TechTalentLab

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

Text Mining: Principles for Automated Qualitative Data Analysis

Dr. Vicenc Fernandez

Edition 2020

Gerd Altmann at Pixabay

Before you get started…

"If you try to boil the ocean, you will never succeed"

- Kaenan Hertz (Ernst & Young) ury

"The simplest algorithms will get better results long term"

- José M Gómez (Pragsis Technologies)

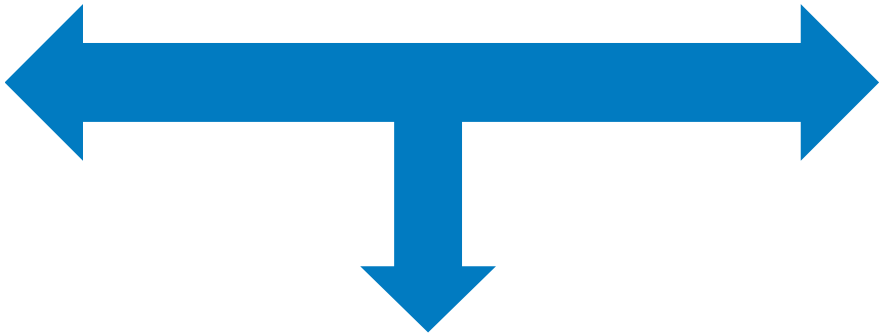
Some Previous Considerations

Qualitative Research

Quantitative Research

Some Previous Considerations

Qualitative Research

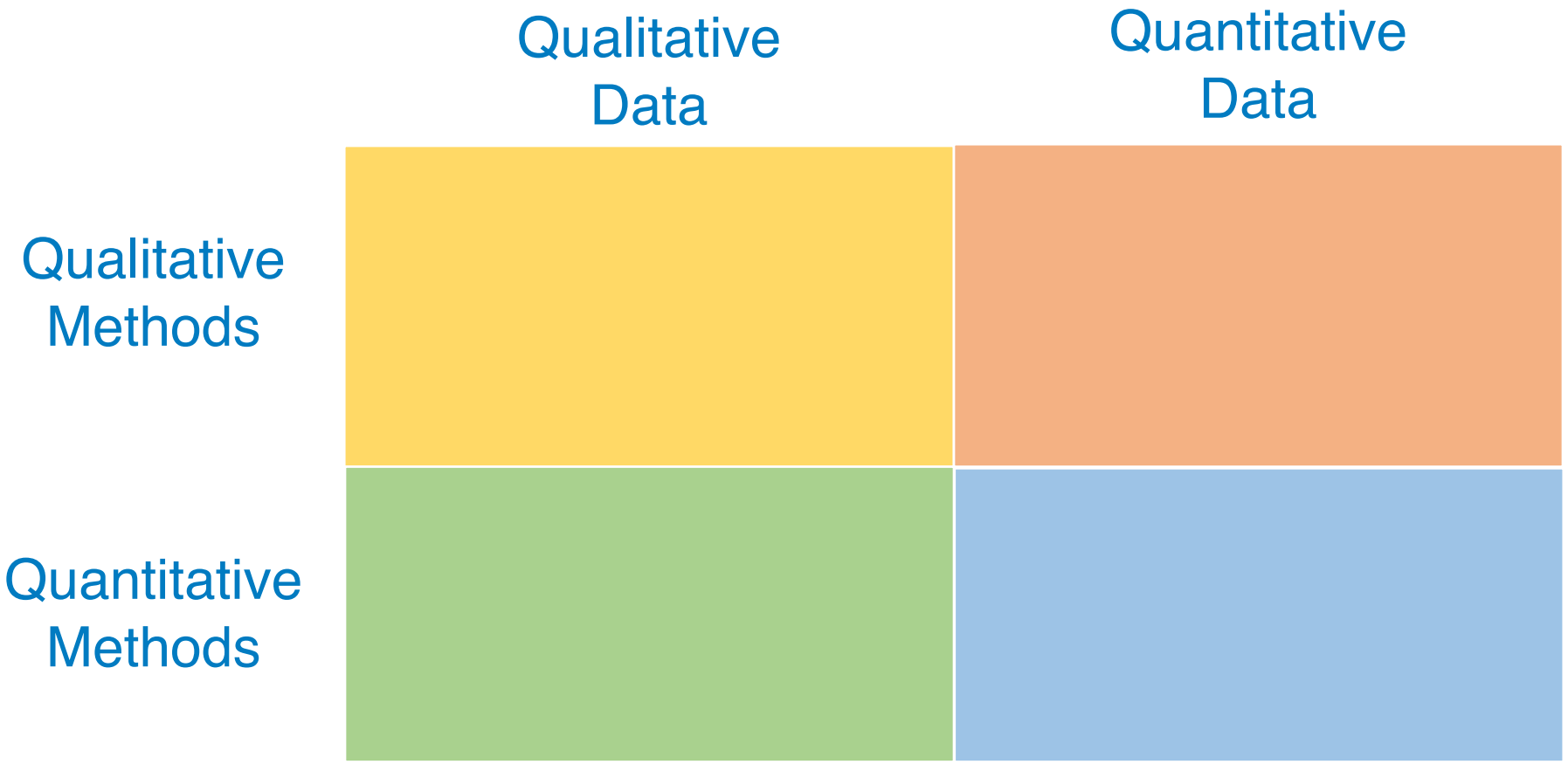


Quantitative Research

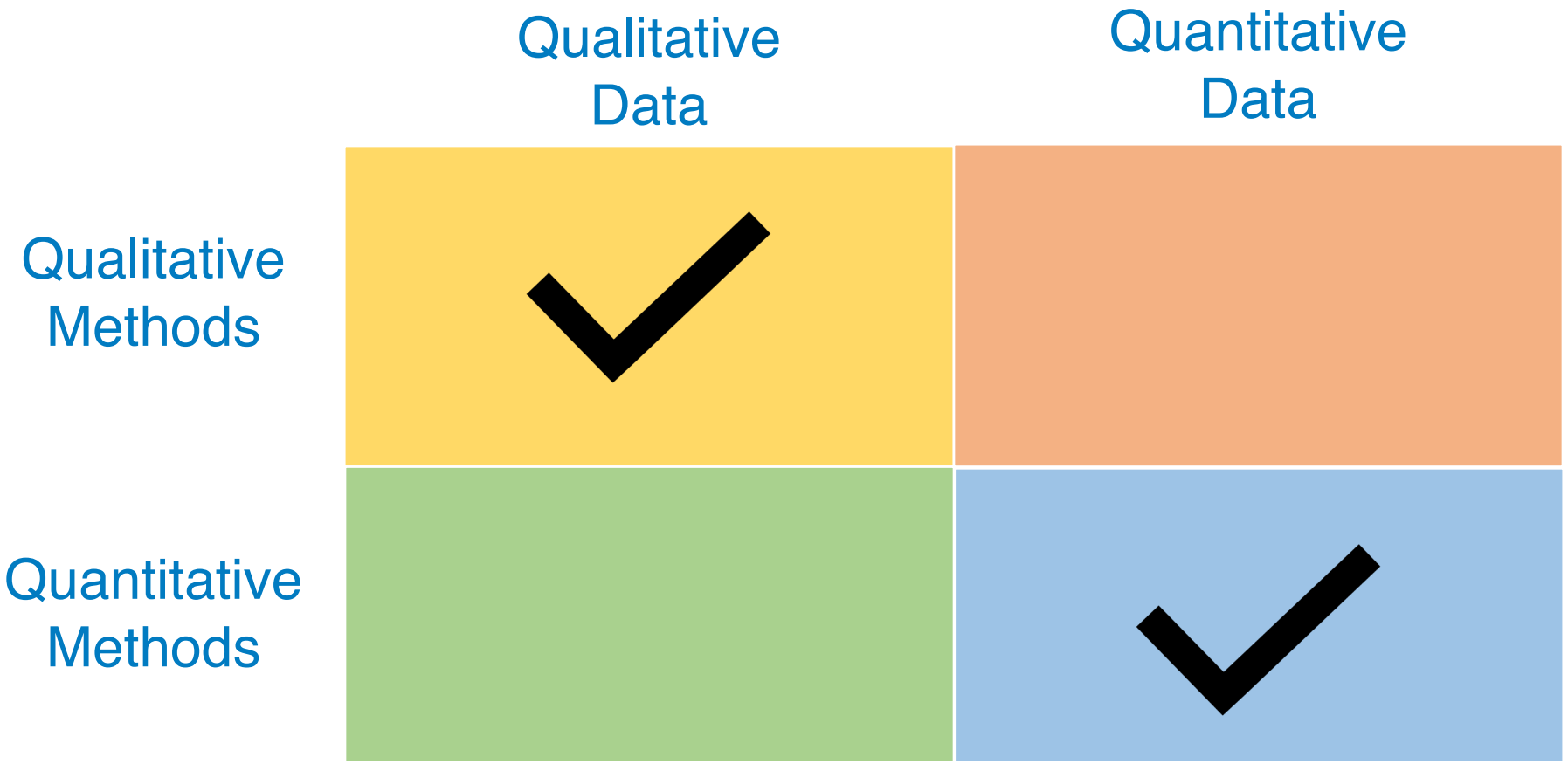
Mix-Methods Research

?

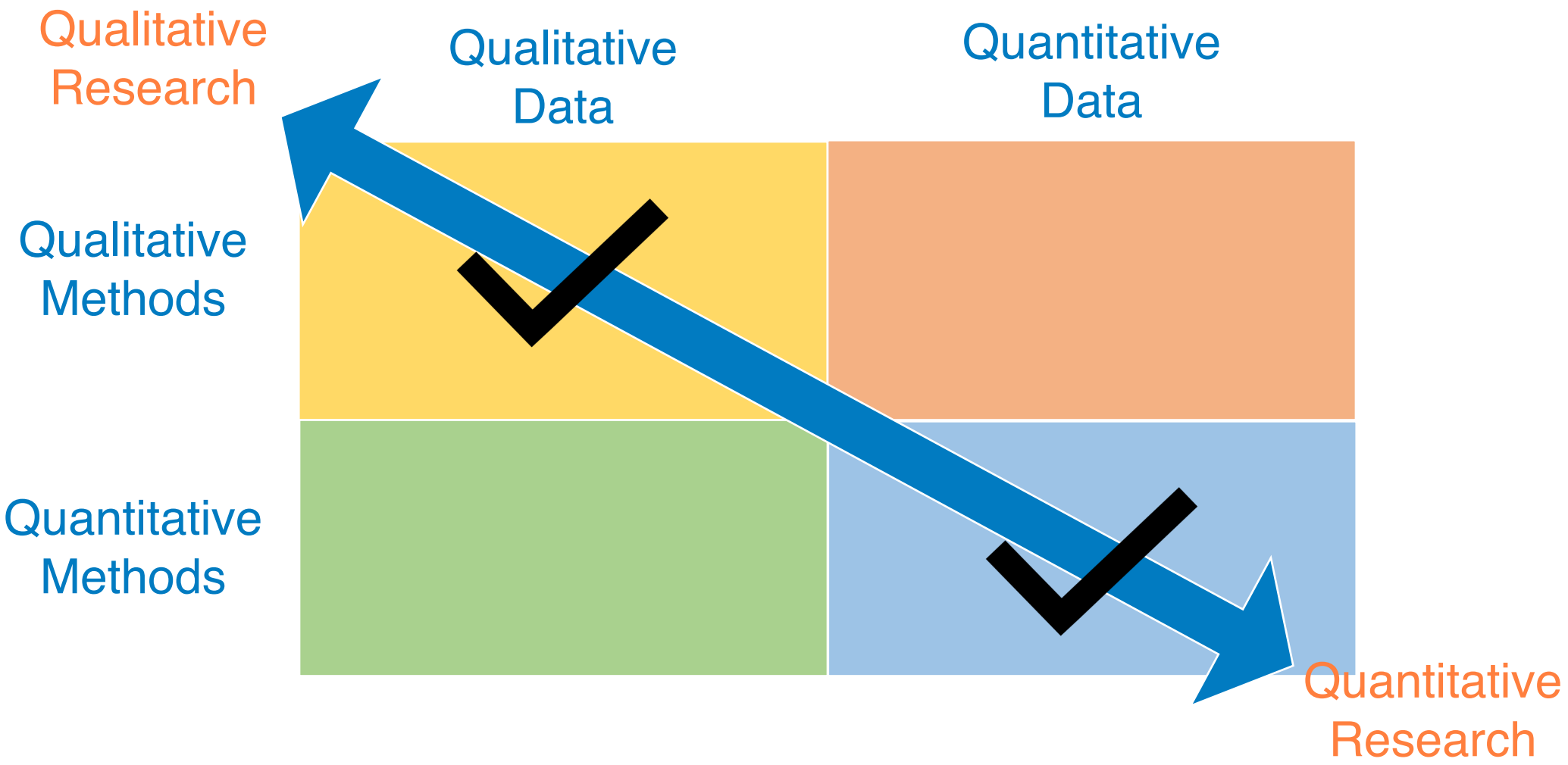
Some Previous Considerations



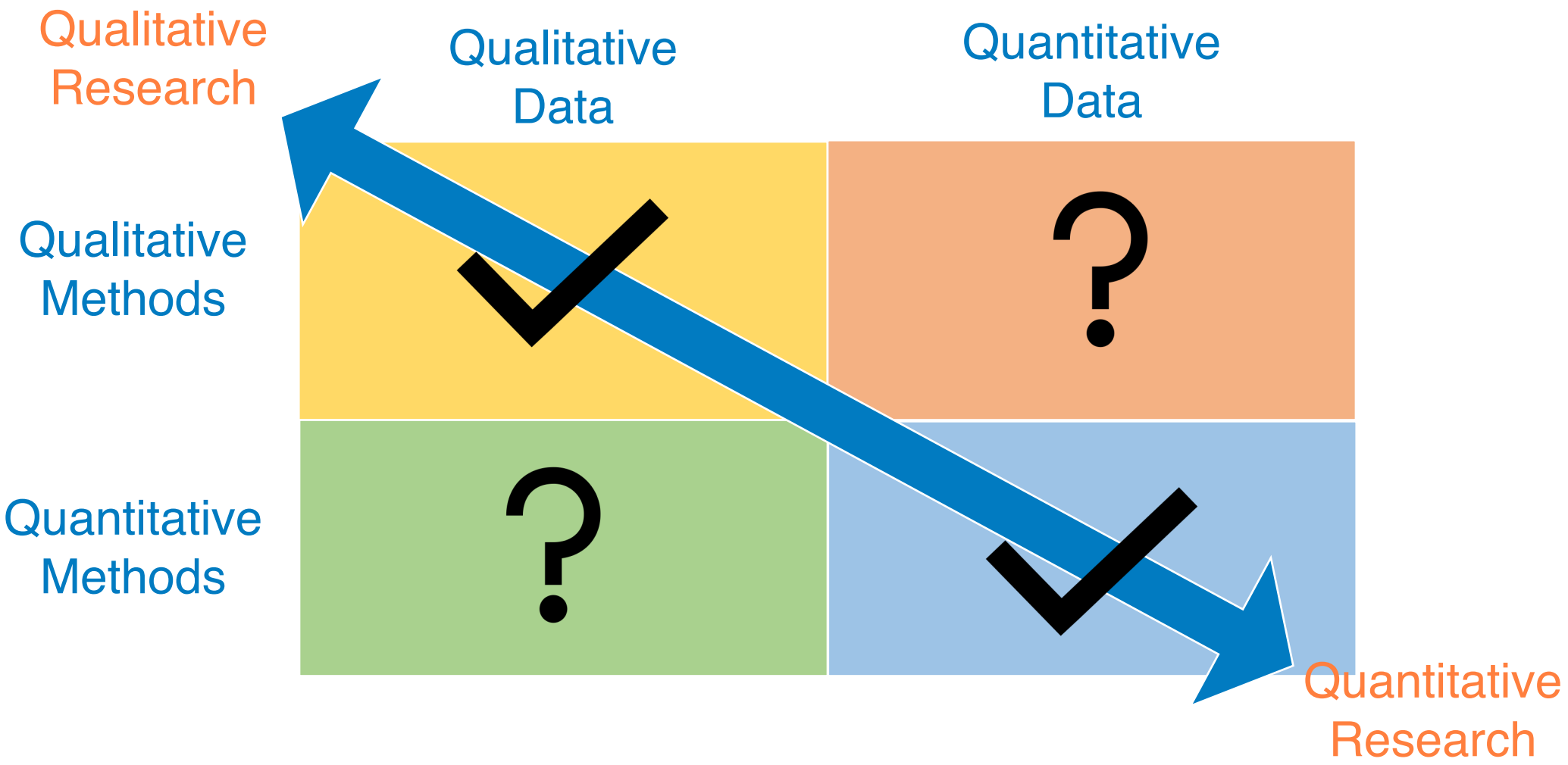
Some Previous Considerations



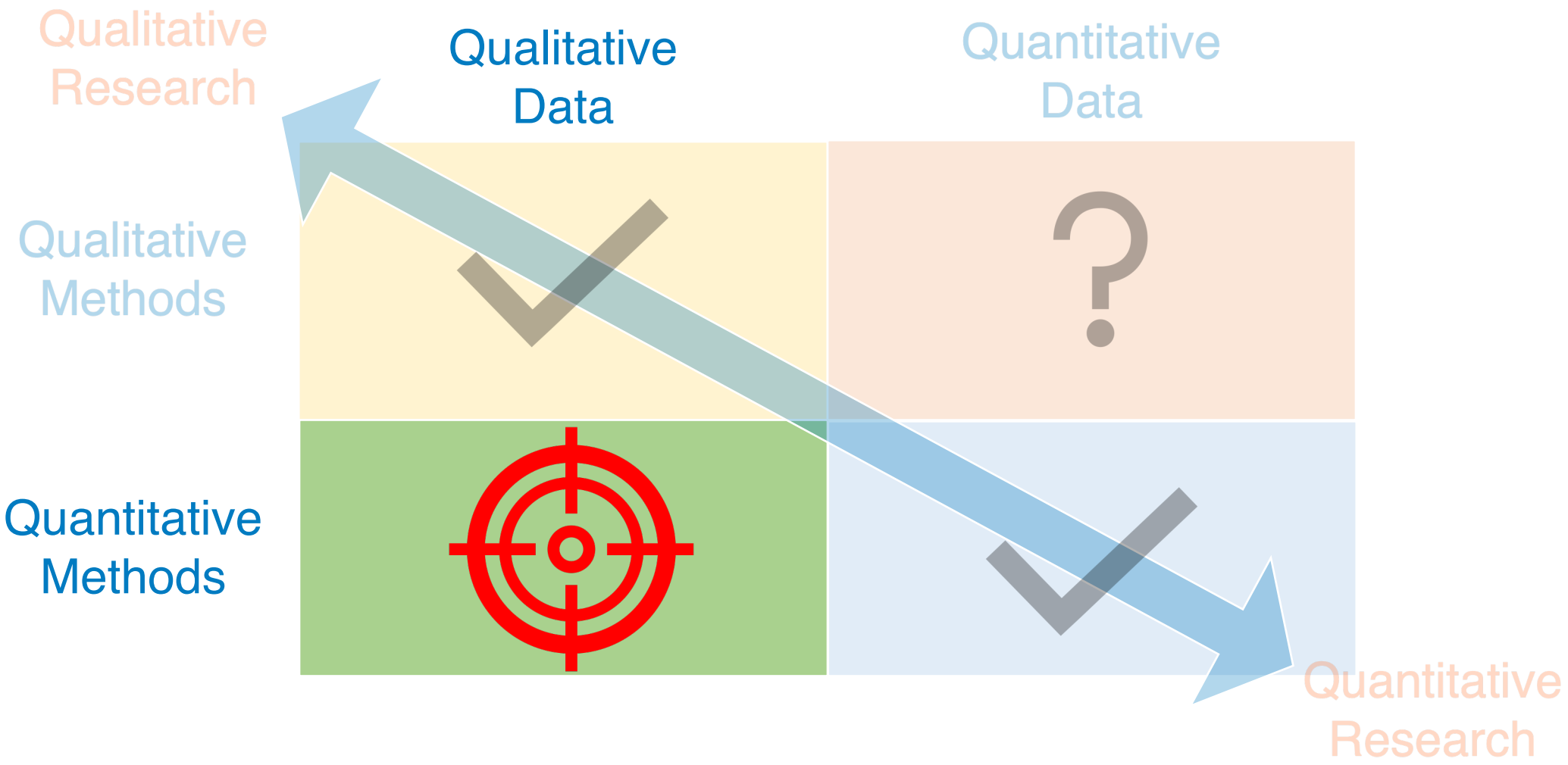
Some Previous Considerations



Some Previous Considerations



Some Previous Considerations



Agenda

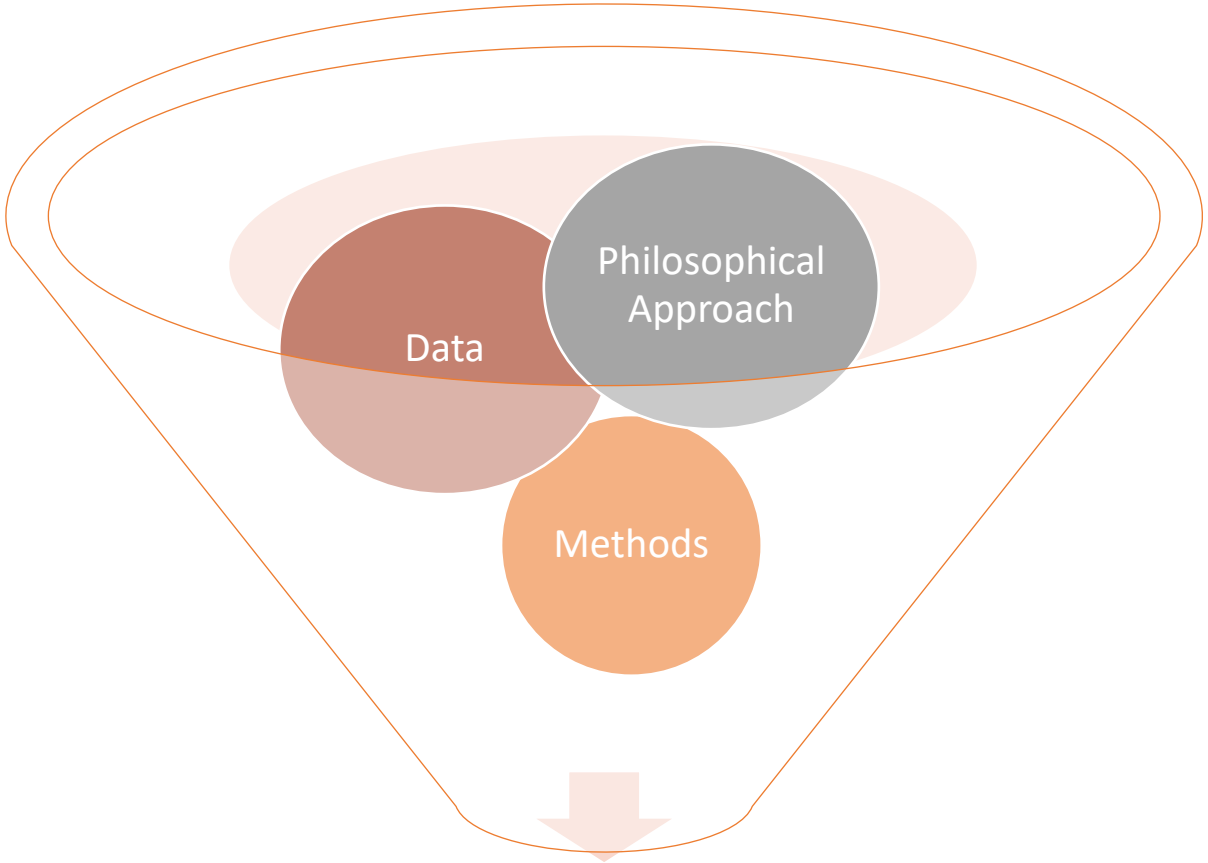
- ✓The Three Branches of Text-Mining
- ✓Philosophical Approach
- ✓Data
- ✓Methods: Tools
- ✓Methods: Analysis

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

TechTalentLab

The Three Branches of Text-Mining

Introduction



Text Mining

Philosophical Approach

Text Analysis Approaches

Conversation Analysts

- How people negotiate the meaning of the conversation.
- How people use language pragmatically to define the situations in which they find themselves.

Analyzing Discourse Positions

- Typical discursive roles that people adopt in their everyday communication practices.
- A way of linking texts to the social spaces in which they have emerged.

Text Analysis Approaches

Critical Discourse Analysis

- A qualitative approach to text analysis.
- Seeking the presence of features from other discourses in the text or discourse to be analyzed.

Content Analysis

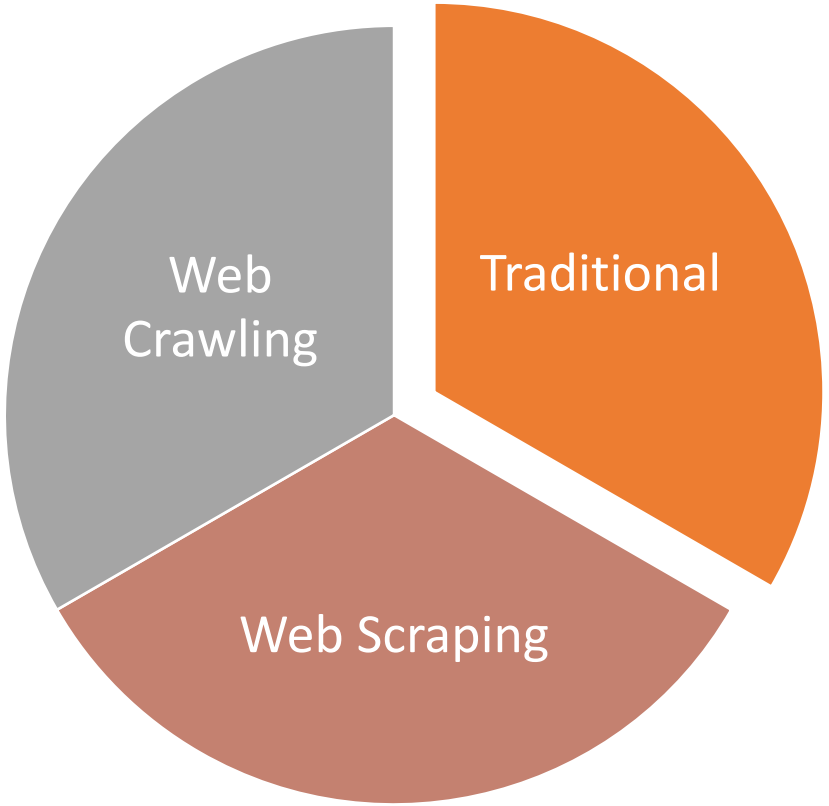
- A quantitative approach to text analysis.
- Focused on texts themselves.

Qualitative Data

Which Qualitative Data Gathered Methods can we use?

?

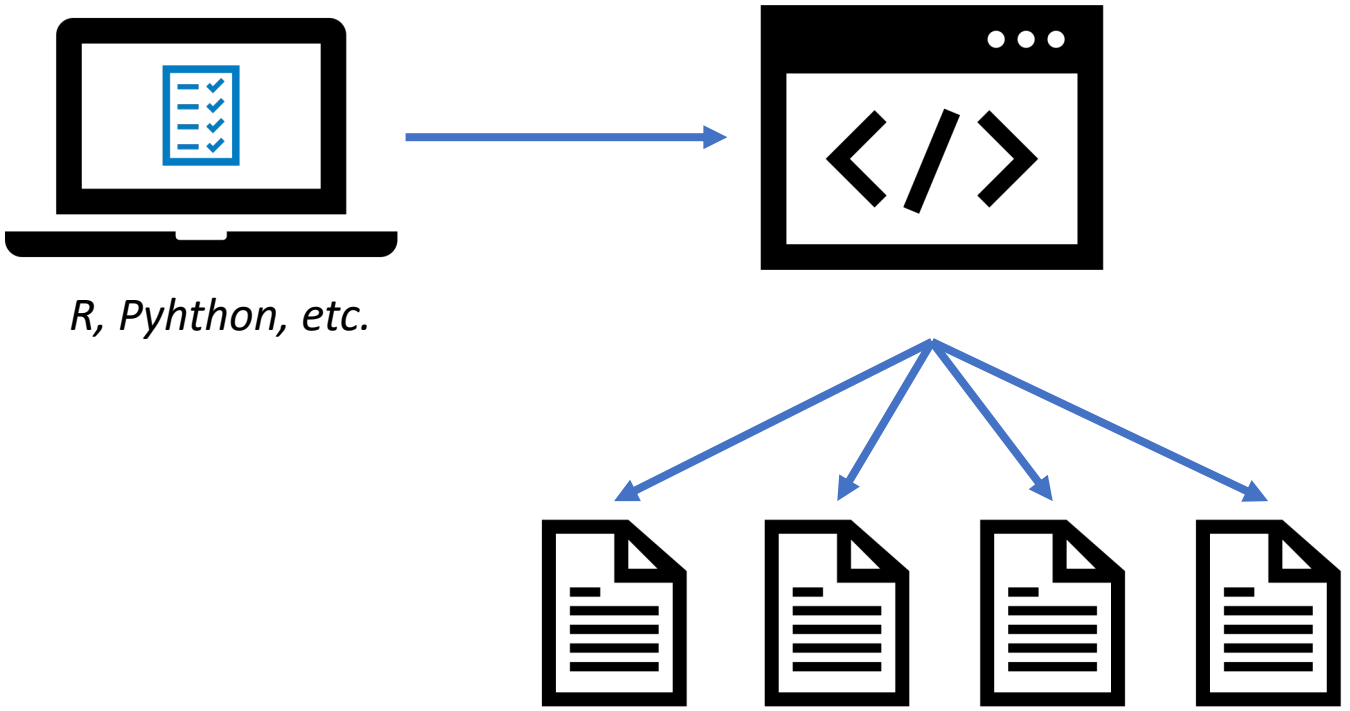
Methods of Acquiring Text Data



Traditional Techniques

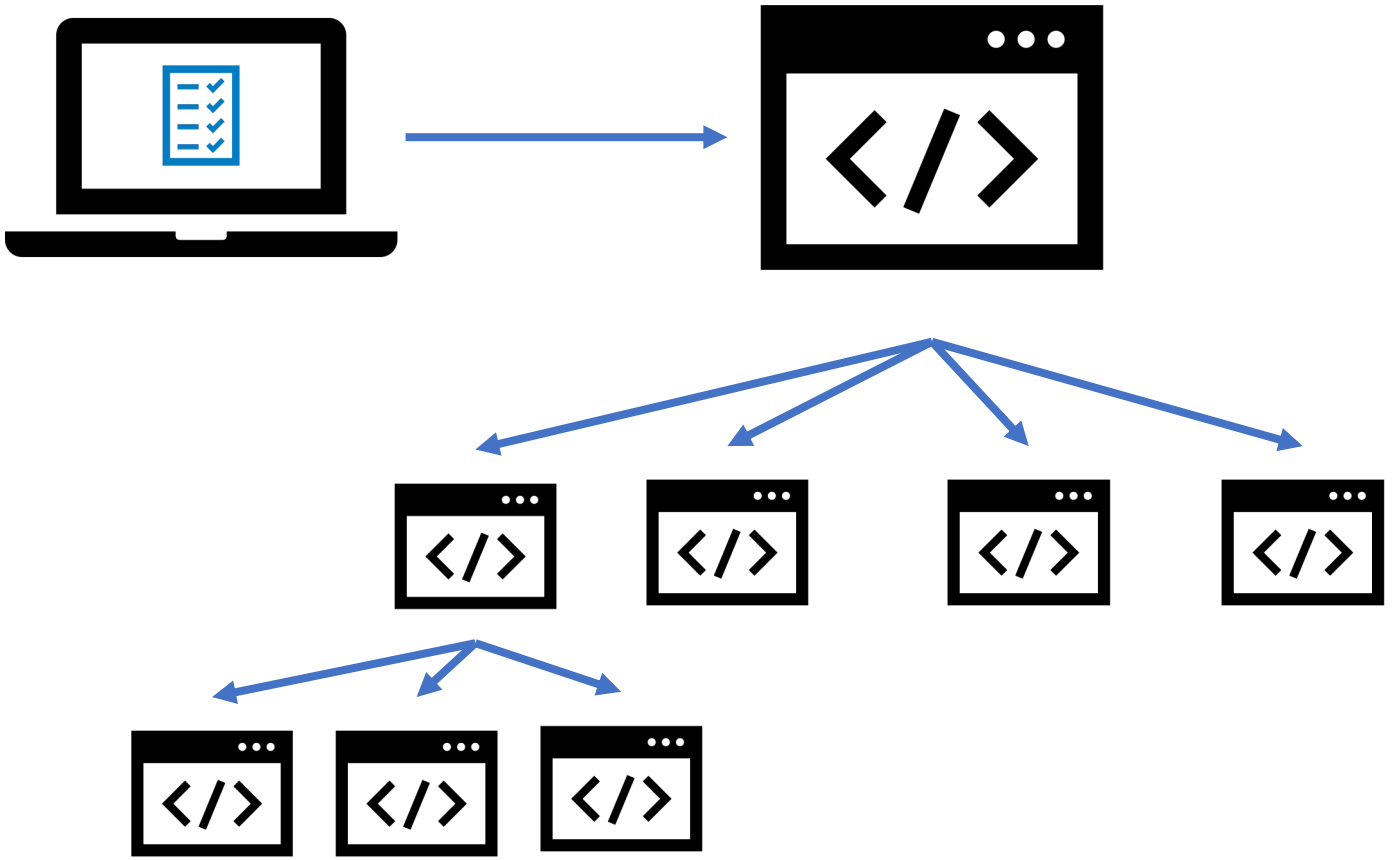
- Interview
- Survey with open questions
- Set of reports
- Newspapers
- Observation

Web Scraping



R, Pyhthon, etc.

Web Crawling



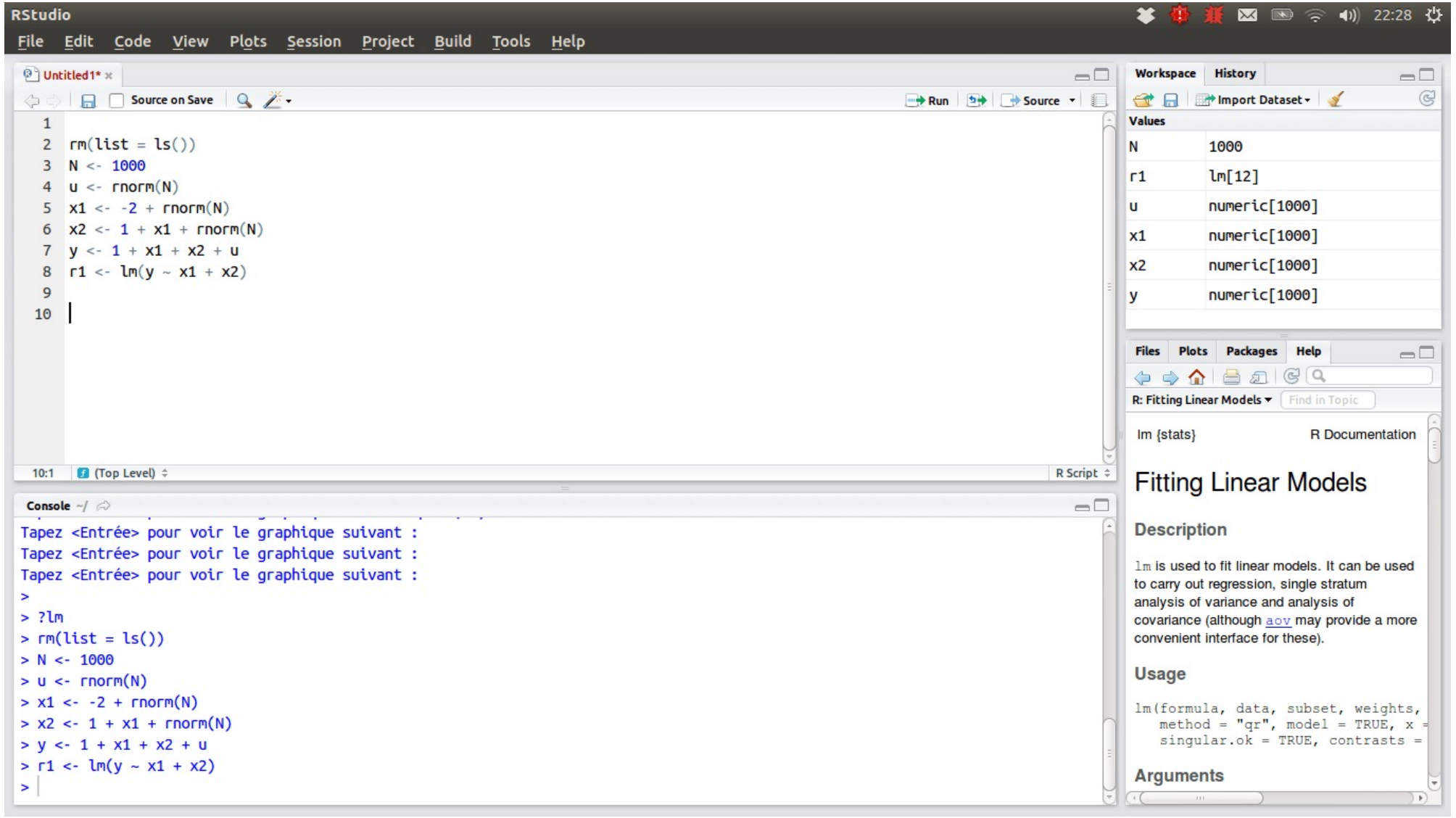
Tools

Tools for Text Mining

https://www.r-project.org/

https://python.org/

Tools for Text Mining



Tools for Text Mining

LEXIMANCER

https://info.leximancer.com

LINGUISTIC INQUIRY AND WORD COUNT

http://liwc.wpengine.com

RAPIDMINER

http://rapidminer.com

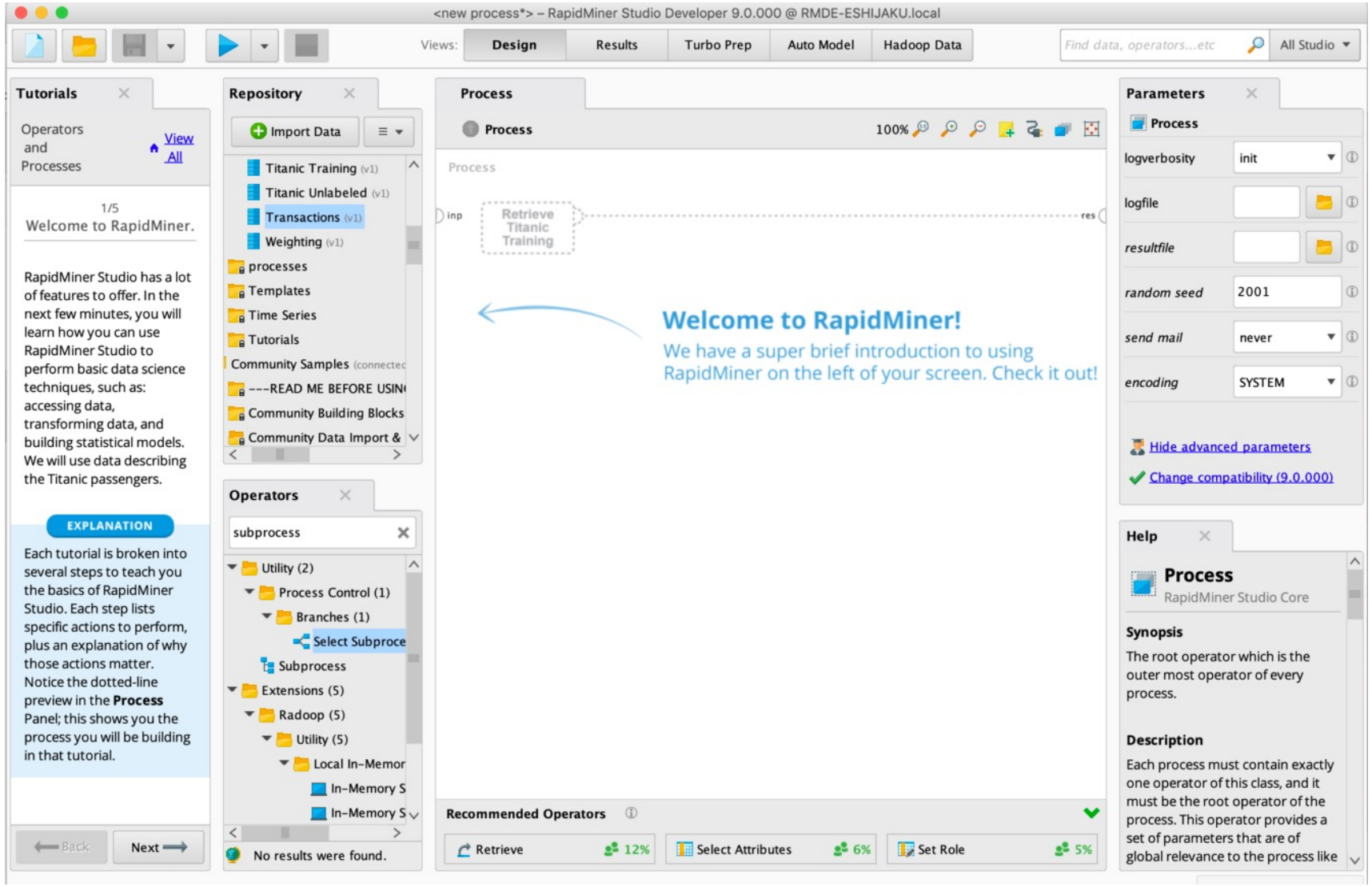
TEXTANALYST

http://megaputer.com/site/textanalyst.php

WORDSTAT

https://provalisresearch.com/products/content-analysis-software

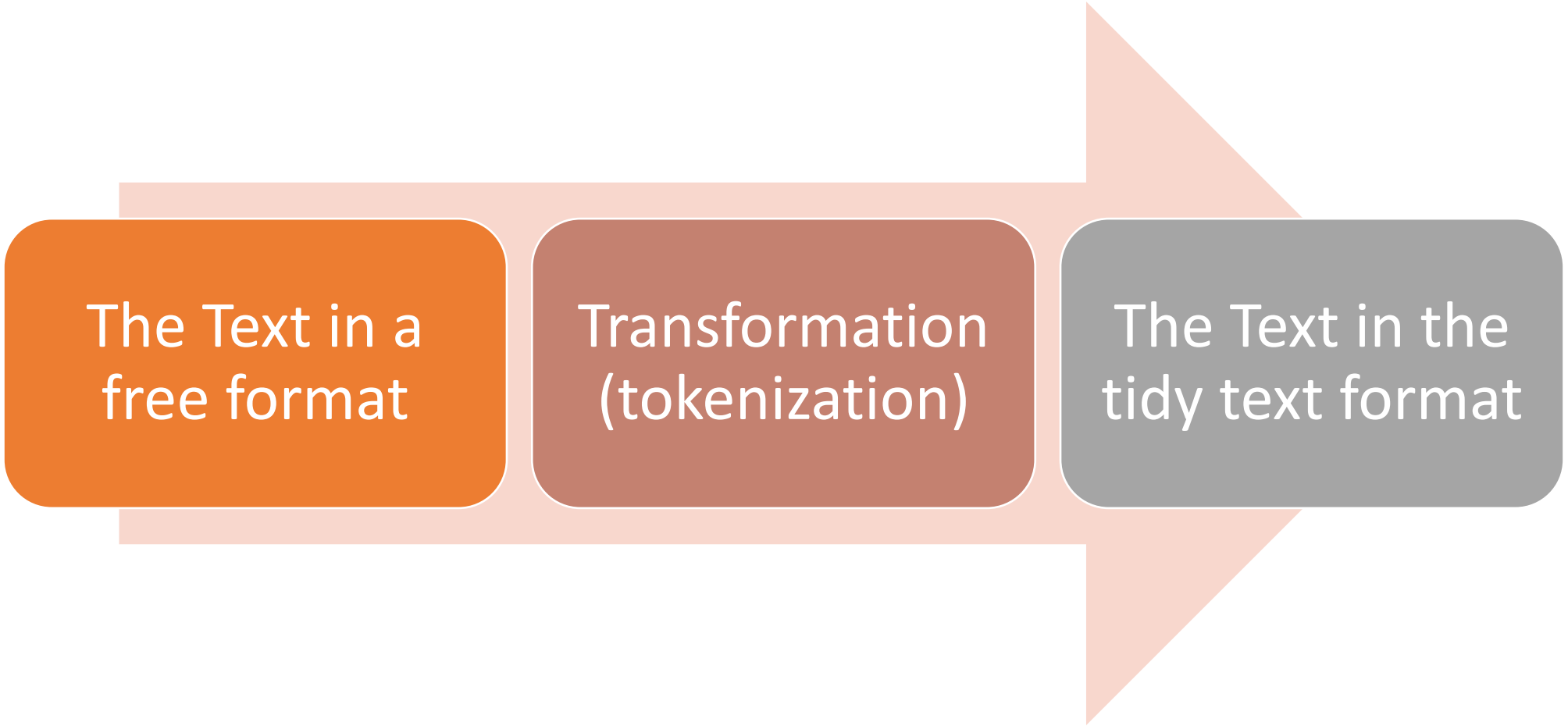
Tools for Text Mining

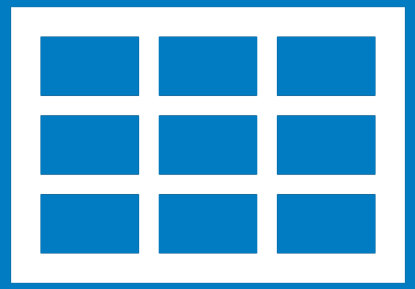


?

Methods

Pre-Process of the Text

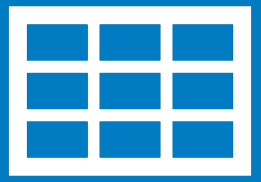




Tidy Data Structure

- Each variable is a column
- Each observation (token) is a row
- Each type of observational unit is a table

Tidy Data Structure



It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us, we were all going direct to Heaven, we were all going direct the other way

tokenization

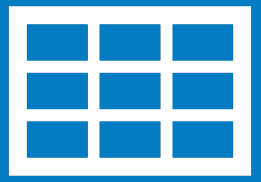
Option 1

Token
It
was
the
best
of
times
it
was
the
worst
of
times
…

Option 2

Token
It was the best of times
it was the worst of times
it was the age of wisdom
it was the age of foolishness
it was the epoch of belief
it was the epoch of incredulity
it was the season of Light
it was the season of Darkness
it was the spring of hope,
it was the winter of despair
we had everything before us
we had nothing before us
…

Tidy Data Structure



It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us, we were all going direct to Heaven, we were all going direct the other way

tokeniz ation

Option 3

Token	Chapter	Page
It	1	1
was	1	1
the	1	1
best	1	1
of	1	1
times	1	1
it	1	1
was	1	1
the	1	2
worst	1	2
of	1	2
times	1	2
…	…	…

Question?

Which are the most common words in any language?

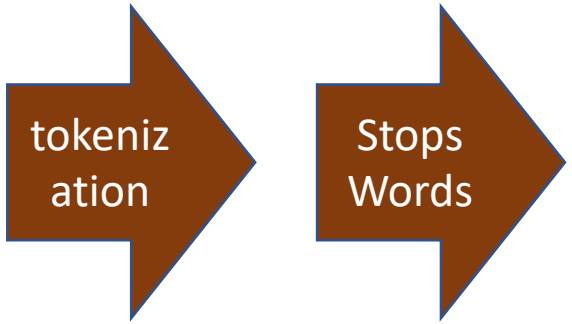
Which is the most common word in a novel?

Removing Stop Words

Stop words are words that are not useful for an analysis, typically extremely common words such as “the”, “of”, “to”, and so forth in English.

Removing Stop Words

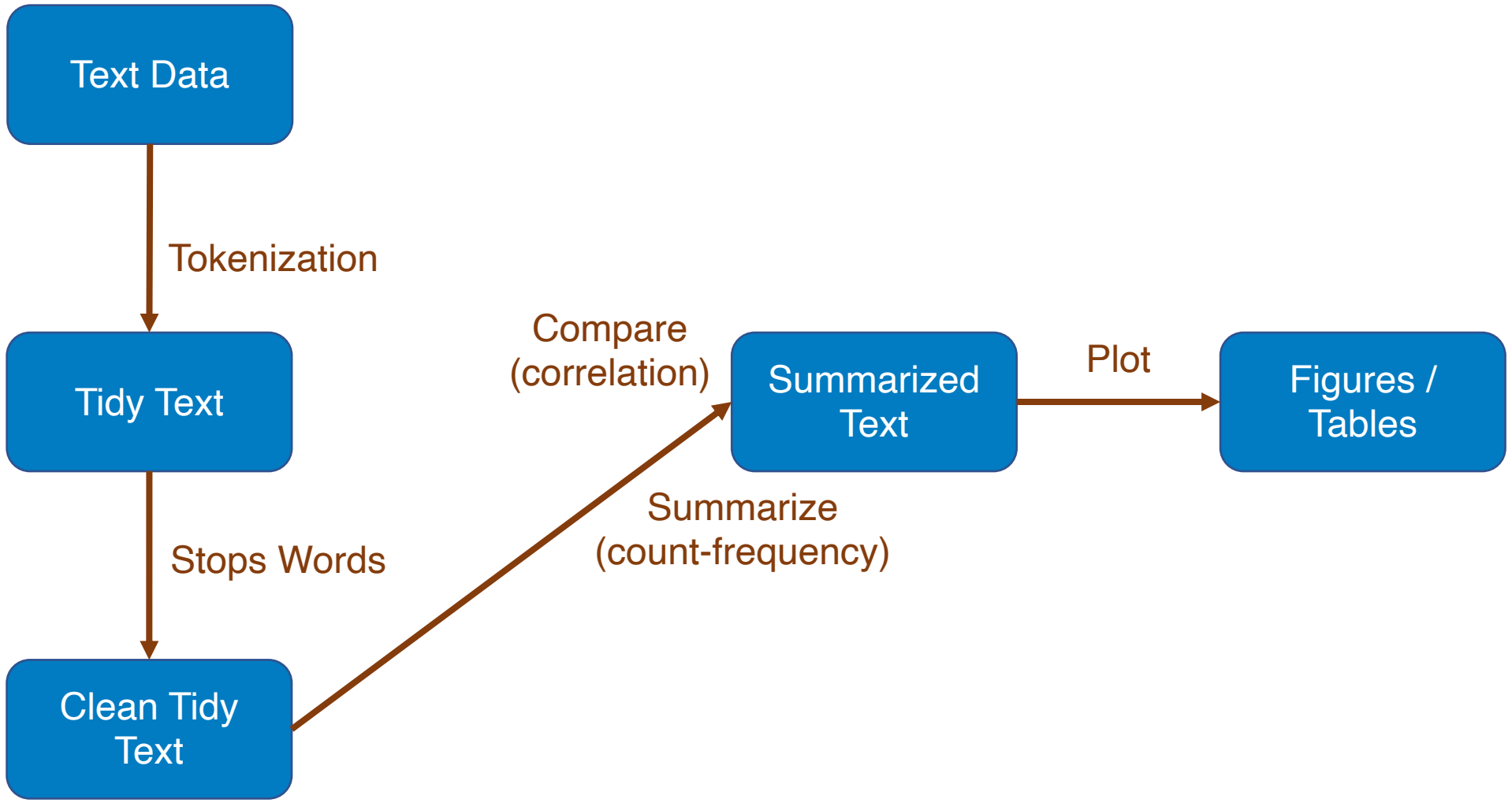
It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us, we were all going direct to Heaven, we were all going direct the other way



Option 4

Token
it
was
best
times
it
was
worst
times
it
was
age
wisdom
…

Basic Analysis



Methods

word and document frequency

Introduction

- ***term frequency (tf),*** how frequently a word occurs in a document

- ***inverse document frequency (idf),*** which decreases the weight for commonly used words and increases the weight for words that are not used very much in a collection of documents.

- ***The statistic tf-idf*** is intended to measure how important a word is to a document in a collection (or corpus) of documents

Example

Document 1
The water bottle is very common in most houses. If we do a study on how many bottles of water there are in each house, we can discover that the average is three bottles of water in each house.

Document 2
Water is, together with air, one of the most important elements we can find in the world. There is no house that has water fountains, as well as water inlets and outlets. In Spanish houses there is always hot water.

Document 3
The main drink in Spain is water. In all the houses it is possible to drink water at any time of the day. In many houses, water is as important as the air they breathe. It is somewhat similar in the rest of the world.

Which is the most important word in these documents?

Example

Document 1
The water bottle is very common in most houses. If we do a study on how many bottles of water there are in each house, we can discover that the average is three bottles of water in each house.

Document 2
Water is, together with air, one of the most important elements we can find in the world. There is no house that has water fountains, as well as water inlets and outlets. In Spanish houses there is always hot water.

Document 3
The main drink in Spain is water. In all the houses it is possible to drink water at any time of the day. In many houses, water is as important as the air they breathe. It is somewhat similar in the rest of the world.

Which is the most distinctive word in document 1 among all documents?

Example

Document 1
The water bottle is very common in most houses. If we do a study on how many bottles of water there are in each house, we can discover that the average is three bottles of water in each house.

tf (water) = ↑↑

Document 2
Water is, together with air, one of the most important elements we can find in the world. There is no house that has water fountains, as well as water inlets and outlets. In Spanish houses there is always hot water.

tf (water) = ↑↑

Document 3
The main drink in Spain is water. In all the houses it is possible to drink water at any time of the day. In many houses, water is as important as the air they breathe. It is somewhat similar in the rest of the world.

tf (water) = ↑↑

Example

Document 1	Document 2	Document 3
The water bottle is very common in most houses. If we do a study on how many bottles of water there are in each house, we can discover that the average is three bottles of water in each house.	Water is, together with air, one of the most important elements we can find in the world. There is no house that has water fountains, as well as water inlets and outlets. In Spanish houses there is always hot water.	The main drink in Spain is water. In all the houses it is possible to drink water at any time of the day. In many houses, water is as important as the air they breathe. It is somewhat similar in the rest of the world.
tf (water) = ↑↑	tf (water) = ↑↑	tf (water) = ↑↑

df (water) = ↑↑ → idf (water) = ↓↓

Example

Document 1	Document 2	Document 3
The water bottle is very common in most houses. If we do a study on how many bottles of water there are in each house, we can discover that the average is three bottles of water in each house.	Water is, together with air, one of the most important elements we can find in the world. There is no house that has water fountains, as well as water inlets and outlets. In Spanish houses there is always hot water.	The main drink in Spain is water. In all the houses it is possible to drink water at any time of the day. In many houses, water is as important as the air they breathe. It is somewhat similar in the rest of the world.
tf (water) = ↑↑	tf (water) = ↑↑	tf (water) = ↑↑
tf-idf (water) = ↓↓	tf-idf (water) = ↓↓	tf-idf (water) = ↓↓

df (water) = ↑↑ → idf (water) = ↓↓

Example

Document 1
The water bottle is very common in most houses. If we do a study on how many bottles of water there are in each house, we can discover that the average is three bottles of water in each house.

tf (bottle) = ↑↑

Document 2
Water is, together with air, one of the most important elements we can find in the world. There is no house that has water fountains, as well as water inlets and outlets. In Spanish houses there is always hot water.

tf (bottle) = ↓↓

Document 3
The main drink in Spain is water. In all the houses it is possible to drink water at any time of the day. In many houses, water is as important as the air they breathe. It is somewhat similar in the rest of the world.

tf (bottle) = ↓↓

Example

Document 1	Document 2	Document 3
The water bottle is very common in most houses. If we do a study on how many bottles of water there are in each house, we can discover that the average is three bottles of water in each house.	Water is, together with air, one of the most important elements we can find in the world. There is no house that has water fountains, as well as water inlets and outlets. In Spanish houses there is always hot water.	The main drink in Spain is water. In all the houses it is possible to drink water at any time of the day. In many houses, water is as important as the air they breathe. It is somewhat similar in the rest of the world.
tf (bottle) = ↑↑	tf (bottle) = ↓↓	tf (bottle) = ↓↓

df (bottle) = ↓↓ → idf (bottle) = ↑↑

Example

Document 1	Document 2	Document 3
The water bottle is very common in most houses. If we do a study on how many bottles of water there are in each house, we can discover that the average is three bottles of water in each house.	Water is, together with air, one of the most important elements we can find in the world. There is no house that has water fountains, as well as water inlets and outlets. In Spanish houses there is always hot water.	The main drink in Spain is water. In all the houses it is possible to drink water at any time of the day. In many houses, water is as important as the air they breathe. It is somewhat similar in the rest of the world.
tf (bottle) = ↑↑	tf (bottle) = ↓↓	tf (bottle) = ↓↓

df (bottle) = ↓↓ → idf (bottle) = ↑↑

tf-idf (bottle) = ↑↑	tf-idf (bottle) = ↓↓	tf-idf (bottle) = ↓↓

Example

Document 1
The water bottle is very common in most houses. If we do a study on how many bottles of water there are in each house, we can discover that the average is three bottles of water in each house.

tf (car) = ↓↓

Document 2
Water is, together with air, one of the most important elements we can find in the world. There is no house that has water fountains, as well as water inlets and outlets. In Spanish houses there is always hot water.

tf (car) = ↓↓

Document 3
The main drink in Spain is water. In all the houses it is possible to drink water at any time of the day. In many houses, water is as important as the air they breathe. It is somewhat similar in the rest of the world.

tf (car) = ↓↓

Example

Document 1	Document 2	Document 3
The water bottle is very common in most houses. If we do a study on how many bottles of water there are in each house, we can discover that the average is three bottles of water in each house.	Water is, together with air, one of the most important elements we can find in the world. There is no house that has water fountains, as well as water inlets and outlets. In Spanish houses there is always hot water.	The main drink in Spain is water. In all the houses it is possible to drink water at any time of the day. In many houses, water is as important as the air they breathe. It is somewhat similar in the rest of the world.
tf (car) = ↓↓	tf (car) = ↓↓	tf (car) = ↓↓

df (car) = ↓↓ → idf (car) = ↑↑

Example

Document 1	Document 2	Document 3
The water bottle is very common in most houses. If we do a study on how many bottles of water there are in each house, we can discover that the average is three bottles of water in each house.	Water is, together with air, one of the most important elements we can find in the world. There is no house that has water fountains, as well as water inlets and outlets. In Spanish houses there is always hot water.	The main drink in Spain is water. In all the houses it is possible to drink water at any time of the day. In many houses, water is as important as the air they breathe. It is somewhat similar in the rest of the world.
tf (car) = ↓↓	tf (car) = ↓↓	tf (car) = ↓↓

df (car) = ↓↓ → idf (car) = ↑↑

tf-idf (car) = ↓↓	tf-idf (car) = ↓↓	tf-idf (car) = ↓↓

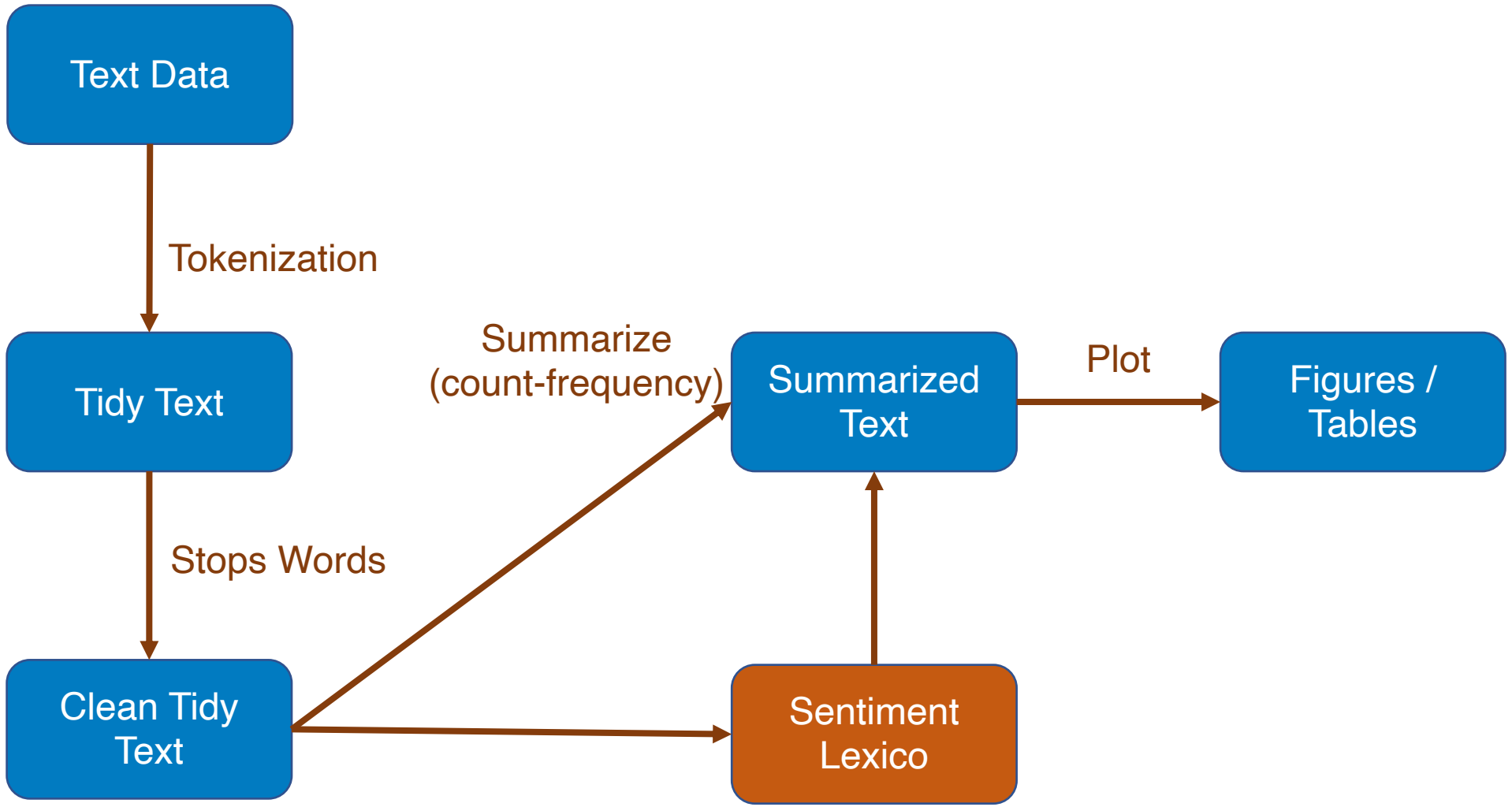
Methods

Sentiment Analysis

Introduction

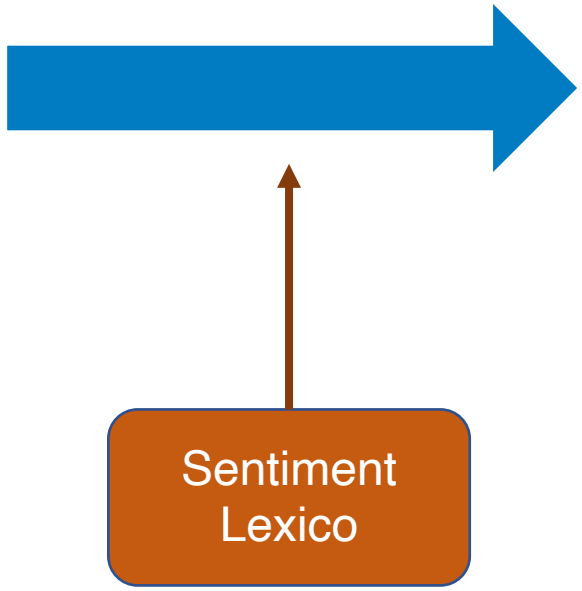
The use the tools of text mining to approach the emotional content of text programmatically

Basic Analysis



Basic Analysis

word
I
am
happy
and
I
get
a
gift
TOTAL



word	value
I	-
am	-
happy	positive
and	-
I	-
get	-
a	-
gift	positive
TOTAL	Positive (+)

Lexical Resources: Linguistic Inquiry and Word Count

Four LIWC categories, classes and sample words

Category	Class	Sample words
We	Linguistic processes	our, ourselves, we, we'd, we'll, us, let's, we've, we're, let's
Optimism	Psychological processes	accept, best, bold, certain, confidence, daring, determined, glorious, hope
Achievement	Personal concerns	better, award, ahead, advance, achieve, motivate, lose, honor, climb, first, fail
Nonfluencies	Spoken categories	er, umm, uh, um, zz

Source: LIWC (Linguistic Inquiry and Word Count), **https://liwc.wpengine.com/**

Pennebaker, Francis, and Booth (2001)

Lexical Resources: General Inquirer

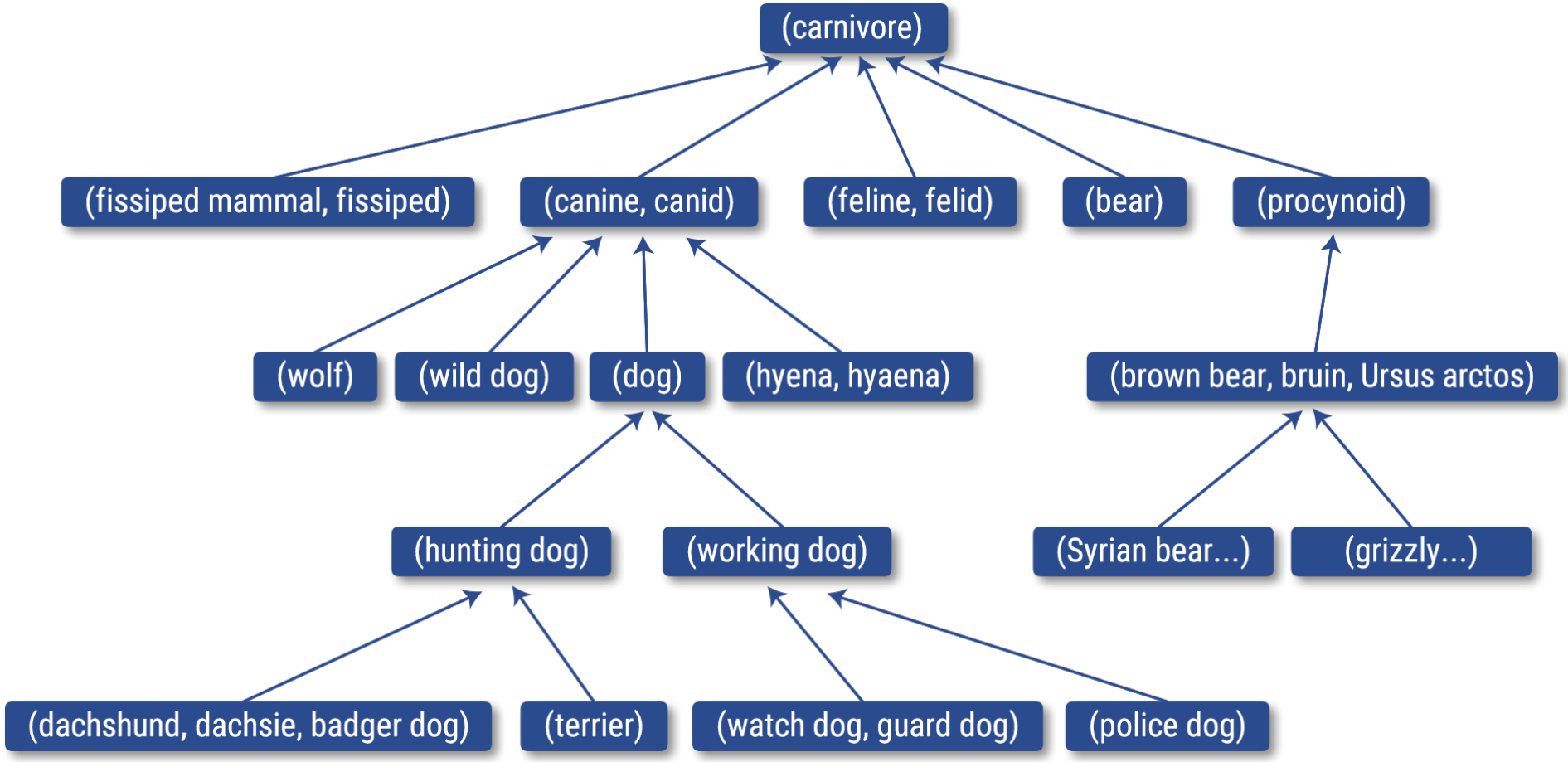
Three General Inquirer Classes

Category	Sample words
Academ[ic]	academic, astronomy, biology, chemistry, credit, dean, degree, physician, library
Ritual	ambush, appointment, affair, bridge, census, commemorate, debut, demonstration
Female	aunt, feminine, girl, goddess, her, heroine, grandmother, mother, queen, she

Source: General Inquirer dictionary program, **http://www.wjh.harvard.edu/~inquirer/**

Stone & Hunt (1963)

Lexical Resources: Wordnet



Lexical Resources:
The three most commons

- ***AFINN*** by Finn Årup Nielsen

  yes/no: positive, negative, anger, anticipation, disgust, fear, joy, sadness, surprise, and trust

- ***bing*** by Bing Liu and collaborators

  yes/no: positive, negative

- ***nrc*** by Saif Mohammad and Peter Turney

  between -5 and 5: negative - positive

Lexical Resources: The three most commons

AFINN by Finn Årup Nielsen

word	*value*
abacus	trust
abandon	fear
abandon	negative
abandon	sadness
abandoned	anger
abandoned	fear
abandoned	negative
abandoned	sadness
abandonment	anger
abandonment	fear

Lexical Resources: The three most commons

bing by Bing Liu and collaborators

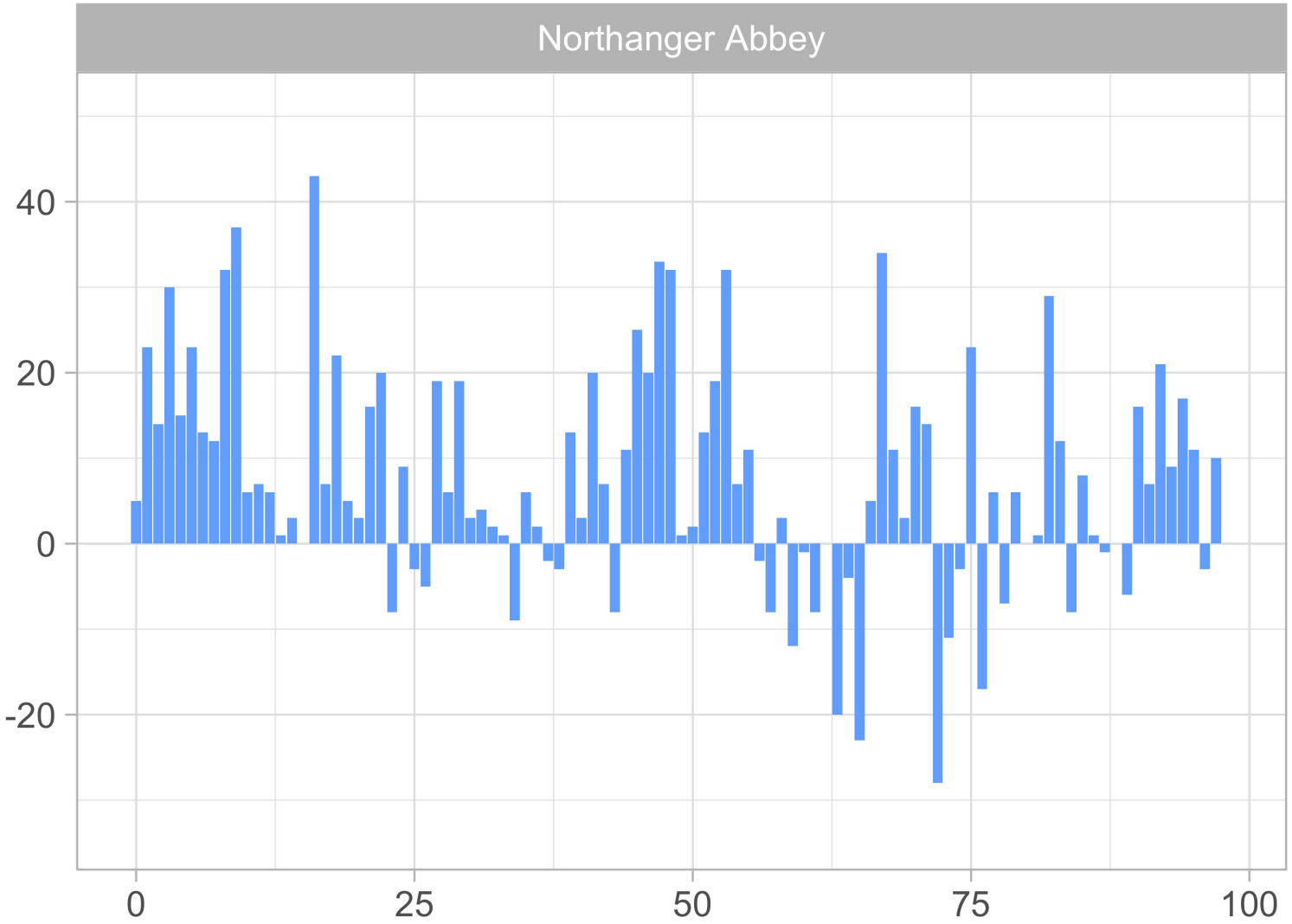
word	*value*
2-faces	negative
abnormal	negative
abolish	negative
abomidable	negative
anominably	negative
abominate	negative
abomination	negative
abort	negative
aborted	negative
aborts	negative

Lexical Resources:
The three most commons

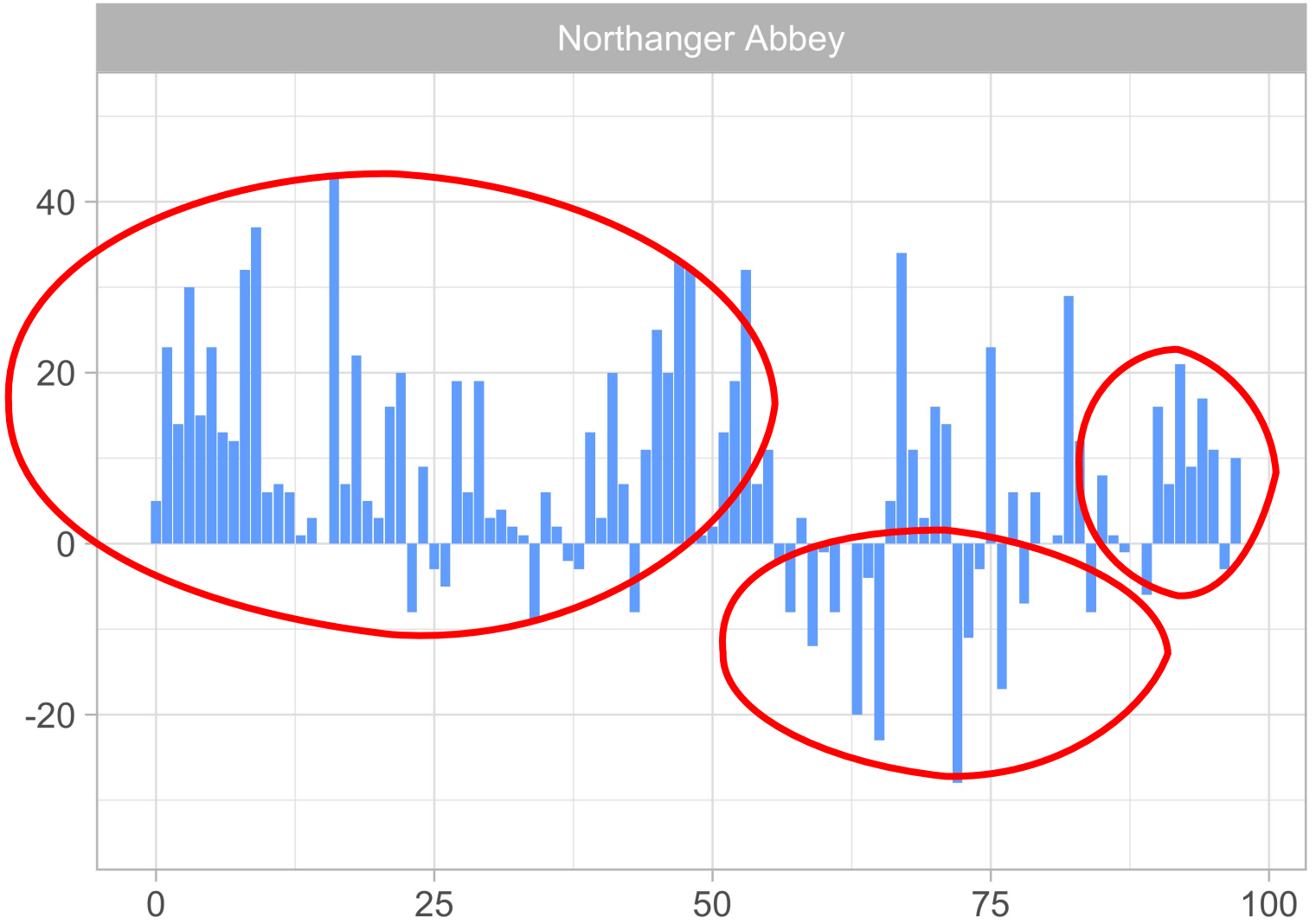
nrc by Saif Mohammad and Peter Turney

word	*value*
abandon	-2
abandoned	-2
abandons	-2
abducted	-2
abduction	-2
abductions	-2
abhor	-3
abhorred	-3
abhorrent	-3
abhors	-3

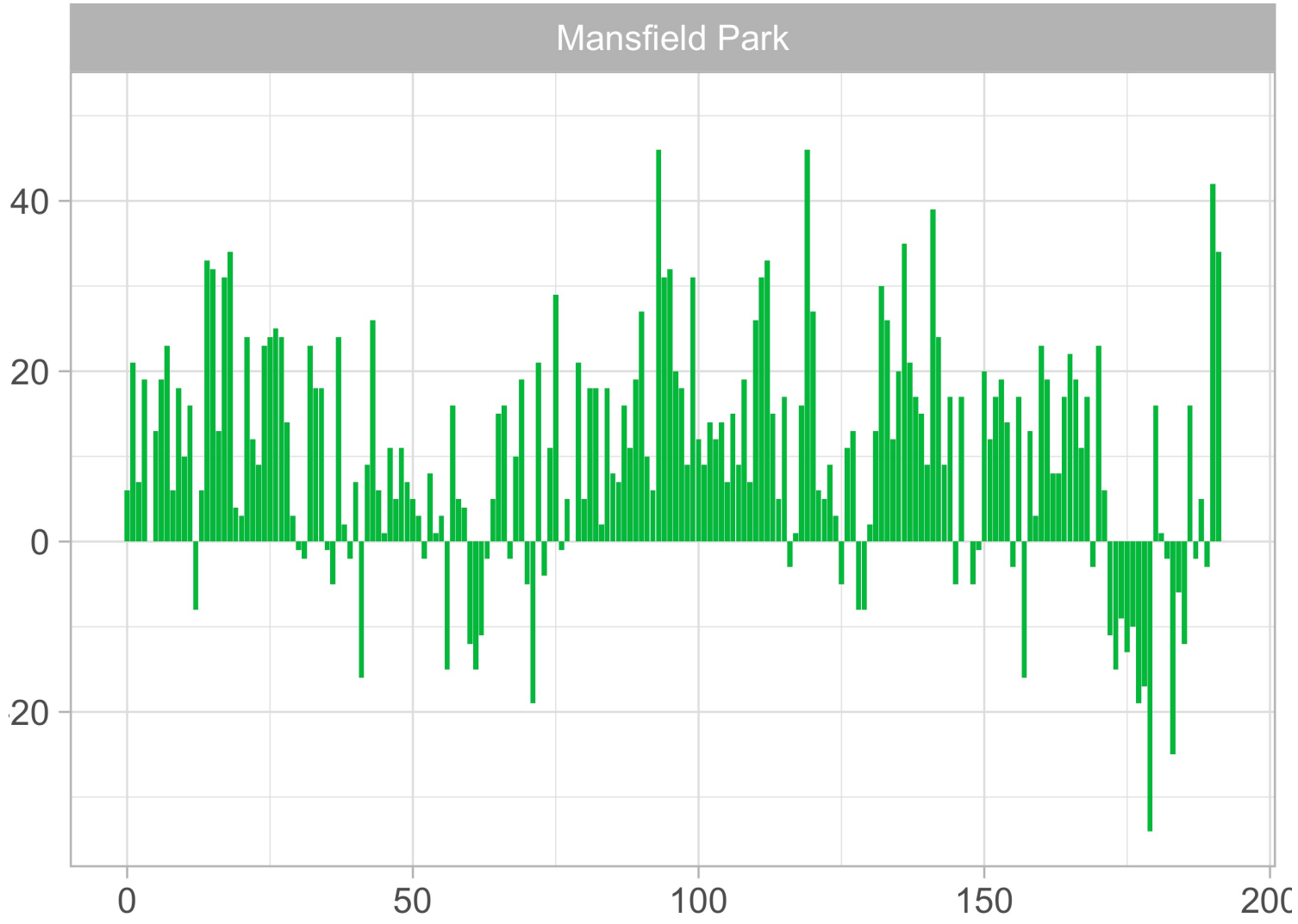
Example



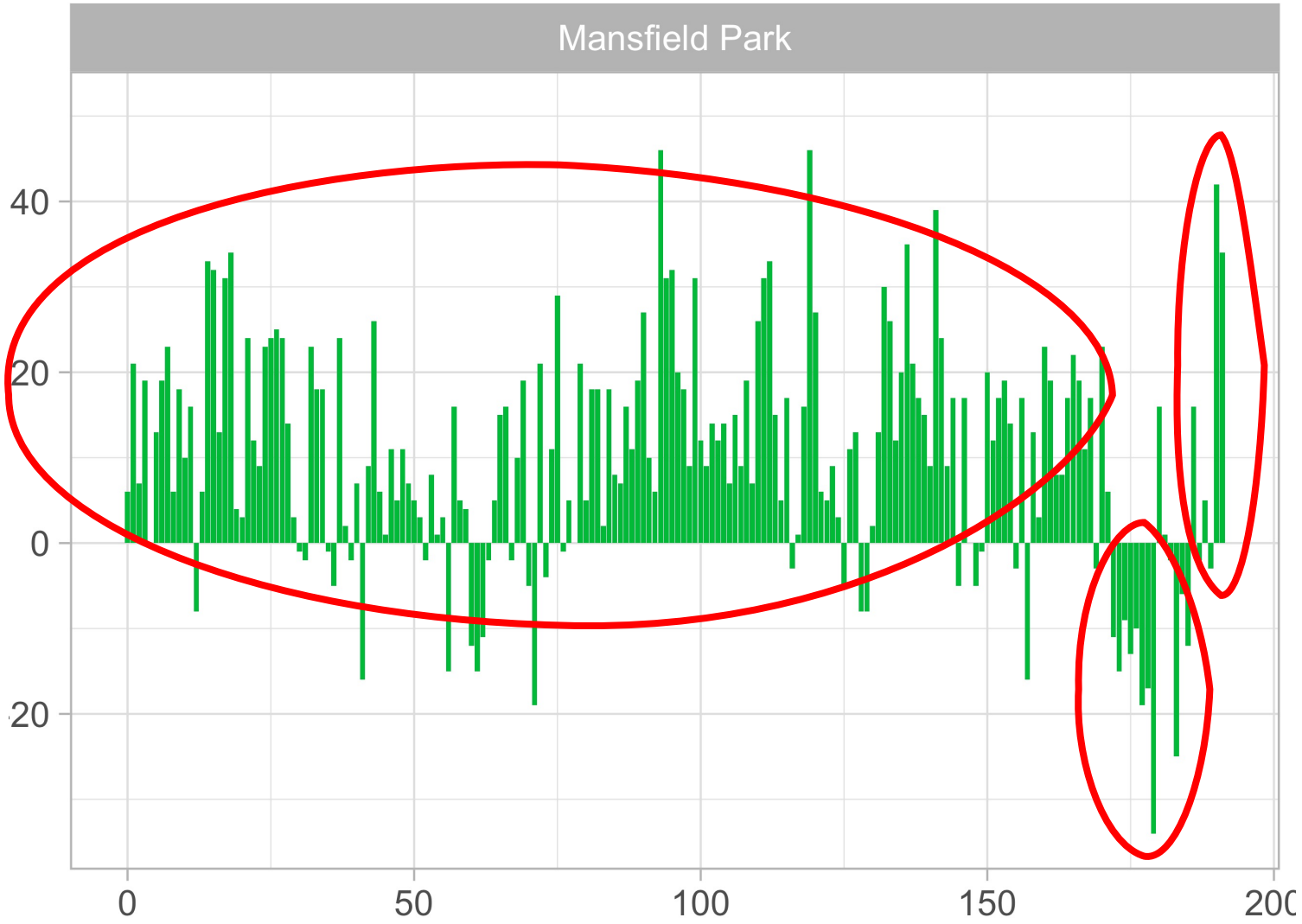
Example



Example



Example

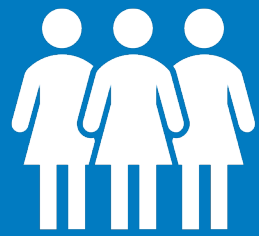


Methods

Relationships between Words

Tokenizing by n-gram

We can analyze tokens with two, three or more words, which we call n-grams



Pairs of consecutive words might capture structure that isn't present when one is just counting single words, and may provide context that makes tokens more understandable

Example: Bi-gram

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us, we were all going direct to Heaven, we were all going direct the other way

Bi-grams

Option A	*Option B (stop words)*
Token	Token
It was	it was
was the	was best
the best	best times
best of	times it
of times	it was
times it	was worst
it was	worst times
was the	times it
the worst	it was
worst of	was age
of times	age wisdom
times it	wisdom it
…	…

Example: Tri-gram

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity, it was the season of Light, it was the season of Darkness, it was the spring of hope, it was the winter of despair, we had everything before us, we had nothing before us, we were all going direct to Heaven, we were all going direct the other way

tokeniz ation

Tri-grams

Option A	*Option B (stop words)*
Token	Token
It was the	it was best
was the best	was best times
the best of	best times it
best of times	times it was
of times it	it was worst
times it was	was worst times
it was the	worst times it
was the worst	times it was
the worst of	it was age
worst of times	was age wisdom
of time it	age wisdom it
times it was	wisdom it was
…	…

Sentiment Analysis with n-grams

We can improve the results using 2-grams, and 3-grams, because we are giving context to the words

Example

I am not ***happy*** and I don't ***like*** it!

Example

I am not ***happy*** and I don't ***like*** it!

Word	***Feeling***
I	
am	
not	
happy	
and	
I	
don't	
like	
it	

Example

I am not ***happy*** and I don't ***like*** it!

Word	***Feeling***
I	-
am	-
not	Negative (-1)
happy	Positive (+1)
and	-
I	-
don't	Negative (-1)
like	Positive (+1)
it	-
	Neutral (0)

Example

I am not ***happy*** and I don't ***like*** it!

Word	***Feeling***
I	-
am	-
not	Negative (-1)
happy	Positive (+1)
and	-
I	-
don't	Negative (-1)
like	Positive (+1)
it	-
	Neutral (0)

Word	***Feeling***
I am	
am not	
not happy	
happy and	
and I	
I don't	
don't like	
like it	

Example

I am not ***happy*** and I don't ***like*** it!

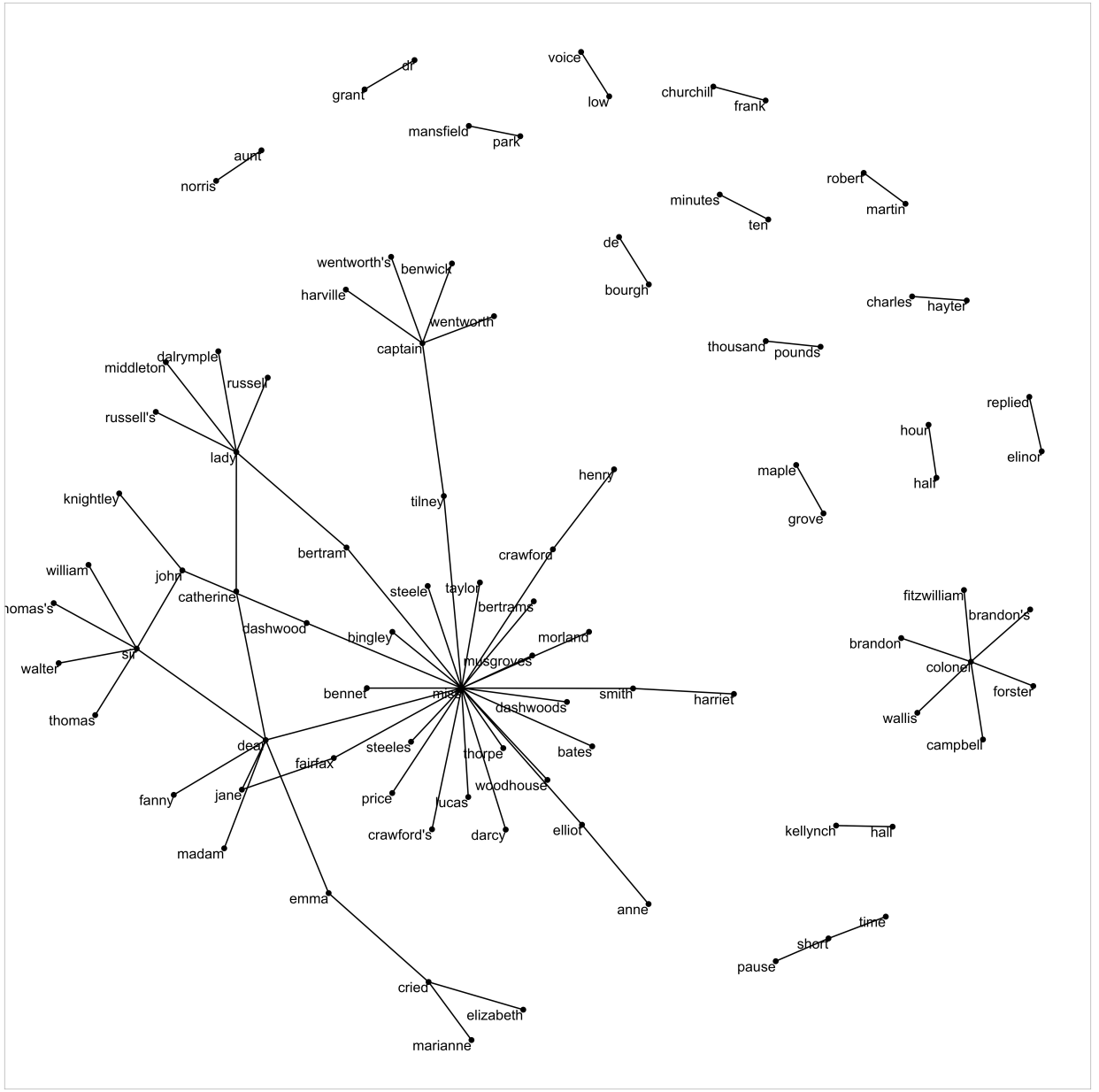
Word	***Feeling***
I	-
am	-
not	Negative (-1)
happy	Positive (+1)
and	-
I	-
don't	Negative (-1)
like	Positive (+1)
it	-
	Neutral (0)

Word	***Feeling***
I am	-
am not	Negative (-1)
not happy	Negative (-1)
happy and	Positive (+1)
and I	-
I don't	Negative (-1)
don't like	Negative (-1)
like it	Positive (+1)
	Negative (-2)

Networks

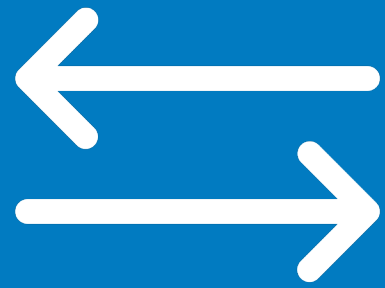
The visualization all of the relationships among words into a network, or “graph.” where there are three variables:

- ***from***: the node an edge is coming from
- ***to***: the node an edge is going towards
- ***weight***: A numeric value associated with each edge



Pairwise correlation

We may instead want to examine ***correlation*** among words, which indicates how often they appear together relative to how often they appear separately.



phi coefficient - measure of association for two binary variables

References

- Robinson, David. 2016. gutenbergr: Download and Process Public Domain Works from Project Gutenberg. https://cran.rstudio.com/package=gutenbergr.
- Robinson, David. 2017. broom: Convert Statistical Analysis Objects into Tidy Data Frames. https://CRAN.R-project.org/package=broom
- Silge, Julia. 2016. janeaustenr: Jane Austen's Complete Novels. https://CRAN.R-project.org/package=janeaustenr.
- Silge, Julia, and David Robinson. 2016. "tidytext: Text Mining and Analysis Using Tidy Data Principles in R." JOSS 1 (3). https://doi.org/10.21105/joss.00037.

UNIVERSITAT POLITÈCNICA DE CATALUNYA BARCELONATECH

TechTalentLab

vicenc.fernandez@upc.edu