CHRIST
(DEEMED TO BE UNIVERSITY)
BANGALORE · INDIA

TEXT ANALYTICS

Unit-2

Unit-2

- Text Representation- tokenization, stemming, stop words, TF-IDF, NER, N-gram modeling. Mining Textual Data: Text Clustering, Text Classification,

- **LabExercises:**

3. Implementation of tokenization, stemming, stop words
4. Implementation of text classification and clustering with TF-IDF and N-gram.

Text analysis

- Text mining or natural language processing (NLP), is a field of study focuses on extracting meaningful insights and patterns from unstructured text data
- With the proliferation of digital content, including social media posts, customer reviews, news articles, and scientific literature, and etc.

Key Concepts in Text Analysis

- **Text Preprocessing**
  - tokenization (splitting text into words or sentences),
  - removing stopwords (common words like "the", "and", "is"),
  - stemming or lemmatization (reducing words to their root form)
  - handling special characters and punctuation.
- **Text Representation**: Text data converted into a numerical format for analysis.
  - Bag of Words (BoW)
  - Term Frequency-Inverse Document Frequency (TF-IDF): A numerical statistic that reflects the importance of a word in a document relative to a corpus.
  - Word Embedding: Dense vector representations of words in a high-dimensional space, capturing semantic relationships between words.

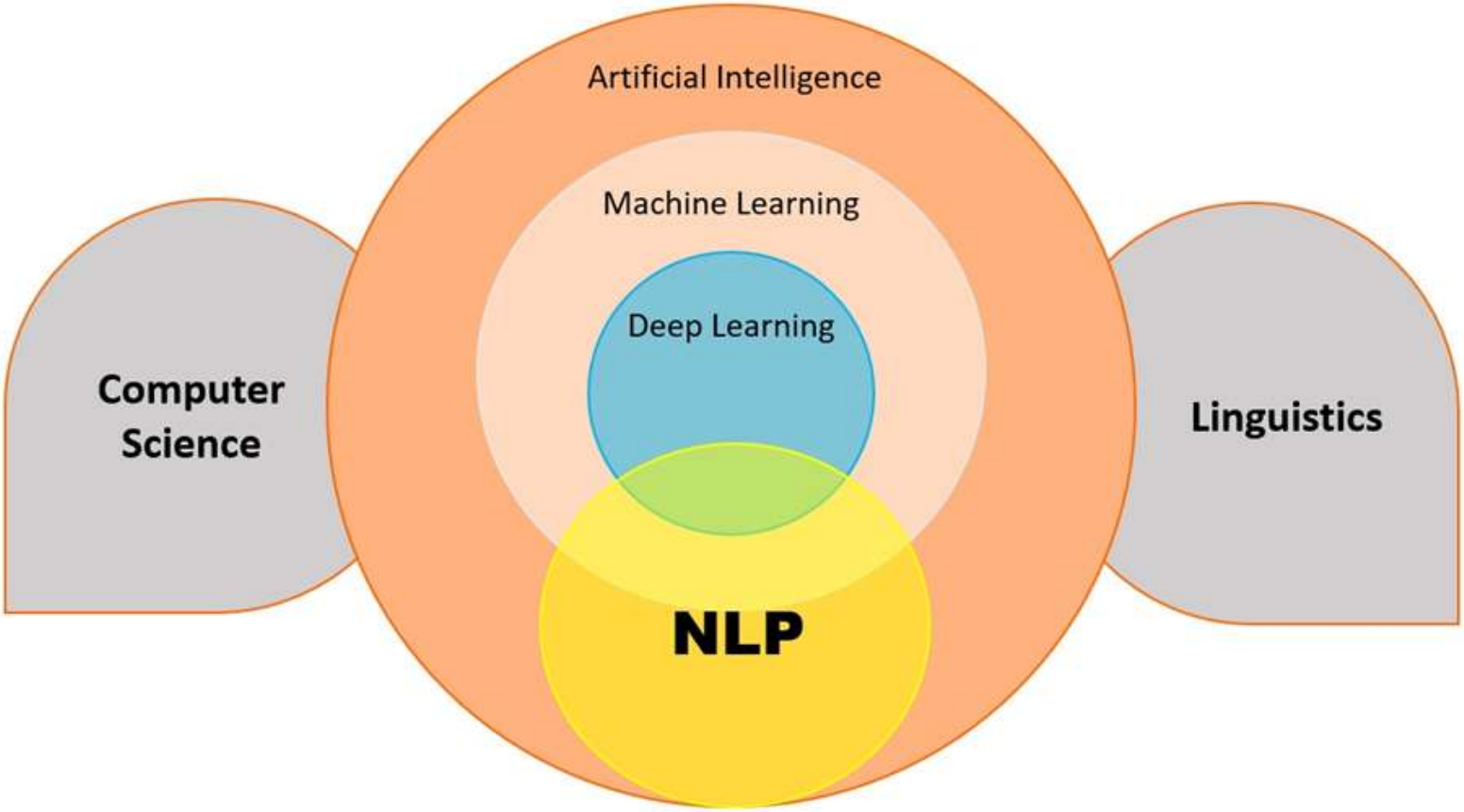
- **Text Analysis Techniques**:
  - Sentiment Analysis: Determining the sentiment or opinion expressed in text (e.g., positive, negative, neutral).
  - Topic Modeling: Discovering latent topics or themes present in a collection of documents.
  - Named Entity Recognition (NER): Identifying and classifying named entities such as persons, organizations, locations, and dates mentioned in text.
  - Text Classification: Categorizing text documents into predefined classes or categories (e.g., spam detection, topic classification).
  - Information Extraction: Extracting structured information from unstructured text (e.g., extracting product names and prices from customer reviews).
- **Tools and Libraries**: Several libraries and tools are available for text analysis in various programming languages. Some popular ones include:
  - Python: NLTK (Natural Language Toolkit), spaCy, scikit-learn, gensim.
  - R: tm (Text Mining), quanteda, tidytext.
  - Java: Apache OpenNLP, Stanford NLP.
  - Commercial platforms: IBM Watson Natural Language Understanding, Google Cloud Natural Language API, Microsoft Azure Text Analytics.

Applications of Text Analysis:

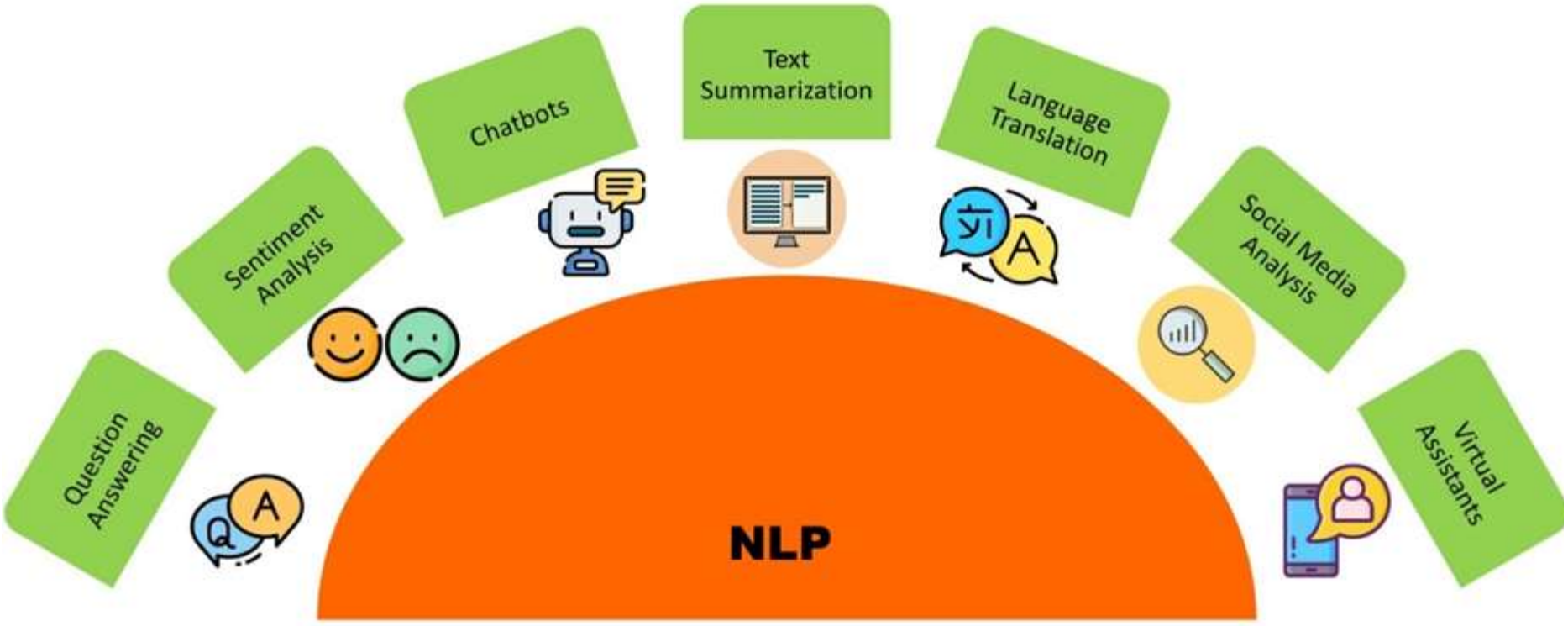
- **Business Intelligence**: Analyzing customer feedback, social media posts, and product reviews to gain insights into customer preferences, sentiment, and market trends.
- **Information Retrieval**: Building search engines that can understand user queries and retrieve relevant documents from a large corpus of text.
- **Healthcare**: Analyzing electronic health records, clinical notes, and biomedical literature for disease diagnosis, drug discovery, and personalized medicine.
- **Social Media Monitoring**: Tracking public opinion, trends, and events by analyzing social media conversations and posts.
- **Legal and Regulatory Compliance**: Analyzing legal documents, contracts, and regulatory texts to extract relevant information and ensure compliance.

NLP-Natural Language Processing

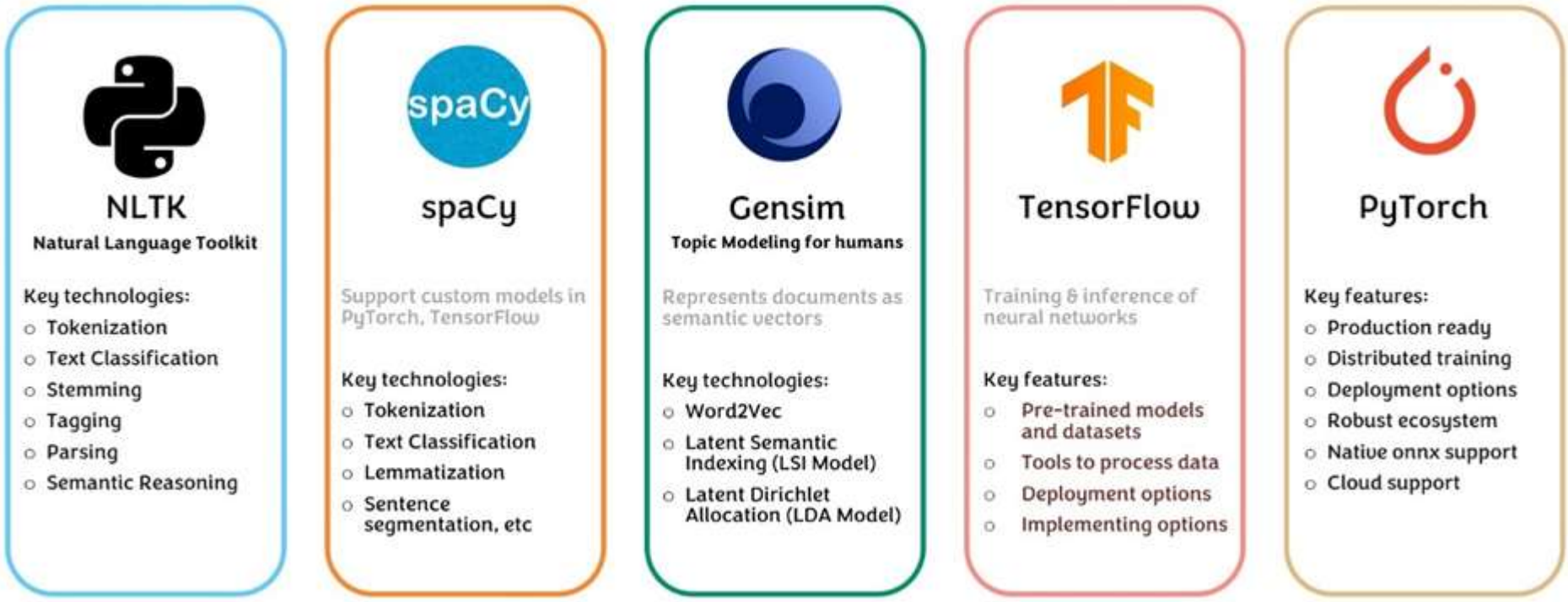
- NLP first appeared as machine translation in the 1950s as decoding messages during World War II for Russian into English
- Prior 1980s, the main driving force behind NLP was a convoluted system of manual set of instructions



Most widely used real-word applications of NLP



Most popular Python libraries for Text Processing



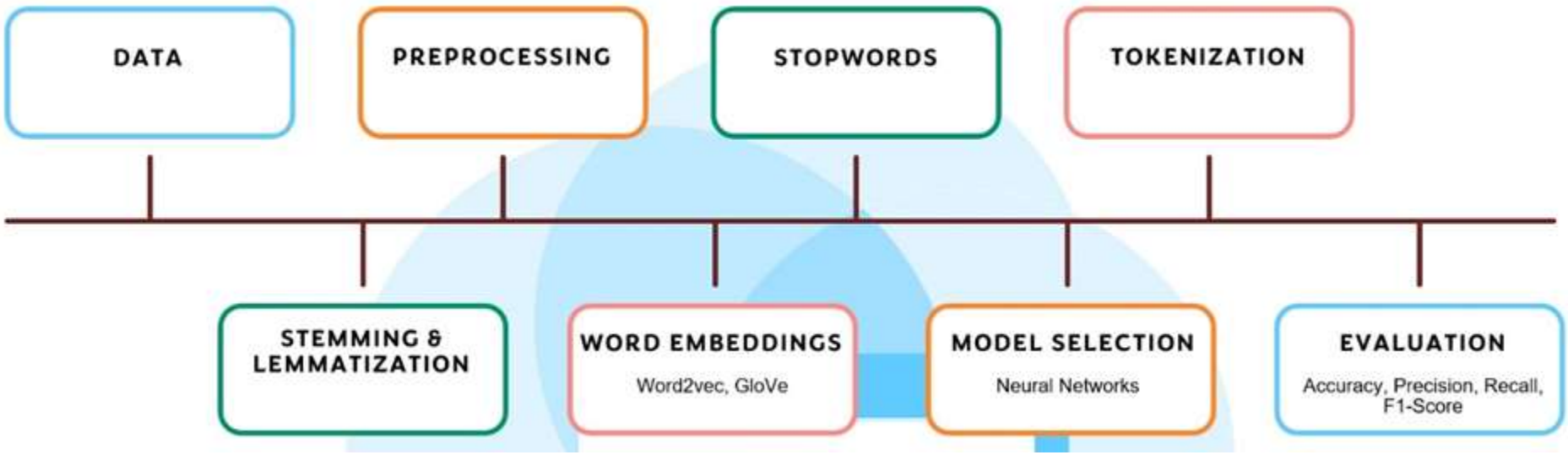
Named Entity Recognition (NER)

- The ISIS has claimed responsibility for a suicide bomb blast in the Tunisian capital earlier this week, the militant group's Amaq news agency said on Thursday. A militant wearing an explosive belt blew himself up in Tunis.

NLP Projects Ideas - Others

- Text-to-Speech (TTS) and Speech-to-Text (STT)
- Search Autocorrect and Autocomplete
- Language Translator
- Hiring and Recruitment
- Targeted Advertising
- Survey Analysis
- Social Media Monitoring
- Emotion Detection
- Inspiring Quote Generator
- Email Filtering

Building blocks of an NLP application



Text Preprocessing

- **Tokenization**: Splitting text into individual words, phrases, symbols, or other meaningful elements called tokens. This can be done at the word level or sentence level.
- **Lowercasing**: Converting all characters in the text to lowercase to ensure uniformity, as "Word" and "word" should be treated the same.
- **Removing Punctuation**: Eliminating punctuation marks such as periods, commas, and exclamation points to focus on the actual words.
- **Removing Stop Words**: Removing common words that do not add significant meaning to the text, such as "and," "the," "is," etc., to reduce noise in the data.
- **Stemming**: Reducing words to their base or root form. For example, "running," "runner," and "ran" are all stemmed to "run.“
- **Lemmatization**: Similar to stemming, but more sophisticated, lemmatization reduces words to their base or dictionary form, considering the context. For example, "better" is lemmatized to "good.

Text Preprocessing(Cont..)

- **Removing Special Characters**: Eliminating special characters, symbols, or numbers that are not relevant to the analysis.
- **Removing Whitespace**: Trimming leading, trailing, and excessive whitespace within the text to maintain consistency.
- **Normalizing Accents**: Converting accented characters to their unaccented counterparts, e.g., "café" to "cafe.“
- **Text Correction**: Correcting spelling and grammatical errors to improve the quality of the text.
- **Removing HTML Tags**: Stripping HTML tags from web-scraped text data to retain only the meaningful content.
- **Removing or Replacing URLs and Email Addresses**: Eliminating or substituting URLs and email addresses with placeholders to avoid irrelevant data.
- **Handling Emojis and Emoticons**: Converting emojis and emoticons to text descriptions if they carry significant meaning for the analysis

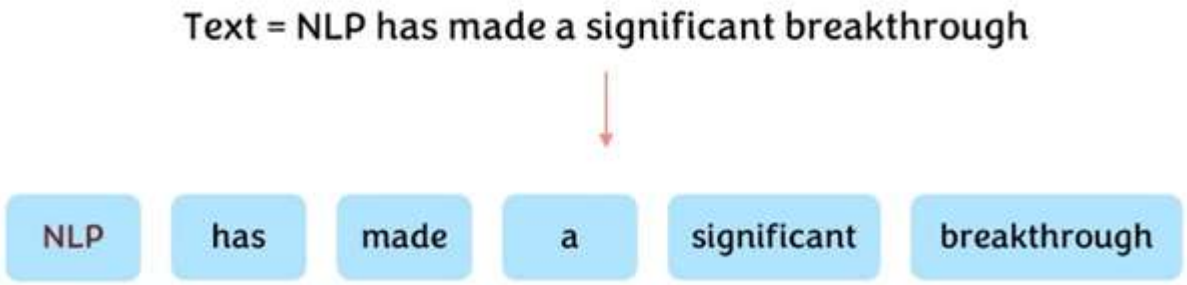
Text Preprocessing(Cont..)

- **Expanding Contractions**: Converting contractions into their full forms, e.g., "don't" to "do not."Text Segmentation: Splitting text into meaningful segments or chunks, such as sentences or paragraphs, for more granular analysis
- **Text Segmentation**: Splitting text into meaningful segments or chunks, such as sentences or paragraphs, for more granular analysis.
- **Noise Removal**: Filtering out irrelevant data such as boilerplate text, headers, footers, and advertisements from the text.
- **Feature Extraction**: Converting text into numerical features for machine learning models. Common methods include Bag of Words (BoW), Term Frequency-Inverse Document Frequency (TF-IDF), and word embeddings (e.g., Word2Vec, GloVe, BERT).Handling Emojis and Emoticons: Converting emojis and emoticons to text descriptions if they carry significant meaning for the analysis

Preprocessing Example-1

- Proper nouns and names (e.g., "John," "Jane," "New York")
- Punctuation (e.g., commas, exclamation points)
- Emojis (e.g., 😊, 🧀)
- URLs (e.g., "www.example.com")
- Dates (e.g., "10/12/2023")
- Email addresses (e.g., "john.doe@example.com")
- Contractions (e.g., "you're," "don't")
- Common stop words (e.g., "the," "it's")
- Mixed case text (e.g., "I hope you're doing well!")
- Special characters (e.g., "😊," "🧀")
- Whitespace (leading, trailing, and within the text)
- Numbers (e.g., "10/12/2023," "1")
- Accented characters (e.g., "cafe's")
- Multiple sentences and phrases that can be tokenized
- Text suitable for stemming and lemmatization (e.g., "visiting," "meeting")
- A mention of a date for segmentation
- Examples of contractions for expansion (e.g., "you're," "I'll")

Text Data Processing - Tokenization

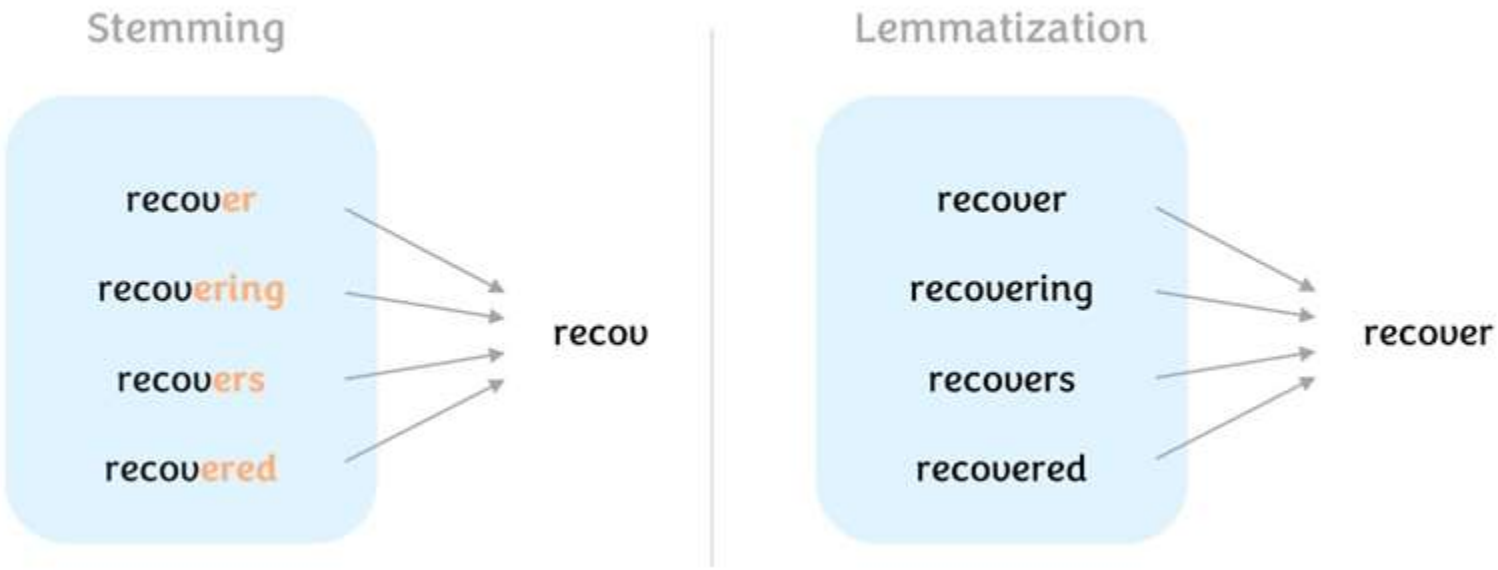


1. Tokenization

2. Stop Words Removal

Text Data Processing Cont..

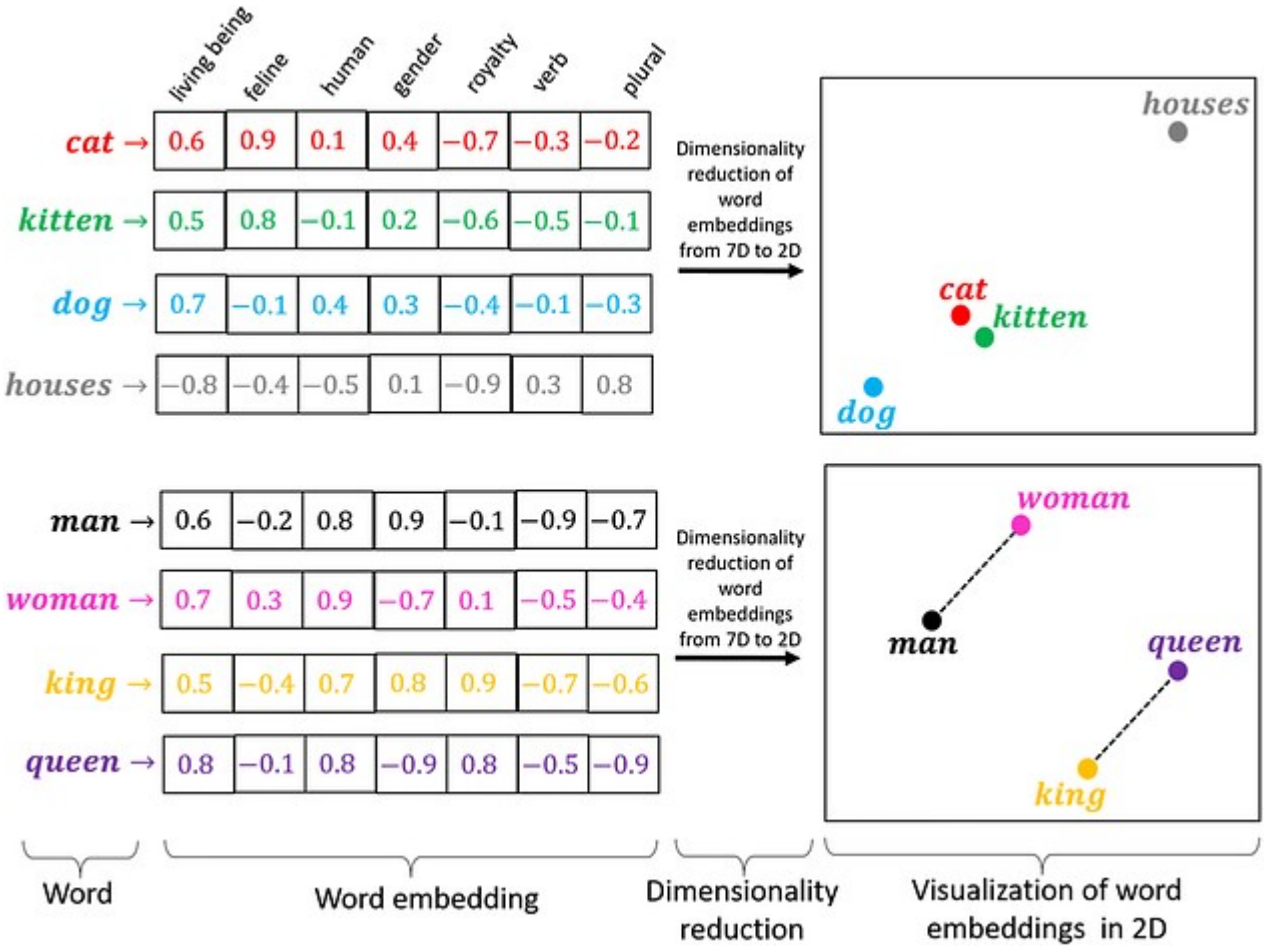
3. Stemming & Lemmatization



Feature Extraction

- Converting text into numerical features for machine learning models. Common methods include
  - Bag of Words (BoW)
  - Term Frequency-Inverse Document Frequency (TF-IDF)
  - word embeddings (e.g., Word2Vec, GloVe, BERT).

Word Embeding



Vectorization or Word Embedding in NLP

- Word embedding methodologies
  - TF-IDF (Term Frequency Inverse Document Frequency)
  - BOW (Bag-of-Words), Count Vectorization, N-grams Vectorization, Word2Vec, GloVe.
- Word2Vec: The process of ‘word embeddings’ involves turning each word into a numerical representation of the word (a vector).
  - Each word is converted into a single vector, which is then trained in a manner resembling a neural network.
  - Based on a word's usage in the text, Word2Vec can infer a word's meaning with a high degree of accuracy.
- GloVe : represents a global vector, which is an unsupervised learning technique that generates word vector representations.
  - The advantage of GloVe is that it integrates global statistics to build word vectors, while Word2Vec depends on local statistics (local context knowledge about words) for generation of word vectors.
  - It is based on word-context matrix factorization algorithms. In GloVe, a co-occurrence matrix is used to determine the semantic relationship between words.

- Sentiment Analysis: Perform sentiment analysis on social media data.
- Text Summarization: Implement extractive and abstractive summarization techniques.
- Information Retrieval: Develop a basic search engine using TF-IDF or embeddings.
- Text Generation: Build a simple text generation model using transformers.
- Advanced NLP Application: Develop a question-answering system or a named entity linking application.

Supervised Learning Approaches &Heuristic Methods

Supervised learning methods involve training a model to identify important sentences based on labeled training data:

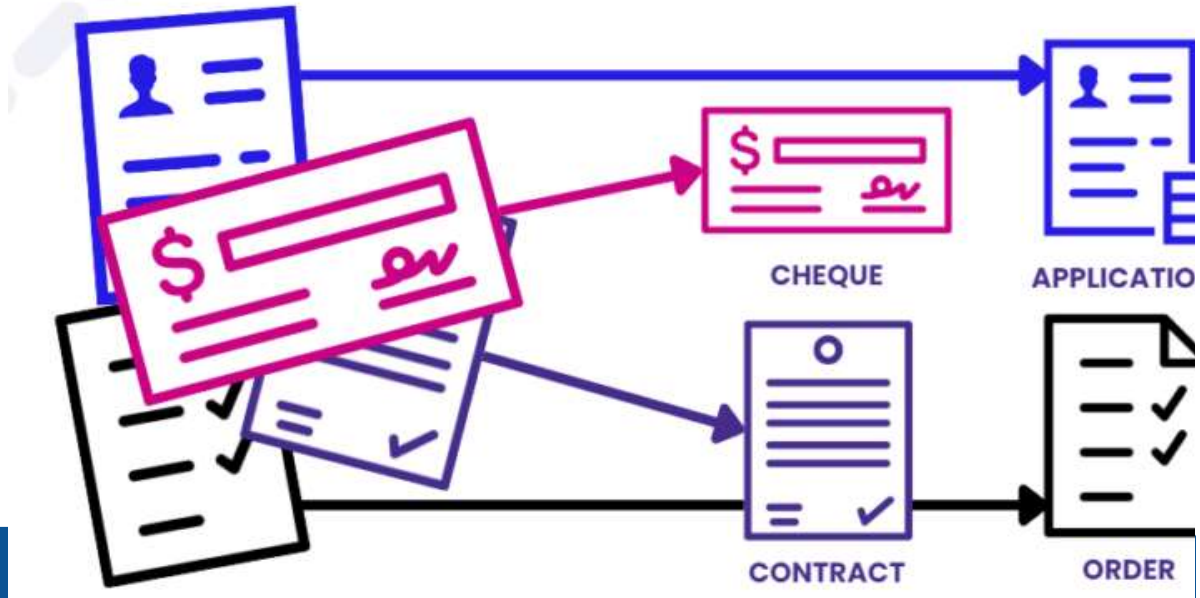
- **Feature Extraction**: Extract features such as sentence position, length, presence of keywords, etc.
- **Model Training**: Train a classifier (e.g., logistic regression, SVM, neural networks) to predict sentence importance.
- **Sentence Scoring**: Score sentences using the trained model.
- **Sentence Selection**: Select the top-scoring sentences for the summary.

Simple heuristic methods rely on predefined rules to select important sentences:

- **Lead-Based Summarization**: Select the first few sentences of the document as the summary (effective for news articles).
- **Title and Heading Matching**: Select sentences that contain keywords from the title or headings.

Document Classification & Clustering

- **Document classification (document categorization)** refers to recognizing a document category based on its content, visual appearance, and other factors.
- **You can classify documents into folders based on labels you create, for example:**
  - **Level of confidentiality:** public, confidential, top secret.
  - **Type of document:** invoice, corrected invoice, receipt.
  - **Language:** English, French, Spanish.



Document Processing by Industries

ACCOUNTS PAYABLE

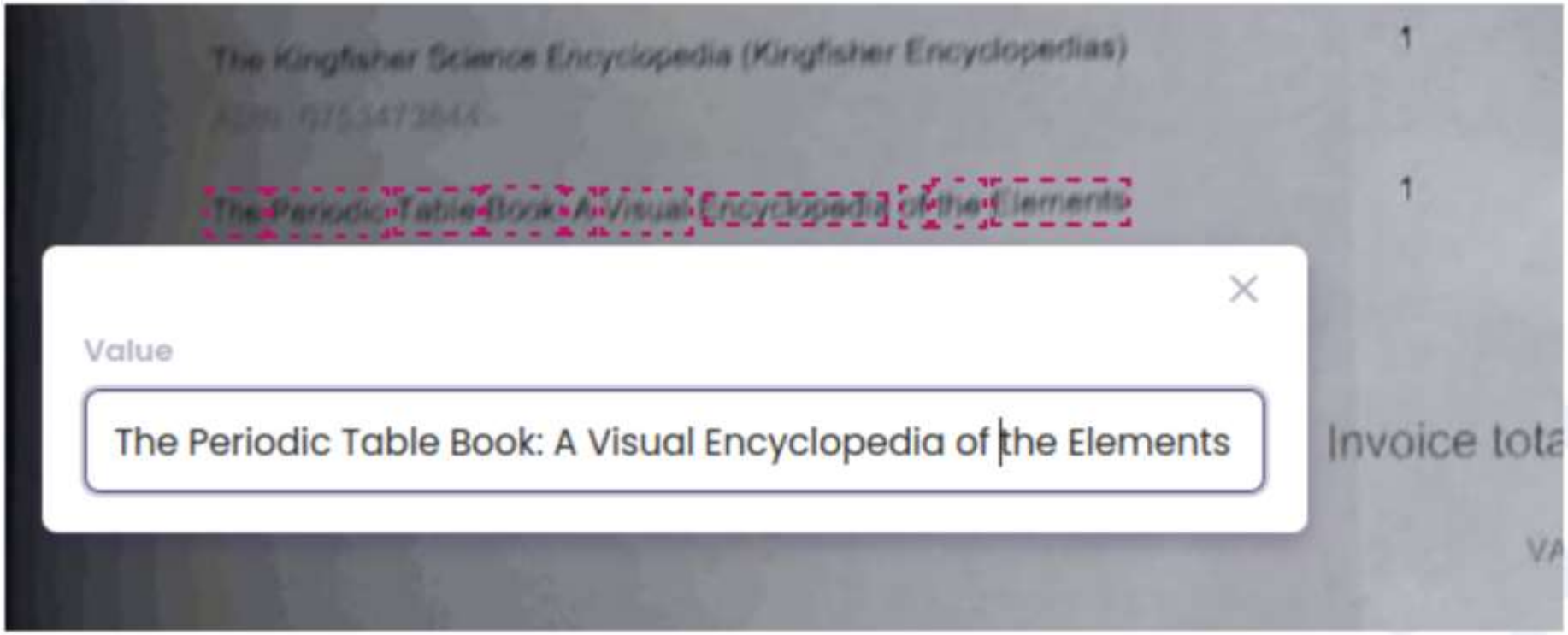
- invoices
- corrected invoices
- receipts
- purchase orders

HEALTHCARE

- health records
- ID scans
- medical certificates

LEGAL

- agreements
- contracts
- notarial deeds
- forms


The Kingfisher Science Encyclopedia (Kingfisher Encyclopedias)
1
The Periodic Table Book: A Visual Encyclopedia of the Elements
1
Value
The Periodic Table Book: A Visual Encyclopedia of the Elements
Invoice tota

prime video

INVOICE

Attn: Olivia James
Amazon Digital UK Limited
1 Principal Place, Worship Street
London, EC2A 2FA
United Kingdom
VAT #: GB882596668

Order date: 27.12.2020
Date of Supply: 27.12.2020
Order #: D01-8134136-6607068

Invoice date / Delivery date: 27.12.2020
Invoice #: AIVGB-INV-GB-2020-97911707

Issued to:
Mathew James
Bournville Garden Centre
Science Park, Milton Road
CAMBRIDGE, CAMBS, CB4 0FY
GB

Qty	Description	Unit price (excl. VAT)	VAT rate	Unit price (incl. VAT)	Total price (incl. VAT)
1	The Chronicles of Narnia - The Lion, the Witch and the Wardrobe	£4.16	20%	£4.99	£4.99
				TOTAL:	£4.99

VAT details:

Subtotal (excl. VAT)	VAT rate	VAT amount
£4.16	20%	£0.83
TOTAL:		£0.83

Correction invoice: **CI-CX-20-00001**
Reference number: **202000001**

Release agency
2020 Melton Road
11716 New York
United States

Yummy & Tasty
1410 Flinderation Road
60605 Chicago
United States, Illinois

Correction
Subject
Advertising materials
Leaflets
Billboards
Total

How Document Classification Works