

# Finding hidden topics in documents and other text analyses in R



Rocío Joo



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	n
<chr>	<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



Danielle Smalls-Perkins @smallperks · Jan 28

Dear Danielle,



Remember these incredibly brilliant, talented, and kind women data scientists whenever you feel like “the only one,” on a team or in a space. Thanks to #rstudioconf2020 for bringing us together! #rladies

▼

## Why participate in R-ladies?

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

*"I like to be part of a group"*

*"We can learn together"*

*"They teach us new things in R"*

## Why participate in R-ladies?

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

*"I like to be part of a group"*

*"We can learn together"*

*"They teach us new things in R"*

## Why participate in R-ladies?

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

*"I like to be part of a group"*

*"We can learn together"*

*"They teach us new things in R"*

Community

Learning

# Topic analysis

## Why participate in R-ladies?

*"I like to be part of a group"*

Community

*"We can learn together"*

Learning

*"They teach us new things in R"*

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

## Why participate in R-ladies?

*"I like to be part of a group"*

Community

*"We can learn together"*

Learning

*"They teach us new things in R"*

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

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

# Topic analysis

## Why participate in R-ladies?

*"I like to be part of a group"*

Community

*"We can learn together"*

Learning

*"They teach us new things in R"*

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**

# Topic analysis

## Why participate in R-ladies?

*"I like to be part of a group"*

Community

*"We can learn together"*

Learning

*"They teach us new things in R"*

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**.

# Topic analysis

## Why participate in R-ladies?

*"I like to be part of a group"*

Community

*"We can learn together"*

Learning

*"They teach us new things in R"*

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



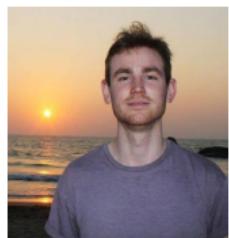
S. Picardi



V. Romero



M. Basille



T. Clay



S. Clusella-Trullas



S. Patrick



# collaborator

**noun** [ C ]

UK /kəlæb.ə.reɪ.tə/ US /kəlæb.ə.reɪ.tə/

---

**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.tə/ US /kəlæb.ə.reɪ.tə/

---

**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*

# collaborator

**noun** [ C ]

UK /kəlæb.ə.reɪ.tə/ US /kəlæb.ə.reɪ.tə/

**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:

- ➊ Define what is a **document**
- ➋ Get the set of documents
- ➌ 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
- ➍ Fit **LDA** with fixed number of topics (you can use a selection criterion)
- ➎ **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

The screenshot shows the Web of Science Core Collection Basic Search page. At the top, there are three tabs: 'Web of Science' (selected), 'InCites', and 'Journal Citation Reports'. Below the tabs, the URL is https://apps.webofknowledge.com/WOS\_GeneralSearch\_Input.do?product=WOS&search\_mode=GeneralSearch&SID=60eUjcw5EcYdsFAUKn&prefere... and the browser title is 'Web of Science [v.5.32] - Web of Science Core Collection Basic Search - Mozilla Firefox'. The main search area has a dropdown 'Select a database' set to 'Web of Science Core Collection'. Below it are four tabs: 'Basic Search' (selected), 'Cited Reference Search', 'Advanced Search', and 'Author Search'. The search bar contains the placeholder 'Example: oil spill\* mediterranean'. To the right of the search bar are dropdowns for 'Topic' (set to 'Topic'), 'Search tips', and buttons for '+ Add row' and 'Reset'. Below the search bar is a 'Timespan' section with a dropdown 'Custom year range' set to '1900 to 2019'. There is also a 'More settings' button. At the bottom of the page, there is a footer with the Clarivate logo, the text 'Accelerating innovation', copyright information (© 2019 Clarivate, Copyright notice, Terms of use, Privacy statement, Cookie policy), and social media links for Twitter and Facebook.

# Topic analysis

In practice, for our case study:

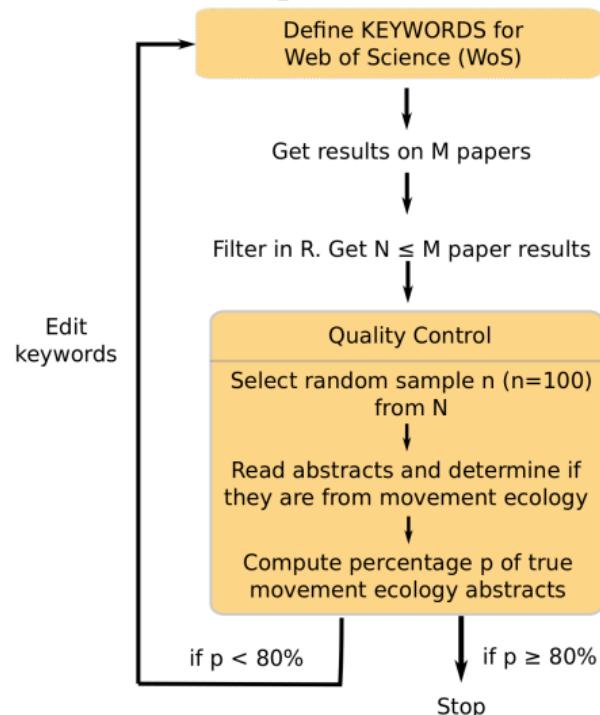
- ① Define what is a **document**: an abstract of a movement ecology paper
- ② Get the set of documents (**example of query result**)

The screenshot shows the Web of Science Core Collection Basic Search page. At the top, there's a navigation bar with tabs for 'Web of Science', 'InCites', 'Journal Citation Reports', 'Essential Science Indicators', 'EndNote', 'Publons', and 'Kopernio'. On the right of the bar are links for 'Rocio', 'Help', and 'English'. The main title 'Web of Science' is at the top left, and the Clarivate Analytics logo is on the right. Below the title, a dropdown menu says 'Select a database: Web of Science Core Collection'. There are four search tabs: 'Basic Search' (which is selected), 'Cited Reference Search', 'Advanced Search', and 'Author Search'. A search bar contains the placeholder 'Example: oil spill\* mediterranean'. To the right of the search bar are dropdown menus for 'Topic' (set to 'Topic') and 'Search tips'. A large blue 'Search' button is next to these. Below the search bar, there's a 'Timespan' section with a 'Custom year range' dropdown set from '1900' to '2019'. A 'More settings' link is also present. At the bottom of the page, there's a footer with the text 'University of Florida', the Clarivate logo ('Accelerating innovation'), and links for 'Copyright notice', 'Terms of use', 'Privacy statement', and 'Cookie policy'. It also includes a newsletter sign-up and social media links for Twitter and Facebook.

# 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](#)