SWINBURNE
UNIVERSITY OF
TECHNOLOGY

COS30082
Applied Machine Learning

Lecture 10
Natural Language Processing

What is Natural Language Processing (NLP)

SWINBURNE UNIVERSITY OF TECHNOLOGY SARAWAK CAMPUS

- Natural Language Processing (NLP) is a branch of artificial intelligence that focuses on the interaction between computers and humans through natural language.
- The goal of NLP is to enable machines to understand, interpret, and respond to human language. For example,
    - **Understand**: NLP systems parse customer feedback to identify key themes and sentiments.
    - **Interpret**: Determine the intent behind queries in virtual assistants.
    - **Respond**: Chatbots generate relevant responses to user inquiries in real-time.

Applications of NLP

- **Text Classification**: Identifies themes like spam or sentiment within texts.
- **Machine Translation**: Converts text or speech across languages, e.g., Google Translate.
- **Speech Recognition**: Transforms spoken words into text, used in devices like Siri or Alexa.
- **Chatbots and Virtual Assistants**: Facilitate interactive dialogues, mimicking human conversation.
- **Information Extraction**: Detects specific data like names or relationships from texts.
- **Text Summarization**: Condenses long documents into essential summaries.
- .....

Challenges in NLP

- **Ambiguity**: Language is inherently ambiguous. For instance,
  - The word "bank" can mean a financial institution or the side of a river.
  - I can see a man with a telescope. (who has the telescope?)
- **Context Understanding**: NLP struggles with context.
  - It is a piece of cake. (a cake or it is easy?)
- **Sarcasm and Irony**: Detecting sarcasm or irony in text is challenging. For example, "Great, another rainy day!" might be sarcastic depending on the speaker's tone.

Challenges in NLP

- **Language Diversity**: Different dialects and slang complicate processing. For example, "soda" vs. "pop" vs. "coke" can all refer to a soft drink, depending on the region.
- **Idioms and Phrases**: Phrases like "raining cats and dogs" (raining very heavily) are difficult to interpret literally for NLP systems.
- **Pronoun Resolution:**

  "Sarah told Emily that she had lost her keys." she refers to whom?
- **Language Variability and Evolution:** Language is constantly evolving, and new words, slang, and usage patterns emerge all the time. "He is salty". "Sick" (very good).

History of NLP

1. **Early Days (1940s-1960s):** Early Models were based on rules of grammar and the formal theory of syntax.
2. **Rise of Machine Learning (1970s-1990s):** there was a shift towards statistical methods for language processing, influenced by the limitations of rule-based systems.
3. **The Internet Era (1990s-2000s):** The explosion of digital text data on the internet provided vast resources for training and improving NLP systems. Significant advances in machine translation during this period, driven by statistical models.

History of NLP

4. Deep Learning Revolution (2010s-present)

   1. **Neural Networks:** The advent of deep learning has significantly transformed NLP, with neural network models achieving state-of-the-art results on many NLP tasks.
   2. **BERT and Transformers:** Introduction of models like BERT (Bidirectional Encoder Representations from Transformers) in 2018 revolutionized NLP, offering vast improvements in understanding context and semantics.

… …

Key Terms/Concepts

Corpus (plural: Corpora) - A corpus refers to a large collection of text data that is used to train language models or perform other types of text analysis.

Bag of Words (BoW) - In the bag of words model, text is represented as a vector of word frequencies, disregarding grammar and word order. Each unique word in the text corresponds to a vector element.

Bag of Words

- The Bag-of-words model is an orderless document representation where only the counts of words matter.
- I love apple and hate orange.

  I also hate banana too.
- I love orange and hate apple.

  I also hate banana too.

The two documents are the same in terms of words and frequency.

Term	frequency
I	2
love	1
apple	1
and	1
hate	2
orange	1
also	1
banana	1
too	1

Key Terms/Concepts

Tokenization - The process of breaking down a text into smaller units, typically words or phrases.

Stop Words - Stop words are common words like “a”, "the" and "is" that are often filtered out in language processing to focus on more meaningful words.

Key Terms/Concepts

Stemming and Lemmatisation

- Both stemming and lemmatization involves reducing the inflectional forms of words to their root forms.
- Inflection forms of words are words that are derived from the root or base form of a word.
- For example, the words jumped, jumping and jumps are inflectional forms of the root word jump.
- **Stemming**: Stemming simplifies words to their base forms by crudely cutting off endings. It is faster but the result may not be a word.
- **Lemmatization**: Lemmatization reduces words to their dictionary form through detailed linguistic analysis, considering context and grammatical rules. It is slower but accurate as the result is a word.

Key Terms/Concepts

Word --- **Porter Stemmer:**
- jumped --- jump
- friends --- friend
- mysteries --- mysteri
- created --- creat
- took --- took

Word --- **Lemmatized word**
- jumped --- jump
- friends --- friend
- mysteries --- mystery
- created --- create
- took --- take

Key Terms/Concepts

Part-of-Speech Tagging (POS Tagging) - The process of marking up a word in a text as corresponding to a particular part of speech, based on both its definition and its context.

For example "She can play the guitar beautifully."

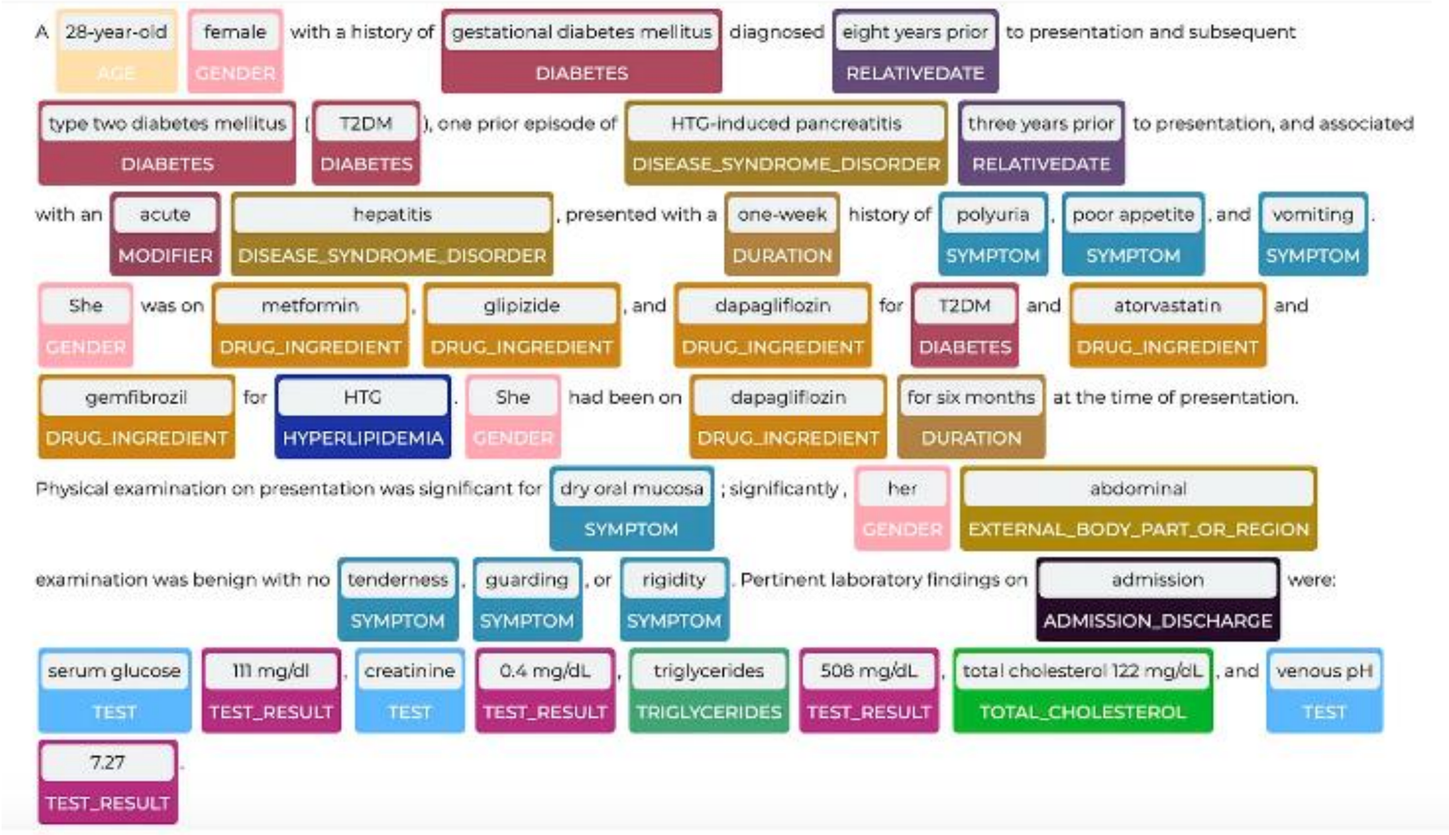
Tagged:

1. "She" (Pronoun)
2. "can" (Modal Verb)
3. "play" (Verb)
4. "the" (Determiner)
5. "guitar" (Noun)
6. "beautifully" (Adverb).

Key Terms/Concepts

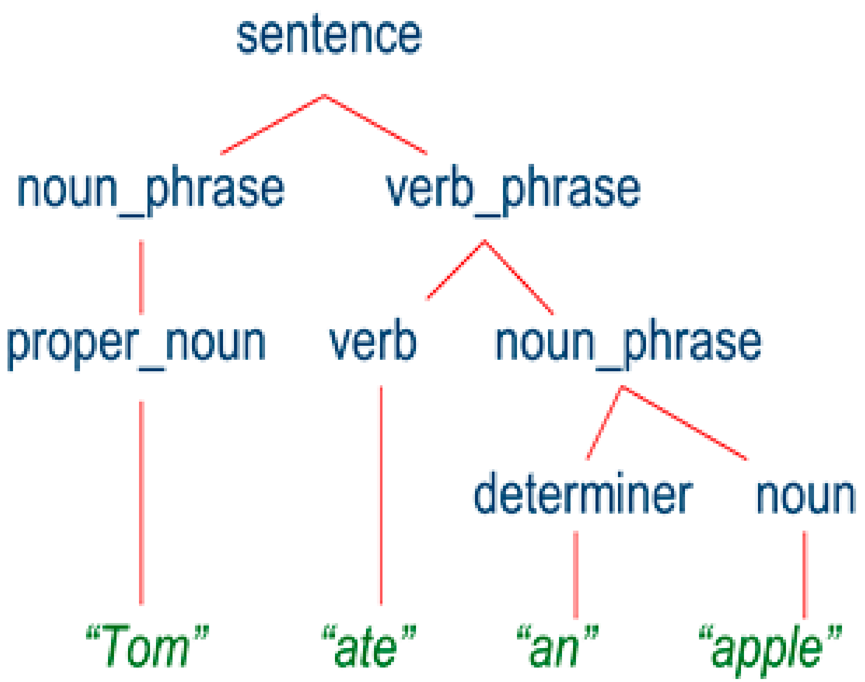
Named Entity Recognition (NER) - A process where an algorithm takes a string of text (sentence or paragraph) and identifies relevant nouns (people, places, and organizations) that are mentioned in that string.

Image from John Snow Lab Medical NER



Key Terms/Concepts

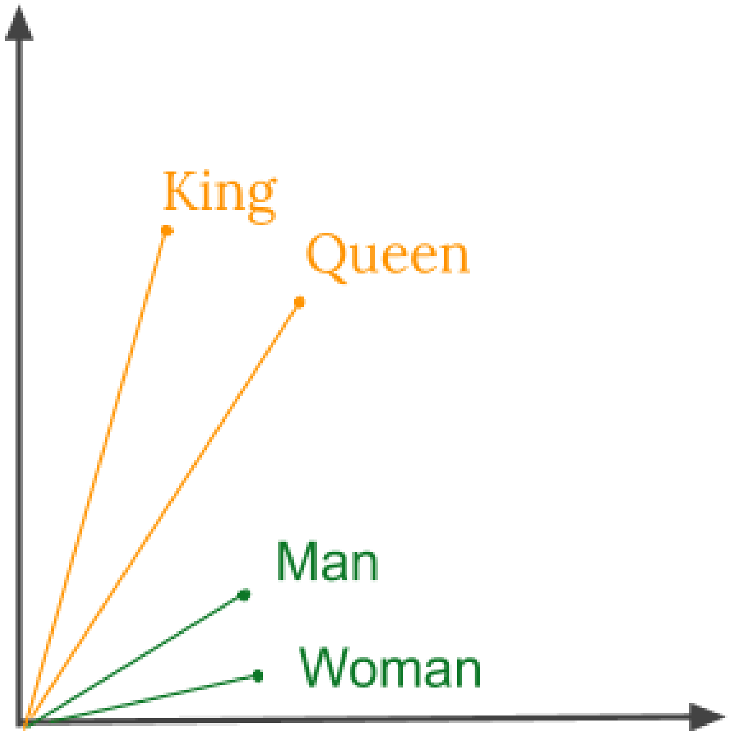
Syntax Tree - A tree representation of syntactic structure of sentences or strings.



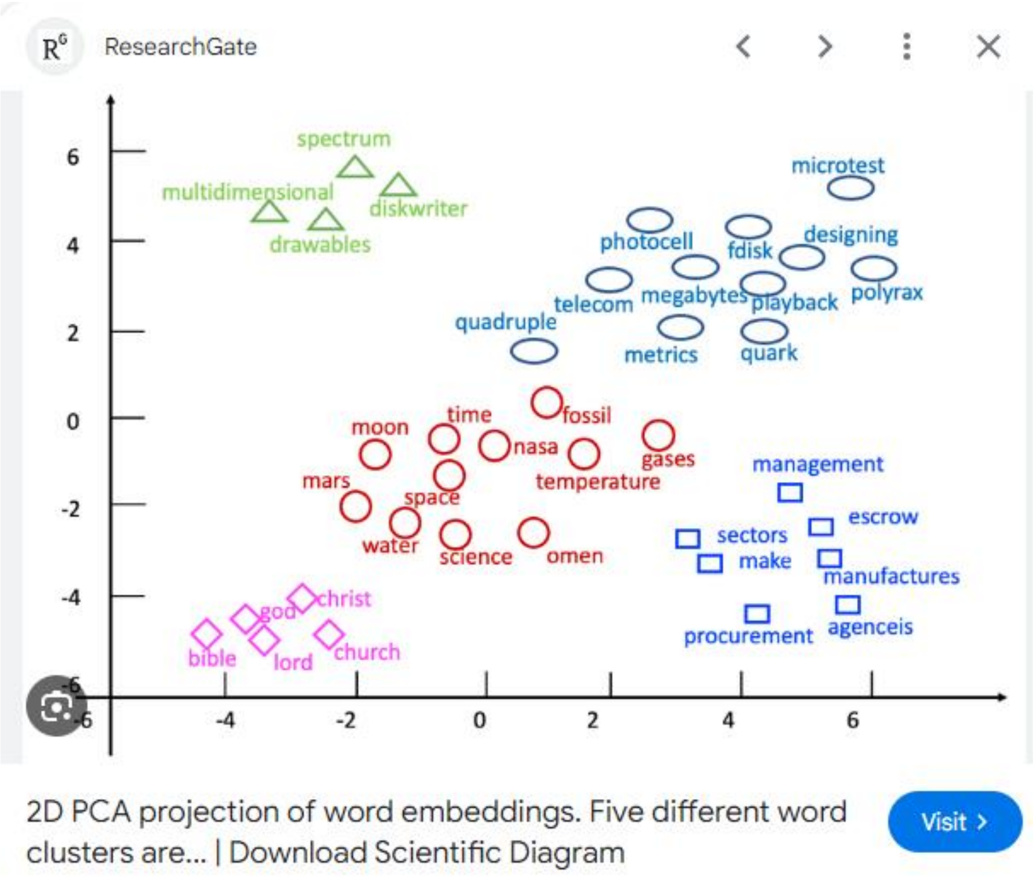
Semantic Analysis - A broad area of natural language processing focused on interpreting meaning from a language and understanding the human communication.

Key Terms/Concepts

Word Embeddings - A type of word representation that allows words with similar meaning to have a similar representation.



Similar words are closely placed in vector space



Key Terms/Concepts

N-grams - A contiguous sequence of n items from a given sample of text or speech. N-grams are calculated based on the frequency of sequences appearing in a corpus of text. The sequences that often appear together are statistically significant.

Example: The cat sleeps in the basket every night.

Unigram (1-gram): [“the”], ["cat"], ["sleeps"] ....

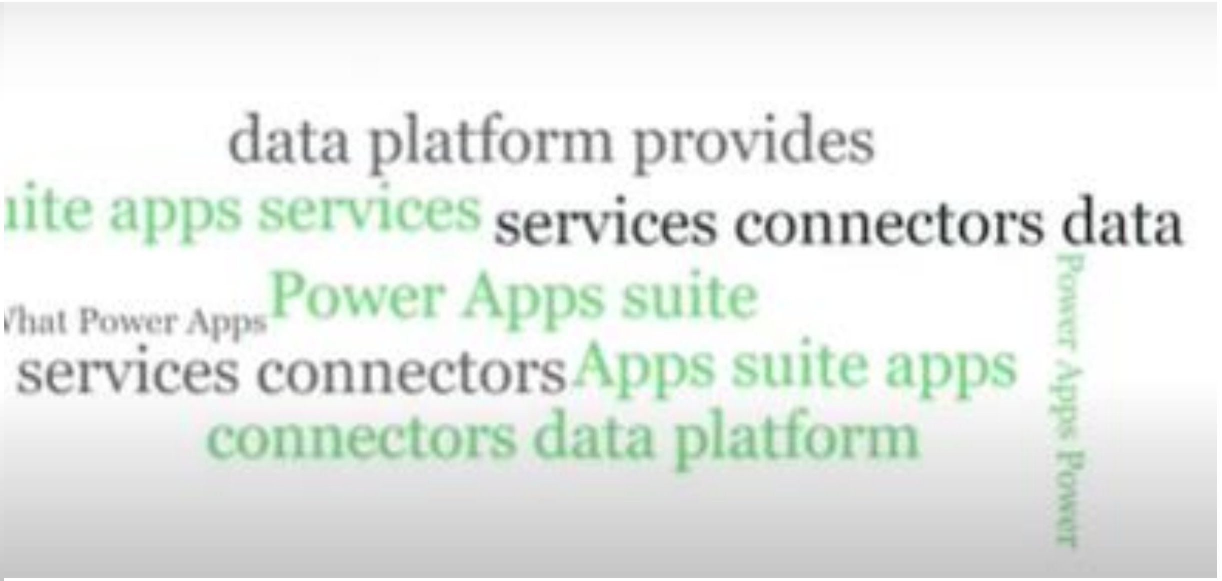
Bigram (2-gram): ["The cat"], ["cat sleeps"], ["the basket"] …

Trigram (3-gram): ["The cat sleeps"], ["in the basket"], …

4-gram: …

Key Terms/Concepts

Word Cloud: A word cloud visually represents text data, highlighting frequently used words in varying sizes and/or color based on their occurrence.



Key Terms/Concepts

TF (Term Frequency): Measures how frequently a term occurs in a document. More occurrences increase the term's relevance in that document.

TF-IDF (Term Frequency-Inverse Document Frequency): Weighs a term's frequency (TF) against its rarity across all documents, enhancing its importance when it is rare.

Key Terms/Concepts

Term Frequency (TF):

$$\text{TF} = \frac{\text{Number of times term appears in a document}}{\text{Total number of terms in the document}}$$

Inverse Document Frequency (IDF):

$$\text{IDF} = \log\left(\frac{\text{Total number of documents}}{\text{Number of documents with the term in it}}\right)$$

Term Frequency - Inverse Document Frequency (TF-IDF):

TF-IDF = TF x IDF

Key Terms/Concepts

Example:

Suppose we have a collection of 100 documents and the word "apple" appears in 5 of them.

- The IDF for "apple" would be:

  $$\text{IDF} = \log\left(\frac{100}{5}\right) = \log(20)$$

If "apple" appears 3 times in a 100-word document:

- The TF for "apple" would be:

  $$\text{TF} = \frac{3}{100} = 0.03$$

Thus, the TF-IDF score for "apple" in that document is:

$$\text{TF-IDF} = 0.03 \times \log(20)$$

Key Terms/Concepts

Document Term Matrix (DTM): This is a matrix where each row represents a document and each column represents a term (word) from the corpus. The entries in the matrix typically contain the frequency of the term in the document.

Term Document Matrix (TDM): This matrix is the transpose of the DTM. Here, each row represents a term and each column represents a document. Similarly, the entries indicate the frequency of the term in the corresponding document.

The entries in the above matrix can include term frequency, TF-IDF, and other similar measures.

Document Term Matrix & Term Document Matrix

Document 1:

I love apple. I hate orange.

Document 2:

I love orange.

TDM (weight is term count)

	doc 1	doc 2
I	2	1
Love	1	1
hate	1	0
apple	1	0
orange	1	1

DTM (weight is term count)

	I	Love	Hate	Apple	Orange
doc 1	2	1	1	1	1
doc 2	1	1	0	0	1