AI AND BIG DATA TERMS AND DEFINITIONS

1	Adversarial Examples: Inputs to a machine learning model that have been intentionally designed to cause it to make an incorrect prediction.
2	Algorithm: A set of instructions that a computer follows to perform a specific task, such as training a machine learning model.
3	Apache Cassandra: A distributed NoSQL database management system designed to handle large amounts of data across multiple servers.
4	Apache Hadoop: An open-source software framework used for distributed storage and processing of big data.
5	Apache HBase: An open-source, distributed, column-oriented database management system that runs on top of Hadoop.
6	Apache Hive: A data warehouse software that enables data querying, summarization, and analysis of large datasets stored in Hadoop.
7	Apache Kafka: A distributed streaming platform used for building real-time data pipelines and streaming applications.
8	Apache Spark: An open-source distributed computing framework for processing large datasets in parallel.
9	Apache Spark: An open-source, distributed computing system used for processing large-scale data.
10	Artificial Intelligence (AI): The ability of machines to perform tasks that typically require human intelligence, such as recognizing speech, understanding language, and making decisions.
11	Artificial Neural Network: A type of machine learning model inspired by the structure and function of the human brain.
12	Association Rule Mining: A data mining technique used to find relationships or associations between variables in a dataset.
13	Association Rule Mining: A technique in data mining that identifies frequent patterns, relationships, or associations in large datasets.
14	Association Rules: The rules that specify the relationships between items in a dataset.
15	Bag of Words (BoW): A simple method for representing text data as a vector by counting the frequency of each word in a corpus.
16	Bag-of-Words: A model that represents text as a collection of individual words, disregarding grammar and word order.
17	Batch Processing: A method of processing data in which large amounts of data are processed in a single batch or job.
18	Batch Processing: A method of processing large volumes of data in scheduled batches, rather than in real-time.
19	Bayesian network: A probabilistic graphical model used to represent the dependencies between different variables.
20	Bias: A term used to describe how well a machine learning model generalizes to new data that is not included in the training set.

21	Big Data Analytics: The use of data analysis techniques to extract insights and knowledge from large and complex datasets.
22	Big Data: A term used to describe large and complex datasets that cannot be easily processed using traditional data processing tools or methods.
23	Bi-gram: A sequence of two adjacent words in a text.
24	Boosting: A machine learning technique that involves combining several weak learners to create a more powerful classifier.
25	Chunking: A process in natural language processing that identifies and extracts noun phrases from text.
26	Classifier: A machine learning algorithm that is used to assign input data to one of several predetermined categories.
27	Clickstream Analysis: The analysis of user clickstream data on websites to gain insights into user behavior and preferences.
28	Cluster Analysis: A statistical technique for grouping similar data points together based on their characteristics or attributes.
29	Clustering: A machine learning technique that involves grouping data points into different clusters based on their similarity.
30	Computer Vision (CV): The ability of machines to interpret and analyze visual information from the world, such as images and videos.
31	Co-reference Resolution: A task in natural language processing that identifies which words in a text refer to the same entity.
32	Corpus: A large collection of written or spoken texts used for linguistic research.
33	Cross-Validation: A method for testing the accuracy of machine learning models by using a portion of the data to train the model and a different portion to test the model's accuracy.
34	Data Aggregation: The process of combining and summarizing data from multiple sources into a single dataset.
35	Data Cleansing: The process of detecting and correcting or removing inaccuracies or inconsistencies in datasets.
36	Data Exploration: The process of visualizing and analyzing data to understand its characteristics and relationships.
37	Data Fusion: The process of combining multiple sources of data to produce a more comprehensive and accurate representation of a phenomenon.
38	Data Governance: The management of data assets to ensure they are accurate, secure, and compliant with regulations and policies.
39	Data Integration: The process of combining data from different sources into a single dataset for analysis.
40	Data Integration: The process of combining data from multiple sources into a single, unified dataset.
41	Data Mining: The process of discovering patterns, relationships, and insights from large datasets using statistical and machine learning techniques.
42	Data Mining: The process of extracting valuable insights and knowledge from large, complex datasets.

43	Data Preprocessing: A stage in machine learning where the input data is cleaned, transformed, and formatted to improve model performance.
44	Data Preprocessing: The process of cleaning, transforming, and preparing data for analysis.
45	Data Science: A multidisciplinary field that combines statistics, computer science, and domain expertise to extract insights and knowledge from data.
46	Data Visualization: The graphical representation of data to enable insights and understanding.
47	Data Visualization: The process of representing data in visual formats such as charts, graphs, and maps to make it easier to understand.
48	Decision Boundary: The boundary that separates the different classes in a machine learning problem.
49	Decision Tree: A machine learning model that consists of a tree-like structure, where each node represents a decision based on a specific feature.
50	Deep Learning (DL): A subset of ML that involves the use of artificial neural networks to enable a system to learn and improve its performance on a specific task.
51	Dependency Parsing: A task in natural language processing that identifies the grammatical relationships between words in a sentence.
52	Dimension Reduction: A technique in data mining that reduces the number of variables in a dataset while preserving its information content.
53	Dimensionality Reduction: A process that involves reducing the number of features in a dataset while retaining the important information.
54	Dimensionality Reduction: The process of reducing the number of variables or features in a dataset to improve the performance of machine learning models.
55	Discourse Analysis: The study of how language is used in larger units of text, such as conversations and narratives.
56	Document Classification: A task in natural language processing that assigns a category or label to a document based on its content.
57	Embedding: A representation of data in a lower-dimensional space that preserves its essential properties, often used in natural language processing and other fields.
58	Ensemble Learning: A technique where multiple models are combined to improve performance and reduce overfitting.
59	Entity Recognition: A task in natural language processing that identifies and classifies named entities in a text, such as people, organizations, and locations.
60	Entity Resolution: The process of identifying and linking different records that refer to the same entity in a dataset.
61	Exploratory Data Analysis (EDA): The process of analyzing and summarizing data to gain insights and understand its distribution and structure.
62	Extract, Transform, Load (ETL): The process of extracting data from various sources, transforming it into a consistent format, and loading it into a data warehouse or other destination.

63	Feature Engineering: The process of selecting and transforming the most relevant features or variables in a dataset to improve the performance of machine learning models.
64	Feature Selection: A technique in data mining that selects a subset of features or variables that are most relevant for a particular task.
65	Feature: A characteristic of a data point that is used by a machine learning algorithm to make predictions.
66	Frequency Distribution: A statistical analysis that shows how often different words or phrases appear in a text.
67	Frequent Itemset Mining: A data mining technique used to find frequent itemsets or sets of items that often occur together in a transactional dataset.
68	Generative Adversarial Network (GAN): A type of deep neural network architecture that can generate realistic synthetic data by training a generator and a discriminator network together.
69	Geospatial Analysis: The analysis of geographic or location-based data to gain insights into patterns or relationships.
70	Grammar: The set of rules that governs the structure of a language.
71	Hadoop: An open-source framework for storing and processing large datasets in a distributed computing environment.
72	Hierarchical Clustering: A clustering technique that groups similar data points together in a hierarchical structure.
73	Inference: The process of applying a trained machine learning model to new data in order to make predictions or decisions.
74	Information Extraction: A task in natural language processing that extracts structured information from unstructured text data.
75	Information Retrieval: The process of retrieving relevant information from a large dataset, often using techniques such as search engines or recommender systems.
76	Instance-Based Learning: A type of machine learning that involves comparing new input data to previously seen instances in order to make predictions.
77	K-Means Clustering: A clustering technique that groups similar data points together based on their distance from a central point.
78	K-Means: A popular clustering algorithm that groups data points into k clusters based on their similarity.
79	K-Nearest Neighbors (KNN): A simple algorithm for classification and regression tasks that predicts the label or value of a data point based on the majority or average of its k nearest neighbors.
80	Language Model: A statistical model that predicts the probability of a sequence of words in a language.
81	Lemmatization: A process in natural language processing that reduces words to their base form, or lemma.
82	Lexicon: A collection of words and their meanings.
83	Linear Regression: A type of supervised learning algorithm used to predict the value of a continuous variable based on one or more input variables.

84	Linguistics: The scientific study of language and its structure.
85	Log-Likelihood: A measure of the likelihood of observing a set of data given a statistical model, often used as a loss function in machine learning.
86	Loss Function: A function used to measure the difference between the predicted output of a model and the actual output.
87	Machine Learning (ML): A subset of AI that involves the use of algorithms and statistical models to enable a system to improve its performance on a specific task by learning from data.
88	Machine Learning Pipeline: The sequence of steps used to build and deploy a machine learning model, from data preparation to model deployment.
89	Machine Translation: The task of automatically translating text from one language to another using a computer program.
90	MapReduce: A programming model and algorithm used for processing large datasets in a distributed computing environment.
91	Model Selection: The process of choosing the best machine learning model for a given dataset and problem.
92	Naive Bayes: A probabilistic algorithm used for classification tasks that assumes features are independent, given the class.
93	Named Entity Recognition (NER): A task in natural language processing that identifies and classifies named entities in a text, such as people, organizations, and locations.
94	Natural Language Generation: The process of generating natural language text based on a set of rules or a machine learning model.
95	Natural Language Processing (NLP): The ability of computers to understand, interpret, and generate human language.
96	Natural Language Understanding (NLU): The ability of machines to understand and interpret human language, often using techniques like parsing and named entity recognition.
97	Neural Network: A machine learning model that simulates the behavior of the human brain using interconnected nodes or "neurons".
98	N-gram: A sequence of n adjacent words in a text.
99	Non-negative Matrix Factorization (NMF): A type of unsupervised learning algorithm that factorizes a matrix into two matrices, one of which has non-negative entries.
100	NoSQL: A class of database management systems that do not use the traditional relational model.
101	Object Detection: A computer vision technique that involves identifying and localizing objects within an image or video.
102	One-Hot Encoding: A technique used to represent categorical data as binary vectors, with each element representing a different category.
103	Online Analytical Processing (OLAP): A technique for analyzing multidimensional data, often used for business intelligence and reporting.
104	Online Learning: A machine learning technique in which models are trained on incoming data in real time.

105	Online Transaction Processing (OLTP): A method of processing real-time transactions in a database system, often used for e-commerce and financial applications.
106	Optimization: The process of finding the best values for the model parameters to minimize the loss function, often done using iterative algorithms like gradient descent.
107	Outlier Detection: A technique in data mining that identifies unusual, anomalous or abnormal data points that deviate from the norm.
108	Overfitting: The phenomenon where a machine learning model performs well on the training data but poorly on new data, due to capturing noise or irrelevant patterns.
109	Parse Tree: A visual representation of the grammatical structure of a sentence.
110	Part-of-Speech Tagging: A task in natural language processing that assigns a grammatical category to each word in a text.
111	PCA (Principal Component Analysis): A dimensionality reduction technique that involves projecting high-dimensional data onto a lower-dimensional space.
112	Phrase: A grammatical unit that contains one or more words.
113	Precision: A metric used to evaluate the performance of a classification model, measuring the proportion of true positive predictions among all positive predictions.
114	Predictive Analytics: The use of statistical and machine learning techniques to make predictions about future events based on historical data.
115	Predictive Modeling: The process of using statistical techniques and machine learning models to make predictions or forecasts based on historical data.
116	Principal Component Analysis (PCA): A technique used for dimensionality reduction by finding the directions of maximum variance in a dataset.
117	Random Forest: A machine learning algorithm that combines multiple decision trees to create a more powerful classifier.
118	Real-Time Analytics: The analysis of data in real-time or near real-time, often used for applications such as fraud detection or predictive maintenance.
119	Recommendation Systems: Systems that use machine learning algorithms to recommend products or content to users based on their preferences and behavior.
120	Recurrent Neural Network (RNN): A type of neural network commonly used in natural language processing that can process sequences of inputs and maintain an internal state.
121	Regression Analysis: A statistical method for modeling and analyzing the relationship between variables, often used for making predictions.
122	Regular Expression: A sequence of characters used to match patterns in text.
123	Reinforcement Learning: A type of ML that involves an agent learning to interact with an environment to maximize a reward signal.
124	Sampling: A technique used to select a subset of data from a larger dataset, often used to balance imbalanced data.
125	Semi-Supervised Learning: A type of machine learning that uses a combination of labeled and unlabeled data to improve the accuracy of predictions.

126	Sentiment Analysis: A task in natural language processing that identifies the emotional tone of a piece of text, often used in social media and customer feedback analysis.
127	Singular Value Decomposition (SVD): A matrix factorization technique commonly used for dimensionality reduction and feature extraction in machine learning.
128	Social Computing: The study of how people interact with each other and with technology in online social networks, forums, and other online communities.
129	Social Media Analytics: The analysis of data from social media platforms to gain insights into user behavior, preferences, opinions, sentiment, and trends.
130	Social Media Mining: The process of extracting and analyzing data from social media platforms to gain insights into user behavior and preferences.
131	Social Network Analysis: The process of analyzing social networks to understand the relationships and interactions between individuals or groups.
132	Speech Recognition: The process of converting spoken language into text.
133	SQL: Short for Structured Query Language, a programming language used for managing and querying relational databases.
134	Statistical Language Modeling: A technique in natural language processing that estimates the probability of word sequences in a language.
135	Stemming: A process in natural language processing that reduces words to their root or stem form.
136	Stop Words: Words that are commonly used in a language, but usually have little semantic value in text analysis and are often removed during preprocessing.
137	Stream Processing: A method of processing data in real-time as it is generated, often used for applications such as fraud detection or sensor data analysis.
138	Streaming Analytics: The analysis of real-time data streams, such as those generated by sensors, social media, or other sources.
139	Supervised Learning: A type of machine learning that uses labeled data to train models to make predictions or classifications.
140	Support Vector Machine (SVM): A type of supervised learning algorithm used for classification and regression that finds the best hyperplane to separate data into different classes.
141	Syntax: The set of rules that govern the structure of sentences in a language.
142	Synthetic Data: Data that has been artificially generated to supplement or replace real-world data.
143	Test Set: A subset of data used to evaluate the performance of a machine learning model after it has been trained on the training set.
144	Text Analytics: The process of extracting insights and meaning from unstructured text data, often using techniques such as natural language processing and machine learning.
145	Text Classification: A task in natural language processing that assigns a label or category to a piece of text.
146	Text Classification: A type of NLP task that involves assigning labels to pieces of text, such as categorizing emails as spam or not spam.

147	Text Generation: A task in natural language processing that generates new text based on a given input or context.
148	Text Mining: The process of analyzing unstructured text data, such as emails or social media posts, to extract useful information and insights.
149	Text Normalization: The process of transforming text data into a standardized format for easier analysis.
150	Text Preprocessing: The process of cleaning and transforming raw text data in preparation for analysis, including tasks such as tokenization and stemming.
151	Text-to-Speech: The process of converting text into spoken language.
152	Time Series Analysis: A statistical method for analyzing and modeling time-dependent data, such as stock prices or weather patterns.
153	Time Series Analysis: A statistical technique used for analyzing time series data to identify trends, patterns, and anomalies.
154	Time Series Forecasting: The process of making predictions about future values of a time series using statistical and machine learning techniques.
155	Token: An individual word or unit of meaning in a text.
156	Tokenization: A process in natural language processing that breaks text into individual words or tokens.
157	Topic Modeling: A technique in natural language processing that identifies themes or topics in a corpus of text documents.
158	Training Set: A set of labeled data used to train a machine learning model.
159	Transfer Learning: A technique where a pre-trained model is used as a starting point for a new task, and then fine-tuned with new data.
160	Transferability: The ability of a machine learning model to perform well on new, unseen data that comes from a different but related distribution than the training data.
161	Translation Memory: A database of previously translated sentences or phrases that can be used to assist in the translation of new text.
162	Treebank: A collection of parsed sentences used for developing and evaluating natural language processing algorithms.
163	Underfitting: A problem that occurs when a machine learning model is too simple and does not capture the underlying patterns in the data, resulting in poor performance on new data.
164	Unigram: A single word or token in a text.
165	Unstructured Data: Data that is not organized or easily machine-readable, such as text, images, or audio files.
166	Unsupervised Learning: A type of machine learning that involves identifying patterns or structure in data without the use of labeled examples.
167	User Behavior Analytics: The analysis of user behavior data to gain insights into patterns or anomalies, often used for security or fraud detection.
168	Validation Set: A subset of data used to tune the hyperparameters of a machine learning model and prevent overfitting.

169	Vector Space Model: A model that represents documents as vectors in a high-dimensional space, used in information retrieval and natural language processing.
170	Vector: A mathematical representation of a word or document in a high-dimensional space.
171	Web Analytics: The analysis of website data to gain insights into user behavior and website performance, often used for website optimization and digital marketing.
172	Web Mining: The process of extracting useful information from web data, including web pages, hyperlinks, and user behavior.
173	Web Scraping: The process of automatically extracting data from websites, often used for data mining or competitive intelligence.
174	Weight: A parameter in a machine learning model that is learned during the training process and determines the contribution of a feature to the output.
175	Word Frequency: The number of times a word appears in a text or corpus.
176	Word Sense Disambiguation: A task in natural language processing that determines the correct meaning of a word in context.
177	WordNet: A large lexical database of English words and their semantic relationships.
178	WordVec: A specific type of word embedding algorithm that learns vector representations of words based on their co-occurrence with other words in a corpus.
179	Workflow Management: The automation and management of business processes, often using data analytics and machine learning to optimize workflows and improve efficiency.
180	Zero-shot Learning: A type of machine learning where a model can recognize new classes that were not seen during the training phase by leveraging the similarity between existing and new classes.
181	Zipf's Law: A statistical law that states that the frequency of a word in a language is inversely proportional to its rank in a frequency table.