Finding hidden topics in documents and other text analyses in R

Rocío Joo

UF UNIVERSITY of FLORIDA

R-Ladies

rocio.joo@ufl.edu

@rocio_joo

Extracting information from text

- Word frequency
- Word clouds
- Sentiment analysis
- Topic modeling

Extracting information from text

- **Word frequency**
- Word clouds
- Sentiment analysis
- Topic modeling

bigram <chr>	n <int>
professor mcgonagall	578
uncle vernon	386
harry potter	349
death eaters	346
harry looked	316
harry ron	302
aunt petunia	206
invisibility cloak	192
professor trelawney	177
dark arts	176

1-10 of 89,120 rows Previous 1 2 3 4 5 6 … 100 Next

Source: https://bit.ly/3gOmLTu

* I do not endorse J.K.R.'s transphobic opinions

Extracting information from text

- Word frequency
- **Word clouds**
- Sentiment analysis
- Topic modeling

Wordcloud of Harry Potter and the Sorcerer's Stone

Extracting information from text

- Word frequency
- Word clouds
- **Sentiment analysis**
- Topic modeling

- Word frequency
- Word clouds
- Sentiment analysis
- **Topic modeling**

Why participate in R-ladies?

"I like to be part of a group"

"We can learn together"

"They teach us new things in R"

- Word frequency
- Word clouds
- Sentiment analysis
- **Topic modeling**

Why participate in R-ladies?

"I like to be part of a group"

"We can learn together"

"They teach us new things in R"

- Word frequency
- Word clouds
- Sentiment analysis
- **Topic modeling**

Why participate in R-ladies?

"I like to be part of a group"

"We can learn together"

"They teach us new things in R"

Community

Learning

Why participate in R-ladies?

"I like to be part of a group"

"We can learn together"

"They teach us new things in R"

Community

Learning

Latent Dirichlet Allocation (LDA) to model **documents**

Why participate in R-ladies?

"I like to be part of a group"

"We can learn together"

"They teach us new things in R"

Community

Learning

Latent Dirichlet Allocation (LDA) to model **documents**

- Behind the **documents**, there can be a number of latent **topics**

Why participate in R-ladies?

"I like to be part of a group"

"We can learn together"

"They teach us new things in R"

Community

Learning

Latent Dirichlet Allocation (LDA) to model **documents**

- Behind the **documents**, there can be a number of latent **topics**
- The choice of **words** in the **documents** are related to the **topics**

Why participate in R-ladies?

"I like to be part of a group"

"We can learn together"

"They teach us new things in R"

Community

Learning

Latent Dirichlet Allocation (LDA) to model **documents**

- Behind the **documents**, there can be a number of latent **topics**
- The choice of **words** in the **documents** are related to the **topics**
- A **topic** is composed of a mixture of **words**.

Why participate in R-ladies?

"I like to be part of a group"

"We can learn together"

"They teach us new things in R"

Community

Learning

Latent Dirichlet Allocation (LDA) to model **documents**

- Behind the **documents**, there can be a number of latent **topics**
- The choice of **words** in the **documents** are related to the **topics**
- A **topic** is composed of a mixture of **words**.
- Interpret **topics** by their **words** composition.

Topic analysis

Steps to follow in practice:

1. Define what is a **document**
2. Get the set of documents
3. Preprocess the documents:
   - **Filter out** non informative words (e.g. prepositions) and lemmatize
   - Prepare input for LDA: list of words per document and their frequency
4. Fit **LDA** with fixed number of topics (you can use a selection criterion)
5. **Interpret** topics

Case-study: movement ecology

Identify the main topics in this field of science in the last decade

Case-study: movement ecology

Identify the main topics in this field of science in the last decade

Collaborators:

M. Boone

UF UNIVERSITY of FLORIDA

S. Picardi

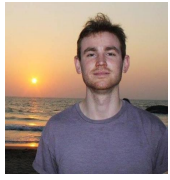
UF UNIVERSITY of FLORIDA

V. Romero

UNIVERSIDAD NACIONAL DE INGENIERÍA

M. Basille

UF UNIVERSITY of FLORIDA

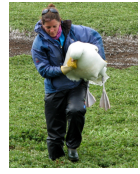


T. Clay

UNIVERSITY OF LIVERPOOL

S. Clusella-Trullas

UNIVERSITEIT IYUNIVESITHI STELLENBOSCH UNIVERSITY



S. Patrick

UNIVERSITY OF LIVERPOOL

HFSP

collaborator

noun [C]

UK /kəlæb.ə.reɪ.təʳ/ **US** /kəlæb.ə.reɪ.ţə-/

collaborator ***noun*** **[C] (ENEMY SUPPORTER)**

disapproving

a person who works with an enemy who has taken control of their country:

- *wartime collaborators*
- *a Nazi collaborator*

collaborator

noun [C]

UK /kəlæb.ə.reɪ.təʳ/ US /kəlæb.ə.reɪ.ţə˞/

collaborator *noun* [C] (ENEMY SUPPORTER)

disapproving

a person who works with an enemy who has taken control of their country:

- *wartime collaborators*
- *a Nazi collaborator*

collaborator *noun* [C] (WORKING WITH)

a person who works together with others for a special purpose:

- *a new production by Andrew Davies and collaborators*

Meaning of **collaborator** in English

collaborator

noun [C]

UK /kəlæb.ə.reɪ.təʳ/ US /kəlæb.ə.reɪ.ţə-/

collaborator *noun* [C] (ENEMY SUPPORTER)

disapproving

a person who works with an enemy who has taken control of their country:

- *wartime collaborators*
- *a Nazi collaborator*

collaborator *noun* [C] (WORKING WITH)

a person who works together with others for a special purpose:

- *a new production by Andrew Davies and collaborators*

Topic analysis

In practice, for our case study:

1. Define what is a **document**
2. Get the set of documents
3. Preprocess the documents:
   - **Filter out** non informative words (e.g. prepositions) and lemmatize
   - Prepare input for LDA: list of words per document and their frequency
4. Fit **LDA** with fixed number of topics (you can use a selection criterion)
5. **Interpret** topics

Topic analysis

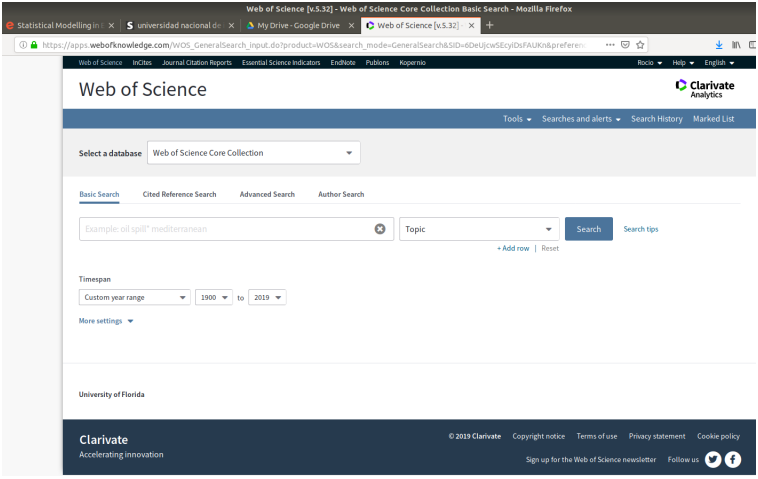
In practice, for our case study:

1. Define what is a **document**: an abstract of a movement ecology paper

Topic analysis

In practice, for our case study:

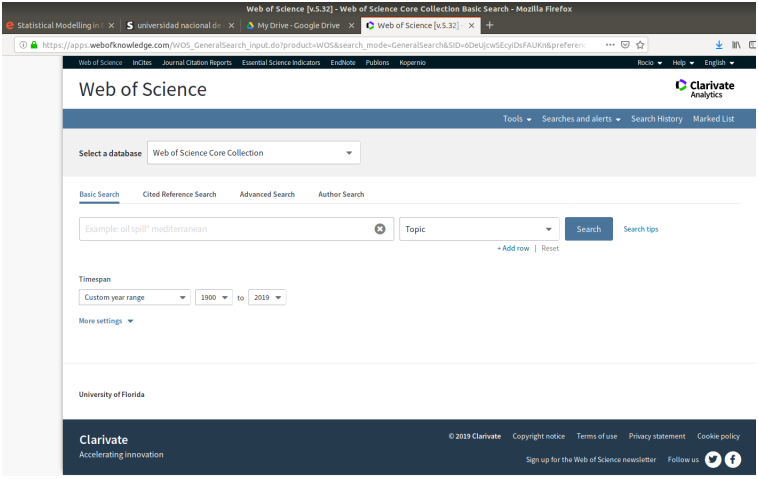
1. Define what is a **document**: an abstract of a movement ecology paper
2. Get the set of documents



Topic analysis

In practice, for our case study:

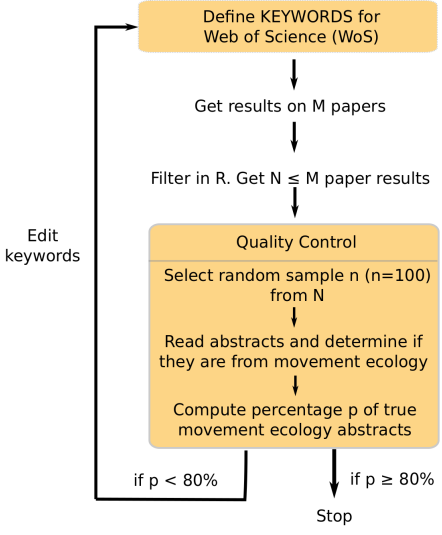
1. Define what is a **document**: an abstract of a movement ecology paper
2. Get the set of documents (example of query result)



Topic analysis

In practice, for our case study:

1. Define what is a **document**: an abstract of a movement ecology paper
2. Get the set of documents: Example on website



Topic analysis

In practice, for our case study:

1. Define what is a document: an abstract of a movement ecology paper
2. Get the set of documents: 8007 abstracts

Topic analysis

In practice, for our case study:

1. Define what is a document: an abstract of a movement ecology paper
2. Get the set of documents: 8007 abstracts
3. Preprocess the documents:
   - **Filter out** non informative words (e.g. prepositions) and lemmatize
   - Prepare input for LDA: list of words per document and their frequency
4. Fit **LDA** with fixed number of topics (you can use a selection criterion)
5. **Interpret** topics

Let's go to the codes! Accessible on website