Bocconi

TEXT ANALYSIS WITH PYTHON

Lesson 2 – 6/10/2021

a.y. 2020-2021

Lesson content

Text mining, Text analytics and NLP

- Tokenization: sentences and words
- Stop words
- Lexicon normalization:
  Stemming versus Lemmatization
- POS tagging
- N-grams

Attendance registration

To track your presence in class (wherever you are), please:

- either use the app on your smartphone or tablet
- or go to this web page → www.unibocconi.it/attendance

using

- your own yoU@B credentials
- today's six-digit code*

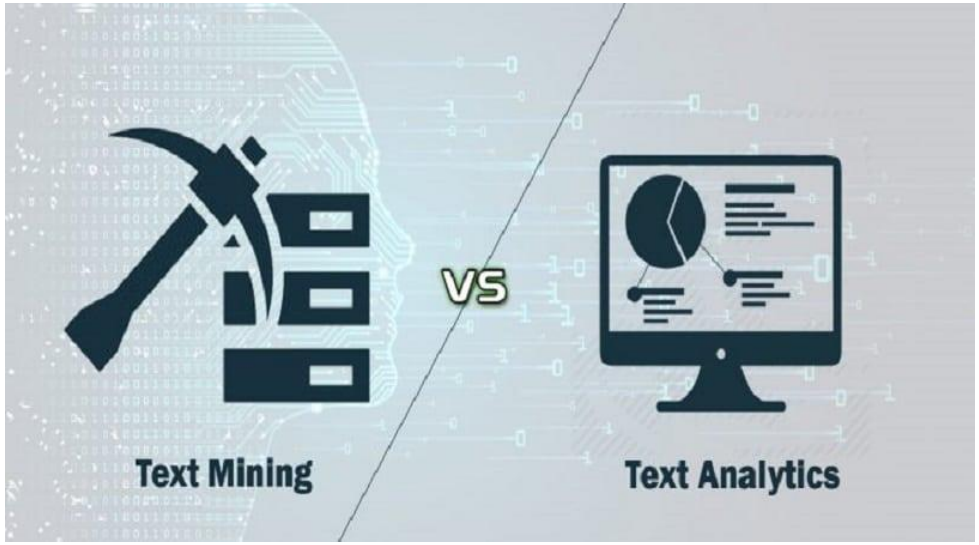
If you have problems with the app:

- try to log out and then log in again
- if the problem persists, notify our tutor via chat who will manually register your presence

((*) It will be shown in the classroom and will remain active only for 10 minutes starting from the first of you who will register

Text mining *versus* Text analytics

- exploration and "excavation" in a deposit of textual materials (corpus) for the recovery and extraction of information



- application of analysis algorithms to structured texts produced by the TM process

Text mining *and* Text analytics

- Steps of Text Mining:
  - Information Retrieval
  - Create text corpus
  - Data Preparation and Cleaning
  - Segmentation
  - Tokenization
  - Stop-word, numbers and punctuation removal
  - Stemming
  - Convert to lowercase
  - POS tagging
  - Term-Document matrix

- Steps of Text Analytics:
  - Modeling (e.g., inferential models, predictive models or prescriptive models)
  - Training and evaluation of models
  - Application of these models
  - Visualizing the models

Sources of textual data

- Websites (social media and more)
- Online databases (e.g., publications, patents or scientific articles)
- Private information sources
- Email
- Opinion surveys
- Newsletters, newsgroups, mailing lists, etc.

Terminology

- **Corpus**: collection of texts
- **Text**: collection of fragments
- **Fragment**: set of words
- **Occurrence (Token):** every appearance of a word in the text; the frequency of a word in a text is given by the number of its occurrences
- **N-Gramma**: series of n elements to be studied in sequence, where n <4/5, to identify important structural elements of the text
- **Document:** term used to refer generically to the unit of text indexed in the system and available for analysis

Preprocessing (not everything, not always!)

- Lower casing
- Removal of Punctuations
- Removal of Stopwords
- Removal of Frequent words
- Removal of Rare words
- Stemming
- Lemmatization

- Removal of emojis
- Removal of emoticons
- Conversion of emoticons to words
- Conversion of emojis to words
- Removal of URLs
- Removal of HTML tags
- Chat words conversion
- Spelling correction

Tokenization algorithms

"La disastrosa campagna di Russia (1812). Il tramonto del suo dominio sull'Europa."

Word & Punctuation

La	disastrosa	campagna	di	Russia	(	1812	)	.	Il	tramonto	del	suo	dominio	sull'Europa	.

Whitespace

La	disastrosa	campagna	di	Russia	(1812).	Il	tramonto	del	suo	dominio	sull'Europa.

Sentence

La disastrosa campagna di Russia (1812).	Il tramonto del suo dominio sull'Europa.

Regexp: \w+ (matches only words, no punctuation)

La	disastrosa	campagna	di	Russia	1812	Il	tramonto	del	suo	dominio	sull	Europa

Regexp: \w{4,} (matches words min long 4 char)

disastrosa	campagna	Russia	1812	tramonto	dominio	sull	Europa

Stemming (Porter algorithm)

	original_word	stemmed_words
0	connect	connect
1	connected	connect
2	connection	connect
3	connections	connect
4	connects	connect

	original_word	stemmed_word
0	trouble	troubl
1	troubled	troubl
2	troubles	troubl
3	troublemsome	troublemsom

Lemmatization (WordNet)

	original_word	lemmatized_word
0	trouble	trouble
1	troubling	trouble
2	troubled	trouble
3	troubles	trouble

	original_word	lemmatized_word
0	goose	goose
1	geese	goose