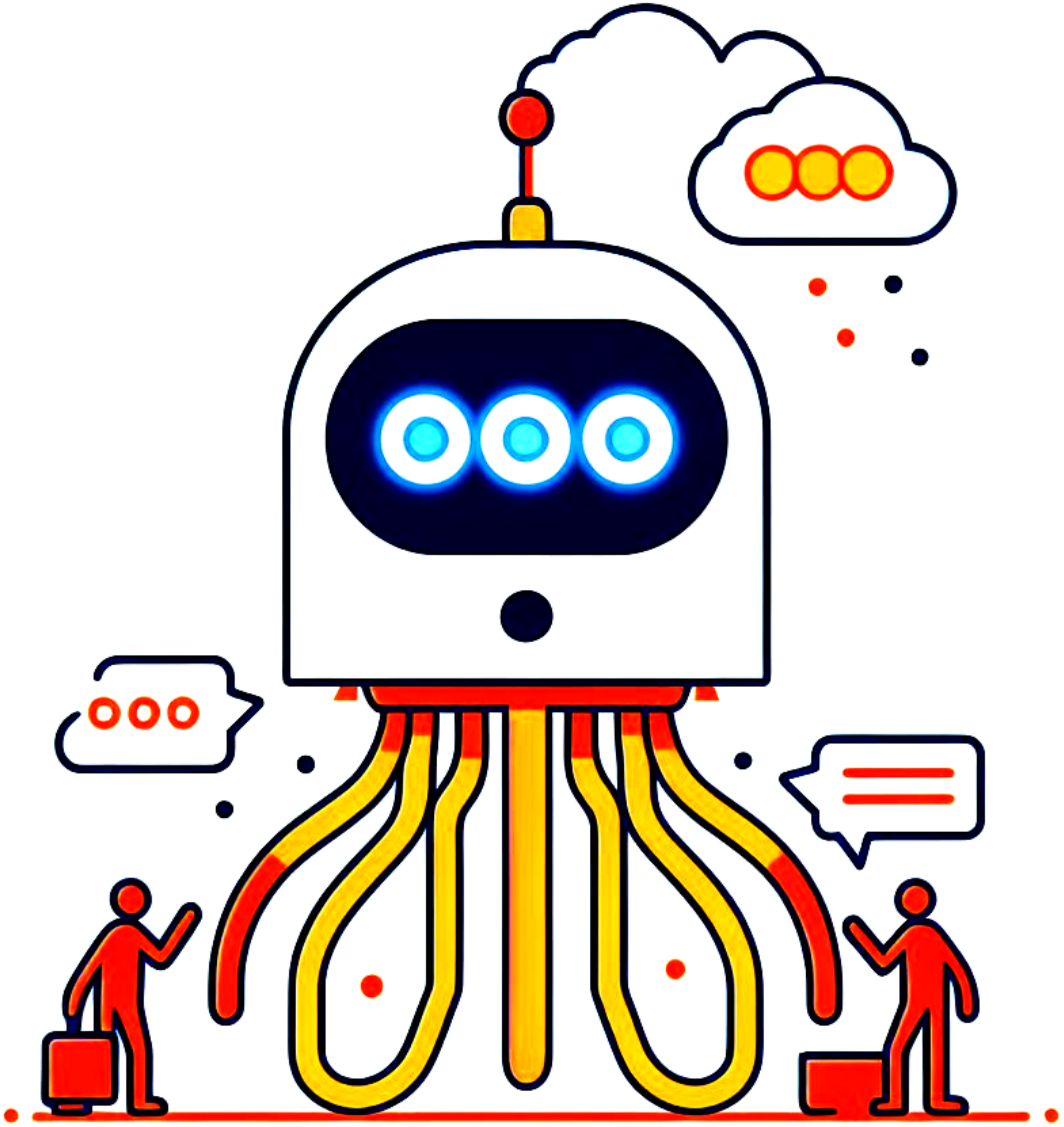
Ünver Çiftçi

NATURAL LANGUAGE PROCESSING



Introduction to the basic concepts
(Early Draft)

CONTENTS

1. Introduction
2. Text Processing
3. Language Models
4. Text Classification
5. Vector Embeddings
6. Neural Networks
7. Sequence
8. Transformer
9. Large Language Models

Chapter 1

INTRODUCTION

- What is NLP?
- What are the NLP tasks?
- How does NLP work?

Natural Language Processing (NLP) is the discipline of building machines that can manipulate human language – or data that resembles human language – in the way that it is written, spoken, and organized.

NLP is an integral part of everyday life and becoming more so as language technology is applied to diverse fields like retailing (for instance, in customer service chatbots) and medicine (interpreting or summarizing electronic health records).

Conversational agents such as Amazon's Alexa and Apple's Siri utilize NLP to listen to user queries and find answers.

The most sophisticated such agents – such as GPT-3, which was recently opened for commercial applications – can generate sophisticated prose on a wide variety of topics as well as power chatbots that are capable of holding coherent conversations.

Google uses NLP to improve its search engine results, and social networks like Facebook use it to detect and filter hate speech.

Chapter 1

INTRODUCTION

- What is NLP?
- What are the NLP tasks?
- How does NLP work?

- **Text generation**, more formally known as natural language generation (NLG), produces text that's similar to human-written text. Such models can be fine-tuned to produce text in different genres and formats – including tweets, blogs, and even computer code. Text generation has been performed using Markov processes, LSTMs, BERT, GPT-2, LaMDA, and other approaches. It's particularly useful for autocomplete and chatbots.
  - Autocomplete predicts what word comes next, and autocomplete systems of varying complexity are used in chat applications like WhatsApp. Google uses autocomplete to predict search queries. One of the most famous models for autocomplete is GPT-2, which has been used to write articles, song lyrics, and much more.
  - Chatbots automate one side of a conversation while a human conversant generally supplies the other side. They can be divided into the following two categories:
    - Database query: We have a database of questions and answers, and we would like a user to query it using natural language.
    - Conversation generation: These chatbots can simulate dialogue with a human partner. Some are capable of engaging in wide-ranging conversations. A high-profile example is Google's LaMDA, which provided such human-like answers to questions that one of its developers was convinced that it had feelings.

Chapter 1

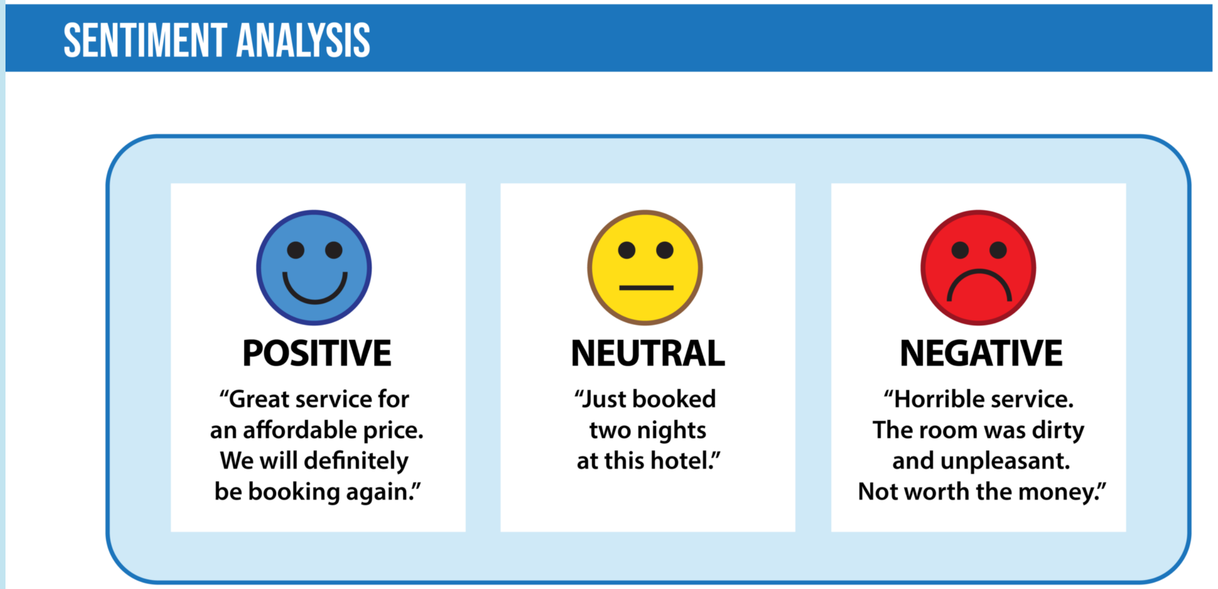
INTRODUCTION

- What is NLP?
- What are the NLP tasks?
- How does NLP work?

NLP is used for a wide variety of language-related tasks, including answering questions, classifying text in a variety of ways, and conversing with users.

Here are 11 tasks that can be solved by NLP:

- **Sentiment analysis** is the process of classifying the emotional intent of text. Generally, the input to a sentiment classification model is a piece of text, and the output is the probability that the sentiment expressed is positive, negative, or neutral. Typically, this probability is based on either hand-generated features, word n-grams, TF-IDF features, or using deep learning models to capture sequential long- and short-term dependencies. Sentiment analysis is used to classify customer reviews on various online platforms as well as for niche applications like identifying signs of mental illness in online comments.



SENTIMENT ANALYSIS

POSITIVE
"Great service for an affordable price. We will definitely be booking again."

NEUTRAL
"Just booked two nights at this hotel."

NEGATIVE
"Horrible service. The room was dirty and unpleasant. Not worth the money."

Given text, sentiment analysis classifies its emotional quality.

Chapter 1

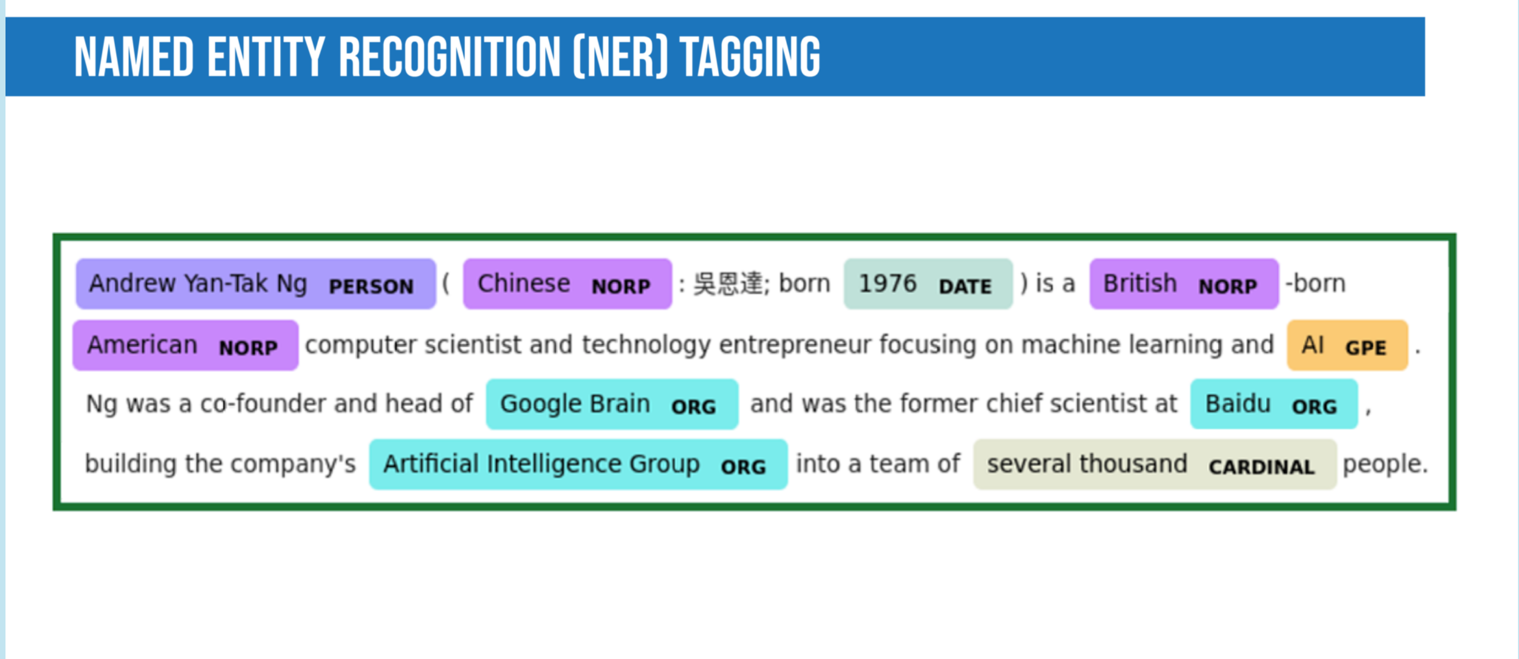
INTRODUCTION

- What is NLP?
- What are the NLP tasks?
- How does NLP work?

- **Toxicity classification** is 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 identities. Toxicity classification models can be used to moderate and improve online conversations by silencing offensive comments, detecting hate speech, or scanning documents for defamation.

- **Machine translation** automates translation between different languages. The input to such a model is text in a specified source language, and the output is the text in a specified target language. Google Translate is perhaps the most famous mainstream application. Such models are used to improve communication between people on social-media platforms such as Facebook or Skype. Some systems also perform language identification; that is, classifying text as being in one language or another.

- **Named entity recognition** aims to extract entities in a piece of text into predefined categories such as personal names, organizations, locations, and quantities. The input to such a model is generally text, and the output is the various named entities along with their start and end positions. Named entity recognition is useful in applications such as summarizing news articles and combating disinformation.



NAMED ENTITY RECOGNITION (NER) TAGGING

Andrew Yan-Tak Ng PERSON (Chinese NORP : 吳恩達; born 1976 DATE) is a British NORP -born American NORP computer scientist and technology entrepreneur focusing on machine learning and AI GPE . Ng was a co-founder and head of Google Brain ORG and was the former chief scientist at Baidu ORG , building the company's Artificial Intelligence Group ORG into a team of several thousand CARDINAL people.

Chapter 1

INTRODUCTION

- What is NLP?
- What are the NLP tasks?
- How does NLP work?

- **Spam detection** is a prevalent binary classification problem in NLP, where the purpose is to classify emails as either spam or not. Spam detectors take as input an email text along with various other subtexts like title and sender's name. They aim to output the probability that the mail is spam. Email providers like Gmail use such models to provide a better user experience by detecting unsolicited and unwanted emails and moving them to a designated spam folder.

- **Grammatical error correction** models encode grammatical rules to correct the grammar within text. This is viewed mainly as a sequence-to-sequence task, where a model is trained on an ungrammatical sentence as input and a correct sentence as output. Online grammar checkers like Grammarly and word-processing systems like Microsoft Word use such systems to provide a better writing experience to their customers. Schools also use them to grade student essays.

- **Topic modeling** is an unsupervised text mining task that takes a corpus of documents and discovers abstract topics within that corpus. The input to a topic model is a collection of documents, and the output is a list of topics that defines words for each topic as well as assignment proportions of each topic in a document. Latent Dirichlet Allocation (LDA), one of the most popular topic modeling techniques, tries to view a document as a collection of topics and a topic as a collection of words. Topic modeling is being used commercially to help lawyers find evidence in legal documents.

Chapter 1

INTRODUCTION

- What is NLP?
- What are the NLP tasks?
- How does NLP work?

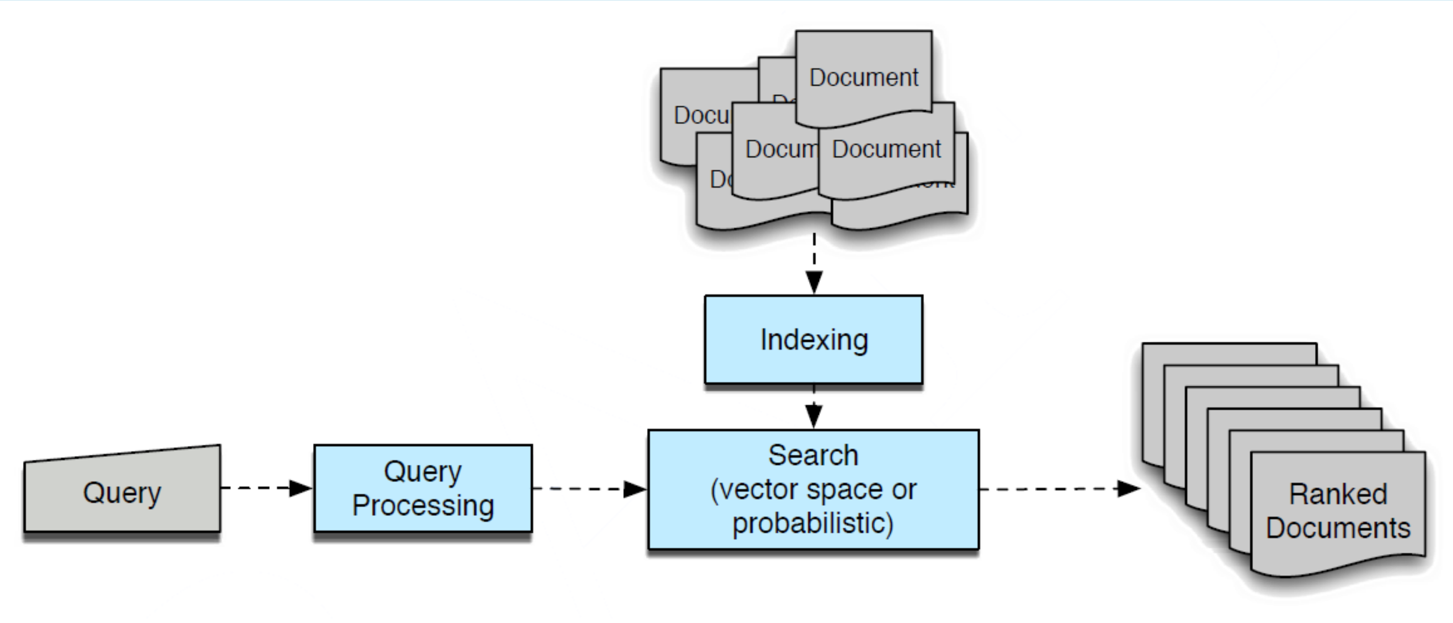
- **Text generation**, more formally known as natural language generation (NLG), produces text that's similar to human-written text. Such models can be fine-tuned to produce text in different genres and formats – including tweets, blogs, and even computer code. Text generation has been performed using Markov processes, LSTMs, BERT, GPT-2, LaMDA, and other approaches. It's particularly useful for autocomplete and chatbots.
    - **Autocomplete** predicts what word comes next, and autocomplete systems of varying complexity are used in chat applications like WhatsApp. Google uses autocomplete to predict search queries. One of the most famous models for autocomplete is GPT-2, which has been used to write articles, song lyrics, and much more.
    - **Chatbots** automate one side of a conversation while a human conversant generally supplies the other side. They can be divided into the following two categories:
        - **Database query:** We have a database of questions and answers, and we would like a user to query it using natural language.
        - **Conversation generation:** These chatbots can simulate dialogue with a human partner. Some are capable of engaging in wide-ranging conversations. A high-profile example is Google's LaMDA, which provided such human-like answers to questions that one of its developers was convinced that it had feelings.

Chapter 1

INTRODUCTION

- What is NLP?
- What are the NLP tasks?
- How does NLP work?

- **Information retrieval** finds the documents that are most relevant to a query. This is a problem every search and recommendation system faces. The goal is not to answer a particular query but to retrieve, from a collection of documents that may be numbered in the millions, a set that is most relevant to the query. Document retrieval systems mainly execute two processes: indexing and matching. In most modern systems, indexing is done by a vector space model through Two-Tower Networks, while matching is done using similarity or distance scores. Google recently integrated its search function with a multimodal information retrieval model that works with text, image, and video data.



Chapter 1

INTRODUCTION

- What is NLP?
- What are the NLP tasks?
- How does NLP work?

- **Summarization** is the task of shortening text to highlight the most relevant information. Researchers at Salesforce developed a summarizer that also evaluates factual consistency to ensure that its output is accurate. Summarization is divided into two method classes:

  - **Extractive summarization** focuses on extracting the most important sentences from a long text and combining these to form a summary. Typically, extractive summarization scores each sentence in an input text and then selects several sentences to form the summary.
  - **Abstractive summarization** produces a summary by paraphrasing. This is similar to writing the abstract that includes words and sentences that are not present in the original text. Abstractive summarization is usually modeled as a sequence-to-sequence task, where the input is a long-form text and the output is a summary.

Question answering deals with answering questions posed by humans in a natural language. One of the most notable examples of question answering was Watson, which in 2011 played the television game-show *Jeopardy* against human champions and won by substantial margins. Generally, question-answering tasks come in two flavors:

- **Multiple choice:** The multiple-choice question problem is composed of a question and a set of possible answers. The learning task is to pick the correct answer.
- **Open domain**: In open-domain question answering, the model provides answers to questions in natural language without any options provided, often by querying a large number of texts.

Chapter 1

INTRODUCTION

- What is NLP?
- What are the NLP tasks?
- How does NLP work?

NLP models work by finding relationships between the constituent parts of language – for example, the letters, words, and sentences found in a text dataset. NLP architectures use various methods for data preprocessing, feature extraction, and modeling. Some of these processes are:

- **Data preprocessing:** Before a model processes text for a specific task, the text often needs to be preprocessed to improve model performance or to turn words and characters into a format the model can understand. Data-centric AI is a growing movement that prioritizes data preprocessing. Various techniques may be used in this data preprocessing:
    - **Stemming and lemmatization**: Stemming is an informal process of converting words to their base forms using heuristic rules. For example, "university," "universities," and "university's" might all be mapped to the base *univers*. (One limitation in this approach is that "universe" may also be mapped to *univers*, even though universe and university don't have a close semantic relationship.) Lemmatization is a more formal way to find roots by analyzing a word's morphology using vocabulary from a dictionary. Stemming and lemmatization are provided by libraries like spaCy and NLTK.
    - **Sentence segmentation** breaks a large piece of text into linguistically meaningful sentence units. This is obvious in languages like English, where the end of a sentence is marked by a period, but it is still not trivial. A period can be used to mark an abbreviation as well as to terminate a sentence, and in this case, the period should be part of the abbreviation token itself. The process becomes even

Chapter 1

INTRODUCTION

- What is NLP?
- What are the NLP tasks?
- How does NLP work?

NLP models work by finding relationships between the constituent parts of language – for example, the letters, words, and sentences found in a text dataset. NLP architectures use various methods for data preprocessing, feature extraction, and modeling. Some of these processes are:

- **Data preprocessing:** Before a model processes text for a specific task, the text often needs to be preprocessed to improve model performance or to turn words and characters into a format the model can understand. Various techniques may be used in this data preprocessing:
    - **Stemming and lemmatization**: Stemming is an informal process of converting words to their base forms using heuristic rules. For example, "university," "universities," and "university's" might all be mapped to the base *univers*.
    - **Sentence segmentation** breaks a large piece of text into linguistically meaningful sentence units.
    - **Stop word removal** aims to remove the most commonly occurring words that don't add much information to the text. For example, "the," "a," "an," and so on.
    - **Tokenization** splits text into individual words and word fragments. The result generally consists of a word index and tokenized text in which words may be represented as numerical tokens for use in various deep learning methods.

Chapter 1

INTRODUCTION

- What is NLP?
- What are the NLP tasks?
- How does NLP work?

- **Feature extraction:** Most conventional machine-learning techniques work on the features – generally numbers that describe a document in relation to the corpus that contains it.
    - **Bag-of-Words:** Bag-of-Words counts the number of times each word or n-gram (combination of n words) appears in a document. For example, below, the Bag-of-Words model creates a numerical representation of the dataset based on how many of each word in the word_index occur in the document.
    - **TF-IDF:** In Bag-of-Words, we count the occurrence of each word or n-gram in a document. In contrast, with TF-IDF, we weight each word by its importance.
    - **Word2Vec**, introduced in 2013, uses a vanilla neural network to learn high-dimensional word embeddings from raw text.
    - **GLoVE** is similar to Word2Vec as it also learns word embeddings, but it does so by using matrix factorization techniques rather than neural learning. The GLoVE model builds a matrix based on the global word-to-word co-occurrence counts.

Chapter 1

INTRODUCTION

- What is NLP?
- What are the NLP tasks?
- How does NLP work?

- **Modeling:** After data is preprocessed, it is fed into an NLP architecture that models the data to accomplish a variety of tasks.

    - Numerical features extracted by the techniques described above can be fed into various models depending on the task at hand. For example, for classification, the output from the TF-IDF vectorizer could be provided to logistic regression, naive Bayes, decision trees, or gradient boosted trees. Or, for named entity recognition, we can use hidden Markov models along with n-grams.

    - **Logistic regression** is 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.
    - **Naive Bayes** is a supervised classification algorithm that finds the conditional probability distribution P(label | text) using the Bayes formula.

    - **Deep neural networks** typically work without using extracted features, although we can still use TF-IDF or Bag-of-Words features as an input.

    - **Language Models**: In very basic terms, the objective of a language model is to predict the next word when given a stream of input words. Probabilistic models that use Markov assumption are one example:

Chapter 1

INTRODUCTION

Programming Languages, Libraries, And Frameworks For NLP

Many languages and libraries support NLP. Here are a few of the most useful.

- **Python** is the most-used programming language to tackle NLP tasks. Most libraries and frameworks for deep learning are written for Python. Here are a few that practitioners may find helpful:
  - **Natural Language Toolkit (NLTK)** is one of the first NLP libraries written in Python. It provides easy-to-use interfaces to corpora and lexical resources such as WordNet. It also provides a suite of text-processing libraries for classification, tagging, stemming, parsing, and semantic reasoning.
  - **spaCy** is one of the most versatile open source NLP libraries. It supports more than 66 languages. spaCy also provides pre-trained word vectors and implements many popular models like BERT.
  - **Deep Learning libraries:** Popular deep learning libraries include TensorFlow and PyTorch, which make it easier to create models with features like automatic differentiation.
  - **Hugging Face** offers open-source implementations and weights of over 135 state-of-the-art models. The repository enables easy customization and training of the models.
  - **Gensim** provides vector space modeling and topic modeling algorithms.

Chapter 1

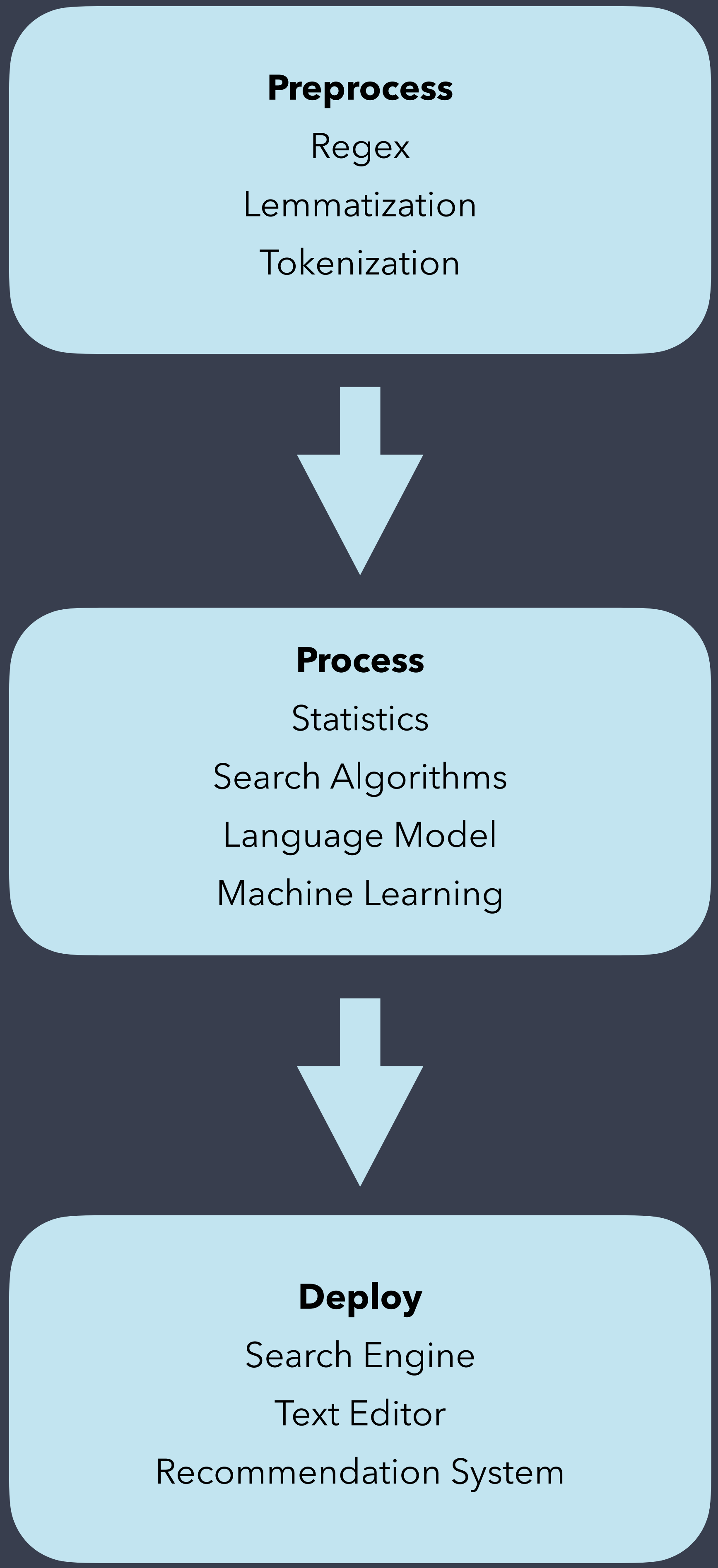
INTRODUCTION

Programming Languages, Libraries, And Frameworks For NLP

- **R:** Many early NLP models were written in R, and R is still widely used by data scientists and statisticians. Libraries in R for NLP include TidyText, Weka, Word2Vec, SpaCyR, TensorFlow, and PyTorch.

- Many other languages including JavaScript, Java, and Julia have libraries that implement NLP methods.

NLP LIFECYCLE



Preprocess

Regex

Lemmatization

Tokenization

Process

Statistics

Search Algorithms

Language Model

Machine Learning

Deploy

Search Engine

Text Editor

Recommendation System

REFERENCES

1. Speech and Language Processing (3rd ed. draft) Dan Jurafsky and James H. Martin, https://web.stanford.edu/~jurafsky/slp3

2. A Complete Guide to Natural Language Processing, DeepLearning.AI, https://www.deeplearning.ai/resources/natural-language-processing