THE UNIVERSITY OF QUEENSLAND
AUSTRALIA
CREATE CHANGE

What is Text Analytics?

Digital Cultures and Societies Hub Workshop (UQ, 23 | 24 May, 2024)

What is Text Analysis?

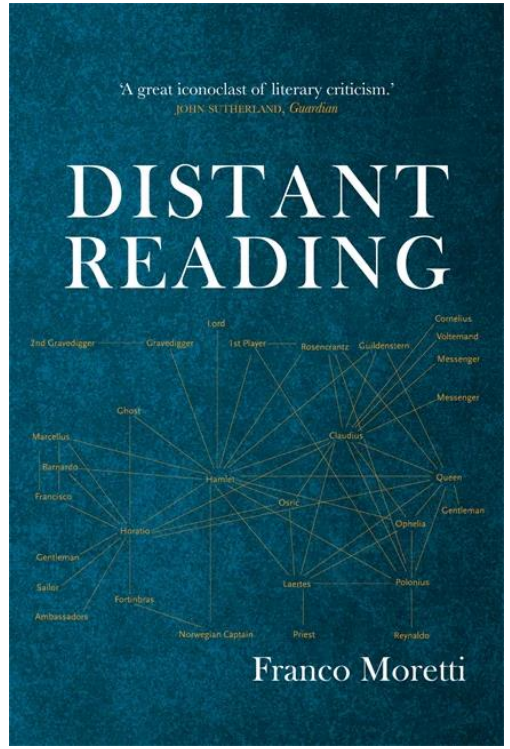
What is Text Analysis?

- Text analysis (TA) refers to the process of examining, processing, and interpreting unstructured data sources (texts) to uncover actionable knowledge using computation
  - Unstructured data (text) can be emails, literary texts, letters, articles, advertisements, official documents, social media content, or product reviews.
  - Actionable knowledge includes insights and patterns used to classify, sort, and extract information or relationships, trends, or sentiments.

What is Text Analysis?

- Sometimes **Text Analysis** is distinguished from **Text Analytics**.
- In this reading Text Analysis refers to manual, close-reading, qualitative interpretative approaches and Text Analytics to refer to quantitative, automated computational analysis of text using natural language processing (NLP) and machine learning (ML).
- However, here, we consider Text Analysis and Text Analytics to be synonymous.

Related concepts



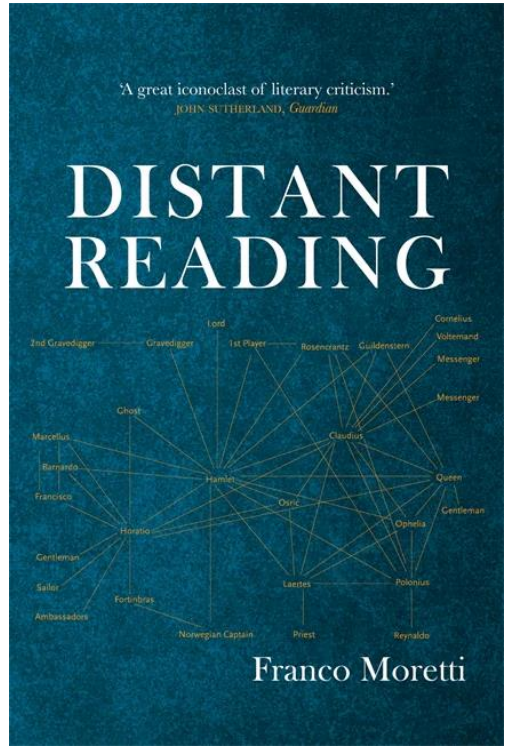
Distant Reading

- Distant reading is an approach to text analysis pioneered by Franco Moretti. It involves analysing large corpora of literary texts, using computational methods to identify broad patterns and trends.
- Distant reading focuses on the quantitative analysis of texts rather than close, qualitative reading.

Corpus Linguistics

- Corpus linguistics (CL) is a branch of linguistics that involves the study of language using large collections of texts known as corpora. It aims to analyse linguistic phenomena by examining patterns and frequencies of words and structures within a corpus.

Related concepts



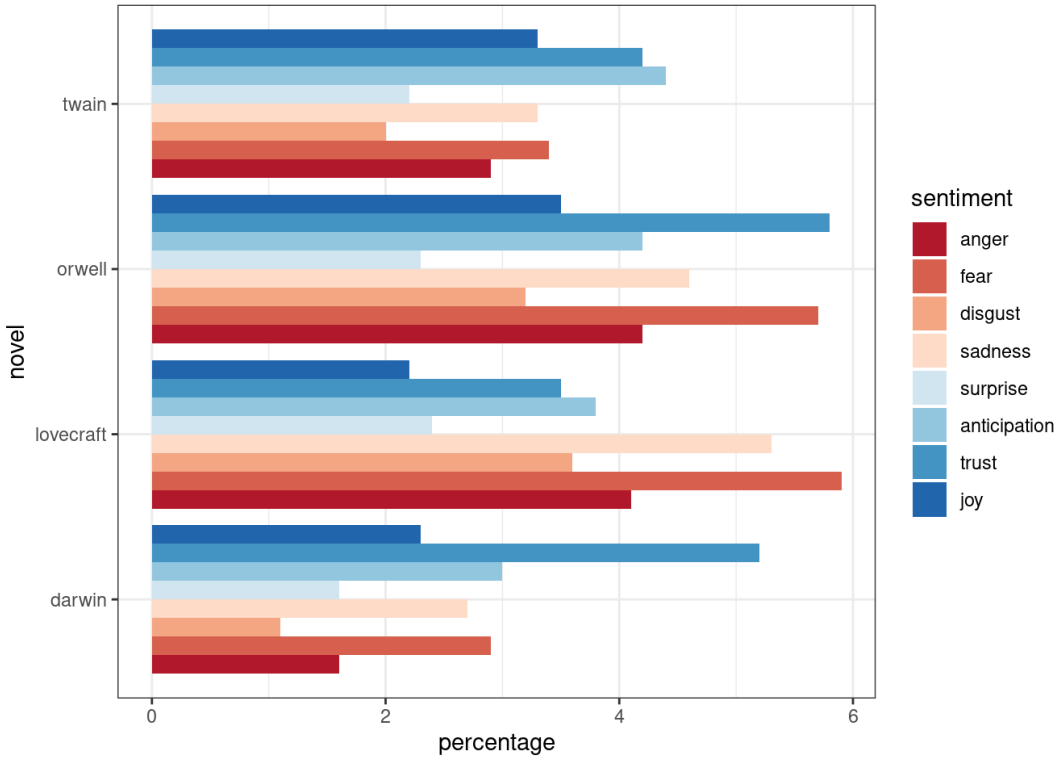
Differences between TA and related concepts

- While TA, distant reading, NLP, and corpus linguistics share the common goal of understanding textual data, they differ in their approaches, methodologies, and objectives.
- TA differs from CL in that CL focuses (exclusively) on linguistic questions, issues, and phenomena to understand language while TA is broader, encompassing linguistic as well as non-linguistic questions, issues, and phenomena.
- TA differs from NLP in that NLP is more focused on the methods themselves (e.g. improving and evaluating classification methods)
- TA can be considered a cover or umbrella term for any type of analysis of textual data.

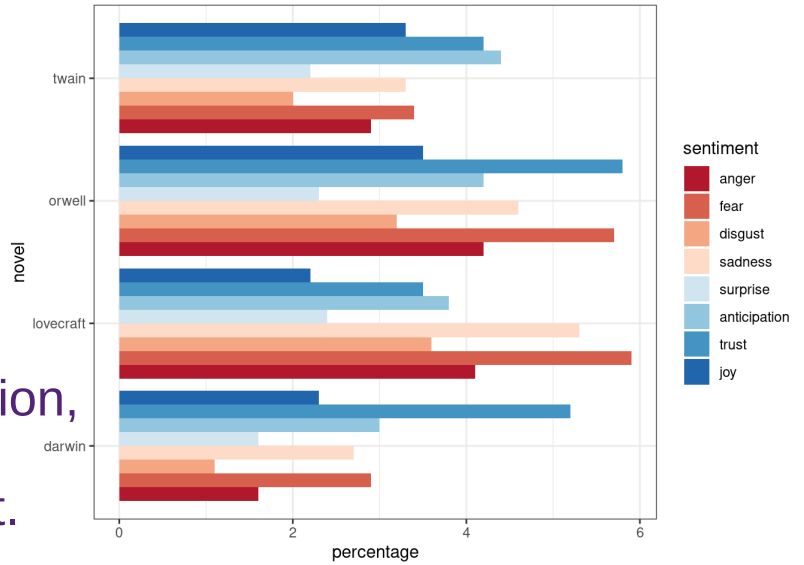
Methods in Text Analysis

What is Sentiment Analysis

- Sentiment analysis, also known as opinion mining, is a computational technique used to analyse and extract subjective information from text data.
- It involves identifying, quantifying, and categorizing the sentiment,
stance, or emotionality expressed in textual content, such as utterances, blog posts, reviews, literary texts, or news articles.



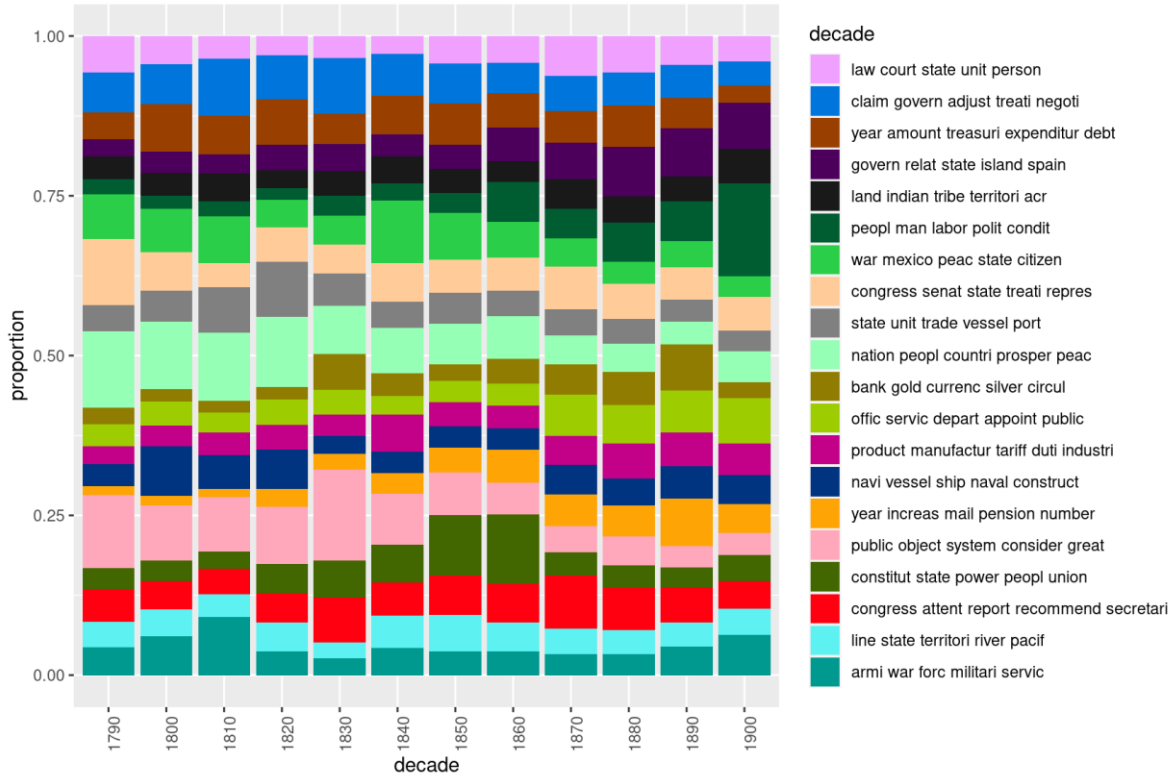
Basic Concepts of Sentiment Analysis



- **Sentiment Polarity**
  Sentiment polarity refers to the classification of (elements of) text(s) into positive, negative, or neutral categories based on the expressed sentiment. Positive sentiment indicates a favourable opinion or emotion, while negative sentiment indicates an unfavourable opinion or emotion. Neutral sentiment signifies the absence of strong sentiment.
- **Sentiment Analysis Techniques**
  Sentiment analysis techniques can be divided into **rule-based methods** and **machine learning-based approaches**.
  Rule-based methods rely on predefined rules and **lexicons** to determine sentiment, while machine learning approaches use algorithms to **learn patterns and classify sentiment** automatically.
- **Aspect-Based Sentiment Analysis**
  Aspect-based sentiment analysis goes beyond overall sentiment polarity to analyze the sentiment towards specific aspects or features mentioned in the text. It identifies sentiments associated with particular entities, attributes, or topics within the text, providing more detailed insights into opinionated content.

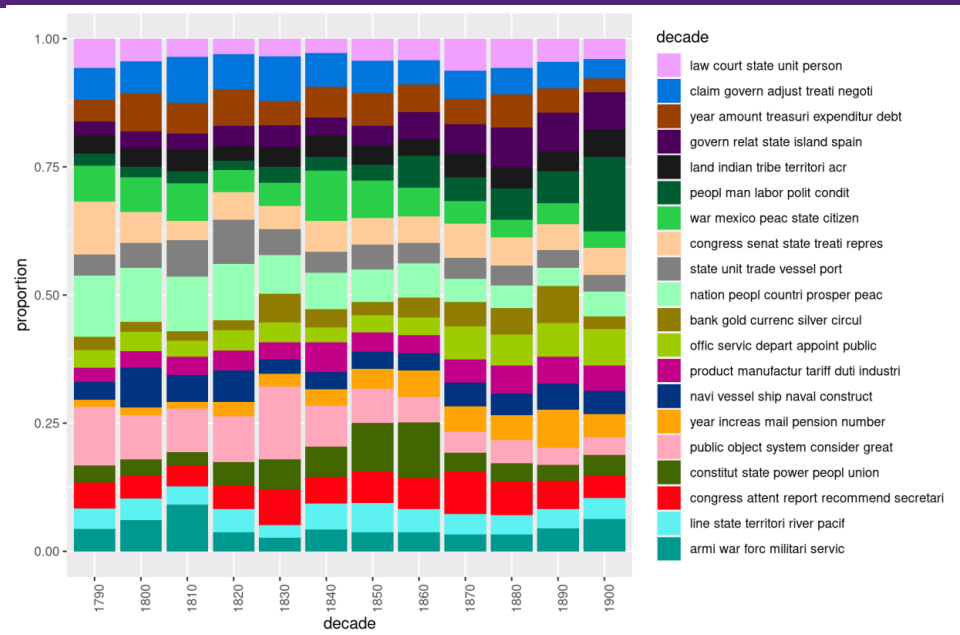
What is Topic Modelling

- Topic modelling (TM) is a computational method used to discover latent thematic structures within a collection of texts.
- It aims to identify recurring topics or themes that characterize the content of textual data.



- Topics represent underlying themes or concepts that are prevalent in a corpus of texts. Each topic consists of a set of words that frequently co-occur within documents in the corpus.

Basic Concepts of Topic Modelling



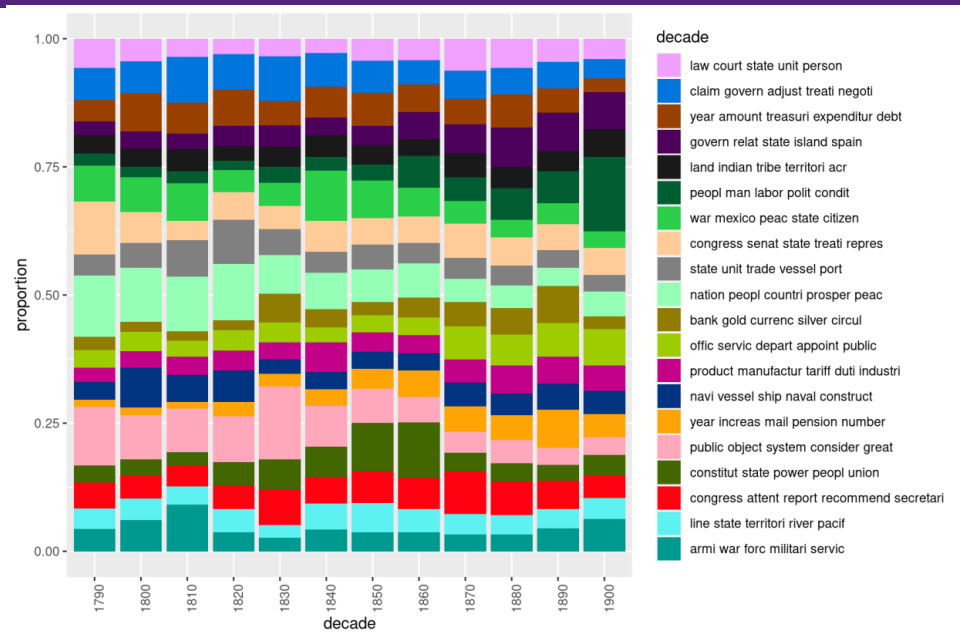
Latent Dirichlet Allocation (LDA)

- LDA is one of the most commonly used algorithms for TM. It assumes that each document in the corpus is a mixture of topics, and each word in the document is attributable to one of the document's topics.

Topic Distribution

- Topic distribution refers to the proportion of each topic present in a document.
- It provides insights into the thematic composition of individual documents within the corpus.

Applications of Topic Modelling



Text Classification

- TM can be used for text classification tasks such as genre classification, authorship attribution, and text categorization.
- By identifying dominant topics in texts, researchers can categorize and classify documents based on their thematic content.

Discourse Analysis

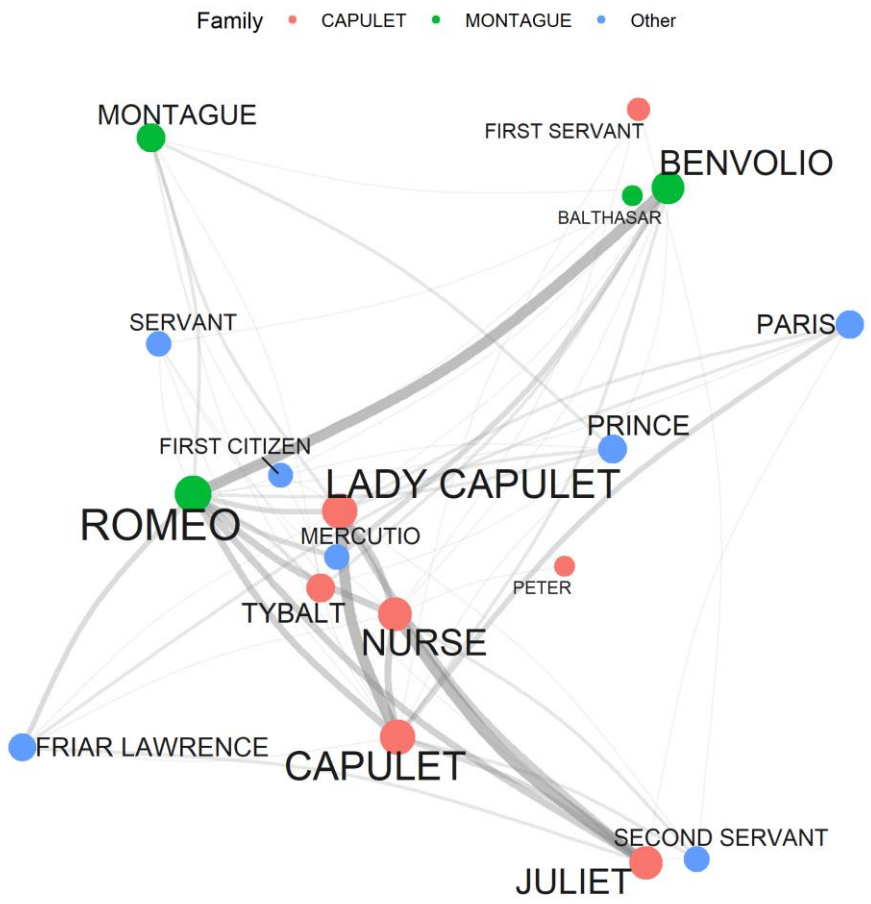
- TM facilitates discourse analysis by uncovering recurring themes and patterns in textual data.
- It helps identify key topics of discussion, discourse markers, and shifts in discourse focus within a corpus of texts.

Language Teaching and Learning

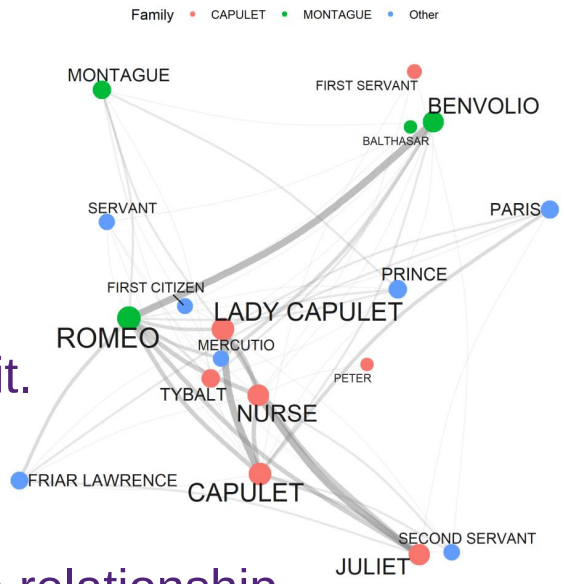
- TM can be applied to analyze language learning materials, textbooks, and learner-generated texts and thereby assist in identifying relevant topics and linguistic features for language instruction and curriculum development.

What is Network Analysis

- Network analysis is a methodological approach used to study the structure, behaviour, and interactions within complex systems represented as networks.
- It involves the analysis of nodes (entities) and edges (connections) to uncover patterns, relationships, and properties of the network.



Basic Concepts of Network Analysis



Nodes

- Nodes represent individual entities within a network. These entities can be people, organizations, speakers, locations, or any other unit of interest.
- Each node in a network typically has attributes or properties associated with it.

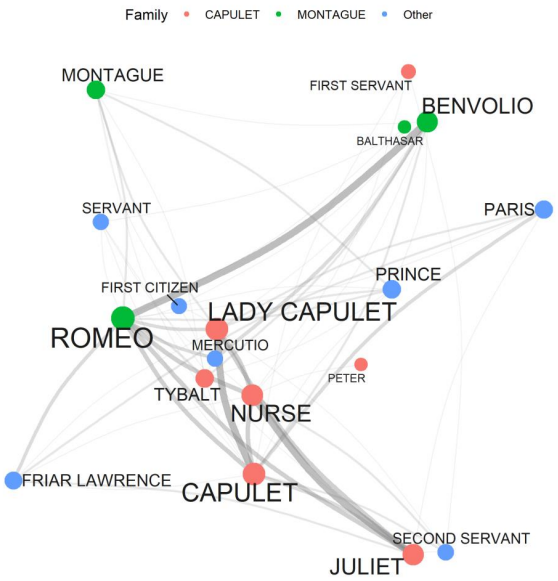
Edges

- Edges (or links) represent the connections between nodes in a network.
- Edges can be directed or undirected, indicating the presence or absence of a relationship between two nodes.
- The strength or weight of an edge may also be considered, reflecting the intensity or frequency of interaction between nodes.

Network Metrics

- Various metrics are used to quantify the structure and properties of a network, such as degree centrality, betweenness centrality, and clustering coefficient.
- These metrics provide insights into the importance, influence, and connectivity of nodes within the network.

Applications of Network Analysis



- **Social Network Analysis**
  Social network analysis examines social structures by analysing the relationships between individuals or groups within a network. It is used to study communication patterns, influence, information flow, and social dynamics in various contexts, including organizations, communities, and online platforms.
- **Sociolinguistic Network Analysis**
  In sociolinguistics, network analysis is used to study social relationships such the type and density of speech communities. It helps uncover the relationships between speakers (network members), identify key participants, and understand the functioning of language variation and change.

Keyword | Keyness Analysis

Keyness refers to a method used to determine how characteristic, unique, or important terms are for a (collection of) text(s).

Keyness is a measure that evaluates frequency of occurrence of a linguistic feature (such as a word or phrase) either in a collection of texts or between two corpora.

Keyness measures

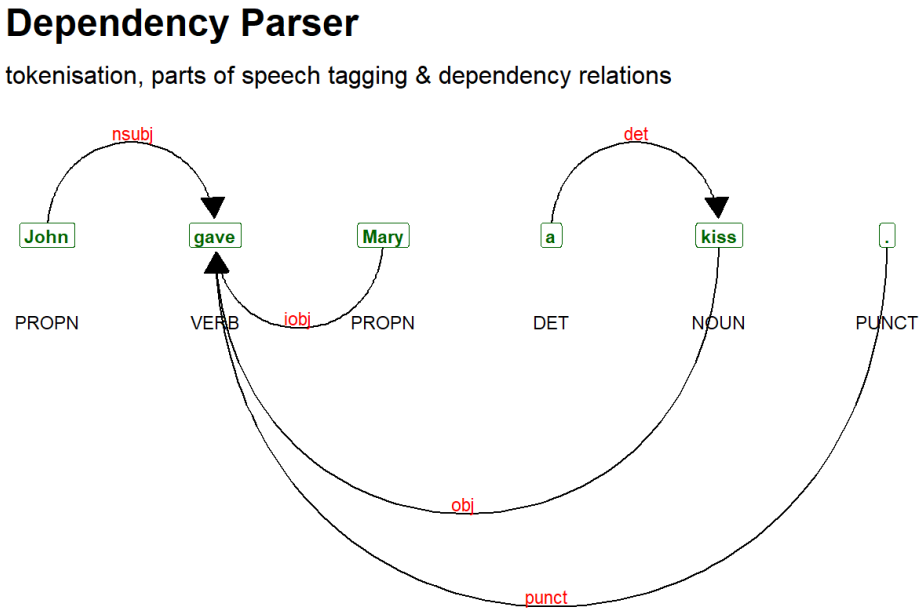
Without a reference corpus

- Term Frequency Inverse Document Frequency (TF-IDF)

With a reference corpus

- Mutual Information (MI), Log Likelihood Ratio (LLR), Relative Frequency Difference (RFD), (X^2 or phi [ϕ])

What is Named Entity Recognition

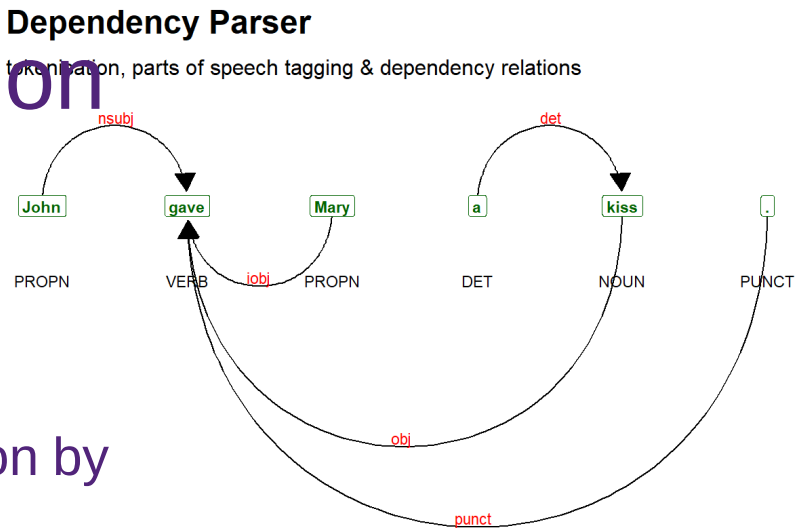


- Named Entity Recognition (NER) is a natural language processing (NLP) task that involves identifying and categorizing named entities in text into predefined categories such as person names, organization names, locations, dates, and more.
- It plays a crucial role in extracting structured information from unstructured text

Named Entities

- NER algorithms typically involve pos-tagging and parsing and aim to identify and classify named entities (e.g., people, organizations, locations, dates, times, quantities, and monetary values) within a text.

Applications of Named Entity Recognition



Language Documentation and Description

- Named Entity Recognition can assist linguists in identifying and categorizing proper nouns and other named entities in linguistic texts.
- It aids in the process of language documentation and description by automatically identifying key entities mentioned in texts.

Cross-Linguistic Studies

- NER techniques can be applied to analyse texts in multiple languages, facilitating cross-linguistic studies and comparative linguistic analysis for instance by identifying similarities and differences in the naming conventions and entity references across languages.

Text Annotation and Corpus Linguistics

- Named Entity Recognition tools are valuable for annotating linguistic corpora with named entity labels, enabling researchers to analyse the distribution and usage of named entities in different textual genres and contexts.

References

Wutich, A., Ryan, G., & Bernard, H. R. (2014). Text Analysis. In H. R. Bernard & C. C. Gravlee (Eds.), *Handbook of Methods in Cultural Anthropology* (2nd ed., pp. 533-559). Lanham, Boulder, New York, Toronto, Plymouth, UK: Rowman & Littlefield.

Jockers, M. L., & Thalken, R. (2020). *Text Analysis with R*. Springer International Publishing.

Welbers, K., Van Atteveldt, W., & Benoit, K. (2017). Text Analysis in R. *Communication Methods and Measures*, 11(4), 245-265.