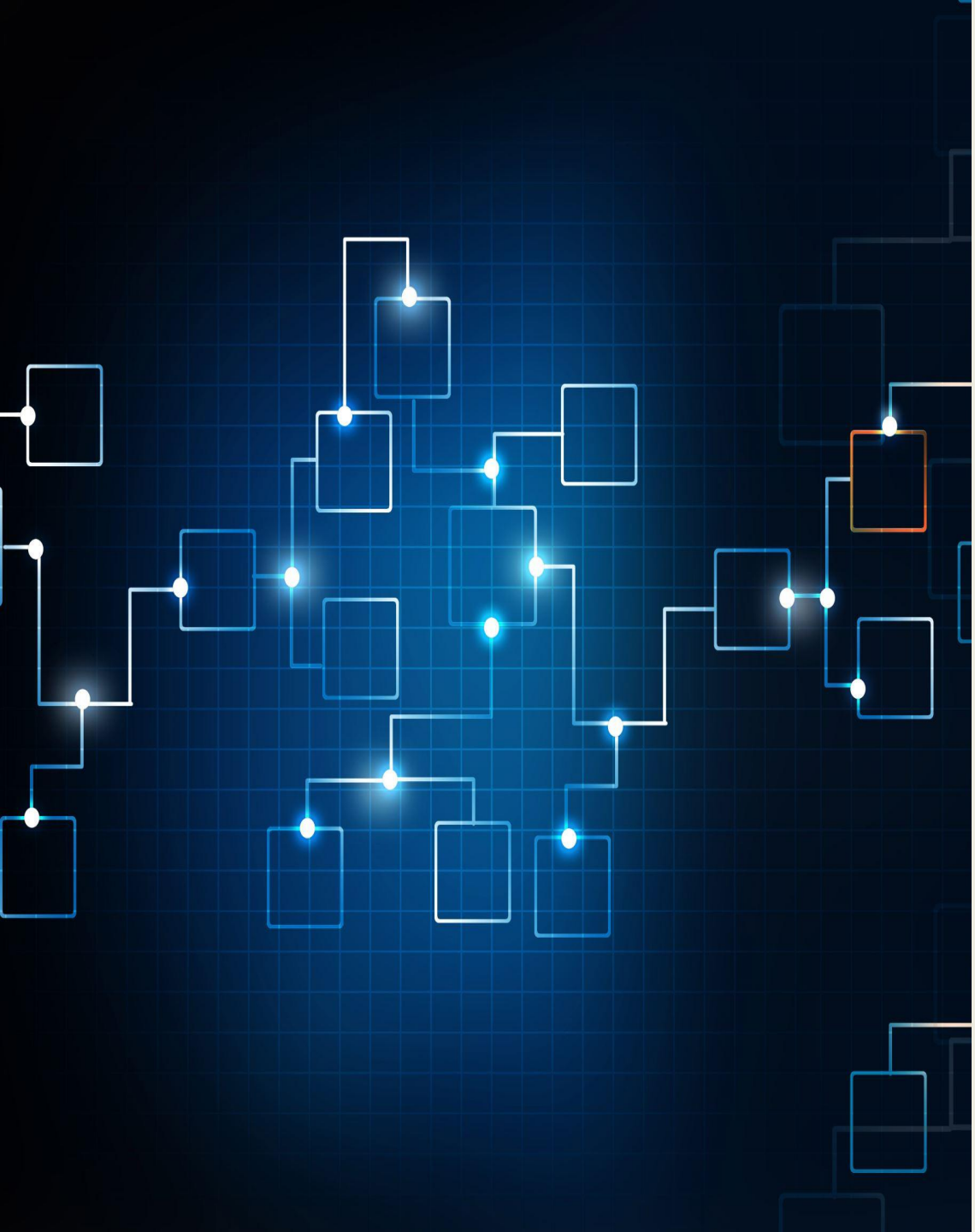


HIT391

MACHINE LEARNING: ADVANCEMENTS AND APPLICATIONS

- Lecturer: Dr. Yan Zhang
- Email: yan.zhang@cdu.edu.au

CHARLES DARWIN UNIVERSITY
AUSTRALIA

Week 10:
Natural Language Processing (NLP)

- **Learning Outcomes**
  - NLP Applications

    - Sentiment analysis, Topic modeling, Text generation, Information retreival
  - Data Pre-processing in NLP

    - Stemming, Tokenization, BoW, TF-IDF
  - Traditional NLP Methods

    - Logistic regression, LDA
  - Deep Learning NLP Methods

    - CNN, RNN, Autoencoders, Seq2Seq

What is Natural Language Processing?

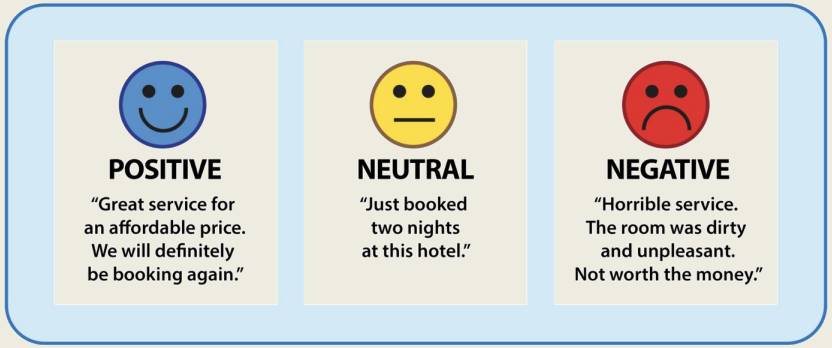
- NLP - Building machines that can manipulate human language.
- Origins and Evolution
  - Evolved from computational linguistics.
  - Computational linguistics: Uses computer science to understand language principles.
  - NLP: An engineering discipline focused on practical applications.
- Subfields of NLP:
  - Natural Language Understanding (NLU): Focuses on semantic analysis and interpreting the intended meaning of text.
  - Natural Language Generation (NLG): Focuses on generating text by a machine.

NLP vs Speech Recognition

- Relation to **Speech Recognition**
  - NLP is separate from but often **used with speech recognition.**
  - Speech recognition: **Parses** spoken language **into text** and vice versa.

What is NLP Used For?

1. Sentiment Analysis - Classifying the emotional intent of text.
   - Process:
     - Input: A piece of text.
     - Output: Probability of sentiment being positive, negative, or neutral.
   - Methods:
     - Hand-Generated Features.
     - Word N-grams.
     - TF-IDF Features.
     - Deep Learning Models: Capture sequential long- and short-term dependencies.



   - Applications:
     - Classifying customer reviews on online platforms.
     - Identifying signs of mental illness in online comments.

What is NLP Used For?

2. Machine translation - automates translation **between different languages**

3. Named entity recognition - extract entities in a piece of text into predefined categories such as **personal names, organizations, locations, and quantities**.

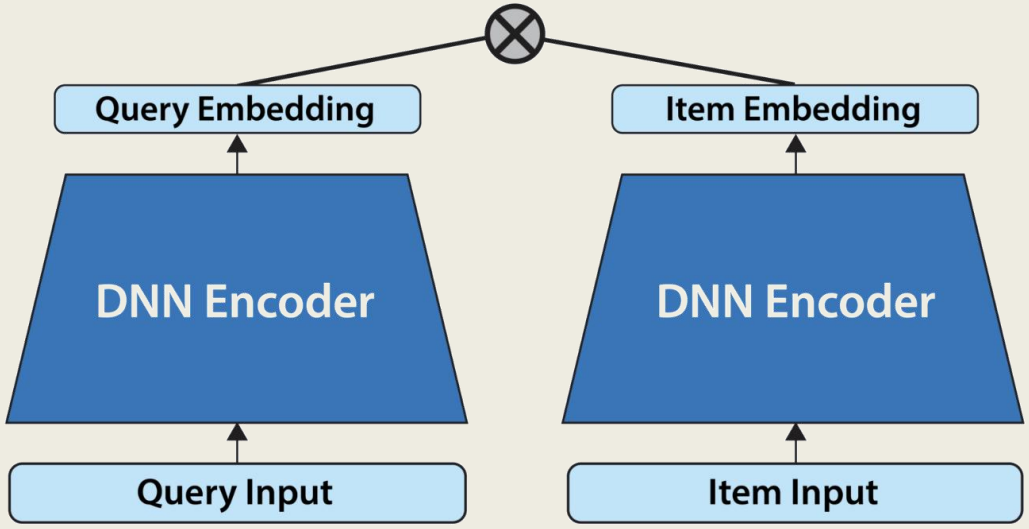
    1. Input: generally text
    2. output: various **named entities** along with their **start and end positions**
    3. Applications: summarizing news articles, combating disinformation

What is NLP Used For?

4. Topic modeling
   - an **unsupervised** text mining task that takes a **corpus** of documents and discovers abstract **topics** within that corpus.
5. Text generation
   - natural language generation (NLG), produces text that's similar to **human-written** text.
   - Text generation has been performed using
     - Markov processes, LSTMs, BERT, GPT-2, LaMDA, etc.
   - Applications
     - Autocomplete
     - Chatbots
6. Information retrieval

6. Information Retrieval

- Information Retrieval - finds **a document set** that are most relevant to a query
  - The goal - retreive **the most relevant set** to the query.
  - Two steps
    1. Indexing: a vector space model through **Two-Tower Networks**
    2. Matching: using similarity or distance scores



Data Preprocessing

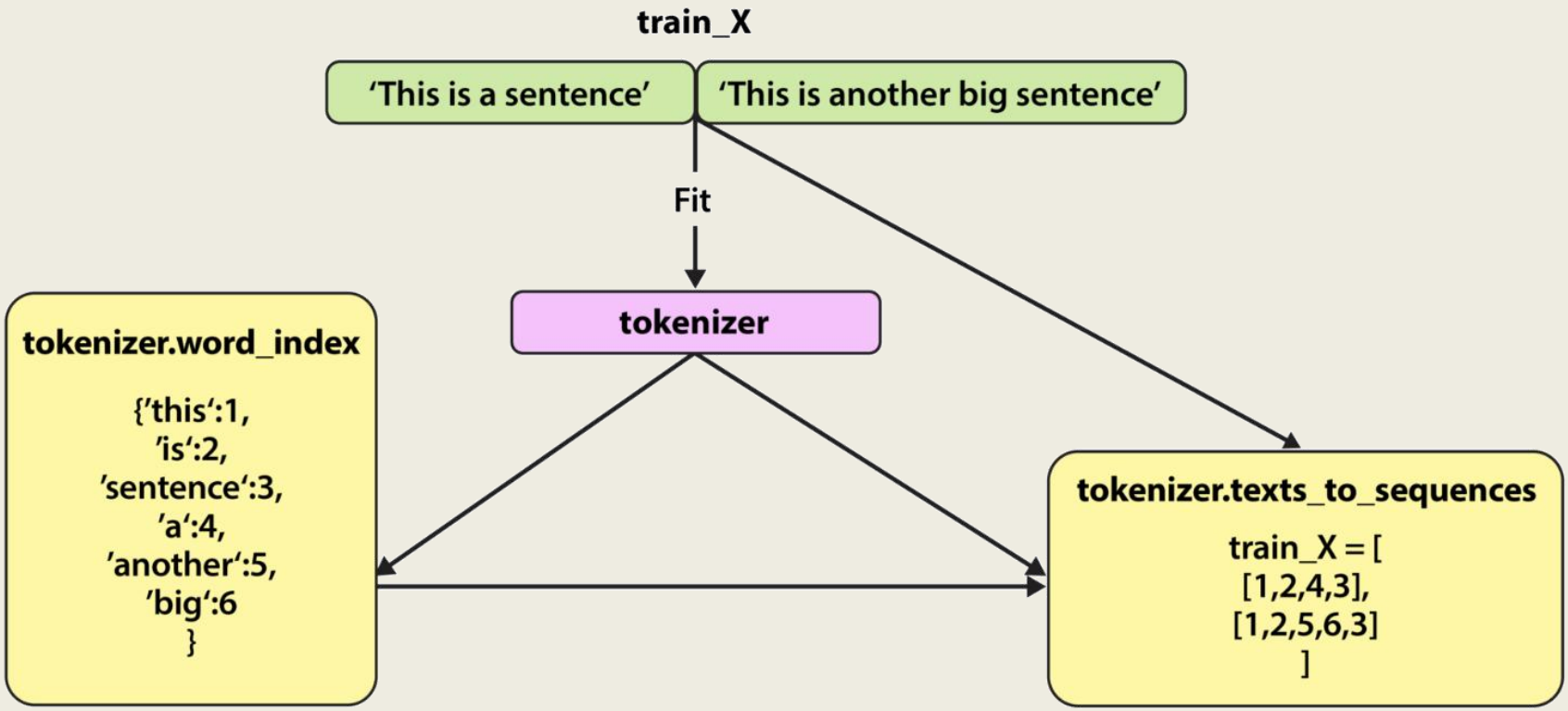
- Before a model processes text **for a specific task**, the text often needs to be preprocessed to improve model performance
  - Stemming and lemmatization
    - converting words to their base forms using heuristic rules:

      e.g., “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**

      e.g., a sentence is marked by **a period**
  - Stop word removal
    - remove the most commonly occurring words that don’t add much information to the text.
    - e.g., “the,” “a,” “an,” and so on.
  - Tokenization
  - Feature extraction

Tokenization

- Split text into **individual words** and **word fragments**

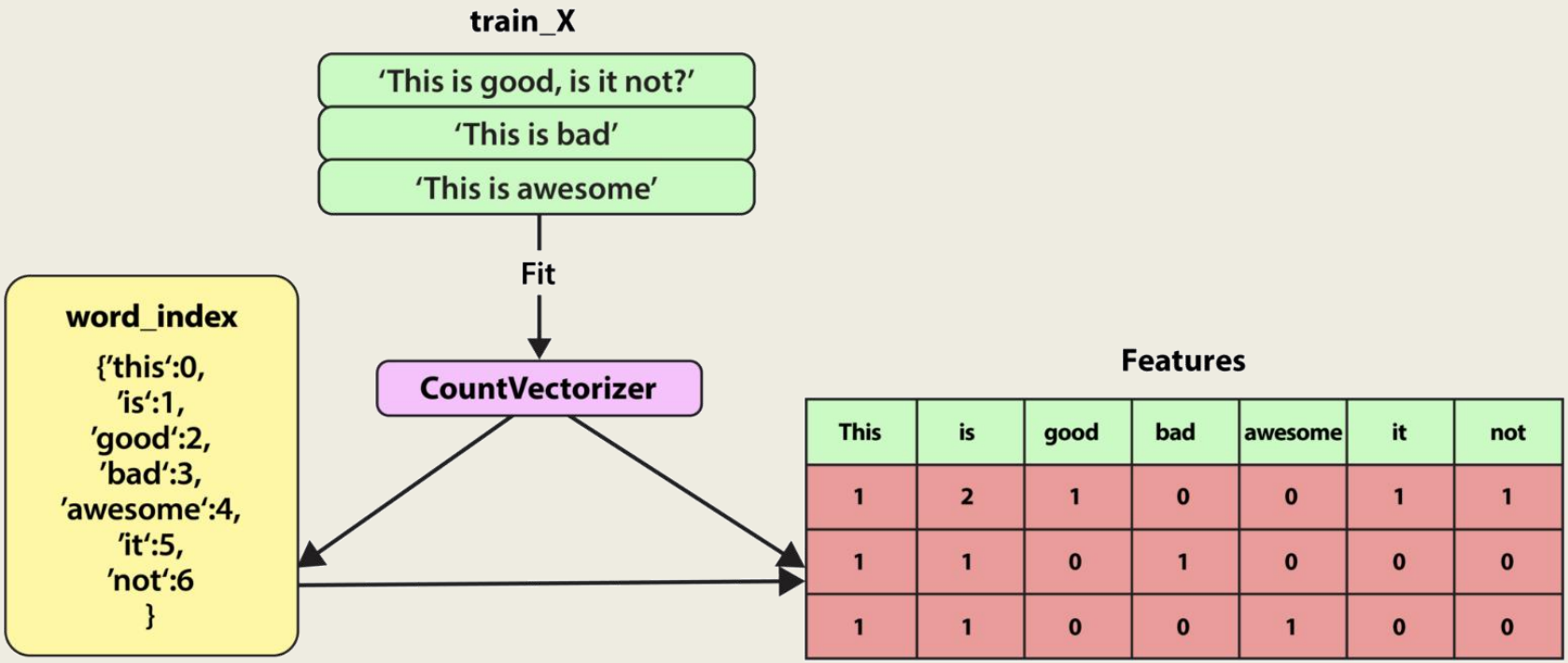


Feature Extraction in NLP

- Bag-of-Words (BoW)
  - Counts the frequency of each word or n-gram in a document.
  - Creates a numerical representation of a dataset based on word occurrences.
- TF-IDF (Term Frequency-Inverse Document Frequency)
  - Adjusts the frequency of words by how common or rare they are across all documents in the corpus
  - Highlights words that are significant in a document but not too common across the corpus

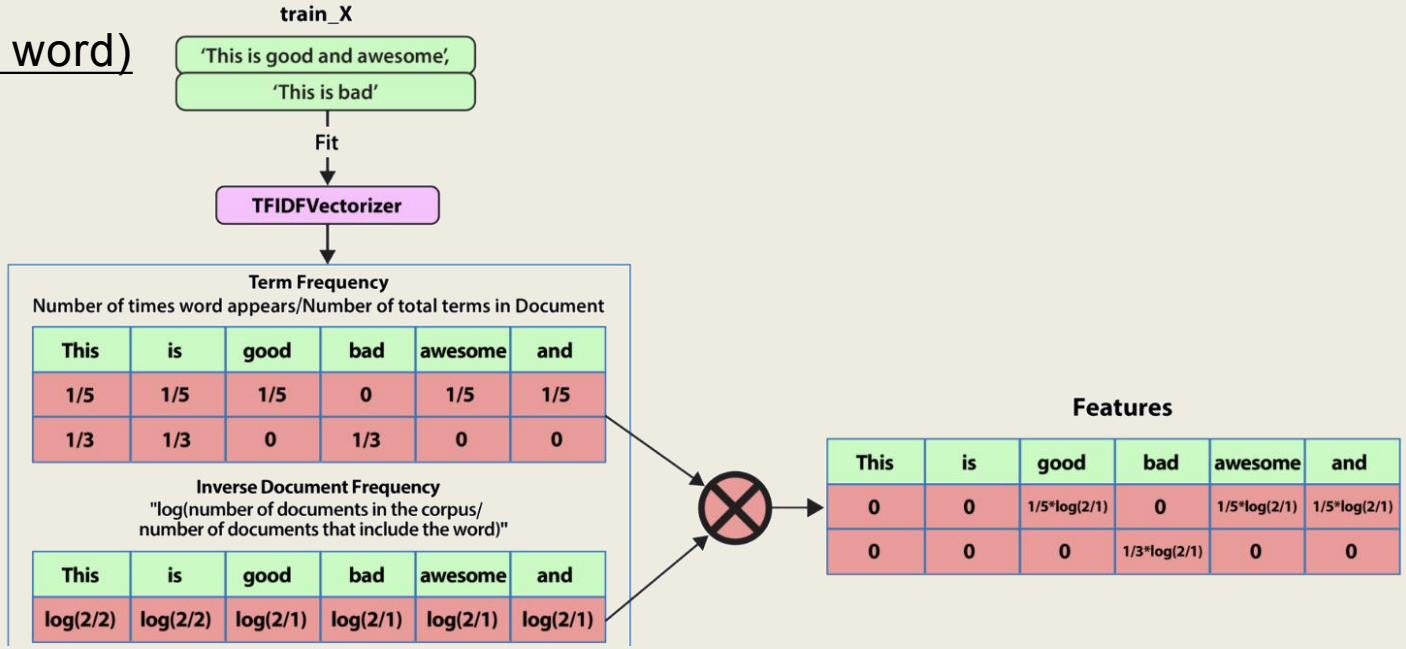
Bag–of–Words (BoW)

- BoW: counts the number of times each word or n-gram (combination of n words) appears in a document.



TF–IDF

- TF-IDF: weight each word by its importance - TF * IDF.
  - Term Frequency(TF): How important is the word in the document?
    - Number of occurrences of that word in document/Number of words in document
  - Inverse Document Frequency(IDF): How important is the term in the whole corpus?
    - log(number of documents in the corpus / number of documents that include the word)

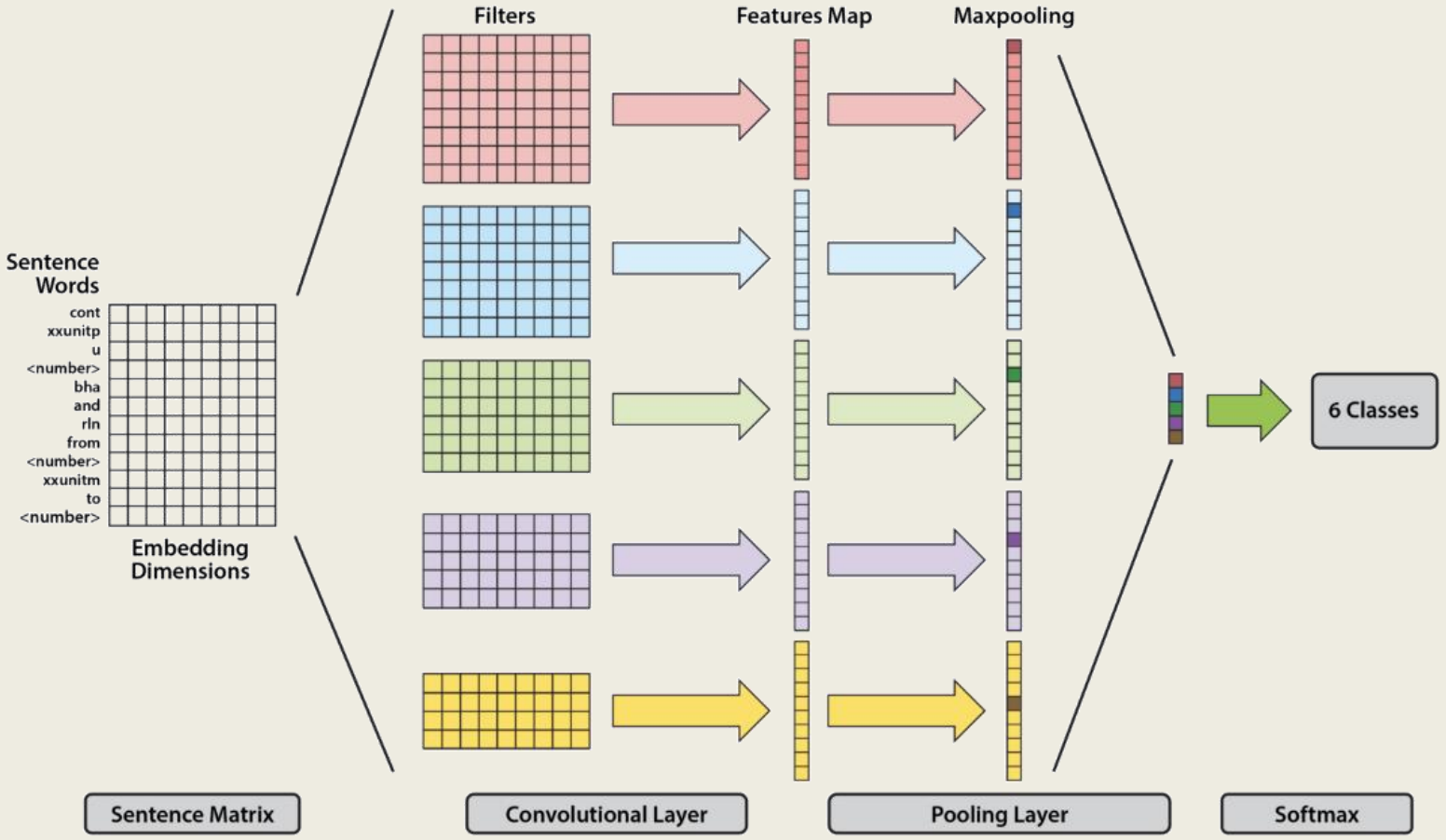


Traditional NLP Algorithms

- Logistic regression (supervised classification)
  - predict the probability that an event will occur based on some input, e.g., sentiment analysis, spam detection, and toxicity classification
- Naive Bayes (supervised)
  - using Bayes formula for spam detection or finding bugs in software code
- Latent Dirichlet Allocation (LDA)
  - is used for topic modeling, we can describe any topic using only a small set of words from the corpus
- Hidden Markov models
  - decide the next state of a system **based on the current state**
    - **suggest the next word** based on the previous word

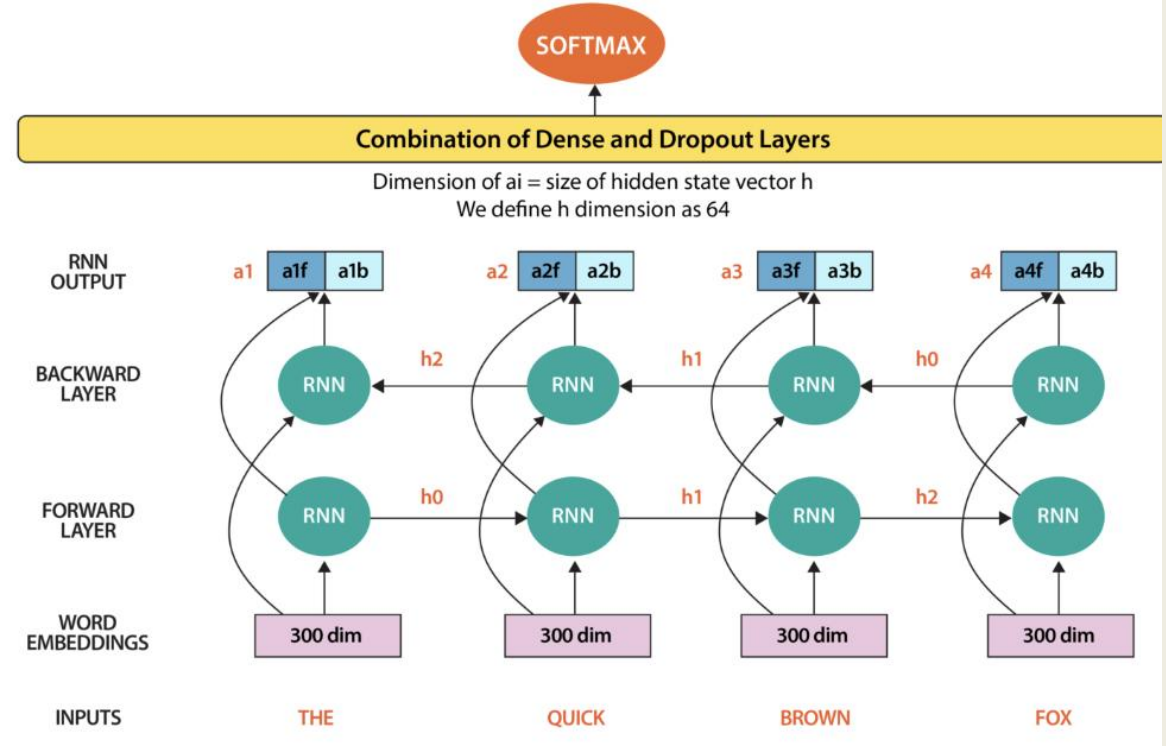
Deep Learning for NLP

- Convolutional Neural Network (CNN)
  - Input consists of sentences or documents represented as a matrix of words (treating each document as if it were an image).



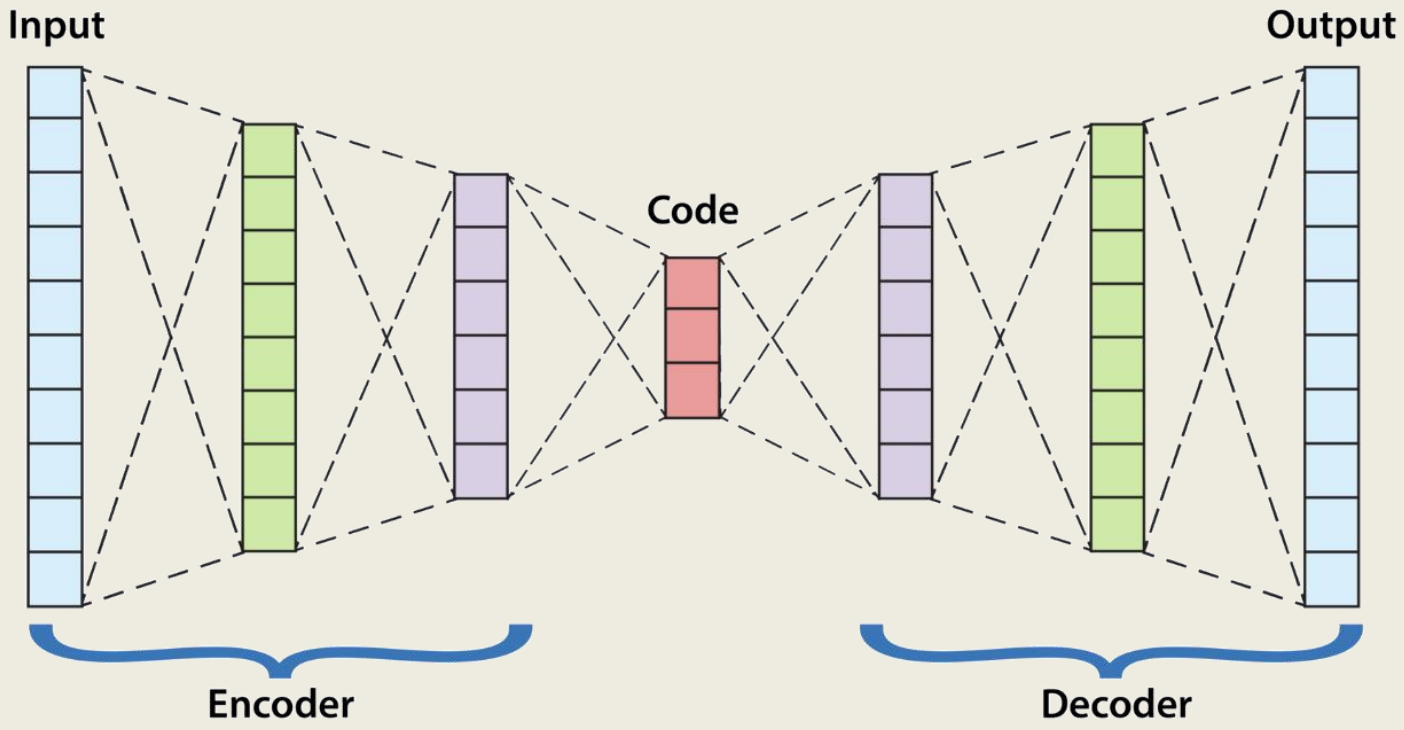
RNN

- RNNs remember previous information using hidden states and connect it to the current task, e.g., RNN-designed models, like
  - Gated Recurrent Unit (GRU)
  - Long short-term memory (LSTM)



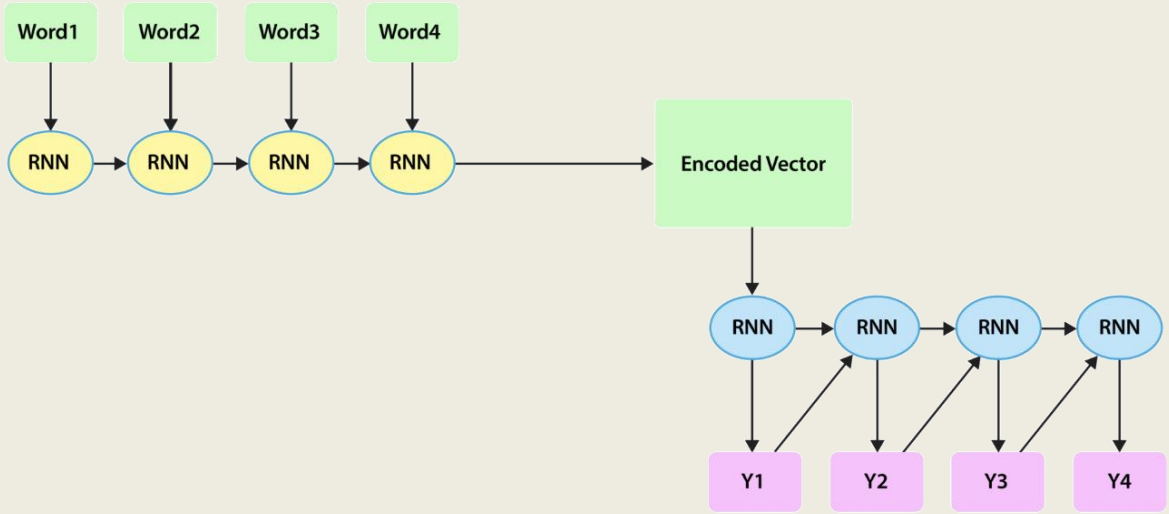
Autoencoders

- Autoencoder that approximates a mapping from X to X, i.e., input=output, for **translation**, **generation** tasks.
  - Encodr: **compress the input features** into a **lower-dimensional representation (code / latent vector)**
  - Decoder: learn to **reconstruct the input**.



Seq2Seq

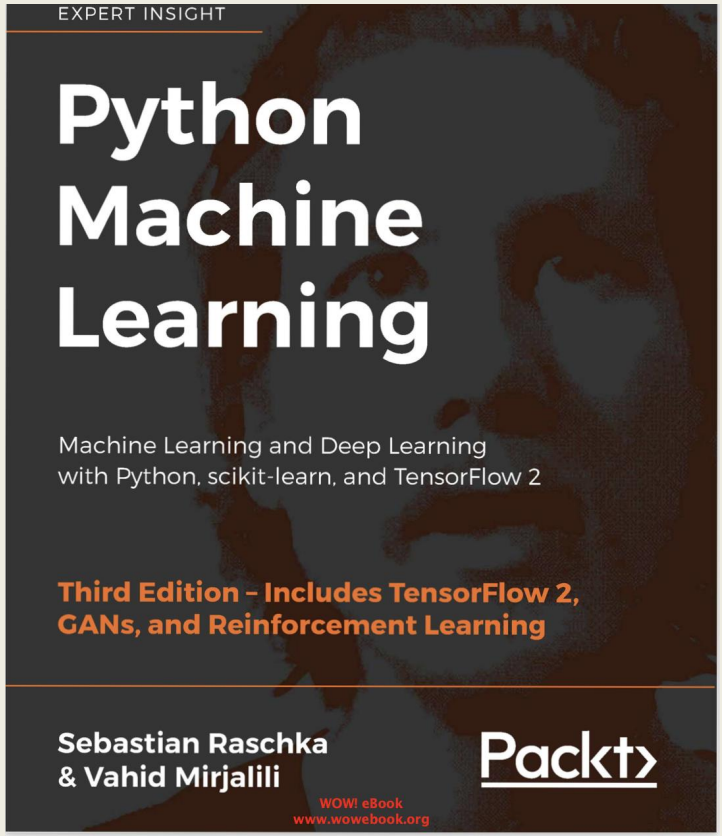
- Encoder-decoder sequence-to-sequence - adaptation to autoencoders specialized for **translation**, **summarization**, and **similar tasks**.
- Encoder: encapsulates/compresses the information in a text into an **encoded vector**
- Decoder: generate a **different desired output**, like a **translation** or **summary.**



NLP Libraries

- Natural Language Toolkit (**NLTK**) - It provides easy-to-use interfaces to corpora and lexical resources such as WordNet
- **spaCy** - supports more than 66 languages, provides pre-trained word vectors and implements many popular models like **BERT**.
- Deep Learning libraries, **TensorFlow** and **PyTorch** - developing NLP models

References



- https://www.deeplearning.ai/resources/natural-language-processing/
- Sebastian Raschka, etl., 'Python machine learning', Third Edition