Hillary Dreyer Bruton
Final Exam Glossary
December 7, 2024
AI 2373 NLP

A

Anaconda - an open-source distribution of Python and R designed for data science, machine learning, and scientific computing. It simplifies package management and deployment, offering tools like Jupyter Notebook and pre-installed libraries such as NumPy, pandas, and TensorFlow.

Artificial Neural Network - A computational system inspired biological neural networks, used for machine learning tasks

Autoencoders - neural networks used to learn efficient data encodings in an unsupervised manner. They consist of an encoder to compress input data into a latent space and a decoder to reconstruct the original input, often used for dimensionality reduction or anomaly detection.

B

Backpropagation - An algorithm for training neural networks by propagating error gradients backward through the layers

Bag-of-Words - a feature extraction tool that counts the number of times each word or n-gram (combination of n words) appears in a document.

Yoshua Bengio - a leading figure in deep learning and artificial intelligence, known for his contributions to representation learning, generative models, and the development of neural network algorithms.

BERT (Bidirectional Encoder Representations from Transformers) - A pre-trained transformer model design for language understanding

Bias-Variance Tradeoff - the balance between underfitting (high bias) and overfitting (high variance) in machine learning models. A well-balanced model minimizes both bias and variance to achieve good generalization on unseen data.

C

Chatbots - automates one side of a conversation while a human conversant generally supplies the other side. Can be divided into two categories - database query or conversation generation.

Classification - A supervised learning task where the goal is to predict discrete labels for input data.

Contextual Embeddings - Word embeddings that change based on the context in which a word appears

Convolutional Neural Network (CNN) - a type of deep learning model designed to process structured data like images by identifying patterns through convolutional layers

D

Data pre-processing - processing text to improve model performance or to turn words and characters into a format that the model can understand, numerical

Decision trees - a class of supervised classification models that split the dataset based on different features to maximize information gain in those splits

Deep learning - A subset of machine learning that involves neural networks with many layers of complex data representations

Dimensionality reduction - Techniques like PCA used to reduce the number of features in a dataset while retaining important information

E

Eliza - a model machine built in the mid-1960s to try to solve the Turing Test and used pattern matching and a series of rules without encoding the context of the language

Encoder-decoder sequence-to-sequence - an adaptation to autoencoders specialized for translation, summarization and similar tasks.

F

Feature Engineering - The process of selecting and transforming variables to improve machine learning model performance

Feature Extraction - the process of transforming raw text data into numerical representations or structured features that machine learning models can use for analysis and predictions. Common methods include word embeddings (e.g., Word2Vec, GloVe), bag-of-words, TF-IDF, and contextual embeddings like BERT.

F1 Score - he harmonic mean of precision and recall, providing a single metric to evaluate a model's performance when dealing with imbalanced datasets. It balances the trade-off between false positives and false negatives.

G

General AI - artificial intelligence systems capable of performing any intellectual task a human can do. Unlike narrow AI, which is specialized for specific tasks, general AI would demonstrate adaptability and general reasoning across diverse domains.

Generative Pre-Trained Transformer 3 (GPT-3) - a 175 billion parameter model that can write original prose with human-equivalent fluency in response to an input prompt.

GitHub - a web-based platform for version control and collaboration, built around Git. It allows developers to host, review, and manage code repositories, making it a key tool for collaborative software development, open-source contributions, and project management.

GLoVE - uses a matrix factorization techniques rather than neural learning to build a matrix based on the global word-to-word co-occurence counts.

Gradient Descent - An optimization algorithm used to minimize the error of machine learning models by updating parameters iteratively.

Grammatical error correction - models that encode grammatical rules to correct the grammar within a text.

H

Geoffrey Hinton - a pioneer in deep learning and artificial neural networks. Known as the "Godfather of Deep Learning," he co-developed backpropagation and contributed significantly to advances in representation learning and neural networks.

Hugging Face - a popular library that provides open source pre-trained NLP models and tools for deployment

I

Intent Recognition - Pinpointing the purpose or goal behind a user’s input, common in chatbot and voice assistant systems

Inverse Document Frequency - answers the question of how important a term is to the whole corpus

J

Jupyter Notebook - An open-source web application for interactive coding and visualization, commonly used in NLP work

Joint probability - A statistical measure in probablistic NLP used to model the likelihood of sequences or events occurring together

K

Kaggle - an online platform for data science competitions, learning resources, and collaborative projects. It provides datasets, code notebooks, and community-driven solutions for machine learning and analytics tasks.

Kernel Methods - Algorithms like SVM that use kernel functions for pattern analysis in NLP and machine learning

K-means Clustering - An unsupervised learning algorithm for partitioning data into K groups based on similarity

L

Language model - A statistical or neural model designed to predict the likelihood of a sequence of words

Yann LeCun - a prominent AI researcher and one of the pioneers of deep learning. He is known for his work on convolutional neural networks (CNNs) and their application in computer vision and other fields.

Lemmatization: The process of reducing words to their base or dictionary form (i.e. running to run)

Logistic regression - a supervised classification algorithm that aims to predict the probability that an event will occur based on some input. In NLP, logistic regression models can be applied to solve problems such as sentiment analysis, spam detection and toxicity classification.

M

Machine Learning - A subset of AI focused on developing algorithms that allow systems to learn and improve from data

Machine Translation - the automation of translation between different languages

N

Naive-Bayes - a supervised classification algorithm that finds the conditional probability distribution P using the Bayes formula

Named entity recognition - the extraction of entities in a piece of text into predefined categories such as personal names, organizaitons, locations and quantities.

Natural Language Processing - the discipline of building machines that can manipulate human language - or data that resemble human language - in teh way that it is written, spoken and organized.

Natural Langauge Toolkit (NLTK) - one of the first NLP libraries written in Python. Provides easy-to-use interfaces to corpora and lexical resources such as WordNet.

O

Overfitting - When a model performs well on training data but poorly on unseen data due to excessive complexity

P

Parsing - The process of analyzing the syntactic structure of a sentence to understand grammatical relationships

Precision - the ratio of true positive predictions to the total positive predictions made by a model. It measures how many of the predicted positives are actually correct, useful in tasks where false positives are costly.

Pre-training - Training a model on a large dataset before fine-tuning it for a specific task

Python - the most-used programming language used to tackle NLP tasks.

Q

Quantization - Reducing the precision of numbers in a model to optimize it for speed and resource efficiency

R

Recall - the ratio of true positive predictions to the total actual positives in the dataset. It measures a model's ability to identify all relevant instances, crucial for applications where false negatives are critical.

Recurrent Neural Network (RNN) - A type of neural network designed to process sequential data, such as text or time series.

Regularization - a technique used in machine learning to prevent overfitting by adding constraints to the model. Common methods include L1 (Lasso) and L2 (Ridge) regularization, which penalize large coefficients in the model.

S

Sentence Segmentation - process of breaking a large piece of text into linguistically meaningful sentence units.

Sentiment Analysis - the process of classifying the emotional intent of text

Spam detection - a prevalent binary classification problem in NLP, where the purpose is to classify emails as either spam or not.

Stemming - The process of reducing words to their base or root form

Stop Word Removal - process to remove the most commonly occurring words that don't add much information to the text. I.e. "The", "a", "an", etc.

Summarization - the task of shortening text to highlight th most relevant information

Support Vector Machines - supervised learning models used for classification and regression tasks. They work by finding the hyperplane that best separates data points into distinct classes while maximizing the margin between them.

T

Text generation - more formally known as natural language generation (NLG), it produces text that's similar to human-written text.TF-IDF (Term Frequency - Inverse Document Frequency) - A statistical method for evaluating the important of a word in a document relative to a corpus

Text representation - the process of converting raw text into structured numerical formats, such as bag-of-words, TF-IDF, or word embeddings, for input into machine learning or NLP models.

TF-IDF - a feature extraction tool that weighs each word by its importance. Two things are considered in a word significance evaluation - term frequency and inverse document frequency.

Tokenization - the process of splitting text into individual words and word fragments

Topic modeling - an unsupervised text mining task that takes a corpus of documents and discovers abstract topics within that corpus.

Toxicity classification - a branch of sentiment analysis where the aim is not just to classify hostile intent but also to classify particular categories such as threats, insults, obscenities, and hatred towards certain identifies

Transformer - A neural network architecture designed for handling sequential data with attention mechanisms.

U

Unsupervised Learning - A machine learning technique where models learn patterns in data without labeled outputs

Underfitting - When a model is too simple to capture the patterns in the data, leading to poor performance

V

Vectorization - Converting text data into numerical representations, such as word embeddings

W

WordPiece Tokenization - A subword tokenization technique used in BERT to handle rare words efficiently.

Word2Vec - uses vanilla neural network to learn high-dimensional word embeddings from raw text. It comes in two variations - skip-gram, which predicts surrounding words based on a given target word and Continuous Bag-of-Words, which tries to predict the target word from surrounding words.

X

XML (Extensible Markup Language) - A markup language used to structure data, often in NLP datasets

Y

YAML (Yet Another Markup Language) - A human-readable data serialization format commonly used in machine learning configuration files.

Z

Zipf's Law - A statistical observation that in language, the frequency of a word is inversely proportional to its rank in the frequency table.