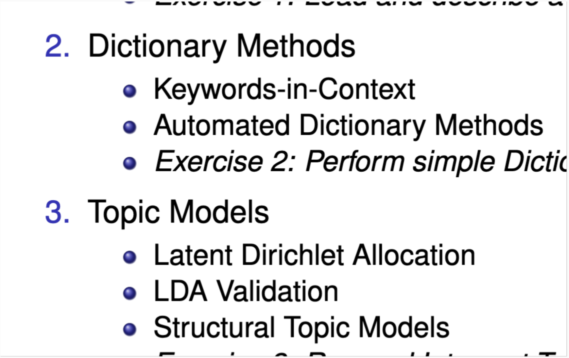
Introduction to R

Toni Rodon

February 9, 2022

R introduction, scraping and text analysis (UB)
Universitat Pompeu Fabra
🌐`www.tonirodon.cat`
🐦@tonirodon

Workflow



- *Exercise 1: Load and describe a*

2. Dictionary Methods
   - Keywords-in-Context
   - Automated Dictionary Methods
   - *Exercise 2: Perform simple Dictic*
3. Topic Models
   - Latent Dirichlet Allocation
   - LDA Validation
   - Structural Topic Models
   - *Exercise 3: Run and Interpret T*

Some key concepts

- (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

Introduction to R

Toni Rodon

February 9, 2022

R introduction, scraping and text analysis (UB)

Universitat Pompeu Fabra
🌍www.tonirodon.cat
🐦@tonirodon