```
import matplotlib.pyplot as plt
import seaborn as sns
import matplotlib as mpl
import matplotlib.pylab as pylab
import numpy as np
%matplotlib inline
from nltk.tokenize import sent_tokenize, word_tokenize
import warnings
warnings.filterwarnings(action = 'ignore')
import gensim
from gensim.models import Word2Vec
import re
import bs4 as bs
import urllib.request
import nltk
```

```
import gensim
print(gensim.__version__)
```

```
4.3.2
```

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

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
True
```

```
scrapped_data=urllib.request.urlopen("https://en.wikipedia.org/wiki/Machine_learning")
article=scrapped_data.read()
paresed_article=bs.BeautifulSoup(article,'lxml')
paragraphs=paresed_article.find_all('p')
article_text=""
for p in paragraphs:
  article_text+=p.text
sentences=article_text
```

```
print(article_text)
```

```
Machine learning (ML) is a field of study in artificial intelligence concerned with the development and study of statistical algo
The mathematical foundations of ML are provided by mathematical optimization (mathematical programming) methods. Data mining is a
ML is known in its application across business problems under the name predictive analytics. Although not all machine learning is
The term machine learning was coined in 1959 by Arthur Samuel, an IBM employee and pioneer in the field of computer gaming and ar
By the early 1960s an experimental "learning machine" with punched tape memory, called Cybertron, had been developed by Raytheon
Tom M. Mitchell provided a widely quoted, more formal definition of the algorithms studied in the machine learning field: "A comp
Modern-day machine learning has two objectives, one is to classify data based on models which have been developed, the other purp
As a scientific endeavor, machine learning grew out of the quest for artificial intelligence (AI). In the early days of AI as an
However, an increasing emphasis on the logical, knowledge-based approach caused a rift between AI and machine learning. Probabili
Machine learning (ML), reorganized and recognized as its own field, started to flourish in the 1990s. The field changed its goal
Machine learning and data mining often employ the same methods and overlap significantly, but while machine learning focuses on p
Machine learning also has intimate ties to optimization: many learning problems are formulated as minimization of some loss funct
The difference between optimization and machine learning arises from the goal of generalization: while optimization algorithms ca
Machine learning and statistics are closely related fields in terms of methods, but distinct in their principal goal: statistics
Conventional statistical analyses require the a priori selection of a model most suitable for the study data set. In addition, on
Leo Breiman distinguished two statistical modeling paradigms: data model and algorithmic model,[28] wherein "algorithmic model" m
Some statisticians have adopted methods from machine learning, leading to a combined field that they call statistical learning.[2
Analytical and computational techniques derived from deep-rooted physics of disordered systems can be extended to large-scale pro
A core objective of a learner is to generalize from its experience.[6][32] Generalization in this context is the ability of a lea
The computational analysis of machine learning algorithms and their performance is a branch of theoretical computer science known
For the best performance in the context of generalization, the complexity of the hypothesis should match the complexity of the fu
In addition to performance bounds, learning theorists study the time complexity and feasibility of learning. In computational lea

Machine learning approaches are traditionally divided into three broad categories, which correspond to learning paradigms, depend
Supervised learning algorithms build a mathematical model of a set of data that contains both the inputs and the desired outputs.
Types of supervised-learning algorithms include active learning, classification and regression.[39] Classification algorithms are
Similarity learning is an area of supervised machine learning closely related to regression and classification, but the goal is t
Unsupervised learning algorithms take a set of data that contains only inputs, and find structure in the data, like grouping or c
Cluster analysis is the assignment of a set of observations into subsets (called clusters) so that observations within the same c
Semi-supervised learning falls between unsupervised learning (without any labeled training data) and supervised learning (with co
In weakly supervised learning, the training labels are noisy, limited, or imprecise; however, these labels are often cheaper to o
Reinforcement learning is an area of machine learning concerned with how software agents ought to take actions in an environment
Dimensionality reduction is a process of reducing the number of random variables under consideration by obtaining a set of princi
The manifold hypothesis proposes that high-dimensional data sets lie along low-dimensional manifolds, and many dimensionality red
Other approaches have been developed which do not fit neatly into this three-fold categorization, and sometimes more than one is
Self-learning, as a machine learning paradigm was introduced in 1982 along with a neural network capable of self-learning, named
The self-learning algorithm updates a memory matrix W =||w(a,s)|| such that in each iteration executes the following machine lear
It is a system with only one input, situation, and only one output, action (or behavior) a. There is neither a separate reinforce
Several learning algorithms aim at discovering better representations of the inputs provided during training.[50] Classic example
Feature learning can be either supervised or unsupervised. In supervised feature learning, features are learned using labeled inp
Manifold learning algorithms attempt to do so under the constraint that the learned representation is low-dimensional. Sparse cod
```

```
Feature learning is motivated by the fact that machine learning tasks such as classification often require input that is mathemat
Sparse dictionary learning is a feature learning method where a training example is represented as a linear combination of basis
In data mining, anomaly detection, also known as outlier detection, is the identification of rare items, events or observations w
In particular, in the context of abuse and network intrusion detection, the interesting objects are often not rare objects, but u
Three broad categories of anomaly detection techniques exist.[62] Unsupervised anomaly detection techniques detect anomalies in a
Robot learning is inspired by a multitude of machine learning methods, starting from supervised learning, reinforcement learning,
Association rule learning is a rule-based machine learning method for discovering relationships between variables in large databa
Rule-based machine learning is a general term for any machine learning method that identifies, learns, or evolves "rules" to stor
Based on the concept of strong rules, Rakesh Agrawal, Tomasz Imieliński and Arun Swami introduced association rules for discoveri



{

o
```

```
sentences="""Alice 23 opened the door and found that it led into a small 90
passage, not much larger than a rat-hole:  she knelt down and
looked along the passage into the loveliest garden you ever saw.
How she longed to get out of that dark hall, and wander about
among those beds of bright flowers and those cool fountains, but
she could not even get her head through the doorway; `and even if
my head would go through,' (thought) $poor Alice, `it would be of
very little use without my shoulders.  Oh, how I wish
I could shut up like a telescope!  I think I could, if I only
know how to begin.'  For, you see, so many out-of-the-way things
had happened lately, that Alice had begun to think that very few
things indeed were really impossible.


"""


sentences = re.sub('[^A-Za-z0-9]+', ' ', sentences)
sentences = re.sub(r'(?:^| )\w(?:$| )', ' ', sentences).strip()
print(sentences)
```

```
Alice 23 opened the door and found that it led into small 90 passage not much larger than rat hole she knelt down and looked along t
```

```
# remove special characters
sentences = re.sub('[^A-Za-z]+', ' ', sentences)

# remove 1 letter words
sentences = re.sub(r'(?:^| )\w(?:$| )', ' ', sentences).strip()

# lower all characters
sentences = sentences.lower()

all_sent=nltk.sent_tokenize(sentences)
all_words=[nltk.word_tokenize(sent) for sent in all_sent]

from nltk.corpus import stopwords
for i in range(len(all_words)):
  all_words[i]=[w for w in all_words[i] if w not in stopwords.words('english')]
data =all_words
data1=data[0]


model1 = gensim.models.Word2Vec(data, min_count = 1,vector_size = 52, window = 5)


vocabulary = model1.wv.index_to_key  # List of words in the vocabulary
print(vocabulary)
```

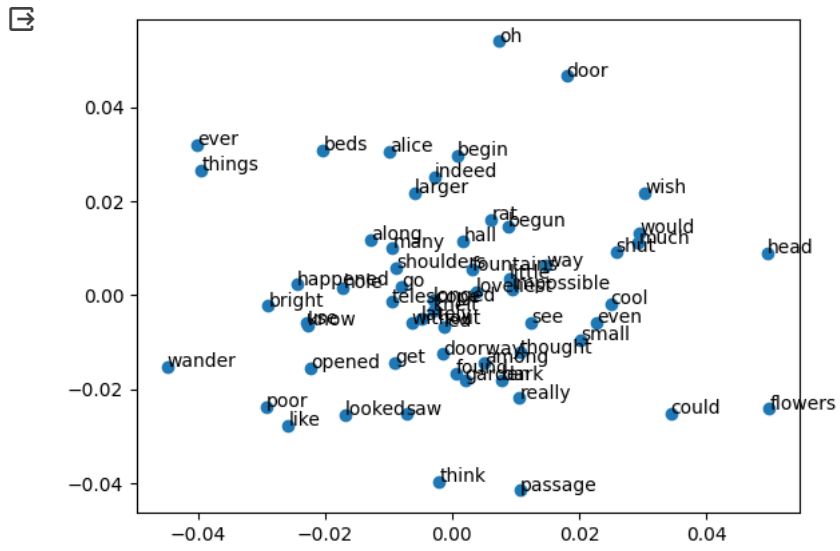
```
['could', 'alice', 'passage', 'think', 'things', 'even', 'head', 'get', 'would', 'ever', 'saw', 'longed', 'indeed', 'dark', 'lovelie
```

```
wrd='door'
#wrd=['subset','machine', 'learning','closely','related']
v1=model1.wv[wrd]
similar_words=model1.wv.most_similar(wrd)
for x in similar_words:
  print(x)
```

```
('beds', 0.36491504311561584)
('much', 0.3305249512195587)
('shut', 0.32979297637939453)
('cool', 0.25908932089805603)
('wish', 0.243193581700325)
('oh', 0.24176433682441711)
('begun', 0.2212926298379898)
```

```
('begin', 0.17681987583637238)
('loveliest', 0.14279094338417053)
('things', 0.13509944081306458)
```

```
from sklearn.decomposition import PCA
import matplotlib.pyplot as plt

# Access the vocabulary (words) and word vectors
words = model1.wv.index_to_key
vectors = model1.wv[words]  # Access vectors for all words in the vocabulary

# Apply PCA to reduce the dimensionality to 2
pca = PCA(n_components=2)
result = pca.fit_transform(vectors)

# Create a scatter plot for the word embeddings
plt.scatter(result[:, 0], result[:, 1])

# Label each point with its corresponding word
for i, word in enumerate(words):
    plt.annotate(word, xy=(result[i, 0], result[i, 1]))

plt.show()
```



Double-click (or enter) to edit

from sklearn.decomposition import PCA import matplotlib.pyplot as plt

Access the vocabulary (words) and word vectors

words = model1.wv.index_to_key vectors = model1.wv[words] # Access vectors for all words in the vocabulary

Apply PCA to reduce the dimensionality to 2

pca = PCA(n_components=2) result = pca.fit_transform(vectors)

Create a scatter plot for the word embeddings

plt.scatter(result[:, 0], result[:, 1])

Label each point with its corresponding word

for i, word in enumerate(words): plt.annotate(word, xy=(result[i, 0], result[i, 1]))

plt.show()

Double-click (or enter) to edit