NLP

Course of DSSC Master degree - University of Trieste

Matilde Trevisani

DEAMS

2020/05/25 (updated: 2020-06-12)

Agenda

1. Introduction a. Areas and terminology b. Corpora and dimensions c. Frequencies d. Data transformations e. Selecting units

2. Content Statistical Analysis

    - Correspondence Analysis
    - Topic analysis

3. Chronological Corpora

Introduction to areas and terminology

Textual Analysis Processing

Textual Data Analysis

The field of textual data analysis (TDA) is used to study any form of written communication (or even oral-transcribed).

How many and what types of texts? In theory, any **collection of texts**.

Examples:

- spoken-transcribed conversations (e.g. qualitative interviews, open-ended questionnaire questions, life stories, psychological interviews);
- written conversations (letters, emails, sms, whatsapp, facebook, twitter, chat);
- written production (diaries, school subjects, essays, stories, novels, poetry, songs, ethnographic diaries, documents, reports);
- media and web writing (press, web pages, news, advertising);
- transcription of public and institutional speeches
- . . .

A problem of terminology

> Analisi quantitative dei fatti di lingua

(De Mauro – Chiari, 2005)

- text mining, content analysis, text analysis,
- opinion mining, web reputation, sentiment analysis
- digital methods (digital humanities), distant reading…
- natural language processing (NLP), information retrieval (IR)
- word mapping, text categorization
- concept extraction, topic detection, document summarization, etc.

(In Italian)

- analisi del testo, analisi dei testi (analisi testuale), analisi automatica dei testi
- analisi dei dati testuali, analisi statistica dei dati testuali, analisi quantitativa dei dati testuali, statistica testuale
- linguistica computazionale, linguistica quantitativa, linguistica dei corpora
- analisi del contenuto
- analisi emozionale del testo (AET), analisi del discorso (discourse analysis), analisi critica del discorso (analisi retorica, ermeneutica)

There are so many different terms because there are so many fields of application.

There are many fields of application because there are many disciplines that use written texts to carry out applied research.

The use of **terminology changes from discipline to discipline**. In this sense, we do not always have terms (biunivocal).

Textual Data Analysis

Text analysis is a bit too general.

Analysis of the text is a very general term, which can refer to very different approaches: qualitative, quantitative, mixed-methods.

Textual data analysis refers to a process of collecting, coding, analyzing and interpreting the information contained in a set of texts.

The focus is on "data".

When statistical (or quantitative) methods are called into question, **statistical analysis of textual data** can be used (quantitative analysis of textual data)

TDA is a research **subject** for some disciplines:

- linguistics, computer science, statistics

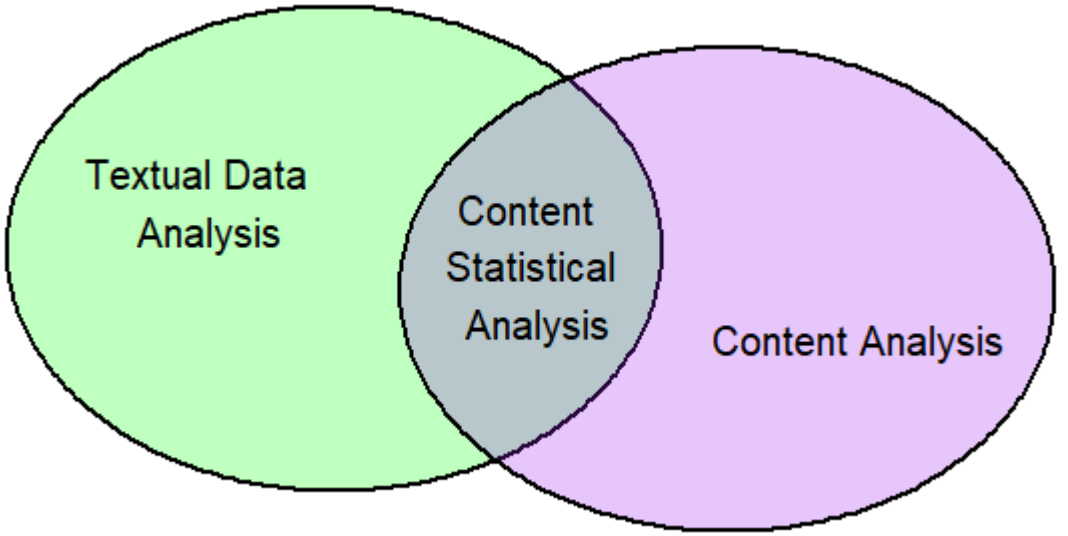
TDA is a research **tool** for other disciplines:

- psychology, sociology, sociolinguistics, history, political science, economics, communication, media studies, etc.

TDA is a very interdisciplinary field

What content analysis?

Between statistics and research methodology



Does it make sense?

The idea that sense and meaning can be automatically extracted from a text is not unanimously accepted.

At present, the tools available are still partly **coarse**.

It is worth reflecting on a crucial point:

> the development of all these tools is dictated by the belief that even human intervention is not free from errors; moreover it is subjective, not reproducible and, fundamentally, too expensive in terms of resources and time.

Strengths

Statistical analysis of textual data guarantees speed and systematicity in operations of search, analysis and synthesis of the information of interest that can hardly be guaranteed - in some cases, impossible - by qualitative analyzes.

It allows you to overcome the obstacles that represent the main limits of qualitative analysis.

An integrated approach

ADT should not be imagined as an alternative to traditional qualitative approaches:

- quantitative methods offer **"upstream"** ideas for qualitative insights;
- quantitative methods offer **"downstream"** tools to verify on a large scale the insights that emerged from a first qualitative analysis.

classical methods

- Sweden, 17th century, Songs of Zion (Lutheran Church)
- Thomas and Znaniecki 1918
- Propaganda Technique in the World War (Lasswell 1927)
- Columbia University School of Journalism
- 1942, sentencing of The Galileian editor, W. D. Peley

modern methods

- 1970, Benzécri (Correspondence analysis)
- quantitative linguistics studies (Yule, Guiraud, Zipf ...)
- lexicon-textual approach
- text mining