Textual Data Analysis

Introduction

Matilde Trevisani
Agenda

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

2. Content Statistical Analysis
   - Content mapping (Correspondence Analysis)

3. Text classification
   - Authorship attribution (Discriminant Analysis)
Areas and terminology
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, questionnaire open-ended questions, life stories, psychological interviews);
- written conversations (letters, emails, sms, whatsapp, facebook, twitter, chat);
- written production (diaries, school compositions, essays, short stories, novels, poetry, songs, ethnographic diaries, documents, reports);
- media and web writing (press, web pages, news, advertising);
- transcription of public and institutional speeches
- scientific literature
A problem of terminology

Quantitative analysis of linguistic objects / Analisi quantitative dei fatti di lingua

-- (De Mauro – Chiari, 2005)

- text mining, text analysis, text analytics, content analysis
- opinion mining, sentiment analysis, web reputation
- digital methods (digital humanities), distant reading...
- natural language processing (NLP), information retrieval (IR)
- word mapping, text categorization
- concept extraction, topic detection, document summarization, etc.
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 emotionale del testo (AET), analisi del discorso (discourse analysis), analisi critica del discorso (analisi retorica, ermeneutica)
Many disciplines, many applications

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 biunivocal terms.
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).
Interdisciplinary field

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
Text mining / analytics

- TDA ≈ Text mining ≈ Text analytics

- Text mining is the process of distilling actionable insights from text

  Turn text data into high-quality information or actionable knowledge

  - Minimizes human effort (on consuming text data)
  - Supplies knowledge for optimal decision making

- Related to text retrieval, which is an essential component in any text mining system

  - Text retrieval can be a preprocessor for text mining
  - Text retrieval is needed for knowledge provenance
Learning workflow

Acquisition → Synthesis → Restitution
Text mining workflow

1 - Problem definition & specific goals

2 - Identify text to be collected

3 - Text organization
4 - Feature extraction
5 - Analysis

6 - Reach an insight, recommendation, or output
Humans as Subjective “Sensors”
The general problem of data mining
The problem of text Mining
Landscape of Text Mining and Analytics

1. Mining knowledge about language
2. Mining content of text data
3. Mining knowledge about the observer
4. Infer other real-world variables (predictive analytics)

Real World → Perceive (Perspective) → Observed World → Express (English) → Text Data + Context → + Non-Text Data
Landscape of Text Mining and Analytics

1. Natural language processing & text representation
2. Word association mining & analysis
3. Topic mining & analysis
4. Opinion mining & sentiment analysis
5. Text-based prediction

Real World

Perceive
(Perspective)

Observed World

Express
(English)

Text Data
What content analysis?

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.

Finally, as for the strengths of statistical analysis of textual data, it 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 analyses.

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

TDA 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.
Corpora and dimensions
Basic concepts

1. constitution of the corpus
2. counting words
3. vocabulary
4. statistical terminology
Constitution of the corpus

A text is an object that has a beginning, a development, an end and has a communicative purpose.

A corpus is not simply a "collection of texts". A corpus must have characteristics of breadth, coherence, homogeneity and, in some cases, exhaustiveness that make it suitable for research purposes.

The corpus is a set of texts that responds to the needs of a specific research question.

You must first formulate a clear research question. Without forgetting that the "good" research questions must be comparative (which refers to the problem of identifying sub-corpora).

Note: many problems can be addressed with statistical tools but the path of definition, construction and analysis of data is often not evident.
Homogeneity and variation

Starting from the research question and from the strategy adopted for the creation of the corpus, it is immediately necessary to distinguish the criteria adopted for the evaluation of the quality of the corpus from those which are, instead, variations that you want to observe because they represent the object of the research study.

Some criteria:

- dimensions
- textual genre
- language, theme, style, the (Muller & Brunet, 1988)
Five dimensions of variation for the language (Berruto, 1987):

1. **diachronic** (variation due to chronological differences),
2. **diatopic** (variation due to differences in geographical location),
3. **diaphasic** (variation due to differences in the communicative situation, typical opposition: formal vs. informal),
4. **diastratic** (variation due to differences in the social groups of reference),
5. **diamesic** (variation linked to the medium of transmission, typical opposition: oral vs. written text).
About the diachronic variation

The language needs long periods of time (centuries) to show significant changes.

Although this consideration is obvious in a linguistic context, it is important to underline that the diachronic variations observable in the short term (decades) almost always concern the lexical level (typical examples are neologisms and foresters), which represents the most superficial part of the language.

A change at the syntactic level would represent a profound change in the language.
Choice of the unit of analysis

- document unit
- granularity level of analysis
- single characters, words, phrases, even groups of phrases.
Representation of the corpus

Corpus – texts – subcorpora – fragments
Some collection methods

1. When you have a large number of texts of limited dimensions, for example interventions in focus groups, short answers to open questions, posts, messages (SMS and Whatsapp), advertising messages ...
   - Excel sheet (a text in a cell)

2. When you have a limited number of medium-large sized texts, eg. speeches, letters, documents, topics, scientific articles ...
   - Word document (one text below the other in the same file)

3. When you have a very large number of texts
   - folder that collects all the files (in .txt format)

4. The "tidy corpus" perspective (new approach)
-> 1. A large number of texts of limited size

<table>
<thead>
<tr>
<th>ID</th>
<th>Day</th>
<th>Month</th>
<th>Year</th>
<th>Representative</th>
<th>Tweet</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>31</td>
<td>12</td>
<td>2013</td>
<td>Renzi</td>
<td>In Palazzo Vecchio, al lavoro per preparare la Giunta di fine anno. Si annuncia bella corposa.</td>
</tr>
<tr>
<td>3</td>
<td>31</td>
<td>12</td>
<td>2013</td>
<td>Renzi</td>
<td>@Fiorillo Nessuno è perfetto, Fiorel Auguri</td>
</tr>
<tr>
<td>5</td>
<td>30</td>
<td>12</td>
<td>2013</td>
<td>Renzi</td>
<td>Oggi a #Firenze, Estacy ha sperto nella via Martelli pedonale. Sono 122 posti di lavoro. E io sono a Casa.</td>
</tr>
<tr>
<td>6</td>
<td>28</td>
<td>12</td>
<td>2013</td>
<td>Renzi</td>
<td>Grazie a tutti, buona serata. Torneremo con #matteorisponde dopo Natale</td>
</tr>
<tr>
<td>7</td>
<td>27</td>
<td>12</td>
<td>2013</td>
<td>Renzi</td>
<td>Pronto per il #matteorisponde. Tra cinque minuti si parte…</td>
</tr>
<tr>
<td>8</td>
<td>27</td>
<td>12</td>
<td>2013</td>
<td>Renzi</td>
<td>Colonne sonora di domani e @pdimaggio: &quot;Resta ribelle&quot; dei Negriti &quot;La tua canzone&quot;. Place?</td>
</tr>
<tr>
<td>9</td>
<td>27</td>
<td>12</td>
<td>2013</td>
<td>Renzi</td>
<td>Insieme a @bobobiagio nel giorno in cui finisce lo sciopero della fame. pic.twitter.com/tsNq6h6o</td>
</tr>
<tr>
<td>10</td>
<td>20</td>
<td>12</td>
<td>2013</td>
<td>Renzi</td>
<td>Caro @beppe_grillo ti rispondo nei prossimi giorni con una #sorpresina che ti sto preparando.</td>
</tr>
<tr>
<td>11</td>
<td>17</td>
<td>12</td>
<td>2013</td>
<td>Renzi</td>
<td>Scusate il ritardo nelle risposte, sono stati giorni intensi. E bellissimi. Ma il meglio dove ancora...</td>
</tr>
<tr>
<td>12</td>
<td>17</td>
<td>12</td>
<td>2013</td>
<td>Renzi</td>
<td>Grazie.</td>
</tr>
<tr>
<td>13</td>
<td>17</td>
<td>12</td>
<td>2013</td>
<td>Renzi</td>
<td>Giornata difficile da dimenticare… Ci vediamo alle 22 all'ObiHall (Firenze Sud) e in streaming su wordpress.com/news</td>
</tr>
<tr>
<td>14</td>
<td>14</td>
<td>12</td>
<td>2013</td>
<td>Renzi</td>
<td>Grazie a tutti i volontari che consentono le primarie e ai cittadini che stanno votando. Buon voto!</td>
</tr>
<tr>
<td>15</td>
<td>14</td>
<td>12</td>
<td>2013</td>
<td>Renzi</td>
<td>La piazza di Empoli mi resterà nel cuore a lungo. Grazie ragazzi, ci vediamo #lavottlabuona</td>
</tr>
<tr>
<td>16</td>
<td>11</td>
<td>12</td>
<td>2013</td>
<td>Renzi</td>
<td>Grazie a tutti i volontari e i cittadini che stanno prendendo freddo ai tavolini nelle mille piazze</td>
</tr>
<tr>
<td>17</td>
<td>11</td>
<td>12</td>
<td>2013</td>
<td>Renzi</td>
<td>Mamma mia, quanto entusiasmo. Grazie a tutti. Adesso inizia la parte difficile: uno per uno, ca</td>
</tr>
<tr>
<td>18</td>
<td>11</td>
<td>12</td>
<td>2013</td>
<td>Renzi</td>
<td>Ultimo giorno e poi si cambiano! Arriviamo a Milano. Questa è la #vottlabuona, ora o mai più</td>
</tr>
<tr>
<td>19</td>
<td>8</td>
<td>12</td>
<td>2013</td>
<td>Renzi</td>
<td>Tra qualche minuto a Tortona con Realacci, Basioli, Angelantoni, Ghisolfi con le proposte su a</td>
</tr>
<tr>
<td>20</td>
<td>8</td>
<td>12</td>
<td>2013</td>
<td>Renzi</td>
<td>Un gigante, #Mandela. Ciao #Madibaba</td>
</tr>
</tbody>
</table>
2. A limited number of medium-large texts

---

2008 election

Hi Chicago! If there is anyone out there who still doubts that America is a place where all things are possible; who still wonders if the dream of our founders is alive in our time; who still questions the power of our democracy, tonight is your answer. [...] This is our chance to answer that call. This is our moment. This is our time – to put our people back to work and open doors of opportunity for our kids; to restore prosperity and promote the cause of peace; to reclaim the American Dream and reaffirm that fundamental truth – that out of many, we are one; that while we breathe, we hope, and where we are met with cynicism, and doubt, and those who tell us that we can't, we will respond with that timeless creed that sums up the spirit of a people: Yes We Can. Thank you, God bless you, and may God Bless the United States of America.

2009 inauguration

All this we can do, all this we will do

My fellow citizens: I stand here today humbled by the task before us, grateful for the trust you have bestowed, mindful of the sacrifices borne by our ancestors. I thank President Bush for his service to our nation, as well as the generosity and cooperation he has shown throughout this transition. [...]
-> 3. A very large number of texts
4. "tidy corpus" perspective

The "tidy data" approach has a specific structure:

- each variable is a column
- each observation is a row
- each type of textual unit observed is a table

The "tidy text" format is a table with one word per row:

<table>
<thead>
<tr>
<th>N</th>
<th>parola</th>
<th>parola2</th>
<th>info</th>
<th>testo</th>
<th>dove</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Hi</td>
<td>hi</td>
<td></td>
<td>2008election</td>
<td>Chicago</td>
</tr>
<tr>
<td>2</td>
<td>Chicago</td>
<td>chicago</td>
<td>NER_place</td>
<td>2008election</td>
<td>Chicago</td>
</tr>
<tr>
<td>3</td>
<td>!</td>
<td>!</td>
<td></td>
<td>2008election</td>
<td>Chicago</td>
</tr>
<tr>
<td>4</td>
<td>If</td>
<td>if</td>
<td></td>
<td>2008election</td>
<td>Chicago</td>
</tr>
<tr>
<td>5</td>
<td>there</td>
<td>there</td>
<td></td>
<td>2008election</td>
<td>Chicago</td>
</tr>
<tr>
<td>6</td>
<td>is</td>
<td>is</td>
<td></td>
<td>2008election</td>
<td>Chicago</td>
</tr>
<tr>
<td>7</td>
<td>anyone</td>
<td>anyone</td>
<td></td>
<td>2008election</td>
<td>Chicago</td>
</tr>
<tr>
<td>8</td>
<td>out</td>
<td>out</td>
<td></td>
<td>2008election</td>
<td>Chicago</td>
</tr>
<tr>
<td>9</td>
<td>there</td>
<td>there</td>
<td></td>
<td>2008election</td>
<td>Chicago</td>
</tr>
<tr>
<td>10</td>
<td>who</td>
<td>who</td>
<td></td>
<td>2008election</td>
<td>Chicago</td>
</tr>
<tr>
<td>11</td>
<td>still</td>
<td>still</td>
<td></td>
<td>2008election</td>
<td>Chicago</td>
</tr>
<tr>
<td>12</td>
<td>doubts</td>
<td>doubts</td>
<td>SENT_negative</td>
<td>2008election</td>
<td>Chicago</td>
</tr>
<tr>
<td>13</td>
<td>that</td>
<td>that</td>
<td></td>
<td>2008election</td>
<td>Chicago</td>
</tr>
<tr>
<td>14</td>
<td>America</td>
<td>america</td>
<td>NER_place</td>
<td>2008election</td>
<td>Chicago</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Size of corpora

The main strength of the statistical analysis of textual data is the opportunity to extract information from large corpora, which represent a source of complex unstructured data, overcoming the obstacles posed by the amount of text which, conversely, represents the main limitation for qualitative analyzes.

The availability of software tools makes it possible to extract information from large corpora and results become more reliable as the size grows.

These software tools are not suitable for studying small corpora and this proves to be a problem in many disciplines where obtaining large text sets is complicated.
Textual data

Textual data that can be analyzed are of a different type. In linguistics there is a distinction between:

- phonetics
- grammar (morphology and syntax)
- lexicon

In this part of the course we will focus on:

- **lexicon** (lexical-based analysis)
- **bag-of-words** approaches
NLP and text representation

Methods that can “naturally” process languages to extract information, methods that “simulate” the process of our brain when we are reading or listening to people speaking.

NLP is a mix of artificial intelligence (AI), computer science and computational linguistics (CL) to extract meaning from documents, recognize text and, ultimately, model and shape text in order to compose original contents.
Natural Language content analysis

1. Natural language processing & text representation
2. Word association mining & analysis
3. Topic mining & analysis
4. Opinion mining & sentiment analysis
5. Text-based prediction

Real World → Perceive (Perspective) → Observed World → Express (English) → Text Data
Mining knowledge about language

Natural Language Content analysis is a very, very difficult task.

The 3 **steps** +4/5 step of NLP:

- **Semantic analysis**
  - Dog(d1).
  - Boy(b1).
  - Playground(p1).
  - Chasing(d1,b1,p1).

- **Syntactic analysis** (Parsing)
  - A person saying this may be reminding another person to get the dog back.

- **Inference**
  - Scared(b1)

- **Lexical analysis** (part-of-speech tagging)
  - A dog is chasing a boy on the playground

- **Pragmatic analysis** (speech act)
  - Scared(x) if Chasing(_,x,_)
NLP is difficult

Computers are far from being able to understand natural language:

- we omit a lot of **common sense** knowledge, which we assume the hearer/reader possesses.
- we keep a lot of **ambiguities**, which we assume the hearer/reader knows how to resolve.

Robust and general NLP tends to be **shallow** while deep understanding doesn’t scale up.

- Deep NLP requires common sense knowledge and inferences, thus only working for very limited domains
- Shallow NLP based on statistical methods can be done in large scale and is thus more broadly applicable
How do we represent text data

A dog is chasing a boy on the playground

Sentence
Verb Phrase
Noun Phrase
Det Noun Aux Verb Det Noun Prep Det Noun
Noun Phrase Complex Verb Noun Phrase Prep Phrase

A dog CHASE A boy ON the playground

Animal Person Location

Dog(d1). Boy(b1). Playground(p1). Chasing(d1,b1,p1).

Speech Act = REQUEST

String of characters
Sequence of words + POS tags
+ Syntactic structures
+ Entities and relations
+ Logic predicates
+ Speech acts

Deeper NLP: requires more human effort; less accurate

Closer to knowledge representation
## Text Representation & Enabled Analysis

<table>
<thead>
<tr>
<th>Text Rep</th>
<th>Generality</th>
<th>Enabled Analysis</th>
<th>Examples of Application</th>
</tr>
</thead>
<tbody>
<tr>
<td>String</td>
<td></td>
<td>String processing</td>
<td>Compression</td>
</tr>
<tr>
<td>Words</td>
<td></td>
<td>Word relation analysis; topic analysis; sentiment analysis</td>
<td>Thesaurus discovery; topic and opinion related applications</td>
</tr>
<tr>
<td>+ Syntactic</td>
<td></td>
<td>Syntactic graph analysis</td>
<td>Stylistic analysis; structure-based feature extraction</td>
</tr>
<tr>
<td>structures</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ Entities &amp;</td>
<td></td>
<td>Knowledge graph analysis; information network analysis</td>
<td>Discovery of knowledge and opinions about specific entities</td>
</tr>
<tr>
<td>relations</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>+ Logic</td>
<td></td>
<td>Integrative analysis of scattered knowledge; logic inference</td>
<td>Knowledge assistant for biologists</td>
</tr>
<tr>
<td>predicates</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Word-based representation

The most common text representation is **word-based** representation

- General and robust: applicable to (almost) any natural language
- No/little manual effort
- “Surprisingly” powerful for many applications (not all!)
- Can be combined with more sophisticated representations
Bag-of-word approach

It refers to a text (document) as **a bag of words** that can be extracted without considering the order.

It assumes that all the terms have the same importance in a document. Even, but not always) it makes no distinction between different parts of speech (verbs, nouns, adjectives, etc.).
N-gram

A possible extension of the analysis (without considering syntax/semantics) may consider N-grams. A **N-gram** is a sequence of \( n \) words (or even of characters). A sort of **rolling window** of size \( n \): by moving this window by one position at the time, we obtain a list of new N-grams.

"The police stopped a vehicle without insurance"

We can build different N-grams by varying the length \( n \):

- \( n = 1 \) (**unigram**) returns {"The", "police", "stopped ", "a", "vehicle", "without ", "insurance"}
- \( n = 2 \) (**bigram**) returns {"The police", "police stopped ", "stopped a", "a vehicle", "vehicle without ", "without insurance"}
- \( n = 3 \) (**trigram**) returns {"The police stopped", "police stopped a", "stopped a vehicle", "a vehicle without ", "vehicle without insurance"}
Using a high value of $n$ means incorporating more context in the units of the document, while a low $n$ value means that the basic unit of data is going to be more granular.

An N-gram model can also be seen as a Markov chain of $m - 1$ elements, where $m$ is the length of the sentence (e.g., number of words).
Basics: corpus, text, ...

- a **corpus** is a collection of texts;
- a **text** is made up of letters, spaces and other symbols (for example: punctuation, numbers, mathematical symbols);
- a **word** is a sequence of letters isolated by means of separators (spaces and punctuation marks);

In order to count the words contained in the corpus it is necessary to distinguish two (or more) concepts ...
Basics: How do we count words?

*common sense is not so common*

\[ N = 6 \text{ word } \text{tokens} \]
\[ \text{(occurrences / total words)} \]
\[ V = 5 \text{ word } \text{types} \]
\[ \text{(different types / words)} \]
an eye for an eye and a tooth for a tooth

\[\begin{array}{llllllllll}
an & eye & for & an & eye & and & a & tooth & for & a & tooth \\
1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 \\
1 & 2 & 3 & 1 & 2 & 4 & 5 & 6 & 3 & 5 & 6 \\
. & 1 & . & . & 1 & . & . & 2 & . & . & 2 \\
\end{array}\]

\[N = 11 \text{ word tokens}\]
\[\mathcal{V} = 6 \text{ word types}\]

No. of content words = 2
A word token is an occurrence of a word type in the text.

Examples:

- "the" is a word type that has numerous word tokens in any English text
- in any text there are word types that have only one occurrence, that is, they occur only once (hapax legomena or simply hapax)

The number \( N \) of word tokens is the **size of the corpus** (size in terms of occurrences or number of total words);

The number \( V \) of word types is the **size of the vocabulary** (size in terms of the number of different words);
The **frequency** of a word type is the number of corresponding word tokens in the corpus (number of repetitions, possibly expressed in the form of relative frequency or rate);

The list of word types with frequencies is the **vocabulary (of frequency)** of the corpus.

When vocabulary is the basis of statistical analysis, the approach is **bag-of-words**.
Basics: vocabulary

*Give a man a fish and you feed him for a day.*
*Teach a man to fish and you feed him for a lifetime.*

*Is it ok?*
*No, it isn't ...*
*(Typically, in pre-processing, upper/lower case must be handled)*

- N = 24 word tokens
- V = 13 word types

For a software tool package "Give" is not the same as "give" --> a pre-processing phase is required that performs a normalization of texts (for example, keep only the relevant capital letters, or transform everything into lowercase)
## Basics: vocabulary

<table>
<thead>
<tr>
<th>word type</th>
<th>frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>(entry)</td>
<td>(occurrences)</td>
</tr>
<tr>
<td>a</td>
<td>5</td>
</tr>
<tr>
<td>and</td>
<td>2</td>
</tr>
<tr>
<td>feed</td>
<td>2</td>
</tr>
<tr>
<td>fish</td>
<td>2</td>
</tr>
<tr>
<td>for</td>
<td>2</td>
</tr>
<tr>
<td>him</td>
<td>2</td>
</tr>
<tr>
<td>man</td>
<td>2</td>
</tr>
<tr>
<td>you</td>
<td>2</td>
</tr>
<tr>
<td>day</td>
<td>1</td>
</tr>
<tr>
<td>give</td>
<td>1</td>
</tr>
<tr>
<td>lifetime</td>
<td>1</td>
</tr>
<tr>
<td>teach</td>
<td>1</td>
</tr>
<tr>
<td>to</td>
<td>1</td>
</tr>
<tr>
<td>CORPUS</td>
<td>24</td>
</tr>
</tbody>
</table>
Basics: Using the language of statistics

The two different ways of counting words correspond to crucial concepts for understanding the logic of statistical analysis of textual data.

\[ N = \text{the number of observations of the corpus (statistical units)} \]

\[ V = \text{the number of modes (or number of categories of the "vocabulary" variable)} \]

The vocabulary is a frequency table (data summary).

The vocabulary is a representation of the corpus.
Basics: Sizes

In the manuals once we read:

A great corpus must exceed 100 thousand words
A corpus of 500 thousand words is a good basis for building a frequency lexicon

Today

corpora of millions of occurrences are quite common

(OK, we are in the era of BIG DATA, digital humanities, etc.)

However ... Are we facing the problem from the right perspective?
Basics: Sizes

Some questions are:

- **How many words** (word tokens) are enough to work with a quantitative approach?
- Is the number of words (word tokens) a good measure of corpus size?
- Is the number of different words (word types) a good measure of the (lexical) richness of the corpus?
Basics: Sizes

To get a very general idea if a corpus is large enough to allow an effective use of quantitative methods, there are empirical measures based on size and lexical richness:

- **TTR - Type-Token Ratio** = ratio of number of word types / number of word tokens (in%)
  - the TTR should be **less than 20%**

- **% of hapax** = number of hapax / number of word types (in%)
  - hapax = word type that only one occurrence in the corpus
  - the hapax% should be **less than 50%**
Basics: Sizes

Warning! These are empirical measures to be taken very carefully. They are largely experienced on some languages (Italian, French), less on others (English).

The basic idea is that the use of statistical tools based on word frequency requires **redundancy**.
Bag-of-words and software

The fundamental difference with computer-assisted qualitative data analysis software (CAQDAS) is that here we work on the frequency of elementary units such as sequences of characters (what we commonly call "words") and not on complex units such as portions of text.

They start from a lexical perspective and not from the idea of "abstraction of concepts" and "labeling" portions of text (sentences / periods / paragraphs of complete meaning).

Examples:
- Alceste
- AntConc
- Iramuteq
- Hyperbase
- KH Coder
- Lexico
- R stylo
- Spad-T (today: Spad)
- Tableau
- Taltac
- Textable (Orange)
- T-Lab
- TXM
- Voyant-tools

+ packages available in statistical environments such as R, SAS, Python.
For a first exploration try:

AntConc: https://www.laurenceanthony.net/software.html

Cite:
Big data: Let's not confuse!

A large dataset (here, corpus) does not fall under the definition of big data.

Big data must be called into question only if V V V (3V)

1. the **volume** of information is large (the dataset must exceed the processing capacity of a personal computer)
2. information is characterized by **velocity** (the dataset must be constantly updated or modified)
3. information is of **various** kinds (the data set contains data, text, audio, video, etc.)

... and after 3V we go on with Veracity, Value, Variability ... complexity ...
But we are working in the "big data era". If one day we will have

- a large corpus (Volume),
- who experiences changes and updates (Velocity)
- linked to other sources of information (Variety)

such a unstructured, large, unstable and heterogeneous dataset might be depicted as Big data and we will also need integrated software which is not currently available.
Introduction to data transformation
Representing text data

From unstructured data (i.e., a collection of text documents) to structured data.

**document x term matrix** (or the reverse, **term x document matrix**): a matrix on whose columns we place every unique term in the corpus and on whose rows the document IDs (or the transpose).

When terms are present/not present in documents, the cells will take the value 1/0 (the simplest weighting).

Three important properties: **Sparsity**, **Non-negativity** and **Side information**.

To reduce the complexity and dimension of corpus representation, we can apply a text **normalization**: - eliminate punctuation, - reduce the number of features, e.g., prepositions, “generic” verbs (“be”, “have”), - and so on.
Tokenization is a process that consists in breaking up texts (documents) into elements called **tokens**, which are used as an input to text mining procedures. Tokens might be similar but stylistically very different from one another, e.g.:

- some have capital letters while others have numbers or punctuation signs in them.
- some tokens have no meaning and thus would add no information to the documents’ representation.

To solve these issues, **normalization** is applied to the tokens to obtain the most meaningful subset of them.

Examples of non-relevant tokens are: case-folding, punctuation and numbers, hyperlinks and spelling. E.g., difference between British English and American **spelling**: e.g., ”Analyze” - ”Analyse” / ”Program” - ”Programme” / ”Center” - ”Centre”
Stop words

Once the tokenization process is completed, we reduce the number of features to select the most significant ones by removing the so-called stop words.

Stop words are basically a set of commonly used words in any language, not just English.

To deal with stop words we use a stop words dictionary, which is a collection of all the terms we consider unnecessary to our specific problem and thus can be removed.
Once the removal of numbers, punctuation, stop words, etc. has been completed, we can further simplify the features using a stemming or a lemmatization process.

In English, as well as in other languages, verbs have different inflectional forms or suffixes to which they are related. For example:

a. see, saw, seen → see
b. insurance, insurer, insure, insured → insur

**Stemming** is merely a heuristic process that truncates the end of every word to reduce it to its common base, also called *root*.
Lemmatization

To improve the accuracy of the truncation process we can use another technique called **lemmatization**.

Basically, instead of defining rules to truncate words, we use a lemmatizer, which carries out a full morphological analysis to accurately identify the lemma for each word.

Lemmatization removes inflection, suffixes and compresses words into the so-called **lemma**.

We define a lemma as the canonical form of a term, deprived of most conjugation suffixes and transformations.

Intuitively, we can think of a lemma as the word we look up in a dictionary when searching for a specific term.
Introduction to vector space models
A document can be seen as a vector whose dimensions are given by the number of features.

This representation is called “Document Term Matrix” (DTM).

Placing documents into a multi-dimensional space requires an accurate coordinate system.

Ideally, documents that are similar are also close to each other, while documents that are semantically different need be distant.

Using only term frequency (tf) may not be enough to capture similarities.

Hence, the tf-idf (term frequency–inverse document frequency) technique is typically used; this is a statistic that reflects the importance of terms in a corpus.
In general, a term is more important than others when it occurs multiple times in a document and when it is also rarer than other terms in the corpus.

The peculiar lexicon is a terminological vocabulary that contains the most significant and representative keywords.

The td-idf statistic combines two quantities:

1. term frequency
2. document frequency
Peculiarity and rarity
Term frequency

Term frequency is the frequency of the term $t$ in the document $d$.

$$tf(t, d) = f_{t,d} = \frac{c(t, d)}{|d|}$$

where $c(t, d)$ is the number of occurrences (or count) of term $t$ in document $d$, while $|d|$ is the total number of occurrences (or count) of terms in document $d$. 
Document Frequency

Document frequency indicates the inverse document frequency of the term $t$ in the collection of $m$ documents. It is equal to the log of the ratio between the number of documents in the collection, $m$, and the number of documents $m_t$ in which the term $t$ appears:

$$\text{idf}(t, m) = \log \left( \frac{m}{m_t} \right)$$
**Term frequency - Inverse document frequency**

\[ tfidf = tf(t, d) \cdot idf(t, m) \]

\( tfidf \) assumes a high value in the case of a high term frequency and a low frequency in the collection.

The statistics enables us to describe documents by considering **peculiar** terms and avoiding a representation based on common terminology.

Words that appear only once/quite rarely in a document will have a high \( tf - idf \) value and influence the position of the document in the multi-dimensional space.
Word Association Mining & Analysis
Word Association Mining

1. Natural language processing and text representation
2. Word association mining and analysis
3. Topic mining and analysis
4. Opinion mining and sentiment analysis
5. Text-based prediction

Real World → Perceive (Perspective) → Observed World → Express (English) → Text Data
Basic word relations: paradigmatic vs syntagmatic

- **Paradigmatic**: A & B have paradigmatic relation if they can be substituted for each other (i.e., A & B are in the same class)
  - E.g., “cat” and “dog”; “Monday” and “Tuesday”

- **Syntagmatic**: A & B have syntagmatic relation if they can be combined with each other (i.e., A & B are related semantically)
  - E.g., “cat” and “sit”; “car” and “drive”

- These two basic and complementary relations can be generalized to describe relations of any items in a language
Why mine word associations

They are useful for improving accuracy of many NLP tasks:

- POS tagging, parsing, entity recognition, acronym expansion
- Grammar learning

They are directly useful for many applications in text retrieval and mining

- Text retrieval (e.g., use word associations to suggest a variation of a query)
- Automatic construction of topic map for browsing: words as nodes and associations as edges
- Compare and summarize opinions (e.g., what words are most strongly associated with “battery” in positive and negative reviews about iPhone 6, respectively?)
Mining paradigmatic associations: idea

Paradigmatic: similar context

My cat eats fish on Saturday
His cat eats turkey on Tuesday
My dog eats meat on Sunday
His dog eats turkey on Tuesday

How similar are context ("cat") and context ("dog")?
How similar are context ("cat") and context ("computer")?
Mining syntagmatic associations: idea

Syntagmatic: correlated occurrences

My cat eats fish on Saturday
His cat eats turkey on Tuesday
My dog eats meat on Sunday
His dog eats turkey on Tuesday
...

What words tend to occur to the left of “eats”? What words to the right?

Whenever “eats” occurs, what other words also tend to occur?
How helpful is the occurrence of “eats” for predicting occurrence of “meat”?
How helpful is the occurrence of “eats” for predicting occurrence of “text”?
Mining word associations: general ideas

Paradigmatic

- Represent each word by its context
- Compute context similarity
- Words with high context similarity likely have paradigmatic relation

Syntagmatic

- Count how many times two words occur together in a context (e.g., sentence or paragraph)
- Compare their co-occurrences with their individual occurrences
- Words with high co-occurrences but relatively low individual occurrences likely have syntagmatic relation

Paradigmatically related words tend to have syntagmatic relation with the same word → joint discovery of the two relations
Discovering paradigmatic relations

Word context as "pseudo-document"

Left1("cat") = {"my", "his", "big", "a", "the", ...}

My ___ eats fish on Saturday
His ___ eats turkey on Tuesday
...

Right1("cat") = {"eats", "ate", "is", "has", ....}

Window8("cat") = {"my", "his", "big", "eats", "fish", ...}

Context = pseudo document = “bag of words”
Context may contain adjacent or non-adjacent words
Discovering paradigmatic relations

Measuring Context Similarity

\[
\text{Sim}("\text{Cat}", "\text{Dog}") = \\
\text{Sim}(\text{Left1}("\text{cat}"), \text{Left1}("\text{dog}")) \\
+ \text{Sim}(\text{Right1}("\text{cat}"), \text{Right1}("\text{dog}")) + \\
\ldots \\
+ \text{Sim}(\text{Window8}("\text{cat}"), \text{Window8}("\text{dog}"))) = ?
\]

\text{High} \text{ sim} (\text{word1}, \text{word2}) \\
\rightarrow \text{word1 and word2 are paradigmatically related}
Bag-of-words and VSM

Each document is represented as a frequency vector in a V-dimensional space.
VSM for paradigmatic relation mining

1. How to compute each vector?
2. How to measure similarity?

Many approaches are possible (most developed originally for text retrieval).
Expected Overlap of Words in Context (EOWC)

Probability that a randomly picked word from \( d_1 \) is \( w_i \)

\[
d_1 = (x_1, \ldots, x_N) \quad x_i = \frac{c(w_i, d_1)}{|d_1|}
\]

\[
d_2 = (y_1, \ldots, y_N) \quad y_i = \frac{c(w_i, d_2)}{|d_2|}
\]

Total counts of words in \( d_1 \)

\[
\text{Sim}(d_1, d_2) = d_1 \cdot d_2 = x_1 y_1 + \ldots + x_N y_N = \sum_{i=1}^{N} x_i y_i
\]

Probability that two randomly picked words from \( d_1 \) and \( d_2 \), respectively, are identical.
Would EOWC Work Well?

Intuitively, EOWC makes sense: The more overlap the two context documents have, the higher the similarity would be. However:

- It **favors** matching one frequent term very well over matching more distinct terms.
- It **treats every word equally** (overlap on “the” isn’t as so meaningful as overlap on “eats”).
Improving EOWC

Improving EOWC with Retrieval Heuristics

- It favors matching one frequent term very well over matching more distinct terms.

  ⇒ **Sublinear transformation of Term Frequency** (TF)

- It treats every word equally (overlap on “the” isn’t as so meaningful as overlap on “eats”).

  ⇒ **Reward matching a rare word**: IDF term weighting
TF Transformation

\[ c(w,d) \rightarrow TF(w,d) \]
TF Transformation: BM25 Transformation

Tune $k \in [0, +\infty)$

$$y = \text{TF}(w,d)$$

$x = c(w,d)$
IDF Weighting: Penalizing Popular Terms

The importance of each term is tuned this way:

$$\text{IDF}(W) = \log\left(\frac{M+1}{k}\right)$$

where $M$ is the total number of documents in the collection, $k$ is the total number of documents containing term $W$, and $\log$ is the logarithm base 10.
Adapting BM25 Retrieval Model for Paradigmatic Relation Mining

\[ d_1 = (x_1, \ldots, x_N) \]

\[
BM25(w_i, d_1) = \frac{(k + 1)c(w_i, d_1)}{c(w_i, d_1) + k(1 - b + b^* \cdot |d_1| / \text{avdl})}
\]

\[
x_i = \frac{BM25(w_i, d_1)}{\sum_{j=1}^{N} BM25(w_j, d_1)}
\]

\[ d_2 = (y_1, \ldots, y_N) \]

\[ y_i \text{ is defined similarly} \]

\[
Sim(d_1, d_2) = \sum_{i=1}^{N} \text{IDF}(w_i)x_i \cdot y_i
\]

\[ b \in [0, 1] \]

\[ k \in [0, +\infty) \]
BM25 can also Discover Syntagmatic Relations

\[ d_1 = (x_1, \ldots, x_N) \]

\[
BM25(w_i, d_1) = \frac{(k + 1)c(w_i, d_1)}{c(w_i, d_1) + k(1 - b + b^* |d_1| / \text{avdl})}
\]

\[
x_i = \frac{BM25(w_i, d_1)}{\sum_{j=1}^{N} BM25(w_j, d_1)}
\]

IDF-weighted \( d_1 = (x_1 * \text{IDF}(w_1), \ldots, x_N * \text{IDF}(w_N)) \)

The highly weighted terms in the context vector of word \( w \) are likely syntagmatically related to \( w \).
Computing similarity

Having represented the documents in a space, it is also interesting to compare pairs of documents to understand document similarities.

The scoring function is obtained as follows: given \( d_h = (x_1, \ldots, x_N) \) and \( d_k = (y_1, \ldots, y_N) \), the similarity between the two contexts is

\[
\text{Sim}(d_h, d_k) = \sum_i \text{IDF}(w_i) x_i y_i
\]

Syntagmatic relations can also be discovered as a “by product”: in

\[
\text{IDF-weighted } d_j = (\text{IDF}(w_1)x_1, \ldots, \text{IDF}(w_N)x_N)
\]

the highly weighted terms in the context vector of word \( w \) are likely syntagmatically related to \( w \).
Cosine and correlation similarity

Another scoring function for computing similarity between documents is the **cosine similarity**.

Given two documents, let $V_1$ and $V_2$ be the two vectors containing the tf-idf values for the $M$ terms in the collection. The cosine similarity is defined as

$$
cos(\theta) = \frac{V_1 \cdot V_2}{|V_1||V_2|}
$$

A further possible function measuring association is the **correlation**. It is more often used for positioning terms. The formula is an analog of the cosine formula, after centering the two vectors.