Istat

CARMA 2024

26-28 June 2024

Topic Modelling Tutorial

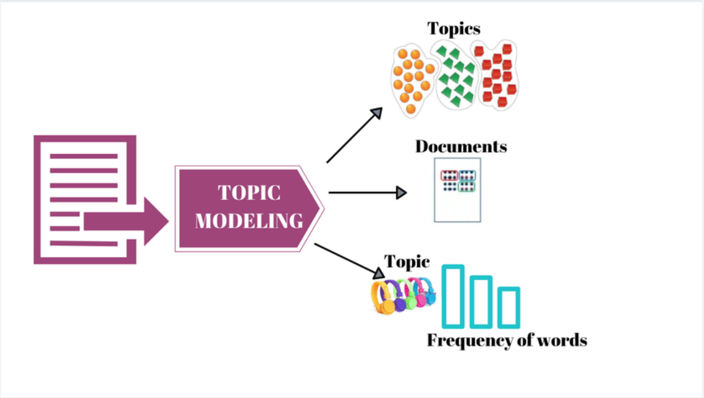
Mauro Bruno, Elena Catanese, Francesco Ortame, Francesco Pugliese,

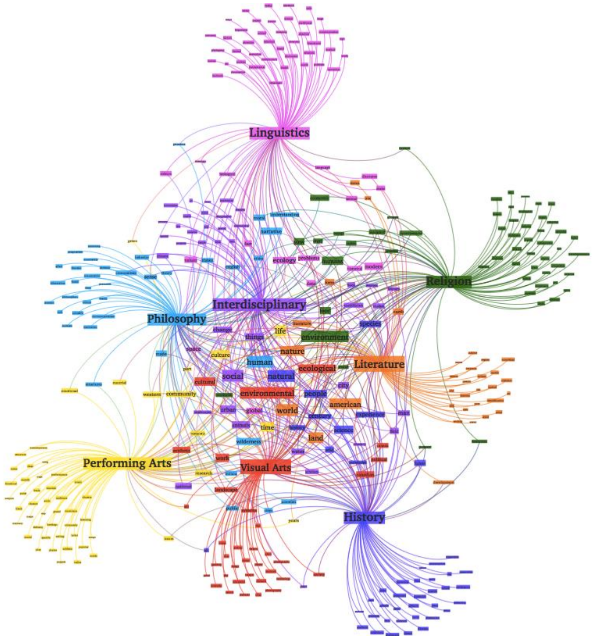
mbruno@istat.it, catanese@istat.it, ortame@istat.it, frpuglie@istat.it

What is the Topic Modelling?

✓ **Topic modeling** is a type of statistical modeling that uses unsupervised Machine Learning to identify **clusters** or **groups** of similar words within a body of text. This **Text Mining** method uses **semantic structures** in text to understand unstructured data without predefined tags or training data.

✓ **Topic modeling** discovers **abstract** topics within a collection of documents. It is widely used in **Natural Language Processing** (NLP) and **Text Mining** to understand the themes or subjects present in large sets of unstructured text data.





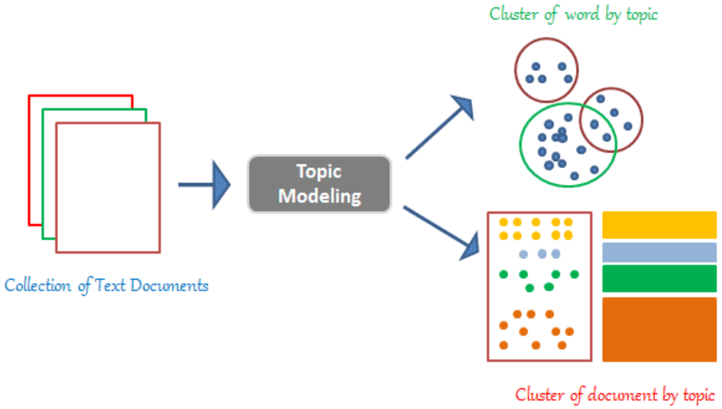
What is the Topic Modelling?

✓ Here’s a more detailed **breakdown** of what topic modeling entails:

- **Documents and Words:** The basic units of topic modeling. Documents are the individual pieces of text (e.g., articles, emails, reviews), and words are the tokens or terms within these documents.

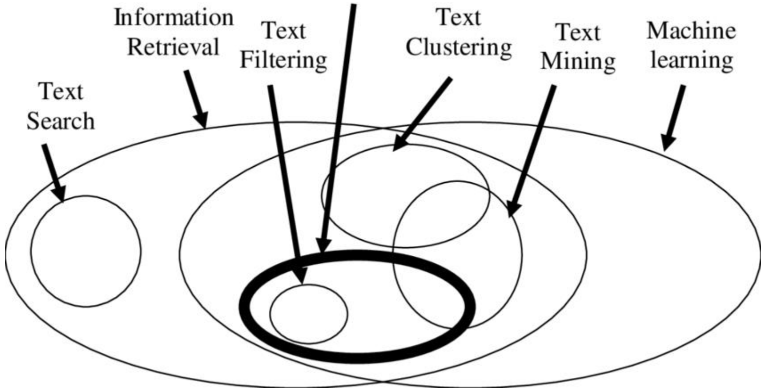
- **Topics:** A topic is a distribution over a fixed vocabulary. It is characterized by a set of words that frequently appear together. Each topic can be seen as a pattern of co-occurrence of words.

- **Latent Variables:** These are the hidden patterns or topics inferred from the observed data (the words in the documents).



Applications of Topic Modelling

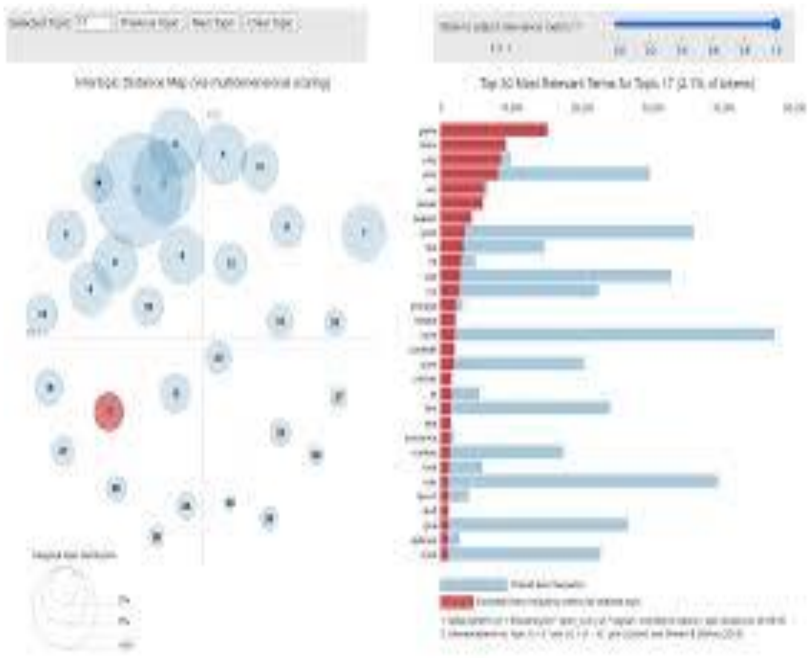
- ✓ **Text Categorization**: Classifying documents into predefined categories based on the identified topics.
- ✓ **Information Retrieval**: Enhancing search engines by indexing documents with topics to improve search relevance.
- ✓ **Recommender Systems**: Suggesting articles, books, or other content based on topics of interest.
- ✓ **Sentiment Analysis**: Understanding public sentiment by analyzing the topics discussed in social media, reviews, or feedback.



Challenges in Topic Modelling

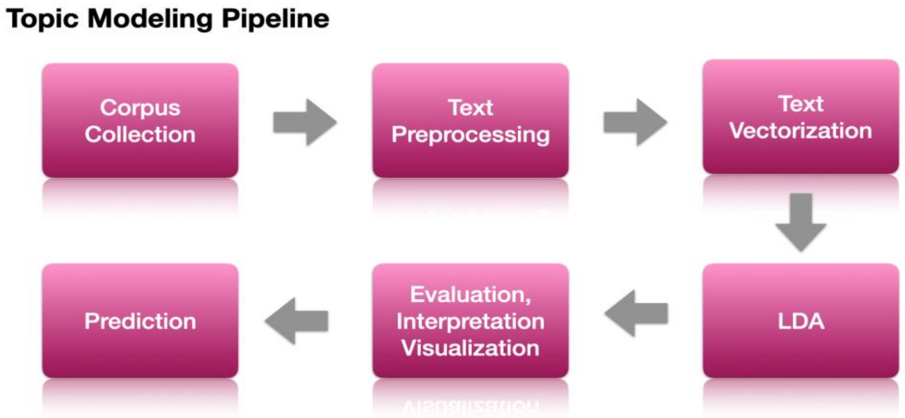
- ✓ Topic modelling is a powerful tool for extracting insights from large text **corpora**, enabling a **deeper** understanding of the underlying themes and patterns in the data.

- ✓ **Choosing the Number of Topics:** Determining the optimal number of topics can be non-trivial and often requires experimentation.
- ✓ **Topic Interpretability:** Ensuring the topics make sense to humans can be challenging.
- ✓ **Scalability:** Handling large datasets efficiently, particularly in terms of computation time and memory usage.



Steps in Topic Modelling

1. **Preprocessing**: Clean the text data (e.g., remove stop words, tokenize, and normalize).
2. **Model Training**: Apply a topic modeling algorithm to the preprocessed data.
3. **Evaluation**: Assess the quality of the topics using coherence scores, human judgment, or other metrics.
4. **Interpretation**: Analyze the topics and assign meaningful labels or descriptions.



Upsides of Topic Modelling

1. **Preprocessing**: Clean the text data (e.g., remove stop words, tokenize, and normalize).
2. **Model Training**: Apply a topic modeling algorithm to the preprocessed data.
3. **Evaluation**: Assess the quality of the topics using coherence scores, human judgment, or other metrics.
4. **Interpretation**: Analyze the topics and assign meaningful labels or descriptions.

Upsides of Topic Modelling

1. **Efficient Data Organization and Summarization:** Topic modeling can automatically categorize vast amounts of text data into meaningful topics, saving significant time and effort.
2. **Insight Generation and Uncovering Hidden Patterns:** Topic modeling has the ability to uncover latent themes and structures within text data that might not be immediately obvious.
3. **Enhanced Information Retrieval:** By indexing documents with topics, search engines can return more relevant results.
4. **Content Recommendation and Personalization:** Topic modeling can be used to recommend content based on the topics of interest, such as articles, books, or multimedia.
5. **Scalability and Versatility:** Its versatility allows it to be applied across various domains, such as academia, business, marketing, etc.

Downsides of Topic Modelling

1. **Choosing the Number of Topics:** Selecting the optimal number of topics is often arbitrary and requires trial and error.
2. **Interpretability of Topics:** The topics generated by models like LDA may not always be easily interpretable. The sets of words associated with each topic might not clearly convey a coherent theme, making it difficult to label and understand them
3. **Vague or Mixed Topics:** Sometimes the words in a topic don't form a clear, coherent idea
4. **Human Judgment Required**: Often needs human interpretation to make sense of the topics.
5. **Scalability and Computational Cost:** Training topic models on large datasets can be computationally expensive and time-consuming

Latent Dirichlet Allocation - LDA

Hierarchical Dirichlet Process - HDP

Non-Negative Matrix Factorization – NMF

Top2Vec

BERTopic

Llama2

BERTopic with Llama2

References

Acknowledgements

Thank You for you Attention