Univerza v Mariboru

Fakulteta za varnostne vede

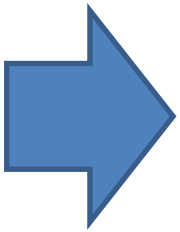
# Automatic text analysis – text mining

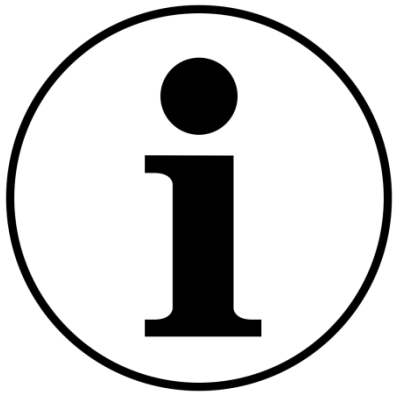
Vanja Ida Erčulj

# Text mining: automatic retrieval of information from text

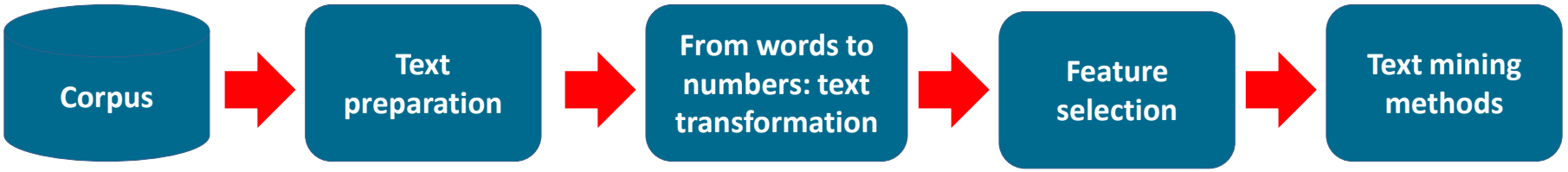

# Basic text analysis online

Costiera Amalfitana

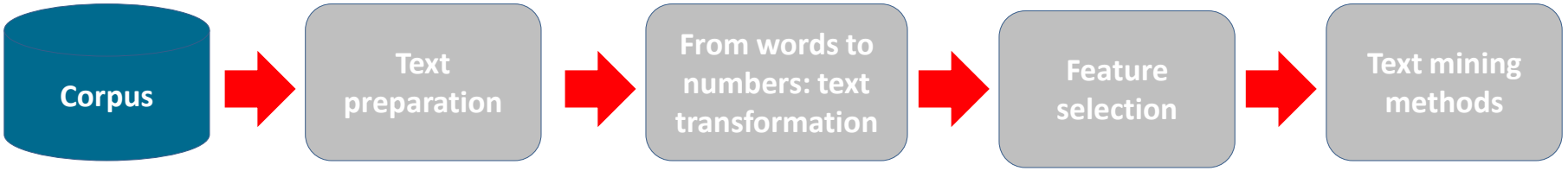




# Text analysis process



# Corpus




A (large) set of (structured) textual documents that were obtained in a certain time frame with objective to adress research questions by employing linguistic analysis, also text mining.

**Example:**

*All the text from web pages (in italian) describing the coast of Amalfi, arranged in a structured list (per web page).*
*OR*
*Text of each of the city on the coast of Amalfi from a web page structured in a list.*

# Corpus: example

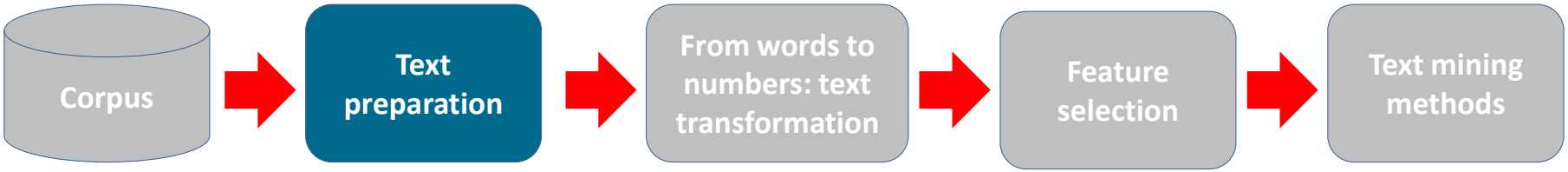
- Document 1

  Amalfi: la Costiera Amalfitana prende il suo nome proprio da questa cittadina che fu la prima delle quattro repubbliche Marinare. Amalfi era considerata una potenza nel traffico commerciale con l'Oriente per via del suo sbocco naturale alla Valle dei Mulini. Appena giunti sul posto si viene accolti dalle case bianche incastonate nella roccia come fossero dei diamanti, collegate tra di loro da suggestive scalinate e vicoli coperti. Il luogo più importante è la piazza dove si trova il Duomo di Sant'Andrea con il campanile e il Chiostro del Paradiso.

- Document 2

  Atrani: è un piccolo borgo di appena 800 anime che si trova all'imbocco della Valle del Dragone. La piazzetta si affaccia direttamente sul mare con scorci mozzafiato, il tutto nel silenzio più completo visto che non si trova negli itinerari del turismo di massa. Atrani è uno dei borghi più belli d'Italia ed è il comune più piccolo per superficie. Da visitare la Chiesa di San Salvatore de' Birecto del X secolo.

# Text preparation




- Normalization of text
  - Lower case words
  - Removal of numbers
  - Removal of punctuation marks
  - Stop words removal (stop words are meaningless words, present in all documents with the same probability)
  - Tokenization: separating text into smaller units, such as words, characters and subwords,…
  - The root of the word eng. stemming
  - Lemmatization (the basic form of the word)

# Text preparation: example

- *Le spiagge della Costiera sono un vero e proprio Paradiso: una gita in barca, oltre allo spettacolo immenso, consente di ammirare le meraviglie della natura direttamente dal mare.*

**Lower case**
**Punctuation removal**
**Stop words removal**

- *le spiagge della costiera sono un vero e proprio paradiso una gita in barca oltre allo spettacolo immenso consente di ammirare le meraviglie della natura direttamente dal mare*

# Text preparation: tokenization

- *le spiagge della costiera sono un vero e proprio paradiso una gita in barca oltre allo spettacolo immenso consente di ammirare le meraviglie della natura direttamente dal mare*

*spiagge*
*costiera*
*vero*
*proprio*
*paradiso*
*gita*
*barca*
*spettacolo*

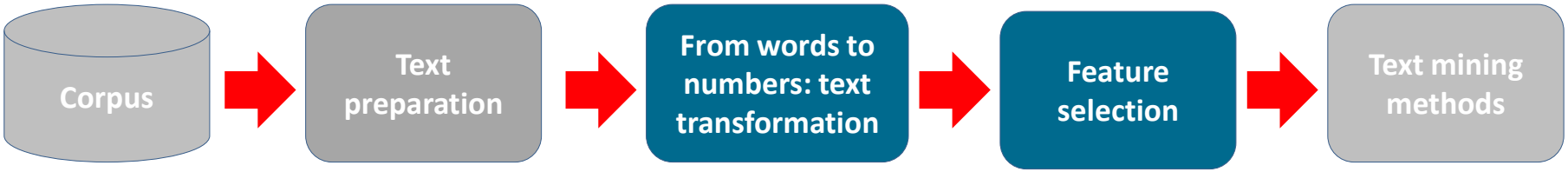
Tokenization

*immenso*
*consente*
*ammirare*
*meraviglie*
*natura*
*direttamente*
*mare*

# Text preparation: stemming and lemmatization

| Stemming | Lemmatization |
| --- | --- |
| *spiagge* → *spiagg* | *spiagge* → *spiaggia* |
| *costiera* → *costier* | *costiera* → *costiera* |
| *vero* → *ver* | *vero* → *vero* |
| *Proprio* → *propri* | *Proprio* → *proprio* |

# Text transformation and feature selection




- From text to numbers
- Bag-of-words model: every word is a feature (similar to a variable)
- Feature weights:
  - Binary
  - Term frequecy (tf)
  - Tf-idf (inverse document frequency)

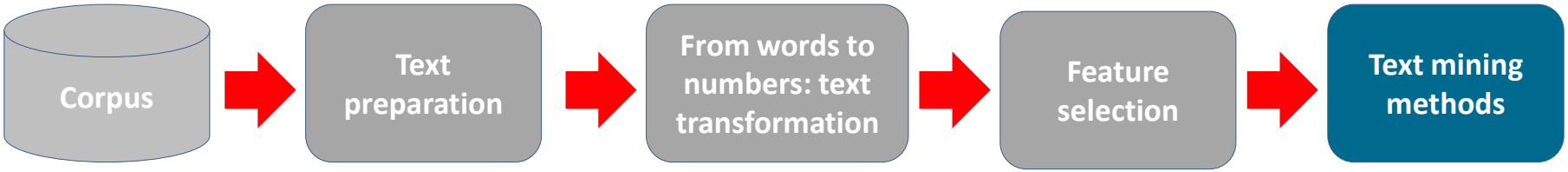
| | text |
|---|---|
| 1 | I like solving interesting problems. |
| 2 | What is machine learning? |
| 3 | I'm not sure. |
| 4 | Machien lerning predicts eveyrthing. |

Binary weight

| | eveyrthing | interesting | learning | lerning | like | Machien | machine | not | predicts | problems | solving | sure | What |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| 2 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| 3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| 4 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |

VIR: http://www.vikparuchuri.com/blog/natural-language-processing-tutorial/

# Text mining methods

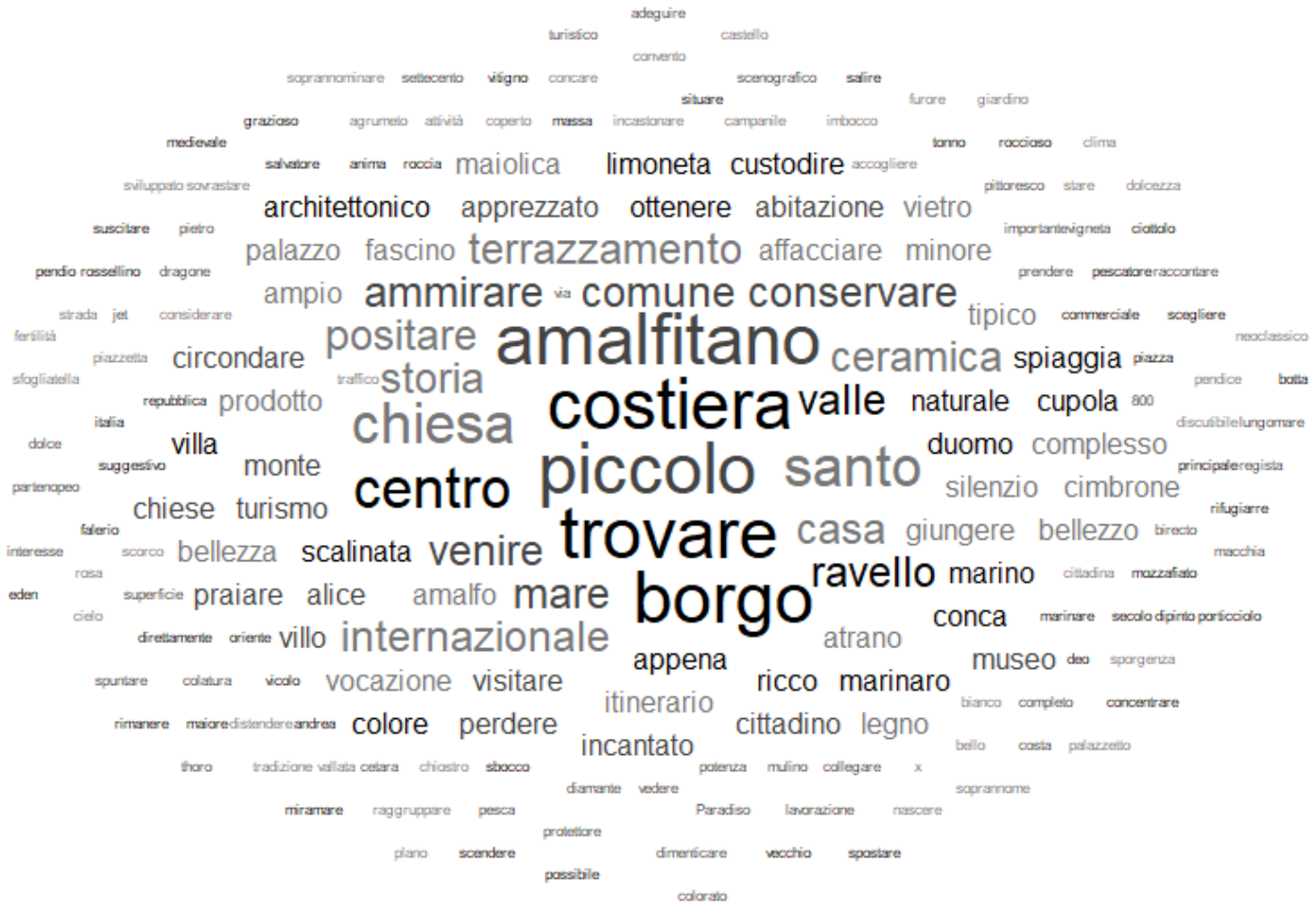



- Keyword extraction
- Text clustering
- Text classification
- Sentiment analysis

# Keyword extraction

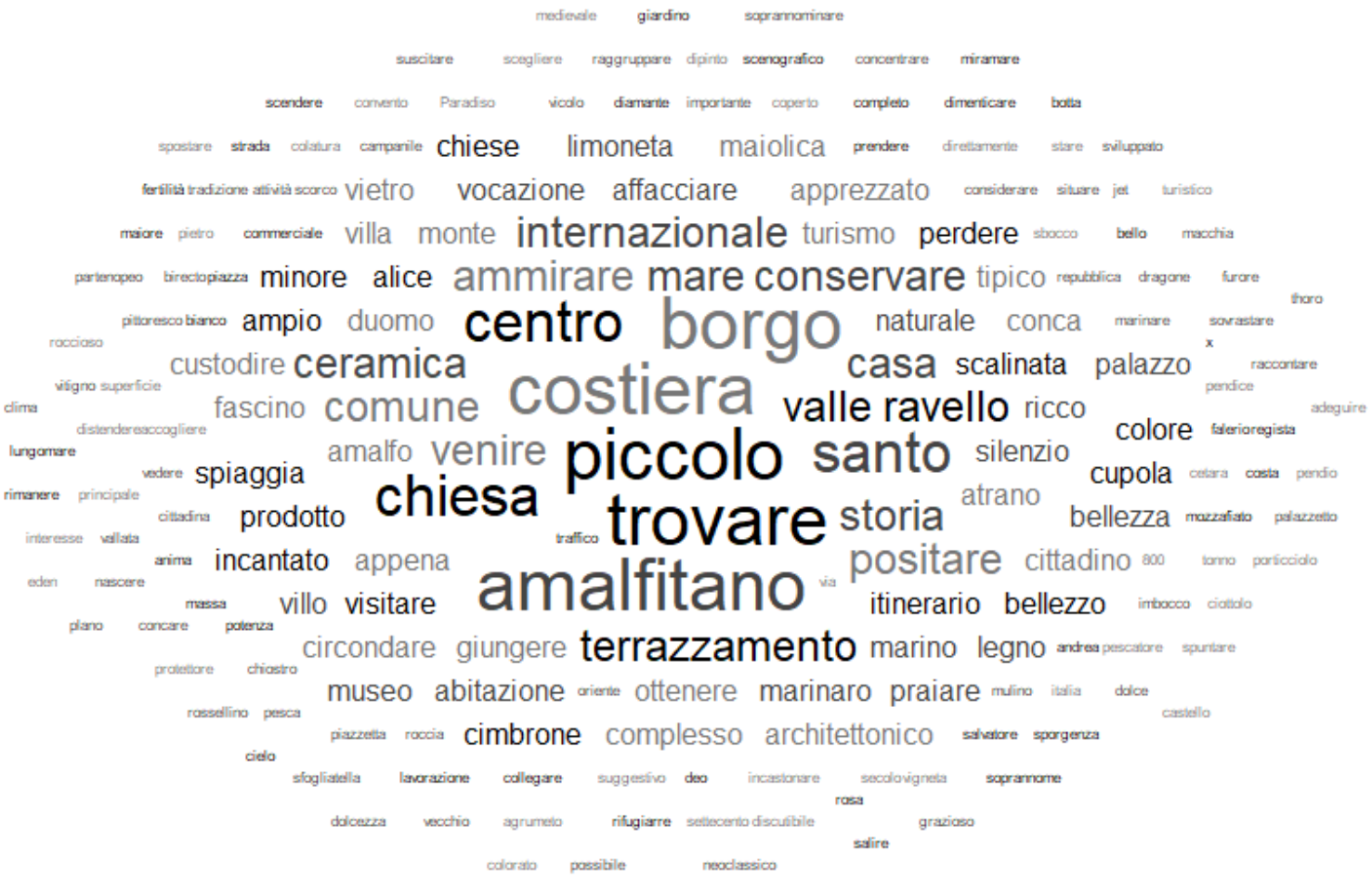
- Term (word) frequency
- n-grams (*n* consequtive words)
- Tf-idf (term frequency – inverse document frequency)
- POS (ang. part-of-speech) tagging
- TextRank
- Rapid automatic Keyword Extraction (RAKE)

# Keywords - tf


adeguire
turistico
castello
convento
soprannominare
settecento
vitigno
concare
scenografico
salire
situare
furore
giardino
grazioso
agrumeto
attività
coperto
massa
incastonare
campanile
imbocco
medievale
tonno
roccioso
clima
salvatore
anima
roccia
maiolica
limoneta
custodire
accogliere
sviluppato
sovrastare
pittoresco
stare
dolcezza
architettonico
apprezzato
ottenere
abitazione
vietro
suscitare
pietro
importantevigneto
ciottolo
palazzo
fascino
terrazzamento
affacciare
minore
pendio
rossellino
dragone
prendere
pescatore
raccontare
ampio
ammirare
via
comune
conservare
strada
jet
considerare
tipico
commerciale
scegliere
fertilità
positare
amalfitano
neoclassico
piazzetta
circondare
ceramica
spiaggia
piazza
sfogliatella
traffico
storia
pendice
botta
repubblica
prodotto
costiera
valle
naturale
cupola
800
italia
chiesa
discutibile
lungomare
dolce
villa
duomo
complesso
suggestivo
monte
piccolo
santo
principale
regista
partenopeo
centro
silenzio
cimbrone
chiese
turismo
rifugiarre
falerio
trovare
casa
giungere
bellezzo
interesse
scorcio
bellezza
scalinata
venire
birecto
macchia
rossa
ravello
marino
cittadina
mozzafiato
eden
superficie
praiare
alice
amalfo
mare
borgo
cielo
conca
marinare
secolo
dipinto
porticciolo
direttamente
oriente
villo
internazionale
atrano
appena
museo
deo
spargenza
spuntare
colatura
vicolo
vocazione
visitare
ricco
marinaro
itinerario
bianco
completo
concentrare
rimanere
maiore
distendere
andrea
colore
perdere
cittadino
legno
incantato
bello
costa
palazzetto
thoro
tradizione
vallata
cetara
chiostro
sbocco
potenza
mulino
collegare
x
diamante
vedere
soprannome
miramare
raggruppare
pesca
Paradiso
lavorazione
nascere
protettore
piano
scendere
dimenticare
vecchio
spostare
possibile
colorato

# Keywords – uni and bigrams

# Keywords – tf-idf (unigrams)



# Keywords – other methods

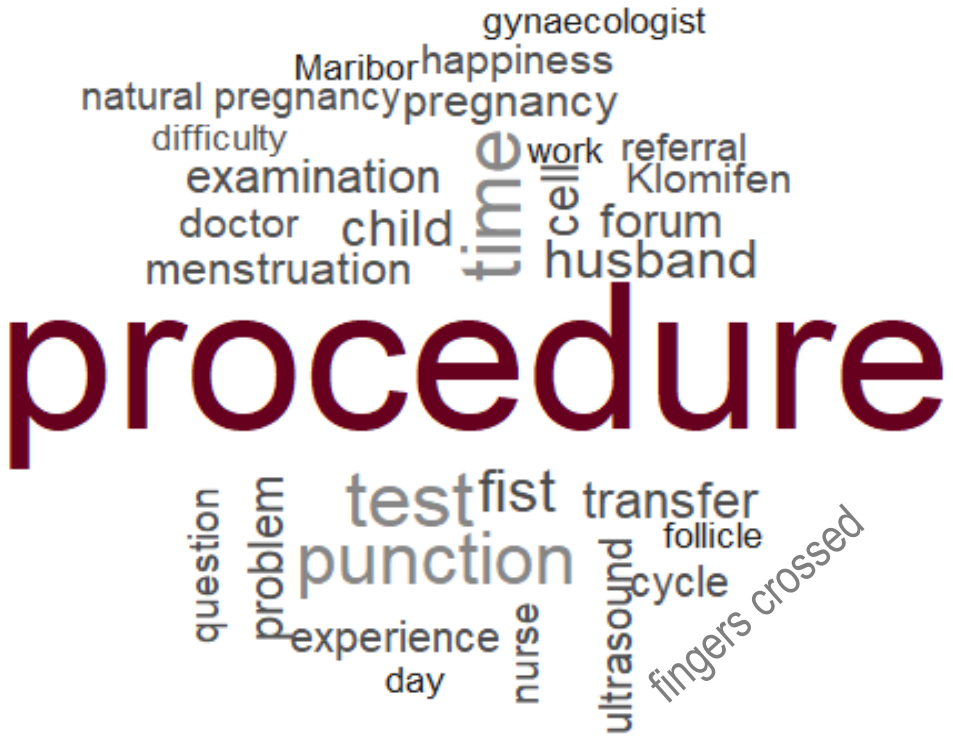
- POS tagging
  - Categories of types of words: nouns, verbs, adjectives, adverbs etc.
  - Use only nouns and adjectives → work best
  - http://linguistic-annotation-tool.italianlp.it/
- Algorithms
  - Textrank (Mihalcea and Tarau, 2004)
    - Words used together in specific window frame (number of consequitive words in text considered) are graphically presented as points (a word) and arrows pointing from a word leading to another word. Weights assigned to words according to the number of words pointing to it and number of words to which a given word is pointing to. Only % (1/3) of words with highest weight are keywords.
  - RAKE (rapid keyword extraction)
    - Potential keywords are words between punctuation marks and stop words. According to occurence of a given word in a text and its coocurence with other words in a text a weight is calculated. 1/3 of words with highest weight are keywords.

# Keywords extraction: example (POS tagging)


Abstract word cloud for Fraud - adjectives and nouns

Source

# Keywords extraction: example (RAKE)




Word cloud of 132.147 posts from the online support group *Infertility*

-

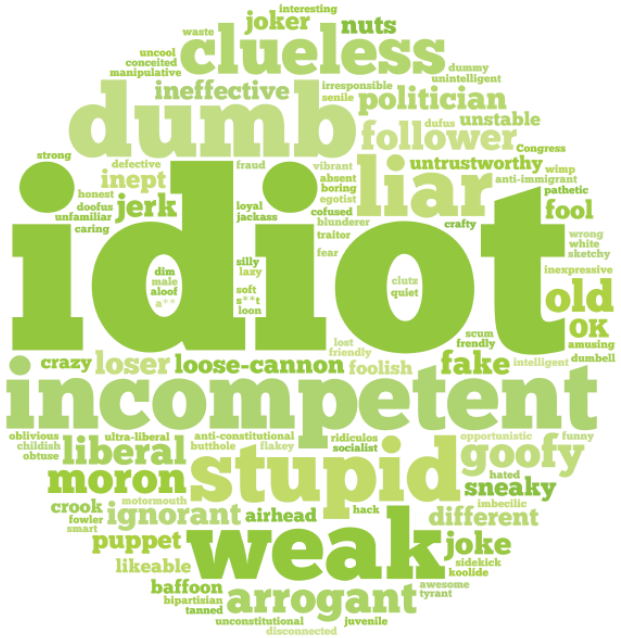
top 30 KW shown

# More examples:
# „Simple“ word cloud at the inauguration of Joe Biden

# More examples: word cloud - group comparison

*What one word would you use to describe Joe Biden?*

| People with a favorable opinion (44%) | People with an unfavorable opinion (45%) | DK (12%) |
|---|---|---|


*Survey conducted March 1-3, 2014. Q15-16. Results may not add up to 100% due to rounding.*

YouGov | yougov.com



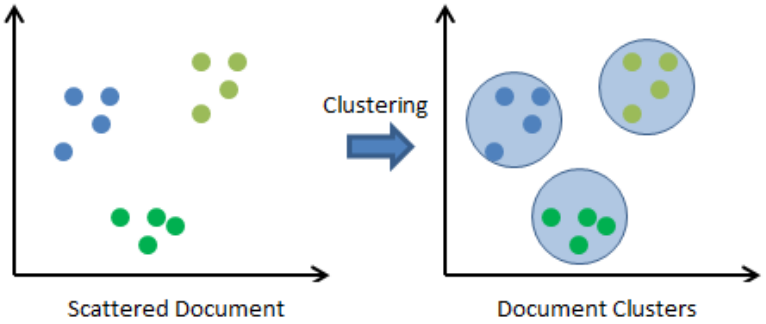
Source: here

# Text clustering

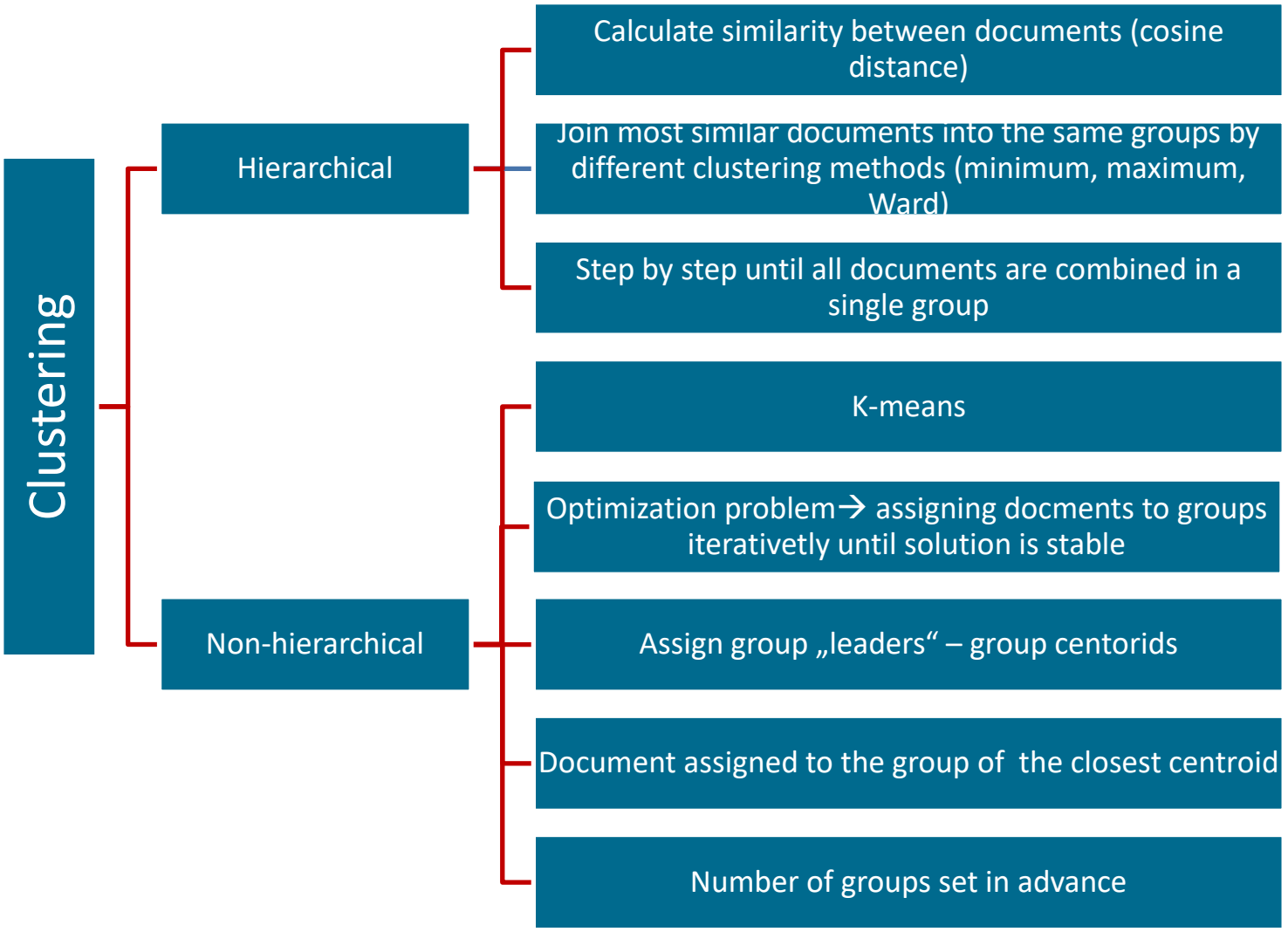
Arrange documents in groups according to similarity

Different methods:

- Hierarchical clustering
- K-means clustering
- Topic modelling - LDA (Latent Dirichlet Allocation)

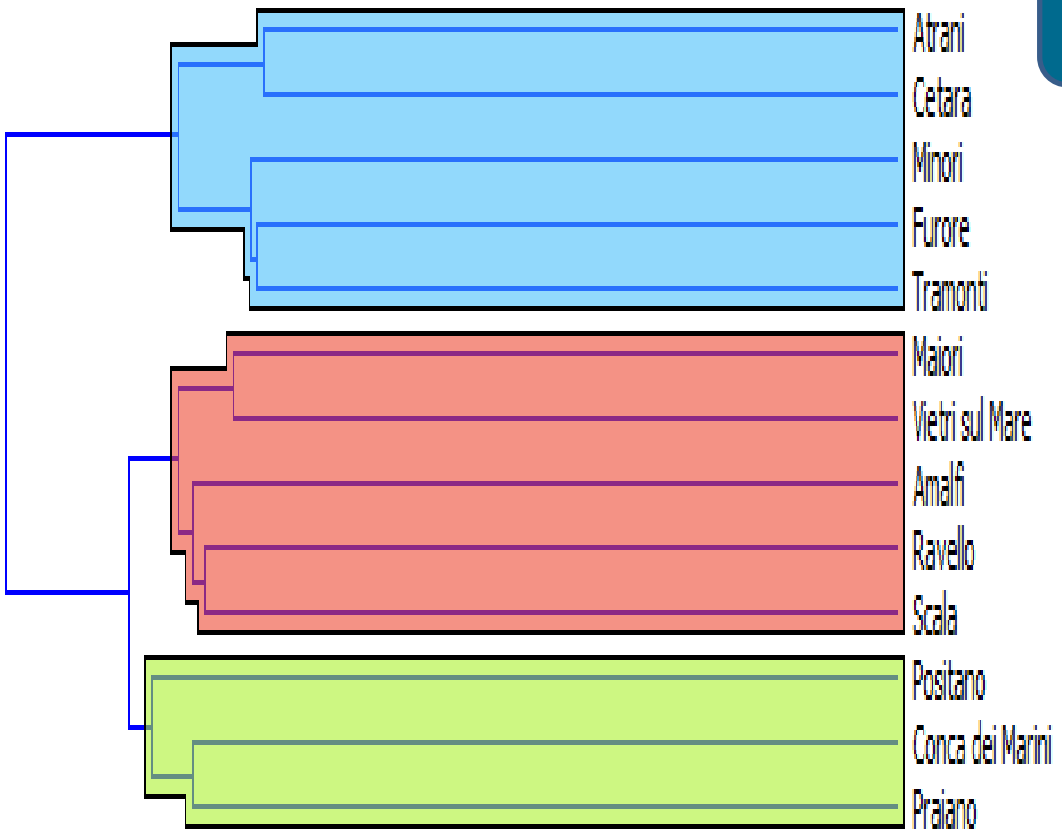


# Text clustering: hierarchical and k-means clustering



# Hierechical clustering: Amalfi coast revisited

Dendrogram




Most similar Amalfi coast villages according to their description

# The first group

Atrani

Cetara

Furore

Minori

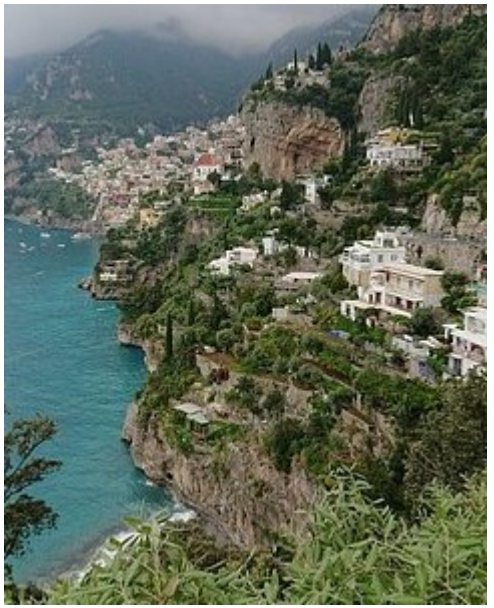
Tramonti

# The second group

Maiori

Amalfi

Scala



Vetri Sul Mare

Ravello

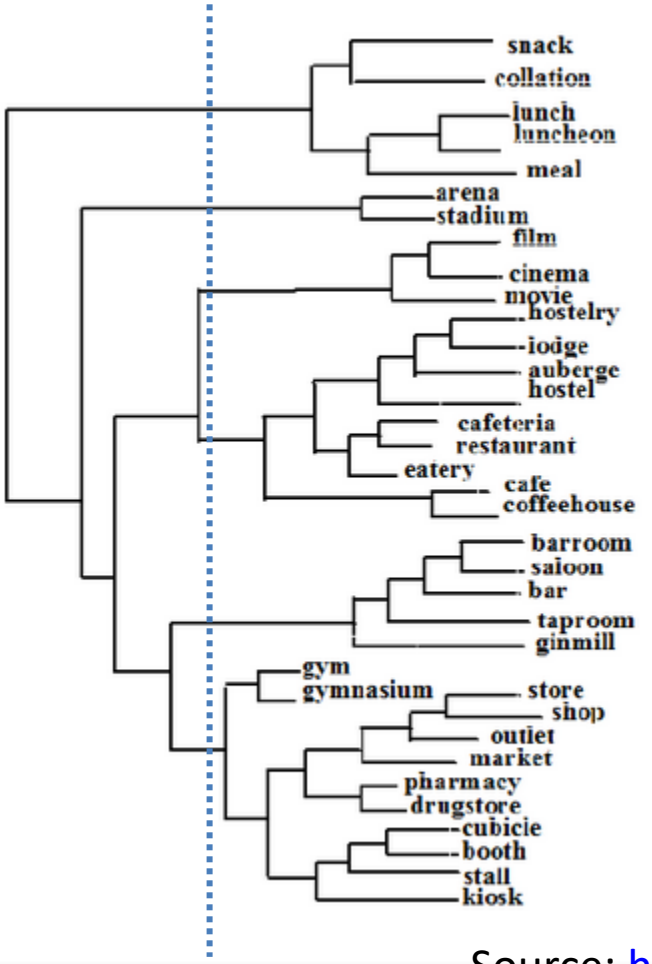
# The third group

Conca dei Marini

Praiano

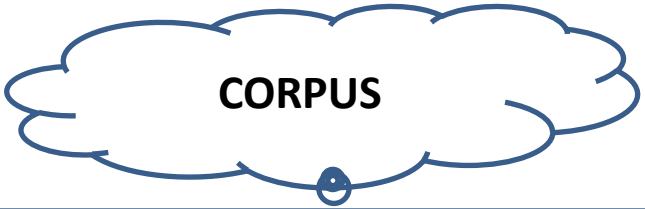
Positano

# Word clustering according to their meaning in WordNet dictionary - example



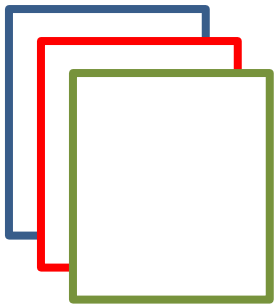

Source: here

# Text clustering: LDA




**Text pre-processing**: lower case, stop words, tokenisation, lemmatisation, part-of-speech tagging (Obeliks)

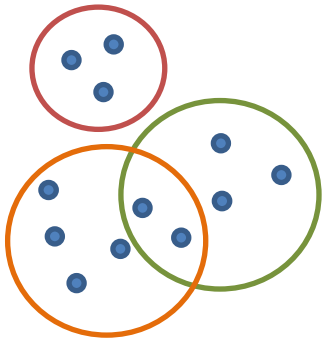
Adjectives, nouns, verbs

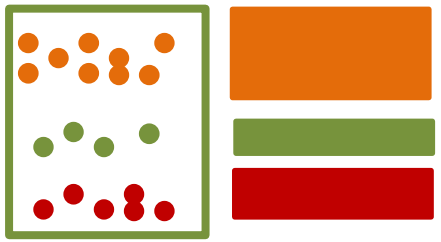


**Collection of documents**

TOPIC MODELING:
LDA

**Cluster of words by topic**

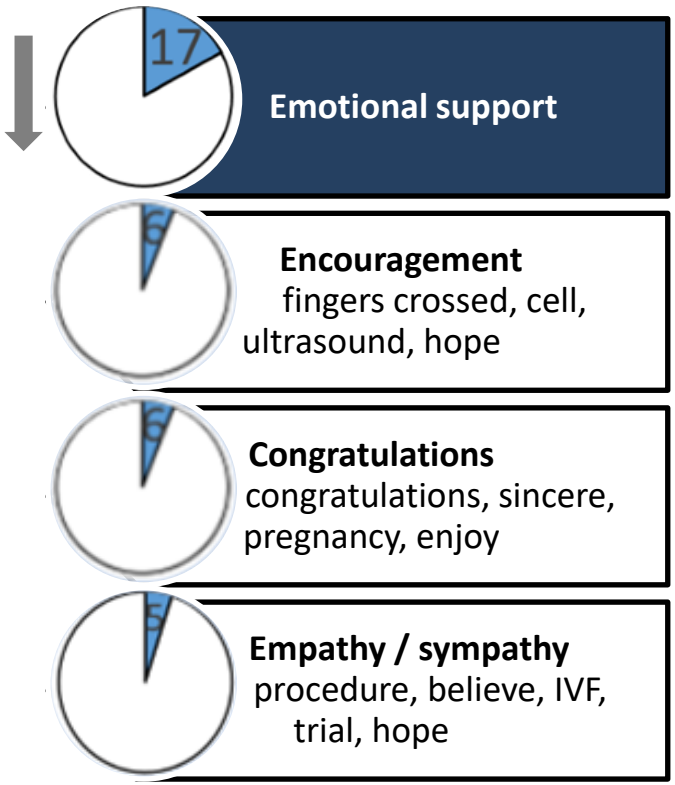




**Distribution of topics**

# LDA: topics of discussion - example

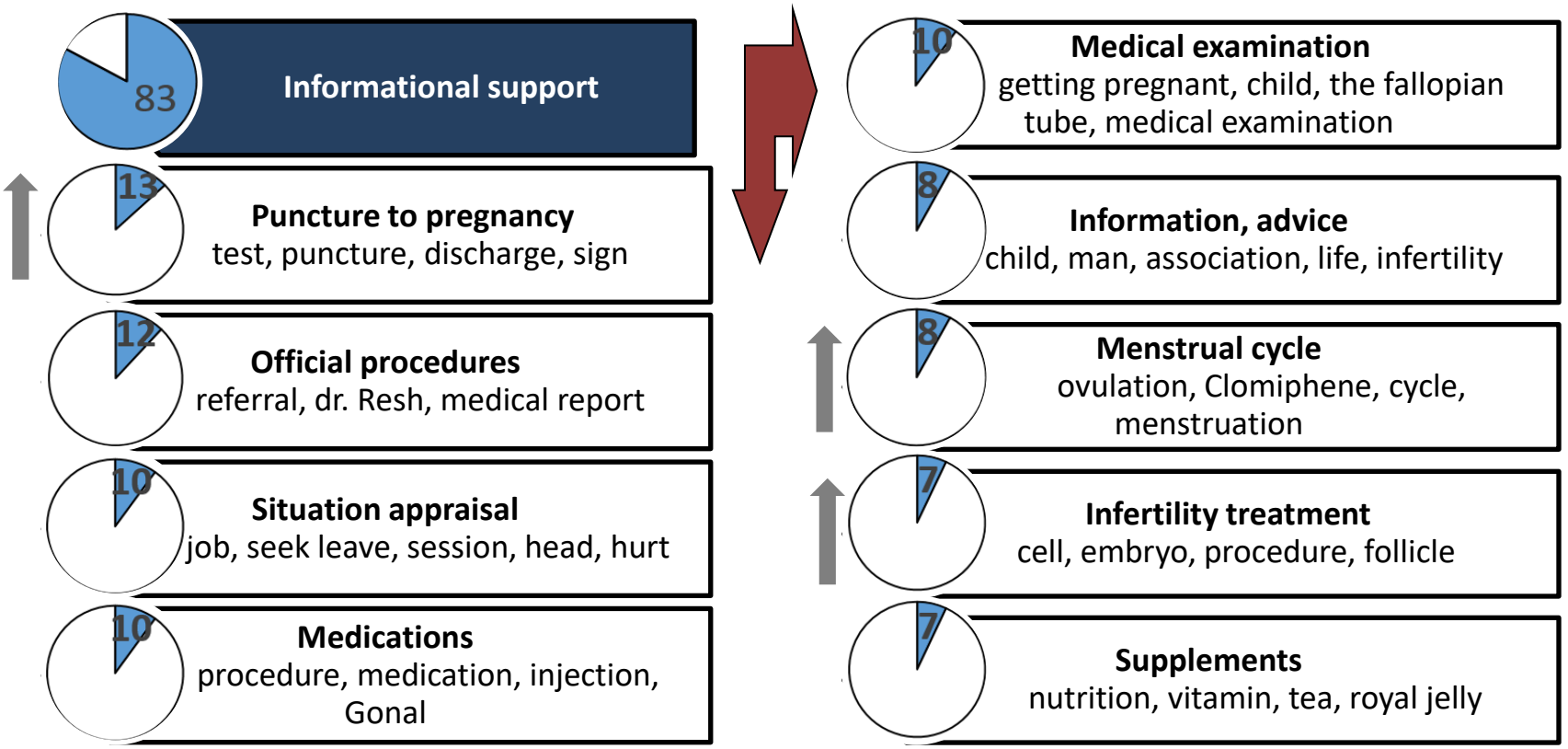
Clustering of 132.147 posts from the online support group *Infertility*




- **Encouragement:** I'm really keeping my fingers crossed and I believe that this time it's really OK. Just keep on warming your little penguin, take care of him and yourself!
- **Congratulations:** My sincere congratulations and wishing you beautiful and peaceful pregnancy.
- **Empathy / sympathy: :** I'm sorry for your loss. Be strong and keep looking ahead. You'll see that your suffering will pay off as you deserve and all this will then seem as a bad dream.

trend through time

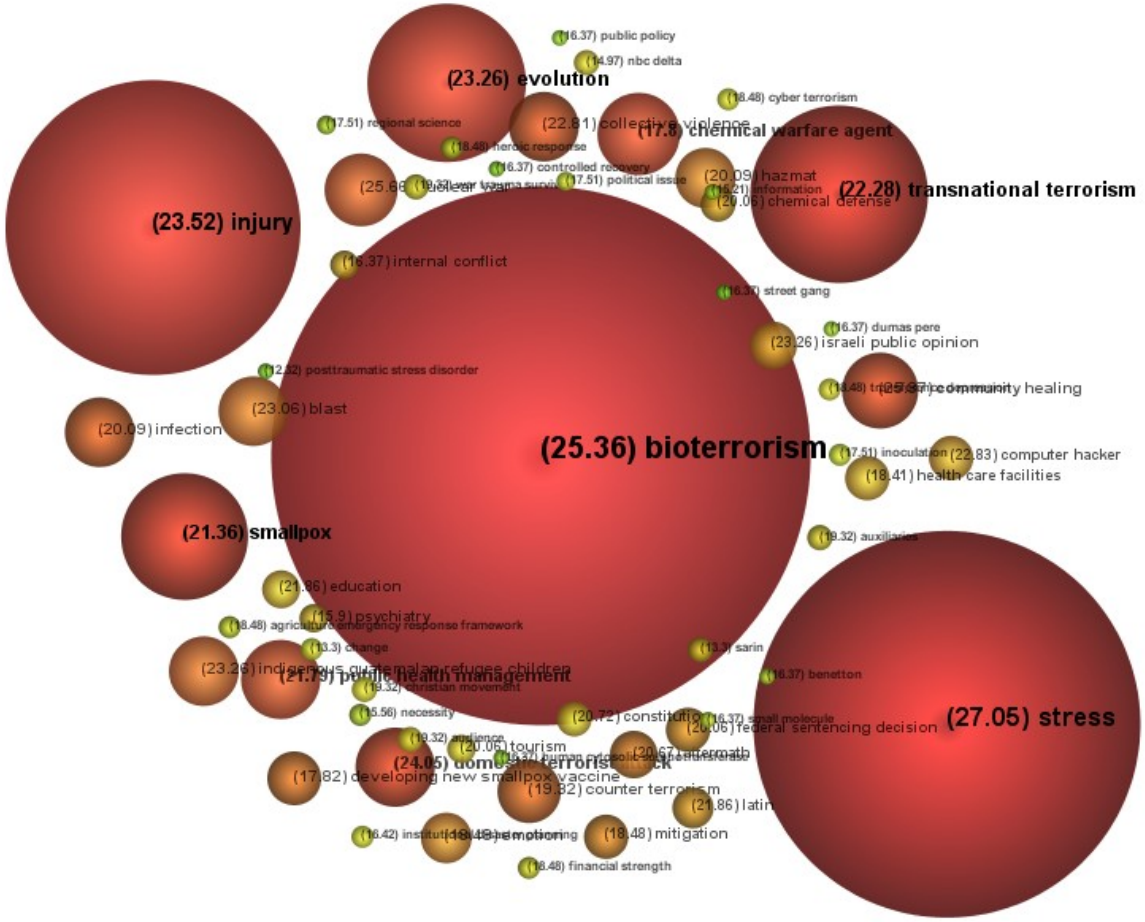
# LDA: topics of discussion - example




trend through time

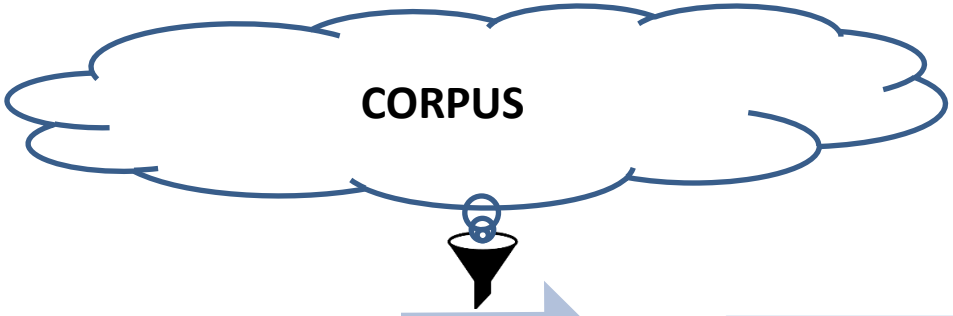
***Lower reliability and validity of measurement: Medications, Menstrual cycle***

# Topic modelling: clustering of research articles on therorism




Source: [here]

# Text classification



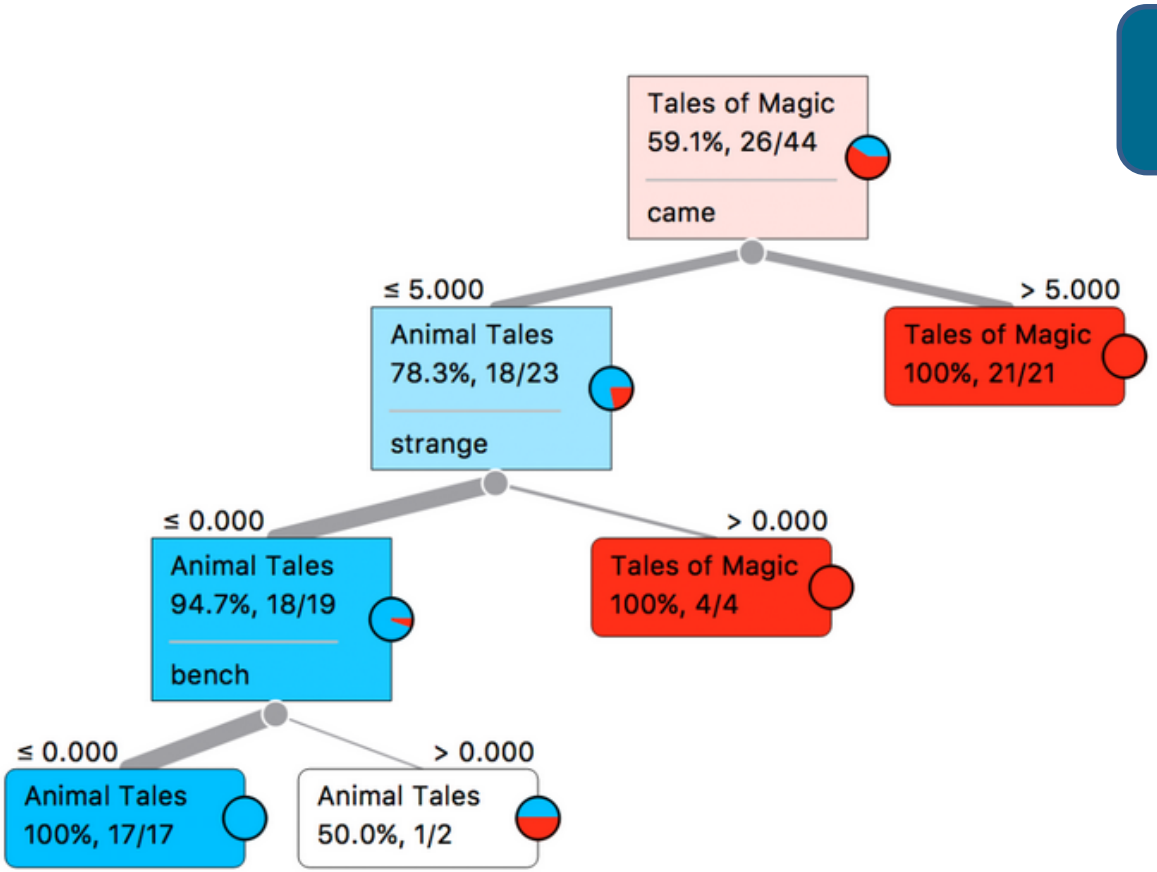

## Classifiers

- Naive Bayes (probabilistic, Bayes' theorem, prior knowledge)
- Support vector machine (decision boundary btw vectors)
- K-nearest neighbours (majority vote by its neighbours)
- Decision tree (information gain of each feature → top-down selection)
- Random forest (averaging decision trees)
- Logistic regression (logit function → probability of a target variable)

## Evalvation

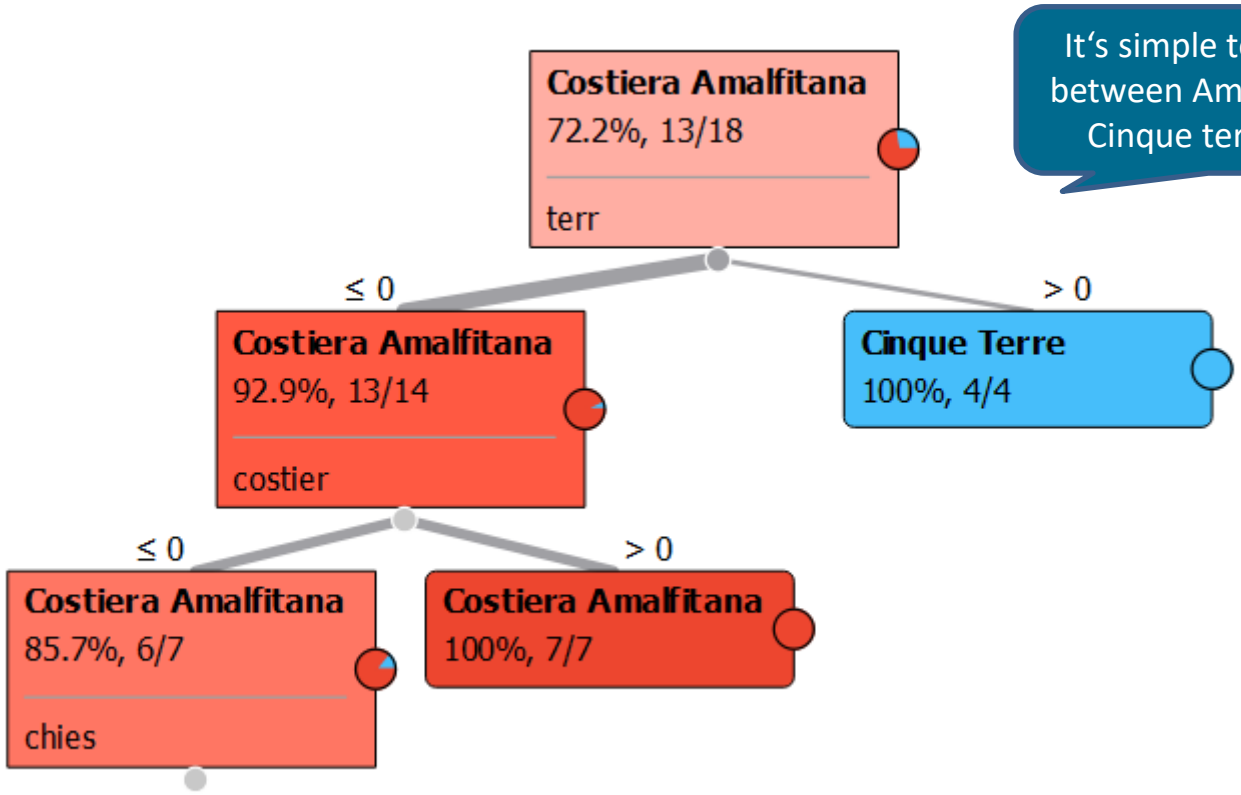
- 10 – fold cross validation (randomly splitting data in 10 folds; each fold for training, remaining for testing)
- Efficacy of the classifier:
  - AUC (> 0.50)
  - Accuracy (% of correctly classified cases)
  - Precision (% of actual positive among predicted positive cases)
  - Recall (% of predicted positive cases among actual positive cases)
  - F – measure (harmonic mean of precision and recall)

# Text classification example: decision tree




Source: Orange

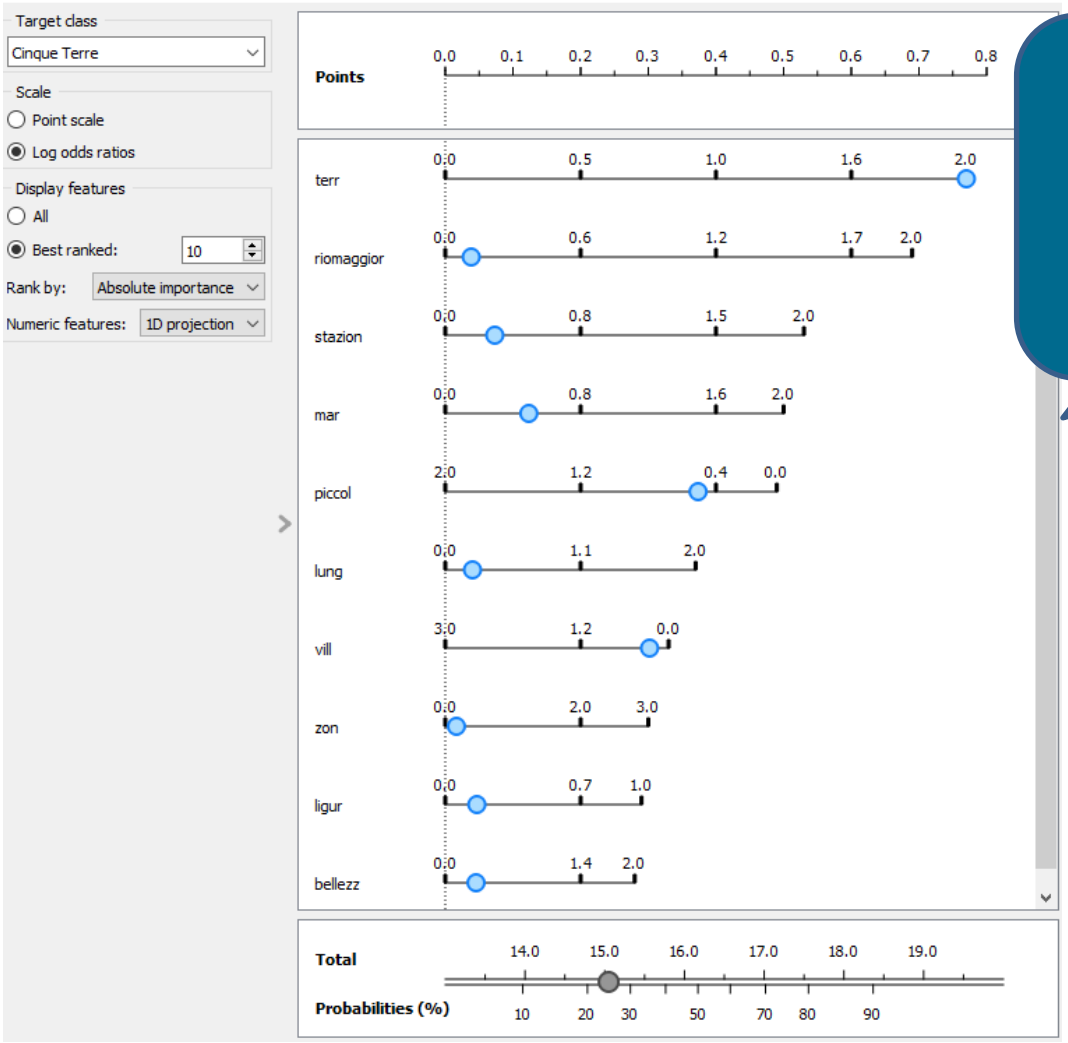
# Text classification: decision tree




It‘s simple to distinguish between Amalfi coast and Cinque terre villages.

https://www.cinqueterre.eu.com/it/

# Text classification: logistic regression




The most important words for the classifier. The more frequent the word „terr“ the higher probability for the Cinque terre village.

# Evaluation of classifiers

Average performance over classes (correctly predicting both, Cinque terre and Amalfi coast)

Evaluation Results

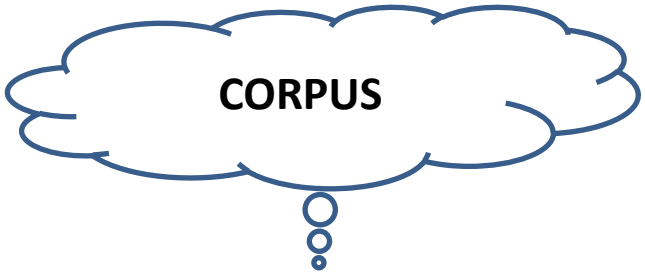
| Model | AUC | CA | F1 | Precision | Recall |
|---|---|---|---|---|---|
| Tree | 0.615 | 0.611 | 0.621 | 0.634 | 0.611 |
| Logistic Regression | 0.800 | 0.778 | 0.719 | 0.830 | 0.778 |

Performance when predicting Amalfi coast villages

| Model | AUC | CA | F1 | Precision | Recall |
|---|---|---|---|---|---|
| Tree | | 0.611 | 0.720 | 0.750 | 0.692 |
| Logistic Regression | | 0.778 | 0.867 | 0.765 | 1.000 |

Logistic regression: 76 % of predicted Amalfi coast villages are indeed Amalfi coast viallges; 100 % of Amalfi coast villages were predicted as being Amalfi coast villages.

# RESEARCH METHODS: sentiment analysis




Lexical approach vs. machine learning approach

**SENTIMENT LEXICON**

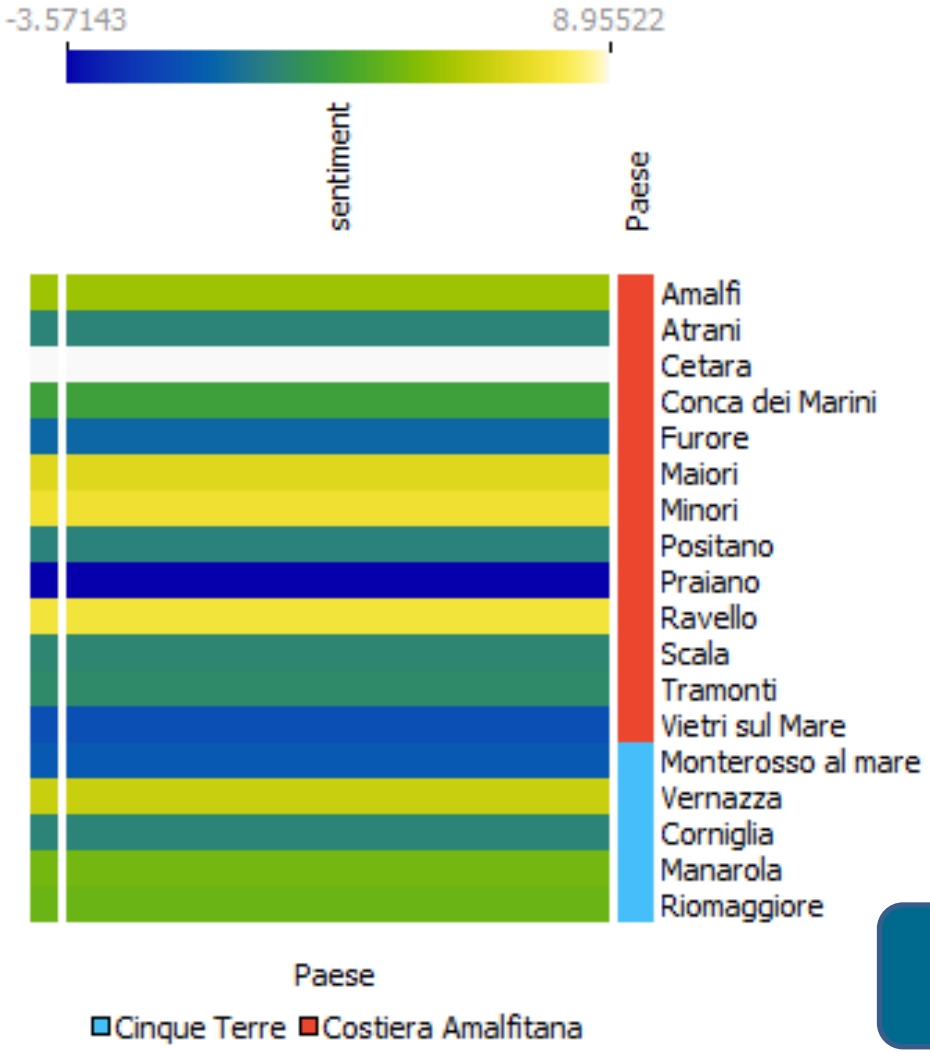
Positive words
Negative words

$$Rpn = \frac{no.\,of\ positive\ words}{no.of\ negative\ words}$$

WordNet-Affect
LIWC (linguistic inqury and word count)
Hu and Liu

Manual annotators
Agreement

# Sentiment analysis: Amalfi coast and Cinque terre revisited

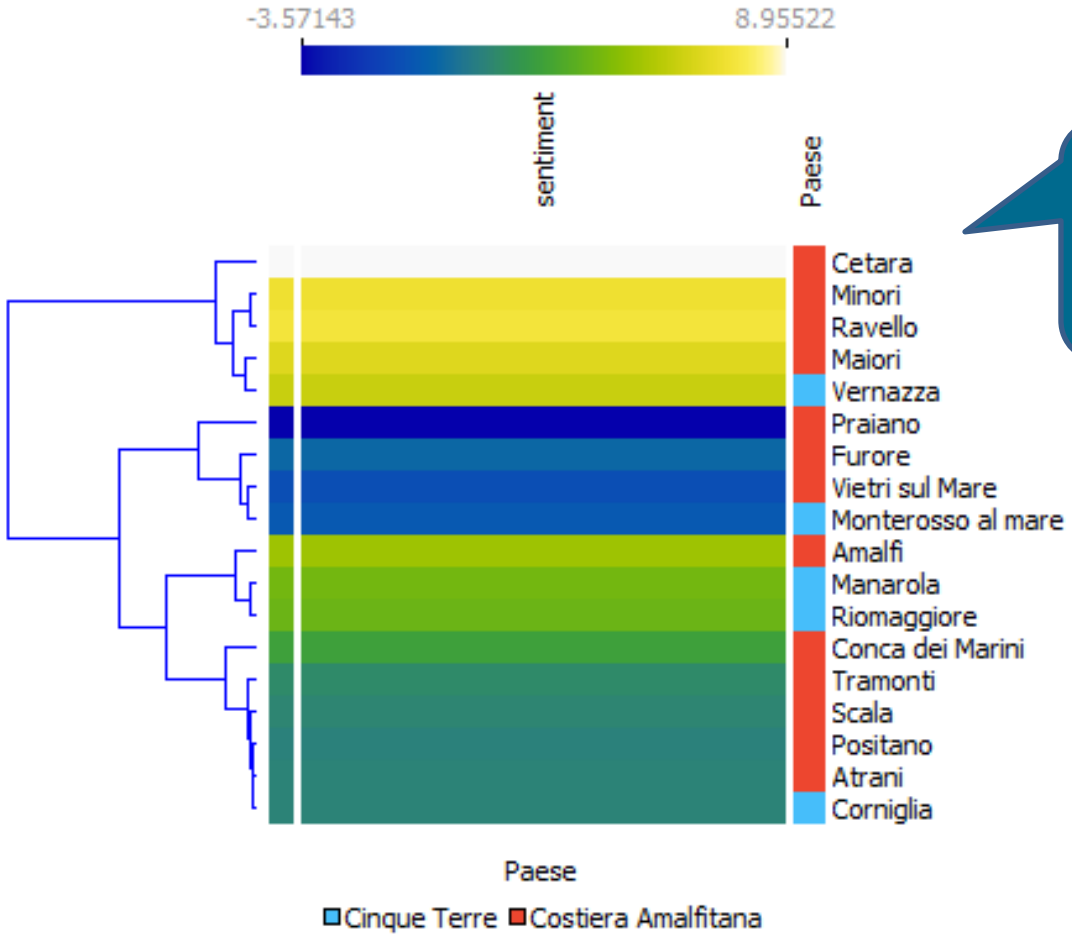



Most negative sentiment when describing Praiano and Vietri sul Mare

Praiano fa parte degli itinerari del turismo di nicchia, lontano dal caos i viaggiatori vengono rapiti e trasportati in una dimensione surreale.

Disadvantage of lexical approach: inability to take into account neighbouring words.

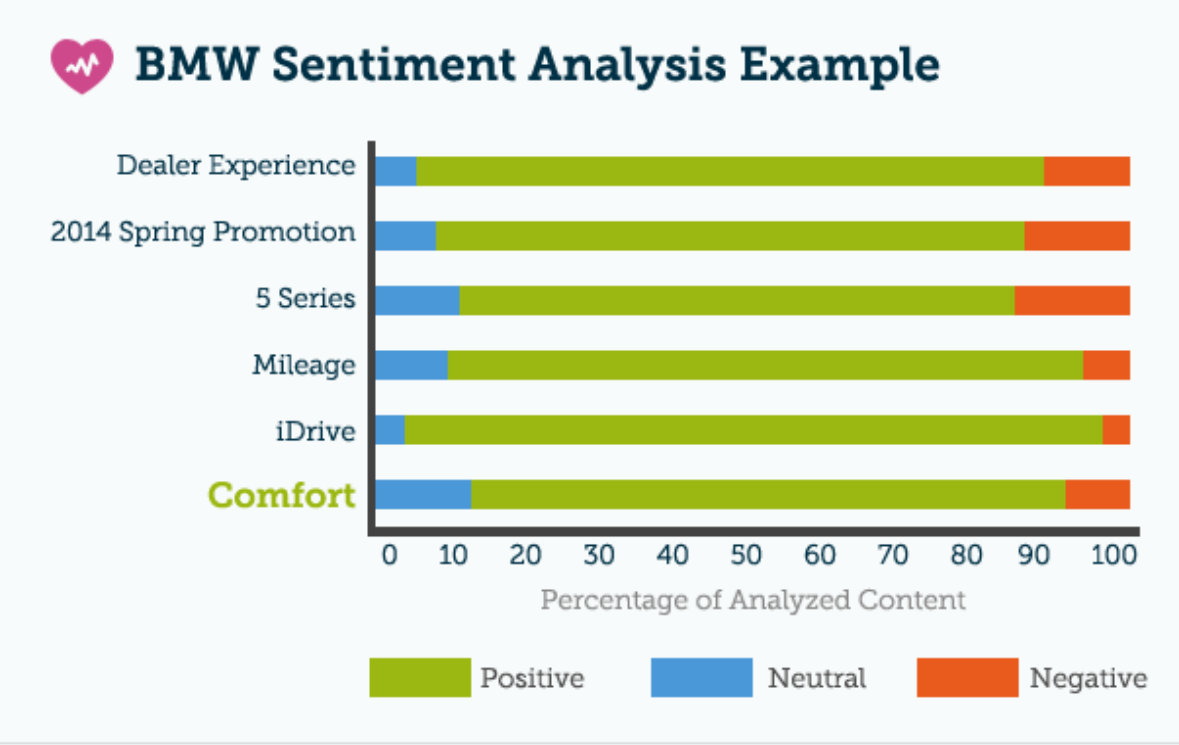
# Sentiment analysis: sentiment clustering




Five groups of villages according to the sentiment of their description.

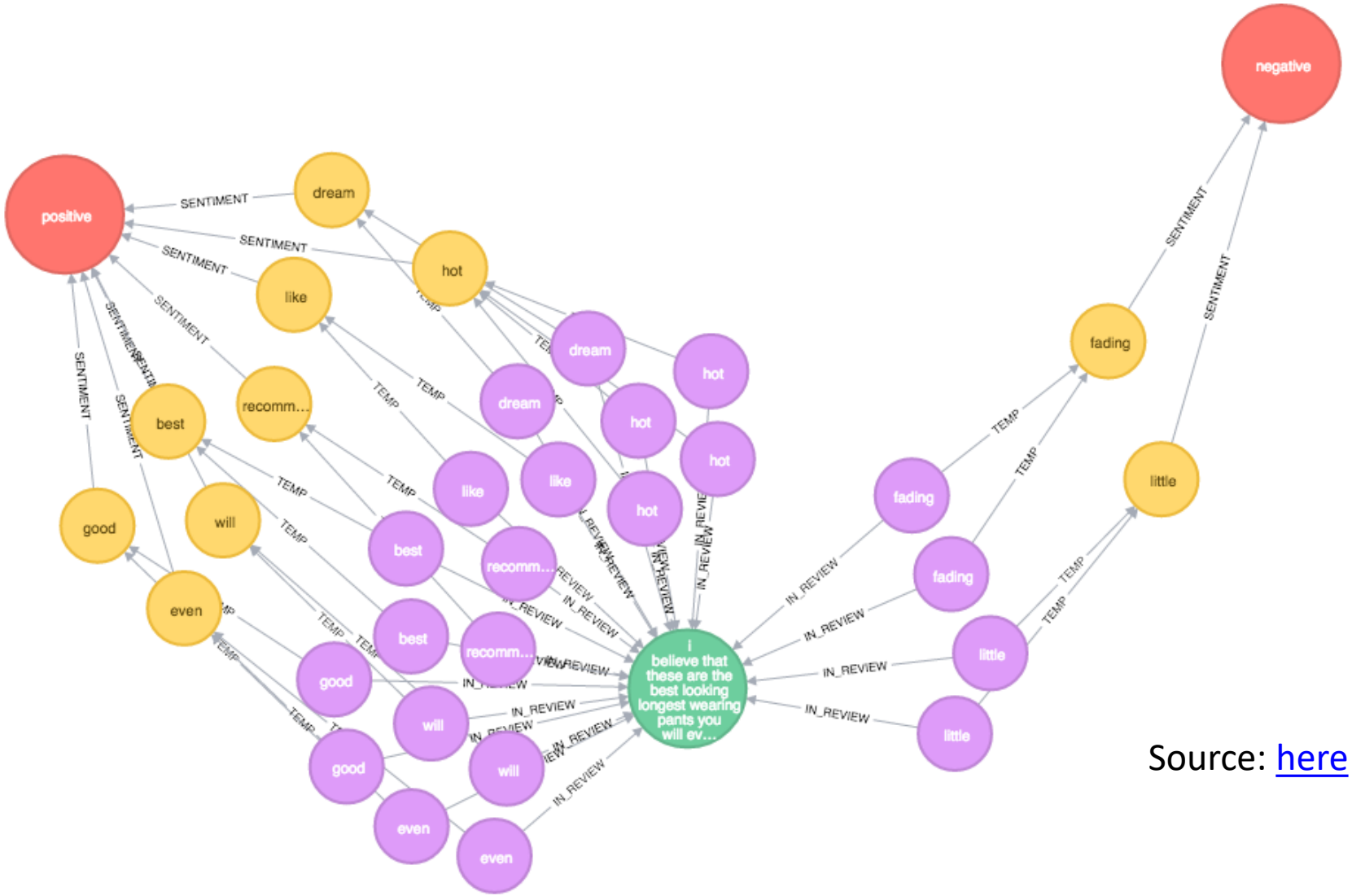
Minori:
soprannominata l'Eden della Costiera Amalfitana per la dolcezza del suo clima e la bellezza e fertilità dei suoi giardini.

# Sentiment analysis example: tweets about BMW



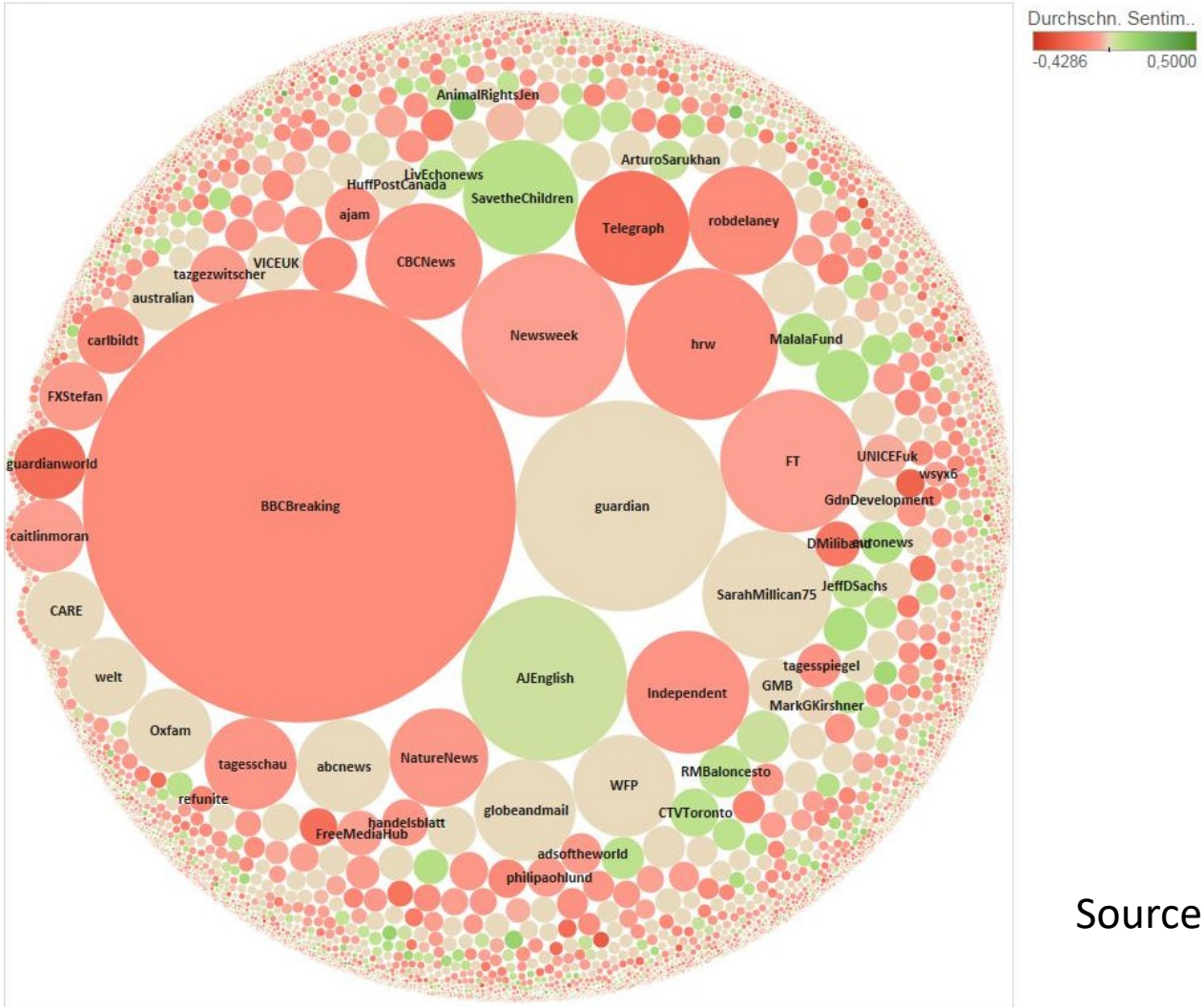

Source: here

# Sentiment analysis example: product opinion on Amazon



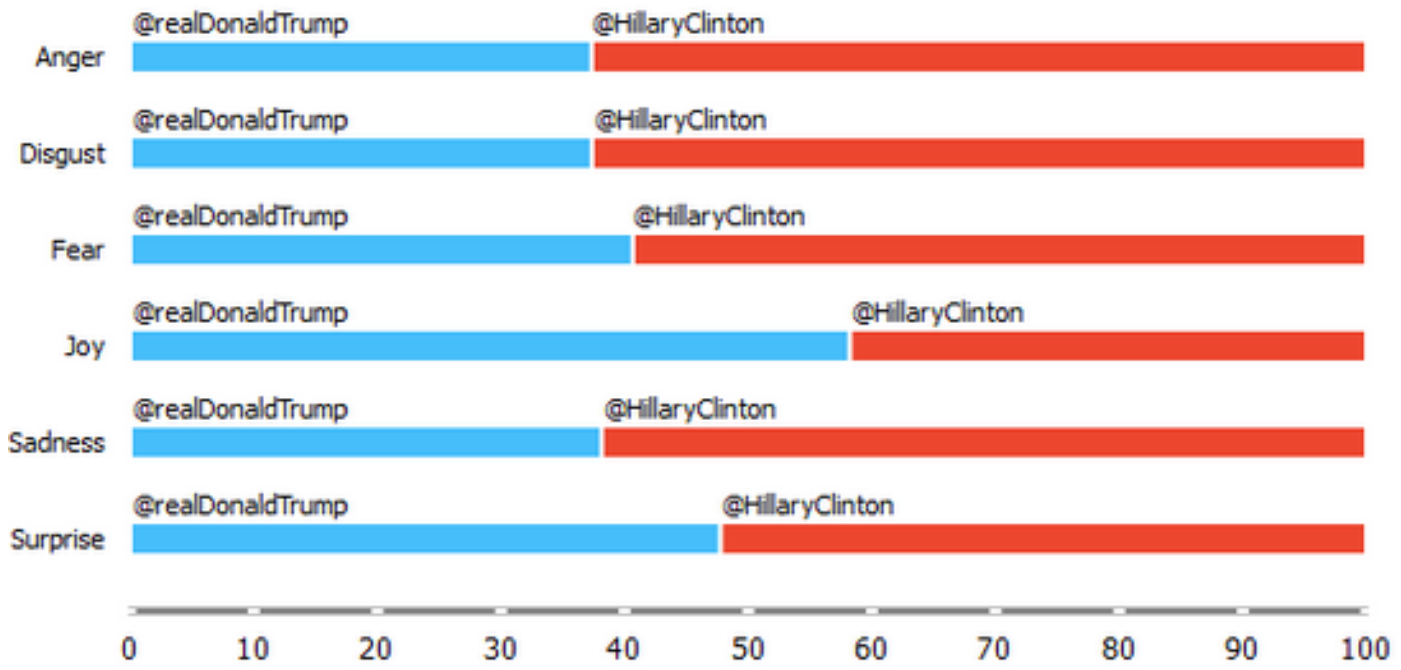

Source: here

# Sentiment analysis example: reporting on migrant crisis




Source: here

# A step forward: analysis of emotions




Source: here

# Sentiment analysis example: president elections and voters´opinion during TV confrontation of president candidates


Source: here

# Conclusions

- Text is a rich source of information.
- Rapidly developing area.
- Some challenges:
  - High assortiment of languages
  - Language development – evolvement of new words.
  - Slang, emoticons.
  - Inability to automatically detect some emotions: for example sarcasm.