AI, NLP, & ChatGPT Vocabulary

BASIC DEFINITIONS

AI (Artificial Intelligence): Decision making capabilities in machines that traditionally required human intelligence.

NLP: Short for Natural Language Processing. It is a branch of AI centered around enabling machines/computers to understand text and spoken language.

Machine Learning: A branch of AI where algorithms work on identifying patterns in data by simulating human learning approaches.

Language Models: A type of statistical/ML model that possesses probability distribution over a sequence of words.

Conversation AI: Branch of AI and NLP that simulates human-human conversation between humans and machines.

GPT: Short for Generative Pre-trained Transformers. A type of language model.

Prompt: A command or an action sentence used to communicate with ChatGPT and other AI.

End-user: Human interacting with the AI tool.

APPLICATIONS

Text classification: a machine learning technique that categorizes a given text into a predefined class.

Sentiment analysis: NLP technique to identify the human emotion from a given text.

Translation: NLP technique to automatically translate a text from one language to another.

Question answering: uses NLP techniques and information retrieval approaches to answer natural language questions by human users.

Natural Language Processing (NLP) BUILDING PROCESS

Corpus: A corpus is a collection of text. A corpus can be a collection of movie reviews, internet comments or conversations between two people.

Vocabulary: The entire set of terms used in a body of text.

Documents: Refers to a body of text. A collection of documents make up a corpus. For instance, a movie review or an email are examples of a document.

Preprocessing: The first step to any NLP task is to preprocess the text. The goal of preprocessing is to "clean" the text by removing as much noise as possible. Some preprocessing techniques are listed below:

Noise:

Parts-of-speech tagging: The syntactic function of a word.

Normalization: The process of reducing similar tokens to a canonical form.

Stop-words: These words are ignored prior to any preprocessing or modeling tasks.

Lemmatisation/Stemming: reduces inflected terms to their base forms.

Tokenization: The process of breaking a large chunk of text into smaller pieces. This is usually done so that each small piece, or token, can be mapped to a meaningful unit of information. If we choose to break our text on the word level, each word becomes its own token.

Word embeddings (vectors): Each token is embedded as a vector before it can be passed to a machine learning model. While generally referred to as word embeddings, embeddings can be created on the character or phrase level as well.

LM: LANGUAGE MODEL

Statistical language models

N-grams: continuous sequence of words or tokens in a text.

LLM: Large language models

Seq2Seq Models (Sequence-to-sequence models): A model that takes a sequence of words/tokens and outputs another sequence of items.
Transformer Model: a deep neural network that learns the context of the given sequential text data by tracking the relationships of the words.
GPT: GPT stands for "generative pre-trained transformer." GPT is a machine learning model that can perform natural language generation (NLG) tasks.

OpenAI developed ChatGPT, and their evolving definitions of various iterations (ChatGPT 2, 3, etc) can be found here.

General Elementary Vocabulary List:

Artificial Intelligence (AI): A field of computer science that focuses on creating machines or software that can perform tasks and make decisions that usually require human intelligence.

Machine Learning (ML): A subset of AI that involves teaching computers how to learn from data and improve their performance without being explicitly programmed.

Algorithm: A set of step-by-step instructions or rules that a computer follows to solve a specific problem or complete a task.

Data: Facts, numbers, or information that is used by computers to learn and make decisions. It can be in the form of text, images, audio, or other types of digital content.

Neural Network: A computer system that is inspired by the structure and function of the human brain. It is designed to recognize patterns and make predictions based on input data.

Automation: The use of machines or software to perform tasks or processes automatically, without the need for constant human intervention.

Chatbot: A computer program designed to simulate conversation with humans, often used to provide information or assistance on websites or messaging platforms.

Natural Language Processing (NLP): The ability of a computer to understand and interpret human language, allowing it to interact with people in a more natural and intuitive way.

Robotics: The branch of technology that deals with the design, construction, and operation of robots. Robots are machines that can perform tasks autonomously or with minimal human input.

Ethics: The moral principles and values that guide human behavior. In the context of AI, ethics refers to considering the potential impact and consequences of AI systems on society, privacy, and fairness.

Works Consulted and For Further Reading:

NLP Glossary for Beginners
Natural Language Processing Key Terms, Explained
What is ChatGPT?
What Is a Transformer Model?
How do Transformers Work?