R introduction, scraping and text analysis

Toni Rodon
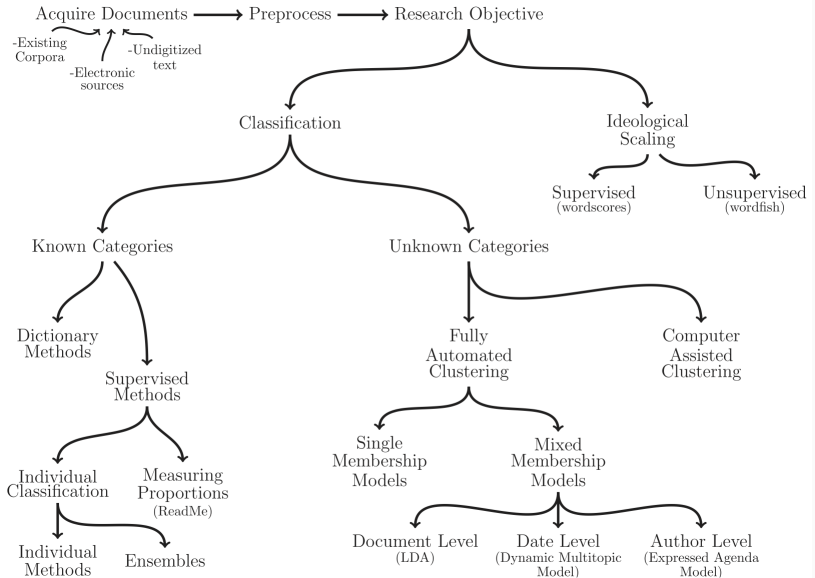
February 10, 2022

Universitat de Barcelona
🌐 www.tonirodon.cat
🐦 @tonirodon

- (text) corpus: a large and structured set of texts for analysis

Some key concepts

- (text) corpus: a large and structured set of texts for analysis
- Document: each of the units of the corpus (e.g. a FB post)

Some key concepts

- (text) corpus: a large and structured set of texts for analysis
- Document: each of the units of the corpus (e.g. a FB post)
- Types: for our purposes, a unique word

Some key concepts

- (text) corpus: a large and structured set of texts for analysis
- Document: each of the units of the corpus (e.g. a FB post)
- Types: for our purposes, a unique word
- Tokens: any word–so token count is total number of words

Some key concepts

- (text) corpus: a large and structured set of texts for analysis
- Document: each of the units of the corpus (e.g. a FB post)
- Types: for our purposes, a unique word
- Tokens: any word–so token count is total number of words
- For example, Doc1: "A corpus is a set of documents" and Doc2: "This is the 2nd document in the corpus".

Some key concepts

- (text) corpus: a large and structured set of texts for analysis
- Document: each of the units of the corpus (e.g. a FB post)
- Types: for our purposes, a unique word
- Tokens: any word–so token count is total number of words
- For example, Doc1: "A corpus is a set of documents" and Doc2: "This is the 2nd document in the corpus".
- This would be a corpus with 2 documents, where each document is a sentence. The first document has 6 types and 7 tokens. The second has 7 types and 8 tokens (we ignore punctuation for now).

Some more basic concepts

- stems: words with suffixes removed (using set of rules)

Some more basic concepts

- stems: words with suffixes removed (using set of rules)
- lemmmas: canonical word form (the base form of a word that has the same meaning even when different suffixes or prefixes are attached)

Some more basic concepts

- stems: words with suffixes removed (using set of rules)
- lemmmas: canonical word form (the base form of a word that has the same meaning even when different suffixes or prefixes are attached)
- be careful: lemmas $\neq$ stems (usage depends on analysis/RD)

word	win	winning	wins	won	winner
stem	win	win	win	won	winner
lemma	win	win	win	win	win

Some more basic concepts

- stems: words with suffixes removed (using set of rules)
- lemmmas: canonical word form (the base form of a word that has the same meaning even when different suffixes or prefixes are attached)
- be careful: lemmas $\neq$ stems (usage depends on analysis/RD)

word	win	winning	wins	won	winner
stem	win	win	win	won	winner
lemma	win	win	win	win	win

- stop words: words that are designated for exclusion from any analysis of a text (e.g. prepositions, uninformative verbs, pronounds...)

Some more basic concepts

- Pre-processing data.

Some more basic concepts

- Pre-processing data.
- Describing.

Some more basic concepts

- Pre-processing data.
- Describing.
- From this on... it depends on the RQ and RD!

Some more basic concepts

- Pre-processing data.
- Describing.
- From this on... it depends on the RQ and RD!
  - Using and developing dictionaries.

- Pre-processing data.
- Describing.
- From this on... it depends on the RQ and RD!
  - Using and developing dictionaries.
  - Supervised methods (word embeddings, sentiment analysis...)

Some more basic concepts

- Pre-processing data.
- Describing.
- From this on... it depends on the RQ and RD!
  - Using and developing dictionaries.
  - Supervised methods (word embeddings, sentiment analysis...)
  - Unsupervised methods (topic model, scaling...)

Some more basic concepts

- Pre-processing data.
- Describing.
- From this on... it depends on the RQ and RD!
  - Using and developing dictionaries.
  - Supervised methods (word embeddings, sentiment analysis...)
  - Unsupervised methods (topic model, scaling...)
  - Manual anotation and crowd coding.

Some more basic concepts

- Pre-processing data.
- Describing.
- From this on... it depends on the RQ and RD!
  - Using and developing dictionaries.
  - Supervised methods (word embeddings, sentiment analysis...)
  - Unsupervised methods (topic model, scaling...)
  - Manual anotation and crowd coding.
  - Regression models.

R introduction, scraping and text analysis

Toni Rodon

February 10, 2022

Universitat de Barcelona
www.tonirodon.cat
@tonirodon