,
'Random': 1,
'forest': 1,
'statisticians': 1,
'adopted': 2,
'combined': 1,
'core': 1,
'objective': 2,
'generalize': 1,

'Generalization': 1,
'context': 3,
'accurately': 1,
'new': 10,
'examples/tasks': 1,
'after': 5,
'experienced': 1,
'generally': 3,
'distribution': 5,
'space': 5,
'occurrences': 1,
'general': 3,
'about': 9,
'enables': 1,
'produce': 2,
'sufficiently': 2,
'accurate': 2,
'branch': 1,
'via': 2,
'Probably': 1,
'Approximately': 1,
'Correct': 1,
'PAC': 1,
'Because': 3,
'sets': 4,
'finite': 2,
'uncertain': 1,
'does': 3,
'yield': 1,
'guarantees': 1,
'Instead': 3,
'probabilistic': 4,
'bounds': 2,
'quite': 1,
'common': 4,
'bias–variance': 1,
'decomposition': 1,
'quantify': 1,
'error': 2,
'best': 4,
'complexity': 5,
'hypothesis': 4,
'match': 1,
'underlying': 2,
'If': 2,
'complex': 4,
'under': 8,
'fitted': 1,
'increased': 2,
'response': 1,
'decreases': 2,
'But': 1,
'subject': 1,
'overfitting': 4,
'poorer': 1,
'addition': 5,
'theorists': 1,
'feasibility': 1,
'computation': 1,
'feasible': 1,
'done': 1,
'polynomial': 4,
'There': 6,
'kinds': 2,
'results': 5,
'Positive': 2,
'show': 2,
'Negative': 2,
'classes': 1,
'traditionally': 1,

'divided': 1,
'three': 1,
'broad': 2,
'categories': 4,
'correspond': 1,
'depending': 2,
'signal': 9,
'feedback': 3,
'•': 12,
'Supervised': 3,
'presented': 1,
'inputs': 12,
'desired': 3,
'outputs': 6,
'teacher': 2,
'rule': 6,
'maps': 1,
'Unsupervised': 4,
'No': 1,
'leaving': 1,
'find': 3,
'structure': 3,
'input': 16,
'itself': 1,
'discovering': 4,
'hidden': 3,
'towards': 1,
'end': 1,
'feature': 9,
'Reinforcement': 4,
'interacts': 1,
'dynamic': 3,
'must': 1,
'driving': 1,
'vehicle': 1,
'playing': 2,
'game': 3,
'against': 3,
'opponent': 2,
'navigates': 1,
'analogous': 1,
'rewards': 2,
'tries': 2,
'contains': 3,
'both': 5,
'36': 1,
'consists': 2,
'Each': 2,
'output': 8,
'supervisory': 1,
'each': 8,
'represented': 10,
'array': 2,
'vector': 3,
'sometimes': 3,
'matrix': 4,
'Through': 1,
'iterative': 1,
'associated': 4,
'37': 1,
'An': 6,
'optimal': 1,
'allow': 1,
'over': 2,
'Types': 1,
'supervised-learning': 1,
'include': 7,
'regression': 10,
'Classification': 2,
'when': 5,

'restricted': 1,
'limited': 3,
'values': 3,
'any': 7,
'numerical': 1,
'value': 3,
'range': 1,
'filters': 1,
'emails': 1,
'would': 6,
'incoming': 1,
'name': 1,
'folder': 1,
'file': 1,
'Similarity': 1,
'area': 4,
'similarity': 4,
'measures': 1,
'similar': 2,
'objects': 4,
'ranking': 1,
'recommendation': 4,
'visual': 1,
'identity': 1,
'tracking': 1,
'face': 1,
'verification': 2,
'speaker': 1,
'grouping': 1,
'clustering': 3,
'points': 5,
'therefore': 1,
'test': 7,
'labeled': 6,
'classified': 1,
'categorized': 1,
'responding': 1,
'identify': 4,
'commonalities': 2,
'react': 1,
'presence': 2,
'absence': 1,
'piece': 1,
'central': 1,
'density': 3,
'estimation': 2,
'finding': 2,
'38': 1,
'Though': 1,
'encompasses': 2,
'involving': 1,
'summarizing': 1,
'explaining': 1,
'features': 12,
'Cluster': 1,
'assignment': 1,
'subsets': 3,
'clusters': 3,
'cluster': 4,
'according': 2,
'predesignated': 1,
'criteria': 1,
'drawn': 2,
'dissimilar': 1,
'Different': 2,
'techniques': 11,
'defined': 2,
'metric': 1,
'internal': 1,
'compactness': 1

'compactness': 1,
'members': 2,
'separation': 1,
'Other': 3,
'estimated': 2,
'graph': 2,
'connectivity': 1,
'Semi-supervised': 2,
'falls': 2,
'completely': 1,
'missing': 1,
'yet': 2,
'machine-learning': 2,
'unlabeled': 3,
'conjunction': 1,
'small': 1,
'amount': 2,
'considerable': 1,
'improvement': 1,
'weakly': 1,
'noisy': 1,
'imprecise': 1,
'however': 1,
'cheaper': 1,
'obtain': 1,
'resulting': 2,
'larger': 1,
'39': 1,
'software': 3,
'agents': 1,
'ought': 1,
'notion': 1,
'cumulative': 1,
'reward': 1,
'Due': 1,
'generality': 1,
'control': 1,
'operations': 1,
'simulation-based': 1,
'multi-agent': 1,
'swarm': 1,
'genetic': 7,
'typically': 5,
'Markov': 1,
'decision': 8,
'process': 12,
'MDP': 2,
'Many': 3,
'assume': 1,
'exact': 2,
'infeasible': 1,
'autonomous': 1,
'vehicles': 1,
'play': 1,
'Dimensionality': 1,
'reduction': 4,
'reducing': 4,
'number': 3,
'random': 3,
'variables': 9,
'consideration': 1,
'obtaining': 1,
'41': 1,
'words': 1,
'dimension': 2,
'Most': 1,
'dimensionality': 3,
'either': 5,
'elimination': 1,
'extraction': 1,
'One': 1

```
 'popular': 3,
 'component': 4,
 'PCA': 2,
 'changing': 3,
 'higher-dimensional': 3,
 'e.g.': 2,
 '3D': 2,
 'smaller': 2,
 '2D': 2,
 'instead': 1,
 'keeping': 1,
 'original': 4,
 '42': 1,
 'manifold': 3,
 'proposes': 1,
 'high-dimensional': 2,
 'lie': 1,
 'along': 2,
 'low-dimensional': 3,
 'manifolds': 1,
 'assumption': 2,
 'regularization': 2,
 "n't": 2,
 'fit': 4,
 'neatly': 1,
 'three-fold': 1,
 'categorization': 1,
 'meta-learning': 2,
 '43': 1,
 '2022': 1,
 'dominant': 2,
 'much': 1,
 'ongoing': 1,
 'Self-learning': 1,
 'paradigm': 1,
 'introduced': 2,
 '1982': 1,
 ...},
{('Study', 'of'): 1,
 ('of', 'algorithms'): 2,
 ('algorithms', 'that'): 3,
 ('that', 'improve'): 1,
 ('improve', 'automatically'): 1,
 ('automatically', 'through'): 1,
 ('through', 'experience'): 1,
 ('Machine', 'learning'): 16,
 (')', 'is'): 5,
 ('is', 'a'): 25,
 ('a', 'field'): 2,
 ('field', 'of'): 9,
 ('of', 'inquiry'): 1,
 ('inquiry', 'devoted'): 1,
 ('devoted', 'to'): 1,
 ('to', 'understanding'): 1,
 ('understanding', 'and'): 1,
 ('and', 'building'): 1,
 ('building', 'methods'): 1,
 ('methods', 'that'): 2,
 ('that', "'learn"): 1,
 (',', 'that'): 3,
 ('that', 'is'): 5,
 (',', 'methods'): 1,
 ('that', 'leverage'): 1,
 ('leverage', 'data'): 1,
 ('data', 'to'): 3,
 ('to', 'improve'): 4,
 ('improve', 'performance'): 1,
 ('performance', 'on'): 1,
 ('on', 'some'): 2,
```

```
('some', 'set'): 1,
('set', 'of'): 17,
('of', 'tasks'): 3,
('It', 'is'): 6,
('is', 'seen'): 1,
('seen', 'as'): 1,
('as', 'a'): 15,
('a', 'part'): 2,
('part', 'of'): 4,
('of', 'artificial'): 2,
('artificial', 'intelligence'): 7,
('learning', 'algorithms'): 23,
('algorithms', 'build'): 2,
('build', 'a'): 3,
('a', 'model'): 6,
('model', 'based'): 2,
('based', 'on'): 9,
('on', 'sample'): 1,
('sample', 'data'): 1,
(',', 'known'): 1,
('known', 'as'): 7,
('as', 'training'): 3,
('training', 'data'): 14,
(',', 'in'): 8,
('in', 'order'): 5,
('order', 'to'): 4,
('to', 'make'): 6,
('make', 'predictions'): 4,
('predictions', 'or'): 1,
('or', 'decisions'): 1,
('decisions', 'without'): 1,
('without', 'being'): 3,
('being', 'explicitly'): 2,
('explicitly', 'programmed'): 2,
('programmed', 'to'): 2,
('to', 'do'): 5,
('do', 'so'): 4,
('algorithms', 'are'): 4,
('are', 'used'): 5,
('used', 'in'): 8,
('in', 'a'): 13,
('a', 'wide'): 1,
('wide', 'variety'): 1,
('variety', 'of'): 4,
('of', 'applications'): 1,
(',', 'such'): 6,
('such', 'as'): 15,
('as', 'in'): 2,
('in', 'medicine'): 1,
(',', 'email'): 1,
('email', 'filtering'): 1,
(',', 'speech'): 2,
('speech', 'recognition'): 3,
(',', 'agriculture'): 1,
(',', 'and'): 42,
('and', 'computer'): 3,
('computer', 'vision'): 4,
(',', 'where'): 3,
('where', 'it'): 2,
('it', 'is'): 3,
('is', 'difficult'): 1,
('difficult', 'or'): 1,
('or', 'unfeasible'): 1,
('unfeasible', 'to'): 1,
('to', 'develop'): 1,
('develop', 'conventional'): 1,
('conventional', 'algorithms'): 1,
('algorithms', 'to'): 1,
('to', 'perform'): 6,
('perform', 'the'): 1,
```

```
('the', 'needed'): 2,
('needed', 'tasks'): 1,
('A', 'subset'): 1,
('subset', 'of'): 1,
('of', 'machine'): 18,
('machine', 'learning'): 86,
('learning', 'is'): 19,
('is', 'closely'): 1,
('closely', 'related'): 3,
('related', 'to'): 6,
('to', 'computational'): 1,
('computational', 'statistics'): 1,
(',', 'which'): 9,
('which', 'focuses'): 1,
('focuses', 'on'): 3,
('on', 'making'): 1,
('making', 'predictions'): 1,
('predictions', 'using'): 1,
('using', 'computers'): 1,
(',', 'but'): 15,
('but', 'not'): 1,
('not', 'all'): 2,
('all', 'machine'): 1,
('is', 'statistical'): 1,
('statistical', 'learning'): 2,
('The', 'study'): 1,
('study', 'of'): 1,
('of', 'mathematical'): 1,
('mathematical', 'optimization'): 1,
('optimization', 'delivers'): 1,
('delivers', 'methods'): 1,
(',', 'theory'): 1,
('theory', 'and'): 1,
('and', 'application'): 1,
('application', 'domains'): 1,
('domains', 'to'): 1,
('to', 'the'): 12,
('the', 'field'): 8,
('Data', 'mining'): 2,
('mining', 'is'): 1,
('a', 'related'): 2,
('related', 'field'): 2,
('of', 'study'): 2,
(',', 'focusing'): 1,
('focusing', 'on'): 1,
('on', 'exploratory'): 1,
('exploratory', 'data'): 1,
('data', 'analysis'): 1,
('analysis', 'through'): 1,
('through', 'unsupervised'): 1,
('unsupervised', 'learning'): 7,
('Some', 'implementations'): 1,
('implementations', 'of'): 1,
('learning', 'use'): 1,
('use', 'data'): 1,
('data', 'and'): 6,
('and', 'neural'): 1,
('neural', 'networks'): 8,
('networks', 'in'): 2,
('a', 'way'): 2,
('way', 'that'): 3,
('that', 'mimics'): 2,
('mimics', 'the'): 2,
('the', 'working'): 1,
('working', 'of'): 1,
('of', 'a'): 16,
('a', 'biological'): 3,
('biological', 'brain'): 3,
('In', 'its'): 1,
('its', 'application'): 1,
```

```
('application', 'across'): 1,
('across', 'business'): 1,
('business', 'problems'): 1,
(',', 'machine'): 11,
('is', 'also'): 2,
('also', 'referred'): 1,
('referred', 'to'): 2,
('to', 'as'): 2,
('as', 'predictive'): 1,
('predictive', 'analytics'): 1,
('Learning', 'algorithms'): 1,
('algorithms', 'work'): 1,
('work', 'on'): 1,
('on', 'the'): 13,
('the', 'basis'): 2,
('basis', 'that'): 1,
('that', 'strategies'): 1,
(',', 'algorithms'): 1,
('and', 'inferences'): 1,
('inferences', 'that'): 1,
('that', 'worked'): 1,
('worked', 'well'): 1,
('well', 'in'): 2,
('in', 'the'): 30,
('the', 'past'): 2,
('past', 'are'): 1,
('are', 'likely'): 3,
('likely', 'to'): 4,
('to', 'continue'): 1,
('continue', 'working'): 1,
('working', 'well'): 1,
('the', 'future'): 2,
('These', 'inferences'): 1,
('inferences', 'can'): 1,
('can', 'be'): 21,
('be', 'obvious'): 1,
('``', 'since'): 1,
('since', 'the'): 1,
('the', 'sun'): 1,
('sun', 'rose'): 1,
('rose', 'every'): 1,
('every', 'morning'): 1,
('morning', 'for'): 1,
('for', 'the'): 6,
('the', 'last'): 2,
('10,000', 'days'): 1,
(',', 'it'): 8,
('it', 'will'): 1,
('will', 'probably'): 1,
('probably', 'rise'): 1,
('rise', 'tomorrow'): 1,
('tomorrow', 'morning'): 1,
('morning', 'as'): 1,
('as', 'well'): 3,
('They', 'can'): 1,
('be', 'nuanced'): 1,
('%', 'of'): 2,
('of', 'families'): 1,
('families', 'have'): 1,
('have', 'geographically'): 1,
('geographically', 'separate'): 1,
('separate', 'species'): 1,
('species', 'with'): 1,
('with', 'color'): 1,
('color', 'variants'): 1,
(',', 'so'): 1,
('so', 'there'): 1,
('there', 'is'): 2,
('%', 'chance'): 1,
('chance', 'that'): 1,
```

```
('that', 'undiscovered'): 1,
('undiscovered', 'black'): 1,
('black', 'swans'): 1,
('swans', 'exist'): 1,
('learning', 'programs'): 1,
('programs', 'can'): 1,
('can', 'perform'): 1,
('perform', 'tasks'): 2,
('tasks', 'without'): 1,
('It', 'involves'): 1,
('involves', 'computers'): 1,
('computers', 'learning'): 1,
('learning', 'from'): 1,
('from', 'data'): 4,
('data', 'provided'): 1,
('provided', 'so'): 1,
('so', 'that'): 3,
('that', 'they'): 2,
('they', 'carry'): 1,
('carry', 'out'): 1,
('out', 'certain'): 1,
('certain', 'tasks'): 1,
('For', 'simple'): 1,
('simple', 'tasks'): 1,
('tasks', 'assigned'): 1,
('assigned', 'to'): 1,
('to', 'computers'): 1,
('is', 'possible'): 1,
('possible', 'to'): 1,
('to', 'program'): 1,
('program', 'algorithms'): 1,
('algorithms', 'telling'): 1,
('telling', 'the'): 1,
('the', 'machine'): 5,
('machine', 'how'): 1,
('how', 'to'): 1,
('to', 'execute'): 1,
('execute', 'all'): 1,
('all', 'steps'): 1,
('steps', 'required'): 1,
('required', 'to'): 1,
('to', 'solve'): 3,
('solve', 'the'): 1,
('the', 'problem'): 3,
('problem', 'at'): 1,
('at', 'hand'): 1,
(';', 'on'): 2,
('the', 'computer'): 2,
('computer', "'s"): 1,
("'s", 'part'): 1,
(',', 'no'): 1,
('no', 'learning'): 1,
('is', 'needed'): 1,
('For', 'more'): 1,
('more', 'advanced'): 1,
('advanced', 'tasks'): 1,
('it', 'can'): 3,
('be', 'challenging'): 1,
('challenging', 'for'): 1,
('for', 'a'): 2,
('a', 'human'): 4,
('human', 'to'): 1,
('to', 'manually'): 1,
('manually', 'create'): 1,
('create', 'the'): 1,
('needed', 'algorithms'): 1,
('In', 'practice'): 1,
('can', 'turn'): 1,
('turn', 'out'): 1,
('out', 'to'): 1,
```

```
('to', 'be'): 11,
('be', 'more'): 1,
('more', 'effective'): 1,
('effective', 'to'): 1,
('to', 'help'): 3,
('help', 'the'): 1,
('machine', 'develop'): 1,
('develop', 'its'): 1,
('its', 'own'): 2,
('own', 'algorithm'): 1,
(',', 'rather'): 2,
('rather', 'than'): 3,
('than', 'having'): 1,
('having', 'human'): 1,
('human', 'programmers'): 1,
('programmers', 'specify'): 1,
('specify', 'every'): 1,
('every', 'needed'): 1,
('needed', 'step'): 1,
('The', 'discipline'): 1,
('discipline', 'of'): 1,
('learning', 'employs'): 1,
('employs', 'various'): 1,
('various', 'approaches'): 1,
('approaches', 'to'): 1,
('to', 'teach'): 1,
('teach', 'computers'): 1,
('computers', 'to'): 1,
('to', 'accomplish'): 1,
('accomplish', 'tasks'): 1,
('tasks', 'where'): 1,
('where', 'no'): 1,
('no', 'fully'): 1,
('fully', 'satisfactory'): 1,
('satisfactory', 'algorithm'): 1,
('algorithm', 'is'): 3,
('is', 'available'): 1,
('In', 'cases'): 1,
('cases', 'where'): 1,
('where', 'vast'): 1,
('vast', 'numbers'): 1,
('numbers', 'of'): 1,
('of', 'potential'): 1,
('potential', 'answers'): 1,
('answers', 'exist'): 1,
(',', 'one'): 4,
('one', 'approach'): 1,
('approach', 'is'): 1,
('is', 'to'): 8,
('to', 'label'): 1,
('label', 'some'): 1,
('some', 'of'): 1,
('of', 'the'): 40,
('the', 'correct'): 1,
('correct', 'answers'): 2,
('answers', 'as'): 1,
('as', 'valid'): 1,
('This', 'can'): 1,
('can', 'then'): 1,
('then', 'be'): 1,
('be', 'used'): 7,
('used', 'as'): 3,
('data', 'for'): 1,
('computer', 'to'): 1,
('improve', 'the'): 3,
('the', 'algorithm'): 3,
('(', 's'): 3,
(')', 'it'): 1,
('it', 'uses'): 1,
('uses', 'to'): 1,
```

```
('to', 'determine'): 2,
('determine', 'correct'): 1,
('For', 'example'): 7,
(',', 'to'): 1,
('to', 'train'): 3,
('train', 'a'): 1,
('a', 'system'): 3,
('system', 'for'): 1,
('the', 'task'): 1,
('task', 'of'): 1,
('of', 'digital'): 1,
('digital', 'character'): 1,
('character', 'recognition'): 1,
(',', 'the'): 23,
('MNIST', 'dataset'): 1,
('dataset', 'of'): 1,
('of', 'handwritten'): 1,
('handwritten', 'digits'): 1,
('digits', 'has'): 1,
('has', 'often'): 1,
('often', 'been'): 1,
('been', 'used'): 5,
('History', 'and'): 1,
('and', 'relationships'): 1,
('relationships', 'to'): 1,
('to', 'other'): 3,
('other', 'fields'): 1,
('The', 'term'): 2,
('term', 'machine'): 1,
('learning', 'was'): 1,
('was', 'coined'): 1,
('coined', 'in'): 1,
('1959', 'by'): 1,
(',', 'an'): 5,
('IBM', 'employee'): 1,
('employee', 'and'): 1,
('and', 'pioneer'): 1,
('pioneer', 'in'): 1,
('of', 'computer'): 1,
('computer', 'gaming'): 1,
('gaming', 'and'): 1,
('and', 'artificial'): 2,
('Also', 'the'): 1,
('the', 'synonym'): 1,
('synonym', 'self-teaching'): 1,
('self-teaching', 'computers'): 1,
('computers', 'were'): 1,
('were', 'used'): 2,
('in', 'this'): 2,
('this', 'time'): 1,
('time', 'period'): 1,
('By', 'the'): 1,
('the', 'early'): 2,
('early', '1960s'): 1,
('1960s', 'an'): 1,
('an', 'experimental'): 1,
('``', 'learning'): 2,
('learning', 'machine'): 2,
("''", 'with'): 1,
('with', 'punched'): 1,
('punched', 'tape'): 1,
('tape', 'memory'): 1,
(',', 'called'): 1,
(',', 'had'): 2,
('had', 'been'): 4,
('been', 'developed'): 3,
('developed', 'by'): 1,
('Company', 'to'): 1,
('to', 'analyze'): 1,
('analyze', 'sonar'): 1,
```

```
('sonar', 'signals'): 1,
(',', 'electrocardiograms'): 1,
('and', 'speech'): 2,
('speech', 'patterns'): 1,
('patterns', 'using'): 1,
('using', 'rudimentary'): 1,
('rudimentary', 'reinforcement'): 1,
('reinforcement', 'learning'): 4,
('It', 'was'): 1,
('was', 'repetitively'): 1,
('``', 'trained'): 1,
("''", 'by'): 1,
('by', 'a'): 5,
('human', 'operator/teacher'): 1,
('operator/teacher', 'to'): 1,
('to', 'recognize'): 3,
('recognize', 'patterns'): 1,
('patterns', 'and'): 1,
('and', 'equipped'): 1,
('equipped', 'with'): 1,
('with', 'a'): 7,
('``', 'goof'): 1,
("''", 'button'): 1,
('button', 'to'): 1,
('to', 'cause'): 1,
('cause', 'it'): 1,
('it', 'to'): 3,
('to', 're-evaluate'): 1,
('re-evaluate', 'incorrect'): 1,
('incorrect', 'decisions'): 1,
('A', 'representative'): 1,
('representative', 'book'): 1,
('book', 'on'): 2,
('on', 'research'): 1,
('research', 'into'): 1,
('into', 'machine'): 1,
('learning', 'during'): 1,
('during', 'the'): 1,
('the', '1960s'): 1,
('1960s', 'was'): 1,
('Nilsson', "'s"): 1,
("'s", 'book'): 1,
(',', 'dealing'): 1,
('dealing', 'mostly'): 1,
('mostly', 'with'): 1,
('with', 'machine'): 1,
('learning', 'for'): 1,
('for', 'pattern'): 1,
('pattern', 'classification'): 1,
('Interest', 'related'): 1,
('to', 'pattern'): 1,
('pattern', 'recognition'): 2,
('recognition', 'continued'): 1,
('continued', 'into'): 1,
('into', 'the'): 1,
('the', '1970s'): 1,
(',', 'as'): 6,
('as', 'described'): 1,
('described', 'by'): 1,
('Duda', 'and'): 1,
('Hart', 'in'): 1,
('1981', 'a'): 1,
('a', 'report'): 1,
('report', 'was'): 1,
('was', 'given'): 1,
('given', 'on'): 1,
('on', 'using'): 1,
('using', 'teaching'): 1,
('teaching', 'strategies'): 1,
('strategies', 'so'): 1,
```

```
('that', 'a'): 5,
('a', 'neural'): 4,
('neural', 'network'): 6,
('network', 'learns'): 1,
('learns', 'to'): 1,
('40', 'characters'): 1,
('26', 'letters'): 1,
('10', 'digits'): 1,
('4', 'special'): 1,
('special', 'symbols'): 1,
(')', 'from'): 1,
('from', 'a'): 4,
('a', 'computer'): 2,
('computer', 'terminal'): 1,
('Mitchell', 'provided'): 1,
('provided', 'a'): 1,
('a', 'widely'): 1,
('widely', 'quoted'): 1,
(',', 'more'): 2,
('more', 'formal'): 1,
('formal', 'definition'): 1,
('definition', 'of'): 3,
('the', 'algorithms'): 2,
('algorithms', 'studied'): 1,
('studied', 'in'): 2,
('learning', 'field'): 1,
('A', 'computer'): 2,
('computer', 'program'): 3,
('program', 'is'): 2,
('is', 'said'): 2,
('said', 'to'): 2,
('to', 'learn'): 5,
('learn', 'from'): 4,
('from', 'experience'): 1,
('E', 'with'): 1,
('with', 'respect'): 3,
('respect', 'to'): 3,
('to', 'some'): 1,
('some', 'class'): 1,
('class', 'of'): 2,
('T', 'and'): 1,
('and', 'performance'): 1,
('performance', 'measure'): 1,
('P', 'if'): 1,
('if', 'its'): 1,
('its', 'performance'): 1,
('performance', 'at'): 1,
('at', 'tasks'): 1,
('tasks', 'in'): 2,
('as', 'measured'): 1,
('measured', 'by'): 1,
(',', 'improves'): 1,
('improves', 'with'): 1,
('with', 'experience'): 1,
('This', 'definition'): 1,
('the', 'tasks'): 1,
('in', 'which'): 9,
('which', 'machine'): 1,
('is', 'concerned'): 2,
('concerned', 'offers'): 1,
('offers', 'a'): 1,
('a', 'fundamentally'): 1,
('fundamentally', 'operational'): 1,
('operational', 'definition'): 1,
('definition', 'rather'): 1,
('than', 'defining'): 1,
('defining', 'the'): 1,
('field', 'in'): 1,
('in', 'cognitive'): 1,
('cognitive', 'terms'): 1,
```

```
('This', 'follows'): 1,
('Turing', "'s"): 1,
("'s", 'proposal'): 1,
('proposal', 'in'): 1,
('in', 'his'): 1,
('his', 'paper'): 1,
('Machinery', 'and'): 1,
('which', 'the'): 2,
('the', 'question'): 2,
('Can', 'machines'): 2,
('machines', 'think'): 1,
("''", 'is'): 1,
('is', 'replaced'): 1,
('replaced', 'with'): 1,
('with', 'the'): 8,
('machines', 'do'): 1,
('do', 'what'): 1,
('what', 'we'): 1,
('(', 'as'): 1,
('as', 'thinking'): 1,
('thinking', 'entities'): 1,
(')', 'can'): 1,
('can', 'do'): 1,
('Modern-day', 'machine'): 1,
('learning', 'has'): 5,
('has', 'two'): 1,
('two', 'objectives'): 1,
('one', 'is'): 3,
('to', 'classify'): 2,
('classify', 'data'): 1,
('data', 'based'): 1,
('on', 'models'): 1,
('models', 'which'): 2,
('which', 'have'): 1,
('have', 'been'): 8,
('the', 'other'): 3,
('other', 'purpose'): 1,
('purpose', 'is'): 1,
('predictions', 'for'): 1,
('for', 'future'): 1,
('future', 'outcomes'): 1,
('outcomes', 'based'): 1,
('on', 'these'): 1,
('these', 'models'): 1,
('A', 'hypothetical'): 1,
('hypothetical', 'algorithm'): 1,
('algorithm', 'specific'): 1,
('specific', 'to'): 1,
('to', 'classifying'): 1,
('classifying', 'data'): 1,
('data', 'may'): 1,
('may', 'use'): 1,
('use', 'computer'): 1,
('vision', 'of'): 1,
('of', 'moles'): 1,
('moles', 'coupled'): 1,
('coupled', 'with'): 1,
('with', 'supervised'): 1,
('supervised', 'learning'): 7,
('learning', 'in'): 4,
('train', 'it'): 1,
('classify', 'the'): 1,
('the', 'cancerous'): 1,
('cancerous', 'moles'): 1,
('A', 'machine'): 2,
('learning', 'algorithm'): 4,
('algorithm', 'for'): 1,
('for', 'stock'): 1,
('stock', 'trading'): 1,
('trading', 'may'): 1,
```

```
('may', 'inform'): 1,
('inform', 'the'): 1,
('the', 'trader'): 1,
('trader', 'of'): 1,
('of', 'future'): 1,
('future', 'potential'): 1,
('potential', 'predictions'): 1,
('As', 'a'): 1,
('a', 'scientific'): 1,
('scientific', 'endeavor'): 1,
('learning', 'grew'): 1,
('grew', 'out'): 1,
('out', 'of'): 2,
('the', 'quest'): 1,
('quest', 'for'): 1,
('for', 'artificial'): 1,
('In', 'the'): 2,
('early', 'days'): 1,
('days', 'of'): 1,
('AI', 'as'): 1,
('as', 'an'): 1,
('an', 'academic'): 2,
('academic', 'discipline'): 1,
(',', 'some'): 1,
('some', 'researchers'): 1,
('researchers', 'were'): 1,
('were', 'interested'): 1,
('interested', 'in'): 1,
('in', 'having'): 1,
('having', 'machines'): 1,
('machines', 'learn'): 1,
('They', 'attempted'): 1,
('attempted', 'to'): 1,
('to', 'approach'): 1,
('approach', 'the'): 1,
('problem', 'with'): 1,
('with', 'various'): 1,
('various', 'symbolic'): 1,
('symbolic', 'methods'): 1,
('well', 'as'): 2,
('as', 'what'): 1,
('what', 'was'): 1,
('was', 'then'): 1,
('then', 'termed'): 1,
('``', 'neural'): 1,
(';', 'these'): 1,
('these', 'were'): 1,
('were', 'mostly'): 1,
('mostly', 'perceptrons'): 1,
('perceptrons', 'and'): 1,
('and', 'other'): 1,
('other', 'models'): 1,
('models', 'that'): 2,
('that', 'were'): 2,
('were', 'later'): 1,
('later', 'found'): 1,
('found', 'to'): 3,
('be', 'reinventions'): 1,
('reinventions', 'of'): 1,
('the', 'generalized'): 1,
('generalized', 'linear'): 1,
('linear', 'models'): 1,
('models', 'of'): 2,
('of', 'statistics'): 1,
('Probabilistic', 'reasoning'): 1,
('reasoning', 'was'): 1,
('was', 'also'): 1,
('also', 'employed'): 1,
(',', 'especially'): 2,
('especially', 'in'): 1
```

```
('especially', 'in'): 1,
('in', 'automated'): 1,
('automated', 'medical'): 1,
('medical', 'diagnosis'): 2,
('an', 'increasing'): 1,
('increasing', 'emphasis'): 1,
('emphasis', 'on'): 1,
('the', 'logical'): 1,
(',', 'knowledge-based'): 1,
('knowledge-based', 'approach'): 1,
('approach', 'caused'): 1,
('caused', 'a'): 1,
('a', 'rift'): 1,
('rift', 'between'): 1,
('AI', 'and'): 2,
('and', 'machine'): 3,
('Probabilistic', 'systems'): 1,
('systems', 'were'): 1,
('were', 'plagued'): 1,
('plagued', 'by'): 1,
('by', 'theoretical'): 1,
('theoretical', 'and'): 1,
('and', 'practical'): 1,
('practical', 'problems'): 1,
('problems', 'of'): 2,
('of', 'data'): 9,
('data', 'acquisition'): 1,
('acquisition', 'and'): 1,
('and', 'representation'): 1,
(',', 'expert'): 1,
('expert', 'systems'): 1,
('systems', 'had'): 1,
('had', 'come'): 1,
('come', 'to'): 1,
('to', 'dominate'): 1,
('and', 'statistics'): 2,
('statistics', 'was'): 1,
('was', 'out'): 1,
('of', 'favor'): 1,
('Work', 'on'): 1,
('on', 'symbolic/knowledge-based'): 1,
('symbolic/knowledge-based', 'learning'): 1,
('learning', 'did'): 1,
('did', 'continue'): 1,
('continue', 'within'): 1,
(',', 'leading'): 4,
('leading', 'to'): 4,
('to', 'inductive'): 1,
('inductive', 'logic'): 1,
('logic', 'programming'): 5,
('but', 'the'): 4,
('the', 'more'): 1,
('more', 'statistical'): 1,
('statistical', 'line'): 1,
('line', 'of'): 1,
('of', 'research'): 1,
('research', 'was'): 1,
('was', 'now'): 1,
('now', 'outside'): 1,
('outside', 'the'): 2,
('AI', 'proper'): 1,
('in', 'pattern'): 1,
('recognition', 'and'): 1,
('and', 'information'): 1,
('information', 'retrieval'): 1,
('Neural', 'networks'): 1,
('networks', 'research'): 1,
('research', 'had'): 1,
('been', 'abandoned'): 1,
('abandoned', 'by'): 1,
('computer', 'science'): 2
```

```
('computer', 'science'): 2,
('science', 'around'): 1,
('around', 'the'): 1,
('the', 'same'): 6,
('same', 'time'): 1,
('This', 'line'): 1,
(',', 'too'): 1,
(',', 'was'): 1,
('was', 'continued'): 1,
('continued', 'outside'): 1,
('AI/CS', 'field'): 1,
('``', 'connectionism'): 1,
(',', 'by'): 3,
('by', 'researchers'): 1,
('researchers', 'from'): 2,
('from', 'other'): 1,
('other', 'disciplines'): 2,
('disciplines', 'including'): 1,
('Their', 'main'): 1,
('main', 'success'): 1,
('success', 'came'): 1,
('came', 'in'): 1,
('the', 'mid-1980s'): 1,
('mid-1980s', 'with'): 1,
('the', 'reinvention'): 1,
('reinvention', 'of'): 1,
('of', 'backpropagation'): 1,
(',', 'reorganized'): 1,
('reorganized', 'as'): 1,
('a', 'separate'): 2,
('separate', 'field'): 1,
(',', 'started'): 1,
('started', 'to'): 1,
('to', 'flourish'): 1,
('flourish', 'in'): 1,
('the', '1990s'): 1,
('The', 'field'): 1,
('field', 'changed'): 1,
('changed', 'its'): 1,
('its', 'goal'): 1,
('goal', 'from'): 1,
('from', 'achieving'): 1,
('achieving', 'artificial'): 1,
('intelligence', 'to'): 2,
('to', 'tackling'): 1,
('tackling', 'solvable'): 1,
('solvable', 'problems'): 1,
('a', 'practical'): 1,
('practical', 'nature'): 1,
('It', 'shifted'): 1,
('shifted', 'focus'): 1,
('focus', 'away'): 1,
('away', 'from'): 1,
('from', 'the'): 15,
('the', 'symbolic'): 1,
('symbolic', 'approaches'): 1,
('approaches', 'it'): 1,
('it', 'had'): 1,
('had', 'inherited'): 1,
('inherited', 'from'): 1,
('and', 'toward'): 1,
('toward', 'methods'): 1,
('methods', 'and'): 2,
('and', 'models'): 2,
('models', 'borrowed'): 1,
('borrowed', 'from'): 1,
('from', 'statistics'): 1,
(',', 'fuzzy'): 1,
('fuzzy', 'logic'): 1,
('and', 'probability'): 1,
('probability', 'theory'): 1
```

```
('probability', 'theory'): 1,
('The', 'difference'): 2,
('difference', 'between'): 3,
('ML', 'and'): 1,
('AI', 'is'): 1,
('is', 'frequently'): 1,
('frequently', 'misunderstood'): 1,
('ML', 'learns'): 1,
('learns', 'and'): 1,
('and', 'predicts'): 1,
('predicts', 'based'): 1,
('on', 'passive'): 1,
('passive', 'observations'): 1,
(',', 'whereas'): 1,
('AI', 'implies'): 1,
('implies', 'an'): 1,
('an', 'agent'): 1,
('agent', 'interacting'): 1,
('interacting', 'with'): 1,
('the', 'environment'): 3,
('environment', 'to'): 1,
('learn', 'and'): 1,
('and', 'take'): 1,
('take', 'actions'): 2,
('actions', 'that'): 1,
('that', 'maximize'): 1,
('maximize', 'its'): 1,
('its', 'chance'): 1,
('chance', 'of'): 1,
('of', 'successfully'): 1,
('successfully', 'achieving'): 1,
('achieving', 'its'): 1,
('its', 'goals'): 1,
('As', 'of'): 2,
(',', 'many'): 1,
('many', 'sources'): 1,
('sources', 'continue'): 1,
('continue', 'to'): 1,
('to', 'assert'): 1,
('assert', 'that'): 1,
('ML', 'remains'): 1,
('remains', 'a'): 1,
('a', 'subfield'): 1,
('subfield', 'of'): 1,
('Others', 'have'): 1,
('have', 'the'): 1,
('the', 'view'): 1,
('view', 'that'): 1,
('that', 'not'): 1,
('ML', 'is'): 1,
('is', 'part'): 1,
('but', 'only'): 1,
('only', 'an'): 1,
('an', "'intelligent"): 1,
("'intelligent", 'subset'): 1,
("'", 'of'): 1,
('ML', 'should'): 1,
('should', 'be'): 1,
('be', 'considered'): 2,
('learning', 'and'): 4,
('and', 'data'): 3,
('data', 'mining'): 7,
('mining', 'often'): 1,
('often', 'employ'): 1,
('employ', 'the'): 1,
('same', 'methods'): 1,
('and', 'overlap'): 1,
('overlap', 'significantly'): 1,
('but', 'while'): 1,
('while', 'machine'): 2,
('learning', 'focuses'): 1
```

```
 ('learning', 'focuses'): 1,
 ('on', 'prediction'): 1,
 (',', 'based'): 1,
 ('on', 'known'): 1,
 ('known', 'properties'): 1,
 ('properties', 'learned'): 1,
 ('learned', 'from'): 3,
 ('the', 'training'): 12,
 (',', 'data'): 3,
 ('mining', 'focuses'): 1,
 ('the', 'discovery'): 2,
 ('discovery', 'of'): 2,
 ('(', 'previously'): 1,
 (')', 'unknown'): 1,
 ('unknown', 'properties'): 1,
 ('properties', 'in'): 1,
 ('the', 'data'): 14,
 ('(', 'this'): 1,
 ('this', 'is'): 1,
 ('is', 'the'): 12,
 ('the', 'analysis'): 1,
 ('analysis', 'step'): 1,
 ('step', 'of'): 1,
 ('of', 'knowledge'): 1,
 ('knowledge', 'discovery'): 2,
 ('discovery', 'in'): 1,
 ('in', 'databases'): 2,
 ('mining', 'uses'): 1,
 ('uses', 'many'): 1,
 ('many', 'machine'): 1,
 ('learning', 'methods'): 3,
 ('but', 'with'): 1,
 ('with', 'different'): 1,
 ('different', 'goals'): 1,
 ('other', 'hand'): 1,
 ('learning', 'also'): 2,
 ('also', 'employs'): 1,
 ('employs', 'data'): 1,
 ('mining', 'methods'): 1,
 ('methods', 'as'): 1,
 ('``', 'unsupervised'): 1,
 ("''", 'or'): 2,
 ('or', 'as'): 1,
 ('a', 'preprocessing'): 1,
 ('preprocessing', 'step'): 1,
 ('step', 'to'): 1,
 ('improve', 'learner'): 1,
 ('learner', 'accuracy'): 1,
 ('Much', 'of'): 1,
 ('the', 'confusion'): 1,
 ('confusion', 'between'): 1,
 ('between', 'these'): 1,
 ('these', 'two'): 1,
 ('two', 'research'): 1,
 ('research', 'communities'): 1,
 ('(', 'which'): 1,
 ('which', 'do'): 2,
 ('do', 'often'): 1,
 ('often', 'have'): 1,
 ('have', 'separate'): 1,
 ...})
```

In [39]:

```
bigramProb = calcBigramProb(listOfBigrams, unigramCounts, bigramCounts)
bigramProb
```

Out[39]:

```
{('Study', 'of'): 1.0,
 ('of', 'algorithms'): 0.008333333333333333,
 ('algorithms', 'that'): 0.06666666666666667,
 ('that', 'improve'): 0.010309278350515464,
```

```
('improve', 'automatically'): 0.16666666666666666,
('automatically', 'through'): 1.0,
('through', 'experience'): 0.25,
('Machine', 'learning'): 0.7619047619047619,
(')', 'is'): 0.06329113924050633,
('is', 'a'): 0.1893939393939394,
('a', 'field'): 0.009174311926605505,
('field', 'of'): 0.5,
('of', 'inquiry'): 0.004166666666666667,
('inquiry', 'devoted'): 1.0,
('devoted', 'to'): 1.0,
('to', 'understanding'): 0.0045662100456621,
('understanding', 'and'): 1.0,
('and', 'building'): 0.005494505494505495,
('building', 'methods'): 1.0,
('methods', 'that'): 0.08695652173913043,
('that', "'learn"): 0.010309278350515464,
(',', 'that'): 0.007352941176470588,
('that', 'is'): 0.05154639175257732,
(',', 'methods'): 0.0024509803921568627,
('that', 'leverage'): 0.010309278350515464,
('leverage', 'data'): 1.0,
('data', 'to'): 0.031578947368421054,
('to', 'improve'): 0.0182648401826484,
('improve', 'performance'): 0.16666666666666666,
('performance', 'on'): 0.08333333333333333,
('on', 'some'): 0.03508771929824561,
('some', 'set'): 0.08333333333333333,
('set', 'of'): 0.5483870967741935,
('of', 'tasks'): 0.0125,
('It', 'is'): 0.46153846153846156,
('is', 'seen'): 0.007575757575757576,
('seen', 'as'): 0.5,
('as', 'a'): 0.22727272727272727,
('a', 'part'): 0.009174311926605505,
('part', 'of'): 0.8,
('of', 'artificial'): 0.008333333333333333,
('artificial', 'intelligence'): 0.3333333333333333,
('learning', 'algorithms'): 0.12169312169312169,
('algorithms', 'build'): 0.044444444444444446,
('build', 'a'): 1.0,
('a', 'model'): 0.027522935779816515,
('model', 'based'): 0.05555555555555555,
('based', 'on'): 1.0,
('on', 'sample'): 0.017543859649122806,
('sample', 'data'): 0.3333333333333333,
(',', 'known'): 0.0024509803921568627,
('known', 'as'): 0.6363636363636364,
('as', 'training'): 0.045454545454545456,
('training', 'data'): 0.2916666666666667,
(',', 'in'): 0.0196078431372549,
('in', 'order'): 0.03424657534246575,

('order', 'to'): 0.6666666666666666,
('to', 'make'): 0.0273972602739726,
('make', 'predictions'): 0.4,
('predictions', 'or'): 0.07692307692307693,
('or', 'decisions'): 0.0196078431372549,
('decisions', 'without'): 0.125,
('without', 'being'): 0.3333333333333333,
('being', 'explicitly'): 0.25,
('explicitly', 'programmed'): 0.6666666666666666,
('programmed', 'to'): 0.6666666666666666,
('to', 'do'): 0.0228310502283105,
('do', 'so'): 0.36363636363636365,
('algorithms', 'are'): 0.08888888888888889,
('are', 'used'): 0.08928571428571429,
('used', 'in'): 0.22857142857142856,
('in', 'a'): 0.08904109589041095,
('a', 'wide'): 0.0045871559633027525,
```

```
('wide', 'variety'): 1.0,
('variety', 'of'): 1.0,
('of', 'applications'): 0.004166666666666667,
(',', 'such'): 0.014705882352941176,
('such', 'as'): 0.7142857142857143,
('as', 'in'): 0.030303030303030304,
('in', 'medicine'): 0.00684931506849315,
(',', 'email'): 0.0024509803921568627,
('email', 'filtering'): 0.3333333333333333,
(',', 'speech'): 0.004901960784313725,
('speech', 'recognition'): 0.6,
(',', 'agriculture'): 0.0024509803921568627,
(',', 'and'): 0.10294117647058823,
('and', 'computer'): 0.016483516483516484,
('computer', 'vision'): 0.25,
(',', 'where'): 0.007352941176470588,
('where', 'it'): 0.16666666666666666,
('it', 'is'): 0.1,
('is', 'difficult'): 0.007575757575757576,
('difficult', 'or'): 0.5,
('or', 'unfeasible'): 0.0196078431372549,
('unfeasible', 'to'): 1.0,
('to', 'develop'): 0.0045662100456621,
('develop', 'conventional'): 0.5,
('conventional', 'algorithms'): 1.0,
('algorithms', 'to'): 0.022222222222222223,
('to', 'perform'): 0.0273972602739726,
('perform', 'the'): 0.08333333333333333,
('the', 'needed'): 0.005865102639296188,
('needed', 'tasks'): 0.2,
('A', 'subset'): 0.05555555555555555,
('subset', 'of'): 0.3333333333333333,
('of', 'machine'): 0.075,
('machine', 'learning'): 0.9247311827956989,
('learning', 'is'): 0.10052910052910052,
('is', 'closely'): 0.007575757575757576,
('closely', 'related'): 1.0,
('related', 'to'): 0.5454545454545454,
('to', 'computational'): 0.0045662100456621,
('computational', 'statistics'): 0.25,
(',', 'which'): 0.022058823529411766,
('which', 'focuses'): 0.04,
('focuses', 'on'): 1.0,
('on', 'making'): 0.017543859649122806,
('making', 'predictions'): 0.5,
('predictions', 'using'): 0.07692307692307693,
('using', 'computers'): 0.07692307692307693,
(',', 'but'): 0.03676470588235294,
('but', 'not'): 0.058823529411764705,
('not', 'all'): 0.07407407407407407,
('all', 'machine'): 0.1,
('is', 'statistical'): 0.007575757575757576,
('statistical', 'learning'): 0.25,
('The', 'study'): 0.034482758620689655,
('study', 'of'): 0.2,
('of', 'mathematical'): 0.004166666666666667,
('mathematical', 'optimization'): 0.14285714285714285,
('optimization', 'delivers'): 0.1111111111111111,
('delivers', 'methods'): 1.0,
(',', 'theory'): 0.0024509803921568627,
('theory', 'and'): 0.07692307692307693,
('and', 'application'): 0.005494505494505495,
('application', 'domains'): 0.2,
('domains', 'to'): 0.3333333333333333,
('to', 'the'): 0.0547945205479452,
('the', 'field'): 0.02346041055718475,
('Data', 'mining'): 0.6666666666666666,
('mining', 'is'): 0.09090909090909091,
('a', 'related'): 0.009174311926605505,
('related', 'field'): 0.18181818181818182,
```

```
('of', 'study'): 0.008333333333333333,
(',', 'focusing'): 0.0024509803921568627,
('focusing', 'on'): 1.0,
('on', 'exploratory'): 0.017543859649122806,
('exploratory', 'data'): 1.0,
('data', 'analysis'): 0.010526315789473684,
('analysis', 'through'): 0.08333333333333333,
('through', 'unsupervised'): 0.25,
('unsupervised', 'learning'): 0.6363636363636364,
('Some', 'implementations'): 0.2,
('implementations', 'of'): 0.5,
('learning', 'use'): 0.005291005291005291,
('use', 'data'): 0.1,
('data', 'and'): 0.06315789473684211,
('and', 'neural'): 0.005494505494505495,
('neural', 'networks'): 0.5,
('networks', 'in'): 0.15384615384615385,
('a', 'way'): 0.009174311926605505,
('way', 'that'): 0.6,
('that', 'mimics'): 0.020618556701030927,
('mimics', 'the'): 1.0,
('the', 'working'): 0.002932551319648094,
('working', 'of'): 0.5,
('of', 'a'): 0.06666666666666667,
('a', 'biological'): 0.013761467889908258,
('biological', 'brain'): 0.75,
('In', 'its'): 0.02857142857142857,
('its', 'application'): 0.058823529411764705,
('application', 'across'): 0.2,
('across', 'business'): 0.5,
('business', 'problems'): 0.5,
(',', 'machine'): 0.02696078431372549,
('is', 'also'): 0.015151515151515152,
('also', 'referred'): 0.05,
('referred', 'to'): 1.0,
('to', 'as'): 0.0091324200913242,
('as', 'predictive'): 0.015151515151515152,
('predictive', 'analytics'): 0.25,
('Learning', 'algorithms'): 0.1111111111111111,
('algorithms', 'work'): 0.022222222222222223,
('work', 'on'): 0.3333333333333333,
('on', 'the'): 0.22807017543859648,
('the', 'basis'): 0.005865102639296188,
('basis', 'that'): 0.3333333333333333,
('that', 'strategies'): 0.010309278350515464,
(',', 'algorithms'): 0.0024509803921568627,
('and', 'inferences'): 0.005494505494505495,
('inferences', 'that'): 0.3333333333333333,
('that', 'worked'): 0.010309278350515464,
('worked', 'well'): 1.0,
('well', 'in'): 0.2857142857142857,
('in', 'the'): 0.2054794520547945,
('the', 'past'): 0.005865102639296188,
('past', 'are'): 0.5,
('are', 'likely'): 0.05357142857142857,
('likely', 'to'): 1.0,
('to', 'continue'): 0.0045662100456621,
('continue', 'working'): 0.3333333333333333,
('working', 'well'): 0.5,
('the', 'future'): 0.005865102639296188,
('These', 'inferences'): 0.5,
('inferences', 'can'): 0.3333333333333333,
('can', 'be'): 0.4666666666666667,
('be', 'obvious'): 0.020833333333333332,
('``', 'since'): 0.02857142857142857,
('since', 'the'): 1.0,
('the', 'sun'): 0.002932551319648094,
('sun', 'rose'): 1.0,
('rose', 'every'): 1.0,
('every', 'morning'): 0.3333333333333333,
```

```
('morning', 'for'): 0.5,
('for', 'the'): 0.1111111111111111,
('the', 'last'): 0.005865102639296188,
('10,000', 'days'): 1.0,
(',', 'it'): 0.0196078431372549,
('it', 'will'): 0.03333333333333333,
('will', 'probably'): 0.14285714285714285,
('probably', 'rise'): 1.0,
('rise', 'tomorrow'): 1.0,
('tomorrow', 'morning'): 1.0,
('morning', 'as'): 0.5,
('as', 'well'): 0.045454545454545456,
('They', 'can'): 0.3333333333333333,
('be', 'nuanced'): 0.020833333333333332,
('%', 'of'): 0.5,
('of', 'families'): 0.004166666666666667,
('families', 'have'): 1.0,
('have', 'geographically'): 0.047619047619047616,
('geographically', 'separate'): 1.0,
('separate', 'species'): 0.2,
('species', 'with'): 0.5,
('with', 'color'): 0.02040816326530612,
('color', 'variants'): 1.0,
(',', 'so'): 0.0024509803921568627,
('so', 'there'): 0.1,
('there', 'is'): 1.0,
('%', 'chance'): 0.25,
('chance', 'that'): 0.5,
('that', 'undiscovered'): 0.010309278350515464,
('undiscovered', 'black'): 1.0,
('black', 'swans'): 0.2,
('swans', 'exist'): 1.0,
('learning', 'programs'): 0.005291005291005291,
('programs', 'can'): 0.25,
('can', 'perform'): 0.022222222222222223,
('perform', 'tasks'): 0.16666666666666666,
('tasks', 'without'): 0.06666666666666667,
('It', 'involves'): 0.07692307692307693,
('involves', 'computers'): 0.25,
('computers', 'learning'): 0.16666666666666666,
('learning', 'from'): 0.005291005291005291,
('from', 'data'): 0.08163265306122448,
('data', 'provided'): 0.010526315789473684,
('provided', 'so'): 0.2,
('so', 'that'): 0.3,
('that', 'they'): 0.020618556701030927,
('they', 'carry'): 0.16666666666666666,
('carry', 'out'): 1.0,
('out', 'certain'): 0.2,
('certain', 'tasks'): 0.2,
('For', 'simple'): 0.09090909090909091,
('simple', 'tasks'): 1.0,
('tasks', 'assigned'): 0.06666666666666667,
('assigned', 'to'): 1.0,
('to', 'computers'): 0.0045662100456621,
('is', 'possible'): 0.007575757575757576,
('possible', 'to'): 1.0,
('to', 'program'): 0.0045662100456621,
('program', 'algorithms'): 0.1111111111111111,
('algorithms', 'telling'): 0.022222222222222223,
('telling', 'the'): 1.0,
('the', 'machine'): 0.01466275659824047,
('machine', 'how'): 0.010752688172043012,
('how', 'to'): 0.14285714285714285,
('to', 'execute'): 0.0045662100456621,
('execute', 'all'): 1.0,
('all', 'steps'): 0.1,
('steps', 'required'): 1.0,
('required', 'to'): 0.5,
('to', 'solve'): 0.0136986301369863,
```

```
('solve', 'the'): 0.25,
('the', 'problem'): 0.008797653958944282,
('problem', 'at'): 0.16666666666666666,
('at', 'hand'): 0.125,
(';', 'on'): 0.2857142857142857,
('the', 'computer'): 0.005865102639296188,
('computer', "'s"): 0.0625,
("'s", 'part'): 0.0625,
(',', 'no'): 0.0024509803921568627,
('no', 'learning'): 0.2,
('is', 'needed'): 0.007575757575757576,
('For', 'more'): 0.09090909090909091,
('more', 'advanced'): 0.07692307692307693,
('advanced', 'tasks'): 1.0,
('it', 'can'): 0.1,
('be', 'challenging'): 0.020833333333333332,
('challenging', 'for'): 1.0,
('for', 'a'): 0.037037037037037035,
('a', 'human'): 0.01834862385321101,
('human', 'to'): 0.125,
('to', 'manually'): 0.0045662100456621,
('manually', 'create'): 1.0,
('create', 'the'): 1.0,
('needed', 'algorithms'): 0.2,
('In', 'practice'): 0.02857142857142857,
('can', 'turn'): 0.022222222222222223,
('turn', 'out'): 1.0,
('out', 'to'): 0.2,
('to', 'be'): 0.0502283105022831,
('be', 'more'): 0.020833333333333332,
('more', 'effective'): 0.07692307692307693,
('effective', 'to'): 0.25,
('to', 'help'): 0.0136986301369863,
('help', 'the'): 0.3333333333333333,
('machine', 'develop'): 0.010752688172043012,
('develop', 'its'): 0.5,
('its', 'own'): 0.11764705882352941,
('own', 'algorithm'): 0.5,
(',', 'rather'): 0.004901960784313725,
('rather', 'than'): 1.0,
('than', 'having'): 0.16666666666666666,
('having', 'human'): 0.25,
('human', 'programmers'): 0.125,
('programmers', 'specify'): 1.0,
('specify', 'every'): 1.0,
('every', 'needed'): 0.3333333333333333,
('needed', 'step'): 0.2,
('The', 'discipline'): 0.034482758620689655,
('discipline', 'of'): 0.5,
('learning', 'employs'): 0.005291005291005291,
('employs', 'various'): 0.5,
('various', 'approaches'): 0.2,
('approaches', 'to'): 0.125,
('to', 'teach'): 0.0045662100456621,
('teach', 'computers'): 1.0,
('computers', 'to'): 0.16666666666666666,
('to', 'accomplish'): 0.0045662100456621,
('accomplish', 'tasks'): 1.0,
('tasks', 'where'): 0.06666666666666667,
('where', 'no'): 0.08333333333333333,
('no', 'fully'): 0.2,
('fully', 'satisfactory'): 0.5,
('satisfactory', 'algorithm'): 1.0,
('algorithm', 'is'): 0.13636363636363635,
('is', 'available'): 0.007575757575757576,
('In', 'cases'): 0.02857142857142857,
('cases', 'where'): 0.3333333333333333,
('where', 'vast'): 0.08333333333333333,
('vast', 'numbers'): 1.0,
('numbers', 'of'): 0.5,
```

```
('of', 'potential'): 0.004166666666666667,
('potential', 'answers'): 0.25,
('answers', 'exist'): 0.3333333333333333,
(',', 'one'): 0.00980392156862745,
('one', 'approach'): 0.06666666666666667,
('approach', 'is'): 0.14285714285714285,
('is', 'to'): 0.06060606060606061,
('to', 'label'): 0.0045662100456621,
('label', 'some'): 0.5,
('some', 'of'): 0.08333333333333333,
('of', 'the'): 0.16666666666666666,
('the', 'correct'): 0.002932551319648094,
('correct', 'answers'): 1.0,
('answers', 'as'): 0.3333333333333333,
('as', 'valid'): 0.015151515151515152,
('This', 'can'): 0.08333333333333333,
('can', 'then'): 0.022222222222222223,
('then', 'be'): 0.1111111111111111,
('be', 'used'): 0.14583333333333334,
('used', 'as'): 0.08571428571428572,
('data', 'for'): 0.010526315789473684,
('computer', 'to'): 0.0625,
('improve', 'the'): 0.5,
('the', 'algorithm'): 0.008797653958944282,
('(', 's'): 0.0379746835443038,
(')', 'it'): 0.012658227848101266,
('it', 'uses'): 0.03333333333333333,
('uses', 'to'): 0.25,
('to', 'determine'): 0.0091324200913242,
('determine', 'correct'): 0.3333333333333333,
('For', 'example'): 0.6363636363636364,
(',', 'to'): 0.0024509803921568627,
('to', 'train'): 0.0136986301369863,
('train', 'a'): 0.3333333333333333,
('a', 'system'): 0.013761467889908258,
('system', 'for'): 0.08333333333333333,
('the', 'task'): 0.002932551319648094,
('task', 'of'): 0.2,
('of', 'digital'): 0.004166666666666667,
('digital', 'character'): 1.0,
('character', 'recognition'): 1.0,
(',', 'the'): 0.056372549019607844,
('MNIST', 'dataset'): 1.0,
('dataset', 'of'): 0.5,
('of', 'handwritten'): 0.004166666666666667,
('handwritten', 'digits'): 1.0,
('digits', 'has'): 0.5,
('has', 'often'): 0.05,
('often', 'been'): 0.058823529411764705,
('been', 'used'): 0.23809523809523808,
('History', 'and'): 1.0,
('and', 'relationships'): 0.005494505494505495,
('relationships', 'to'): 0.25,
('to', 'other'): 0.0136986301369863,
('other', 'fields'): 0.058823529411764705,
('The', 'term'): 0.06896551724137931,
('term', 'machine'): 0.2,
('learning', 'was'): 0.005291005291005291,
('was', 'coined'): 0.05555555555555555,
('coined', 'in'): 1.0,
('1959', 'by'): 1.0,
(',', 'an'): 0.012254901960784314,
('IBM', 'employee'): 0.5,
('employee', 'and'): 1.0,
('and', 'pioneer'): 0.005494505494505495,
('pioneer', 'in'): 1.0,
('of', 'computer'): 0.004166666666666667,
('computer', 'gaming'): 0.0625,
('gaming', 'and'): 1.0,
('and', 'artificial'): 0.01098901098901099,
```

```
('Also', 'the'): 1.0,
('the', 'synonym'): 0.002932551319648094,
('synonym', 'self-teaching'): 1.0,
('self-teaching', 'computers'): 1.0,
('computers', 'were'): 0.16666666666666666,
('were', 'used'): 0.2222222222222222,
('in', 'this'): 0.0136986301369863,
('this', 'time'): 0.1,
('time', 'period'): 0.1,
('By', 'the'): 0.25,
('the', 'early'): 0.005865102639296188,
('early', '1960s'): 0.5,
('1960s', 'an'): 0.5,
('an', 'experimental'): 0.024390243902439025,
('``', 'learning'): 0.05714285714285714,
('learning', 'machine'): 0.010582010582010581,
("''", 'with'): 0.030303030303030304,
('with', 'punched'): 0.02040816326530612,
('punched', 'tape'): 1.0,
('tape', 'memory'): 1.0,
(',', 'called'): 0.0024509803921568627,
(',', 'had'): 0.004901960784313725,
('had', 'been'): 0.4,
('been', 'developed'): 0.14285714285714285,
('developed', 'by'): 0.2,
('Company', 'to'): 1.0,
('to', 'analyze'): 0.0045662100456621,
('analyze', 'sonar'): 1.0,
('sonar', 'signals'): 1.0,
(',', 'electrocardiograms'): 0.0024509803921568627,
('and', 'speech'): 0.01098901098901099,
('speech', 'patterns'): 0.2,
('patterns', 'using'): 0.14285714285714285,
('using', 'rudimentary'): 0.07692307692307693,
('rudimentary', 'reinforcement'): 1.0,
('reinforcement', 'learning'): 0.6666666666666666,
('It', 'was'): 0.07692307692307693,
('was', 'repetitively'): 0.05555555555555555,
('``', 'trained'): 0.02857142857142857,
("''", 'by'): 0.030303030303030304,
('by', 'a'): 0.10638297872340426,
('human', 'operator/teacher'): 0.125,
('operator/teacher', 'to'): 1.0,
('to', 'recognize'): 0.0136986301369863,
('recognize', 'patterns'): 0.3333333333333333,
('patterns', 'and'): 0.14285714285714285,
('and', 'equipped'): 0.005494505494505495,
('equipped', 'with'): 1.0,
('with', 'a'): 0.14285714285714285,
('``', 'goof'): 0.02857142857142857,
("''", 'button'): 0.030303030303030304,
('button', 'to'): 1.0,
('to', 'cause'): 0.0045662100456621,
('cause', 'it'): 1.0,
('it', 'to'): 0.1,
('to', 're-evaluate'): 0.0045662100456621,
('re-evaluate', 'incorrect'): 1.0,
('incorrect', 'decisions'): 1.0,
('A', 'representative'): 0.05555555555555555,
('representative', 'book'): 0.3333333333333333,
('book', 'on'): 0.6666666666666666,
('on', 'research'): 0.017543859649122806,
('research', 'into'): 0.125,
('into', 'machine'): 0.08333333333333333,
('learning', 'during'): 0.005291005291005291,
('during', 'the'): 0.5,
('the', '1960s'): 0.002932551319648094,
('1960s', 'was'): 0.5,
('Nilsson', "'s"): 0.5,
("'s", 'book'): 0.0625,
```

(',', 'dealing'): 0.0024509803921568627,
('dealing', 'mostly'): 0.5,
('mostly', 'with'): 0.5,
('with', 'machine'): 0.02040816326530612,
('learning', 'for'): 0.005291005291005291,
('for', 'pattern'): 0.018518518518518517,
('pattern', 'classification'): 0.25,
('Interest', 'related'): 1.0,
('to', 'pattern'): 0.0045662100456621,
('pattern', 'recognition'): 0.5,
('recognition', 'continued'): 0.16666666666666666,
('continued', 'into'): 0.5,
('into', 'the'): 0.08333333333333333,
('the', '1970s'): 0.002932551319648094,
(',', 'as'): 0.014705882352941176,
('as', 'described'): 0.015151515151515152,
('described', 'by'): 1.0,
('Duda', 'and'): 1.0,
('Hart', 'in'): 1.0,
('1981', 'a'): 0.5,
('a', 'report'): 0.004587155963302725,
('report', 'was'): 0.3333333333333333,
('was', 'given'): 0.05555555555555555,
('given', 'on'): 0.16666666666666666,
('on', 'using'): 0.017543859649122806,
('using', 'teaching'): 0.07692307692307693,
('teaching', 'strategies'): 1.0,
('strategies', 'so'): 0.5,
('that', 'a'): 0.05154639175257732,
('a', 'neural'): 0.01834862385321101,
('neural', 'network'): 0.375,
('network', 'learns'): 0.08333333333333333,
('learns', 'to'): 0.2,
('40', 'characters'): 0.5,
('26', 'letters'): 0.14285714285714285,
('10', 'digits'): 0.5,
('4', 'special'): 0.5,
('special', 'symbols'): 1.0,
(')', 'from'): 0.012658227848101266,
('from', 'a'): 0.08163265306122448,
('a', 'computer'): 0.009174311926605505,
('computer', 'terminal'): 0.0625,
('Mitchell', 'provided'): 1.0,
('provided', 'a'): 0.2,
('a', 'widely'): 0.004587155963302725,
('widely', 'quoted'): 1.0,
(',', 'more'): 0.004901960784313725,
('more', 'formal'): 0.07692307692307693,
('formal', 'definition'): 1.0,
('definition', 'of'): 0.75,
('the', 'algorithms'): 0.005865102639296188,
('algorithms', 'studied'): 0.022222222222222223,
('studied', 'in'): 1.0,
('learning', 'field'): 0.005291005291005291,
('A', 'computer'): 0.1111111111111111,
('computer', 'program'): 0.1875,
('program', 'is'): 0.2222222222222222,
('is', 'said'): 0.015151515151515152,
('said', 'to'): 1.0,
('to', 'learn'): 0.0228310502283105,
('learn', 'from'): 0.3333333333333333,
('from', 'experience'): 0.02040816326530612,
('E', 'with'): 1.0,
('with', 'respect'): 0.061224489795918366,
('respect', 'to'): 1.0,
('to', 'some'): 0.0045662100456621,
('some', 'class'): 0.08333333333333333,
('class', 'of'): 0.2857142857142857,
('T', 'and'): 0.3333333333333333,
('and', 'performance'): 0.005494505494505495

```
('and', 'performance'): [illegible],
('performance', 'measure'): 0.08333333333333333,
('P', 'if'): 0.5,
('if', 'its'): 0.2,
('its', 'performance'): 0.058823529411764705,
('performance', 'at'): 0.08333333333333333,
('at', 'tasks'): 0.125,
('tasks', 'in'): 0.13333333333333333,
('as', 'measured'): 0.015151515151515152,
('measured', 'by'): 1.0,
(',', 'improves'): 0.0024509803921568627,
('improves', 'with'): 0.5,
('with', 'experience'): 0.02040816326530612,
('This', 'definition'): 0.08333333333333333,
('the', 'tasks'): 0.002932551319648094,
('in', 'which'): 0.06164383561643835,
('which', 'machine'): 0.04,
('is', 'concerned'): 0.015151515151515152,
('concerned', 'offers'): 0.3333333333333333,
('offers', 'a'): 1.0,
('a', 'fundamentally'): 0.0045871559633027525,
('fundamentally', 'operational'): 1.0,
('operational', 'definition'): 1.0,
('definition', 'rather'): 0.25,
('than', 'defining'): 0.16666666666666666,
('defining', 'the'): 0.5,
('field', 'in'): 0.05555555555555555,
('in', 'cognitive'): 0.00684931506849315,
('cognitive', 'terms'): 1.0,
('This', 'follows'): 0.08333333333333333,
('Turing', "'s"): 1.0,
("'s", 'proposal'): 0.0625,
('proposal', 'in'): 1.0,
('in', 'his'): 0.00684931506849315,
('his', 'paper'): 1.0,
('Machinery', 'and'): 1.0,
('which', 'the'): 0.08,
('the', 'question'): 0.005865102639296188,
('Can', 'machines'): 1.0,
('machines', 'think'): 0.16666666666666666,
("''", 'is'): 0.030303030303030304,
('is', 'replaced'): 0.007575757575757576,
('replaced', 'with'): 1.0,
('with', 'the'): 0.16326530612244897,
('machines', 'do'): 0.16666666666666666,
('do', 'what'): 0.09090909090909091,
('what', 'we'): 0.3333333333333333,
('(', 'as'): 0.012658227848101266,
('as', 'thinking'): 0.015151515151515152,
('thinking', 'entities'): 1.0,
(')', 'can'): 0.012658227848101266,
('can', 'do'): 0.022222222222222223,
('Modern-day', 'machine'): 1.0,
('learning', 'has'): 0.026455026455026454,
('has', 'two'): 0.05,
('two', 'objectives'): 0.125,
('one', 'is'): 0.2,
('to', 'classify'): 0.0091324200913242,
('classify', 'data'): 0.5,
('data', 'based'): 0.010526315789473684,
('on', 'models'): 0.017543859649122806,
('models', 'which'): 0.08333333333333333,
('which', 'have'): 0.04,
('have', 'been'): 0.38095238095238093,
('the', 'other'): 0.008797653958944282,
('other', 'purpose'): 0.058823529411764705,
('purpose', 'is'): 1.0,
('predictions', 'for'): 0.07692307692307693,
('for', 'future'): 0.018518518518518517,
('future', 'outcomes'): 0.25,
('outcomes', 'based'): 0.5
```

('outcomes', 'based'): 0.5,
('on', 'these'): 0.017543859649122806,
('these', 'models'): 0.08333333333333333,
('A', 'hypothetical'): 0.05555555555555555,
('hypothetical', 'algorithm'): 1.0,
('algorithm', 'specific'): 0.045454545454545456,
('specific', 'to'): 0.2,
('to', 'classifying'): 0.0045662100456621,
('classifying', 'data'): 0.5,
('data', 'may'): 0.010526315789473684,
('may', 'use'): 0.07692307692307693,
('use', 'computer'): 0.1,
('vision', 'of'): 0.2,
('of', 'moles'): 0.004166666666666667,
('moles', 'coupled'): 0.5,
('coupled', 'with'): 1.0,
('with', 'supervised'): 0.02040816326530612,
('supervised', 'learning'): 0.5384615384615384,
('learning', 'in'): 0.021164021164021163,
('train', 'it'): 0.3333333333333333,
('classify', 'the'): 0.5,
('the', 'cancerous'): 0.002932551319648094,
('cancerous', 'moles'): 1.0,
('A', 'machine'): 0.1111111111111111,
('learning', 'algorithm'): 0.021164021164021163,
('algorithm', 'for'): 0.045454545454545456,
('for', 'stock'): 0.018518518518518517,
('stock', 'trading'): 1.0,
('trading', 'may'): 1.0,
('may', 'inform'): 0.07692307692307693,
('inform', 'the'): 1.0,
('the', 'trader'): 0.002932551319648094,
('trader', 'of'): 1.0,
('of', 'future'): 0.004166666666666667,
('future', 'potential'): 0.25,
('potential', 'predictions'): 0.25,
('As', 'a'): 0.2,
('a', 'scientific'): 0.0045871559633027525,
('scientific', 'endeavor'): 1.0,
('learning', 'grew'): 0.005291005291005291,
('grew', 'out'): 1.0,
('out', 'of'): 0.4,
('the', 'quest'): 0.002932551319648094,
('quest', 'for'): 1.0,
('for', 'artificial'): 0.018518518518518517,
('In', 'the'): 0.05714285714285714,
('early', 'days'): 0.5,
('days', 'of'): 0.5,
('AI', 'as'): 0.05,
('as', 'an'): 0.015151515151515152,
('an', 'academic'): 0.04878048780487805,
('academic', 'discipline'): 0.5,
(',', 'some'): 0.0024509803921568627,
('some', 'researchers'): 0.08333333333333333,
('researchers', 'were'): 0.2,
('were', 'interested'): 0.1111111111111111,
('interested', 'in'): 1.0,
('in', 'having'): 0.00684931506849315,
('having', 'machines'): 0.25,
('machines', 'learn'): 0.16666666666666666,
('They', 'attempted'): 0.3333333333333333,
('attempted', 'to'): 1.0,
('to', 'approach'): 0.0045662100456621,
('approach', 'the'): 0.14285714285714285,
('problem', 'with'): 0.16666666666666666,
('with', 'various'): 0.02040816326530612,
('various', 'symbolic'): 0.2,
('symbolic', 'methods'): 0.5,
('well', 'as'): 0.2857142857142857,
('as', 'what'): 0.015151515151515152,
('what', 'was'): 0.3333333333333333

```
('what', 'was'): 0.3333333333333333,
('was', 'then'): 0.05555555555555555,
('then', 'termed'): 0.1111111111111111,
('``', 'neural'): 0.02857142857142857,
(';', 'these'): 0.14285714285714285,
('these', 'were'): 0.08333333333333333,
('were', 'mostly'): 0.1111111111111111,
('mostly', 'perceptrons'): 0.5,
('perceptrons', 'and'): 0.5,
('and', 'other'): 0.005494505494505495,
('other', 'models'): 0.058823529411764705,
('models', 'that'): 0.08333333333333333,
('that', 'were'): 0.020618556701030927,
('were', 'later'): 0.1111111111111111,
('later', 'found'): 1.0,
('found', 'to'): 0.375,
('be', 'reinventions'): 0.020833333333333332,
('reinventions', 'of'): 1.0,
('the', 'generalized'): 0.002932551319648094,
('generalized', 'linear'): 1.0,
('linear', 'models'): 0.2,
('models', 'of'): 0.08333333333333333,
('of', 'statistics'): 0.004166666666666667,
('Probabilistic', 'reasoning'): 0.5,
('reasoning', 'was'): 0.5,
('was', 'also'): 0.05555555555555555,
('also', 'employed'): 0.05,
(',', 'especially'): 0.004901960784313725,
('especially', 'in'): 0.3333333333333333,
('in', 'automated'): 0.00684931506849315,
('automated', 'medical'): 0.5,
('medical', 'diagnosis'): 0.4,
('an', 'increasing'): 0.024390243902439025,
('increasing', 'emphasis'): 1.0,
('emphasis', 'on'): 1.0,
('the', 'logical'): 0.002932551319648094,
(',', 'knowledge-based'): 0.0024509803921568627,
('knowledge-based', 'approach'): 1.0,
('approach', 'caused'): 0.14285714285714285,
('caused', 'a'): 1.0,
('a', 'rift'): 0.0045871559633027525,
('rift', 'between'): 1.0,
('AI', 'and'): 0.1,
('and', 'machine'): 0.016483516483516484,
('Probabilistic', 'systems'): 0.5,
('systems', 'were'): 0.05,
('were', 'plagued'): 0.1111111111111111,
('plagued', 'by'): 1.0,
('by', 'theoretical'): 0.02127659574468085,
('theoretical', 'and'): 0.25,
('and', 'practical'): 0.005494505494505495,
('practical', 'problems'): 0.5,
('problems', 'of'): 0.18181818181818182,
('of', 'data'): 0.0375,
('data', 'acquisition'): 0.010526315789473684,
('acquisition', 'and'): 1.0,
('and', 'representation'): 0.005494505494505495,
(',', 'expert'): 0.0024509803921568627,
('expert', 'systems'): 1.0,
('systems', 'had'): 0.05,
('had', 'come'): 0.1,
('come', 'to'): 0.5,
('to', 'dominate'): 0.0045662100456621,
('and', 'statistics'): 0.01098901098901099,
('statistics', 'was'): 0.1,
('was', 'out'): 0.05555555555555555,
('of', 'favor'): 0.004166666666666667,
('Work', 'on'): 1.0,
('on', 'symbolic/knowledge-based'): 0.017543859649122806,
('symbolic/knowledge-based', 'learning'): 1.0,
('learning', 'did'): 0.005291005291005291,
```

```
('learning', 'did'): 0.005291005291005291,
('did', 'continue'): 1.0,
('continue', 'within'): 0.3333333333333333,
(',', 'leading'): 0.00980392156862745,
('leading', 'to'): 1.0,
('to', 'inductive'): 0.0045662100456621,
('inductive', 'logic'): 0.3333333333333333,
('logic', 'programming'): 0.625,
('but', 'the'): 0.23529411764705882,
('the', 'more'): 0.002932551319648094,
('more', 'statistical'): 0.07692307692307693,
('statistical', 'line'): 0.125,
('line', 'of'): 0.3333333333333333,
('of', 'research'): 0.004166666666666667,
('research', 'was'): 0.125,
('was', 'now'): 0.05555555555555555,
('now', 'outside'): 1.0,
('outside', 'the'): 1.0,
('AI', 'proper'): 0.05,
('in', 'pattern'): 0.00684931506849315,
('recognition', 'and'): 0.16666666666666666,
('and', 'information'): 0.005494505494505495,
('information', 'retrieval'): 0.14285714285714285,
('Neural', 'networks'): 1.0,
('networks', 'research'): 0.07692307692307693,
('research', 'had'): 0.125,
('been', 'abandoned'): 0.047619047619047616,
('abandoned', 'by'): 1.0,
('computer', 'science'): 0.125,
('science', 'around'): 0.3333333333333333,
('around', 'the'): 1.0,
('the', 'same'): 0.017595307917888565,
('same', 'time'): 0.16666666666666666,
('This', 'line'): 0.08333333333333333,
(',', 'too'): 0.0024509803921568627,
(',', 'was'): 0.0024509803921568627,
('was', 'continued'): 0.05555555555555555,
('continued', 'outside'): 0.5,
('AI/CS', 'field'): 1.0,
('``', 'connectionism'): 0.02857142857142857,
(',', 'by'): 0.007352941176470588,
('by', 'researchers'): 0.02127659574468085,
('researchers', 'from'): 0.4,
('from', 'other'): 0.02040816326530612,
('other', 'disciplines'): 0.11764705882352941,
('disciplines', 'including'): 0.5,
('Their', 'main'): 1.0,
('main', 'success'): 1.0,
('success', 'came'): 1.0,
('came', 'in'): 1.0,
('the', 'mid-1980s'): 0.002932551319648094,
('mid-1980s', 'with'): 1.0,
('the', 'reinvention'): 0.002932551319648094,
('reinvention', 'of'): 1.0,
('of', 'backpropagation'): 0.004166666666666667,
(',', 'reorganized'): 0.0024509803921568627,
('reorganized', 'as'): 1.0,
('a', 'separate'): 0.009174311926605505,
('separate', 'field'): 0.2,
(',', 'started'): 0.0024509803921568627,
('started', 'to'): 1.0,
('to', 'flourish'): 0.0045662100456621,
('flourish', 'in'): 1.0,
('the', '1990s'): 0.002932551319648094,
('The', 'field'): 0.034482758620689655,
('field', 'changed'): 0.05555555555555555,
('changed', 'its'): 0.5,
('its', 'goal'): 0.058823529411764705,
('goal', 'from'): 0.125,
('from', 'achieving'): 0.02040816326530612,
('achieving', 'artificial'): 0.5
```

('achieving', 'artificial'): 0.5,
('intelligence', 'to'): 0.25,
('to', 'tackling'): 0.0045662100456621,
('tackling', 'solvable'): 1.0,
('solvable', 'problems'): 1.0,
('a', 'practical'): 0.0045871559633027525,
('practical', 'nature'): 0.5,
('It', 'shifted'): 0.07692307692307693,
('shifted', 'focus'): 1.0,
('focus', 'away'): 1.0,
('away', 'from'): 1.0,
('from', 'the'): 0.30612244897959184,
('the', 'symbolic'): 0.002932551319648094,
('symbolic', 'approaches'): 0.5,
('approaches', 'it'): 0.125,
('it', 'had'): 0.03333333333333333,
('had', 'inherited'): 0.1,
('inherited', 'from'): 1.0,
('and', 'toward'): 0.005494505494505495,
('toward', 'methods'): 0.5,
('methods', 'and'): 0.08695652173913043,
('and', 'models'): 0.01098901098901099,
('models', 'borrowed'): 0.041666666666666664,
('borrowed', 'from'): 1.0,
('from', 'statistics'): 0.02040816326530612,
(',', 'fuzzy'): 0.0024509803921568627,
('fuzzy', 'logic'): 1.0,
('and', 'probability'): 0.005494505494505495,
('probability', 'theory'): 0.3333333333333333,
('The', 'difference'): 0.06896551724137931,
('difference', 'between'): 0.75,
('ML', 'and'): 0.14285714285714285,
('AI', 'is'): 0.05,
('is', 'frequently'): 0.007575757575757576,
('frequently', 'misunderstood'): 0.5,
('ML', 'learns'): 0.14285714285714285,
('learns', 'and'): 0.2,
('and', 'predicts'): 0.005494505494505495,
('predicts', 'based'): 0.5,
('on', 'passive'): 0.017543859649122806,
('passive', 'observations'): 1.0,
(',', 'whereas'): 0.0024509803921568627,
('AI', 'implies'): 0.05,
('implies', 'an'): 1.0,
('an', 'agent'): 0.024390243902439025,
('agent', 'interacting'): 1.0,
('interacting', 'with'): 1.0,
('the', 'environment'): 0.008797653958944282,
('environment', 'to'): 0.1,
('learn', 'and'): 0.08333333333333333,
('and', 'take'): 0.005494505494505495,
('take', 'actions'): 0.3333333333333333,
('actions', 'that'): 0.3333333333333333,
('that', 'maximize'): 0.010309278350515464,
('maximize', 'its'): 0.3333333333333333,
('its', 'chance'): 0.058823529411764705,
('chance', 'of'): 0.5,
('of', 'successfully'): 0.004166666666666667,
('successfully', 'achieving'): 1.0,
('achieving', 'its'): 0.5,
('its', 'goals'): 0.058823529411764705,
('As', 'of'): 0.4,
(',', 'many'): 0.0024509803921568627,
('many', 'sources'): 0.07692307692307693,
('sources', 'continue'): 1.0,
('continue', 'to'): 0.3333333333333333,
('to', 'assert'): 0.0045662100456621,
('assert', 'that'): 1.0,
('ML', 'remains'): 0.14285714285714285,
('remains', 'a'): 1.0,

```
('a', 'subfield'): 0.0045871559633027525,
('subfield', 'of'): 1.0,
('Others', 'have'): 1.0,
('have', 'the'): 0.047619047619047616,
('the', 'view'): 0.002932551319648094,
('view', 'that'): 1.0,
('that', 'not'): 0.010309278350515464,
('ML', 'is'): 0.14285714285714285,
('is', 'part'): 0.007575757575757576,
('but', 'only'): 0.058823529411764705,
('only', 'an'): 0.1111111111111111,
('an', "'intelligent"): 0.024390243902439025,
("'intelligent", 'subset'): 1.0,
("'", 'of'): 0.125,
('ML', 'should'): 0.14285714285714285,
('should', 'be'): 0.5,
('be', 'considered'): 0.041666666666666664,
('learning', 'and'): 0.021164021164021163,
('and', 'data'): 0.016483516483516484,
('data', 'mining'): 0.07368421052631578,
('mining', 'often'): 0.09090909090909091,
('often', 'employ'): 0.058823529411764705,
('employ', 'the'): 1.0,
('same', 'methods'): 0.16666666666666666,
('and', 'overlap'): 0.005494505494505495,
('overlap', 'significantly'): 1.0,
('but', 'while'): 0.058823529411764705,
('while', 'machine'): 0.25,
('learning', 'focuses'): 0.005291005291005291,
('on', 'prediction'): 0.017543859649122806,
(',', 'based'): 0.0024509803921568627,
('on', 'known'): 0.017543859649122806,
('known', 'properties'): 0.09090909090909091,
('properties', 'learned'): 0.3333333333333333,
('learned', 'from'): 0.3,
('the', 'training'): 0.03519061583577713,
(',', 'data'): 0.007352941176470588,
('mining', 'focuses'): 0.09090909090909091,
('the', 'discovery'): 0.005865102639296188,
('discovery', 'of'): 0.4,
('(', 'previously'): 0.012658227848101266,
(')', 'unknown'): 0.012658227848101266,
('unknown', 'properties'): 0.25,
('properties', 'in'): 0.3333333333333333,
('the', 'data'): 0.04105571847507331,
('(', 'this'): 0.012658227848101266,
('this', 'is'): 0.1,
('is', 'the'): 0.09090909090909091,
('the', 'analysis'): 0.002932551319648094,
('analysis', 'step'): 0.08333333333333333,
('step', 'of'): 0.25,
('of', 'knowledge'): 0.004166666666666667,
('knowledge', 'discovery'): 0.18181818181818182,
('discovery', 'in'): 0.2,
('in', 'databases'): 0.0136986301369863,
('mining', 'uses'): 0.09090909090909091,
('uses', 'many'): 0.25,
('many', 'machine'): 0.07692307692307693,
('learning', 'methods'): 0.015873015873015872,
('but', 'with'): 0.058823529411764705,
('with', 'different'): 0.02040816326530612,
('different', 'goals'): 0.2,
('other', 'hand'): 0.058823529411764705,
('learning', 'also'): 0.010582010582010581,
('also', 'employs'): 0.05,
('employs', 'data'): 0.5,
('mining', 'methods'): 0.09090909090909091,
('methods', 'as'): 0.043478260869565216,
('``', 'unsupervised'): 0.02857142857142857,
("''", 'or'): 0.06060606060606061,
```

```
 ('or', 'as'): 0.0196078431372549,
 ('a', 'preprocessing'): 0.0045871559633027525,
 ('preprocessing', 'step'): 1.0,
 ('step', 'to'): 0.25,
 ('improve', 'learner'): 0.16666666666666666,
 ('learner', 'accuracy'): 0.3333333333333333,
 ('Much', 'of'): 1.0,
 ('the', 'confusion'): 0.002932551319648094,
 ('confusion', 'between'): 1.0,
 ('between', 'these'): 0.05555555555555555,
 ('these', 'two'): 0.08333333333333333,
 ('two', 'research'): 0.125,
 ('research', 'communities'): 0.125,
 ('(', 'which'): 0.012658227848101266,
 ('which', 'do'): 0.08,
 ('do', 'often'): 0.09090909090909091,
 ('often', 'have'): 0.058823529411764705,
 ('have', 'separate'): 0.047619047619047616,
 ...}
```

In [40]:

```python
inputList = input('Give the sentence to check Probablitiy: ')
```

```
Give the sentence to check Probablitiy: machine learning
```

In [41]:

```python
print("\n Bigrams along with their probability ")

splt=inputList.split()
outputProb1 = 1
bilist=[]
bigrm=[]

for i in range(len(splt) - 1):
  if i < len(splt) - 1:
    bilist.append((splt[i], splt[i + 1]))

print("\n The bigrams in given sentence are ")
print(bilist)

for i in range(len(bilist)):
  if bilist[i] in bigramProb:
    outputProb1 *= bigramProb[bilist[i]]
  else:
    outputProb1 *= 0

print('\n' + 'Probablility of sentence = ' + str(outputProb1))
```

```

 Bigrams along with their probability 

 The bigrams in given sentence are 
[('machine', 'learning')]

Probablility of sentence = 0.9247311827956989
```

In [ ]:

```python

```

In [ ]:

```python

```

In [ ]:

```python

```