MINERÍA DE TEXTOS CIENTÍFICOS
¿Cómo analizar publicaciones sin leerlas?

Rocío Joo

UF UNIVERSITY of FLORIDA

R Ladies Lima #SegundoMeetup

rocio.joo@ufl.edu

@rocio_joo

Colaboradores:

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

Contexto: aplicación en ecología del movimiento

Contexto: aplicación en ecología del movimiento

Contexto: aplicación en ecología del movimiento

QUANTITATIVE
REVIEW

MOV
ECO

Contexto: aplicación en ecología del movimiento

QUANTITATIVE
REVIEW

MOV
ECO

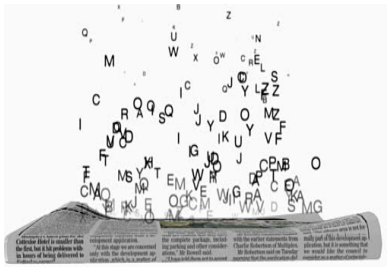
Contexto: aplicación en ecología del movimiento

QUANTITATIVE
REVIEW

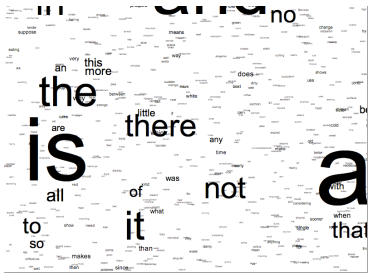
MOV
ECO

Esquema metodológico

Identify
mov-eco papers



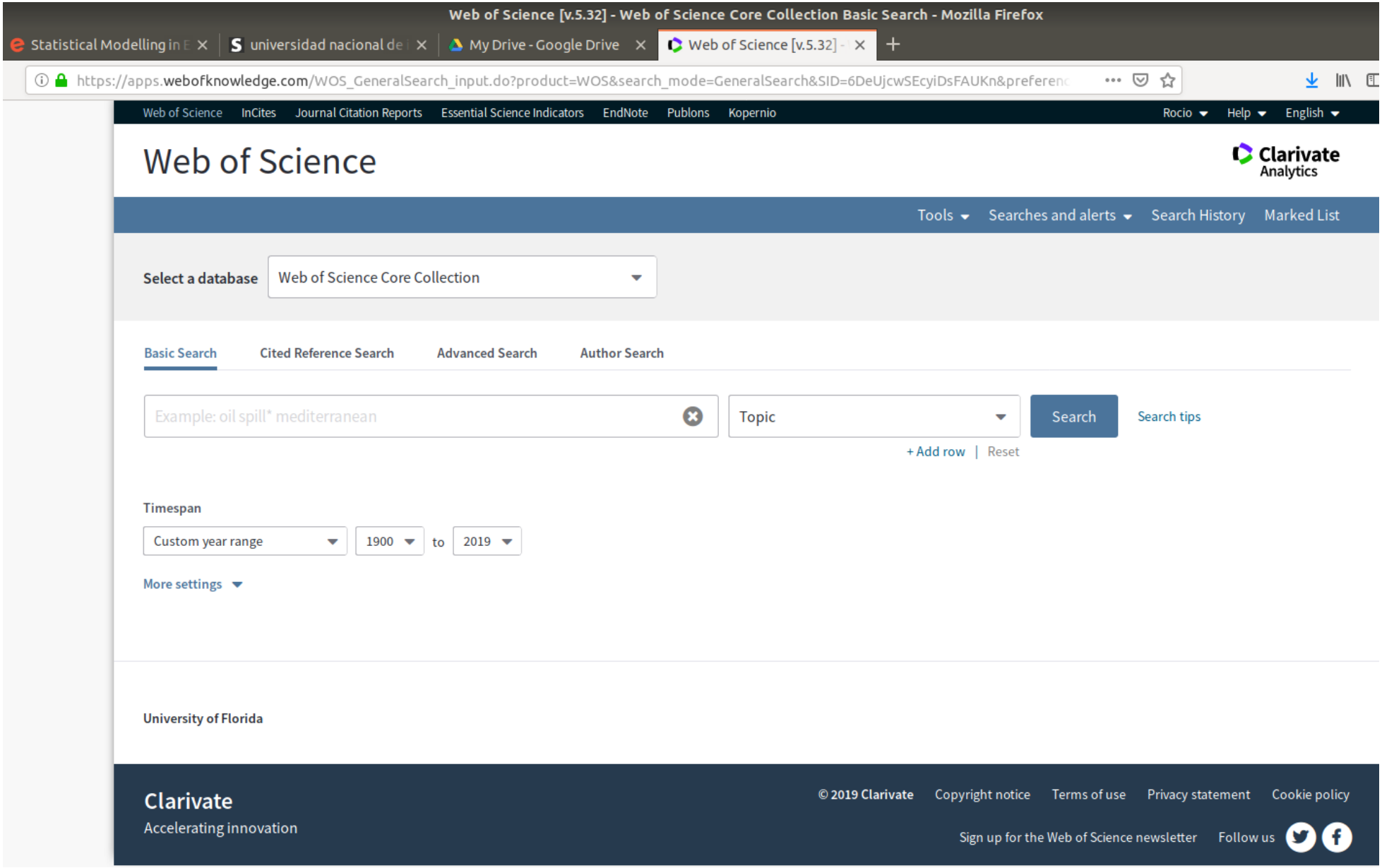
Download &
extract text



Extract key
information

Statistical
analysis

Búsqueda de papers



Palabras claves finales

Papers de ecología del movimiento: Artículos que estudian el movimiento voluntario de uno o más individuos vivientes

Palabras claves finales

Papers de ecología del movimiento: Artículos que estudian el movimiento voluntario de uno o más individuos vivientes

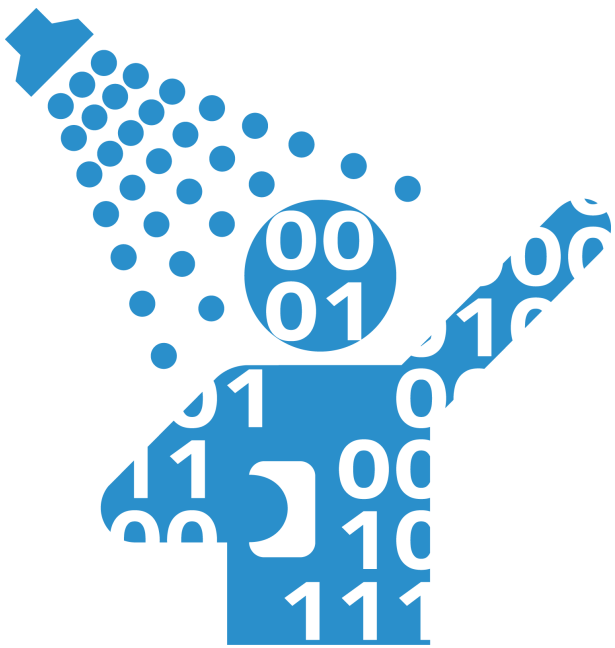
4 grupos de palabras se establecieron para la búsqueda en WoK:

1. **Behavior**: behavio
2. **Movement**: movement, moving, motion, spatiotemporal, kinematics, spatio-temporal
3. **Biologging**: telemetry, geolocat, biologg, accelerom, gps, geo-locat, bio-logg, reorient, vhf, argos, radar, sonar, gls, vms, animal-borne
4. **Individuals**: animal, individual, human, person, people, player, wildlife, fishermen

Un paper de ecología del movimiento debe tener palabras de **al menos 3** de estos grupos.

Limpieza de datos

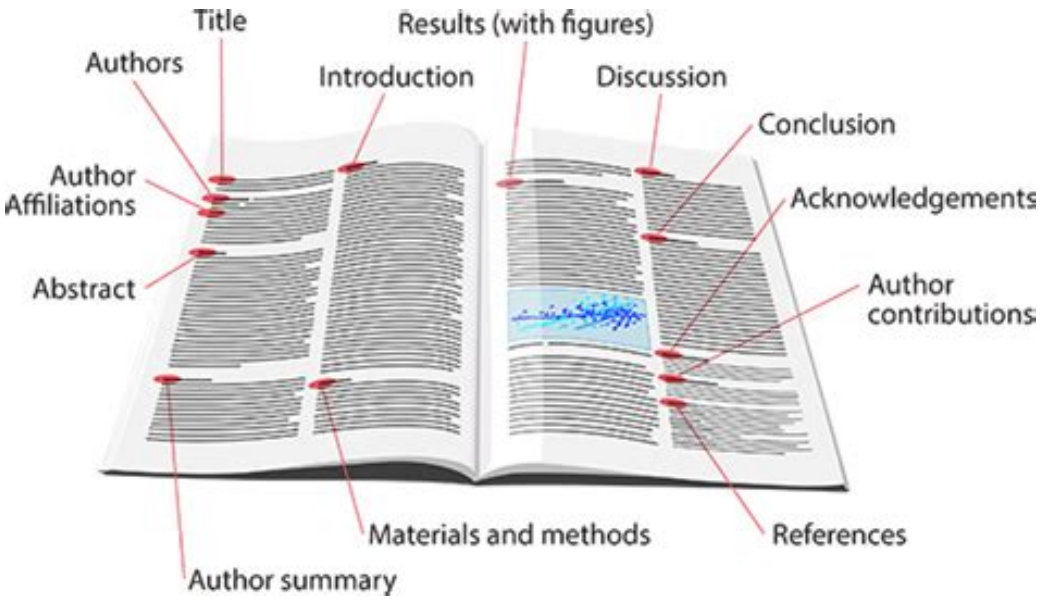
Los resultados de búsqueda son procesados nuevamente en R.



- refsplittr: ordenar las búsquedas en un sólo archivo y separar autores (*devtools* :: *install_github*('*embruna/refsplitr*'))
- Datos: DOI, título, palabras clave, abstract, autores, etc.

Descarga de artículos

Para algunos análisis, necesitaremos la sección de **Materiales y Métodos**



- fulltext: descargar artículos (depende de derechos de acceso); ejm: notebook
- formato xml o pdf

Selección de Materiales y Métodos

- xml → datos organizados en nodos padres e hijos
  - xml2: leer xml en R
  - en base a estructura de nodos, encontrar sección con título 'methods', 'model' o 'data'
  - ejm

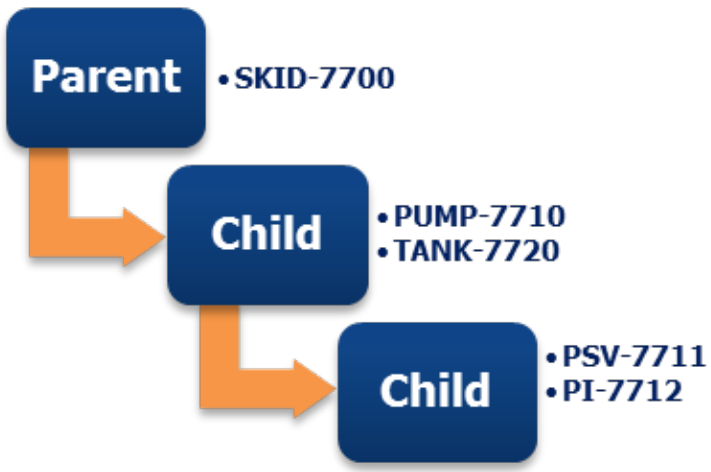
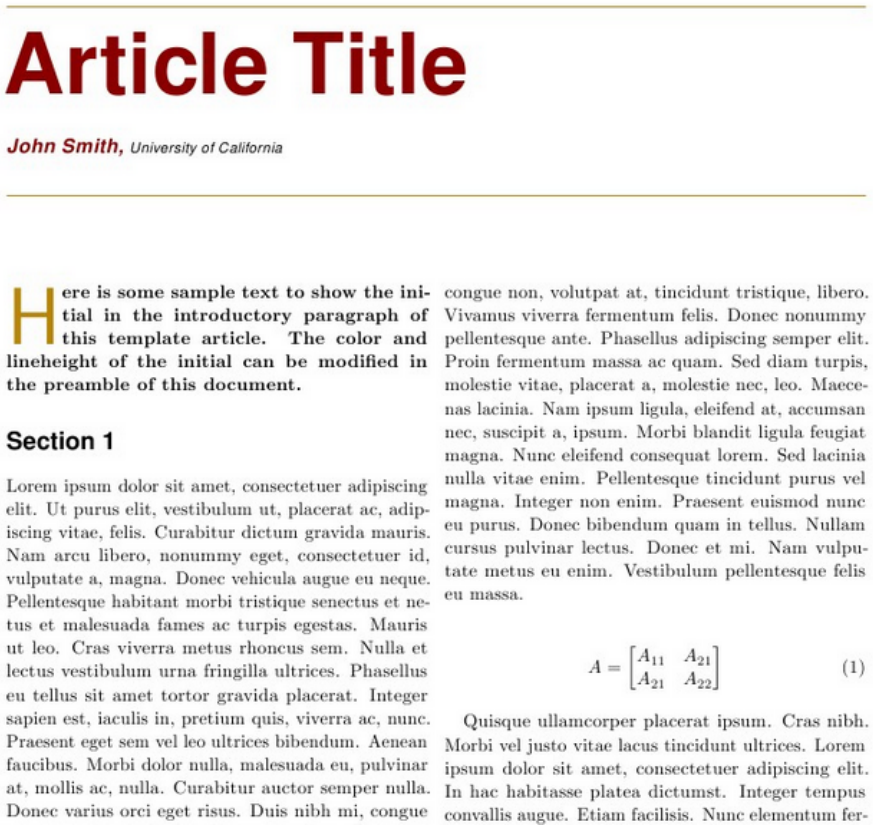


Figure 1: "Parent-Child" Hierarchical Relationship Example

Selección de Materiales y Métodos

- pdf → datos divididos sequencialmente por página

  - ¡Reto mayor!



Article Title

John Smith, University of California

Here is some sample text to show the initial in the introductory paragraph of this template article. The color and lineheight of the initial can be modified in the preamble of this document.

Section 1

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

$$A = \begin{bmatrix} A_{11} & A_{21} \\ A_{21} & A_{22} \end{bmatrix} \quad (1)$$

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fer-

Selección de Materiales y Métodos

- pdf → datos divididos sequencialmente por página
- ¡Reto mayor!

Nils Peters,* Trond Lossius,† and Jan C. Schacher**

*Center For New Music and Audio Technologies
University of California, Berkeley
1750 Arch Street
Berkeley, California 94720, USA
nils@icsi.berkeley.edu
†BEK, Bergen Center for Electronic Arts
C. Sundtsgt. 55
5004 Bergen, Norway
trond.lossius@bek.no
**Institute for Computer Music and Sound Technology
Zurich University of the Arts
Baslerstrasse 30
8048 Zurich, Switzerland
jan.schacher@zhdk.ch

The Spatial Sound Description Interchange Format: Principles, Specification, and Examples

Abstract: SpatDIF, the Spatial Sound Description Interchange Format, is an ongoing collaborative effort offering a semantic and syntactic specification for storing and transmitting spatial audio scene descriptions. The SpatDIF core is a lightweight minimal solution providing the most essential set of descriptors for spatial sound scenes. Additional descriptors are introduced as extensions, expanding the namespace and scope with respect to authoring, scene description, rendering, and reproduction of spatial sound. A general overview presents the principles informing the specification, as well as the structure and the terminology of the SpatDIF syntax. Two use cases exemplify SpatDIF's potential for pre-composed pieces as well as interactive installations, and several prototype implementations that have been developed show its real-life utility.

Introduction

SpatDIF, the Spatial Sound Description Interchange Format, presents a structured approach for working with spatial sound information, one that addresses the different tasks involved in creating and performing spatial sound.

A major problem when working on spatial sound is that the methods and the resulting works are often tied to a specific system or infrastructure—for example, with regards to the software used for composition and reproduction or the available speaker arrangement and the characteristics of the physical venue. The lack of flexibility this produces impedes the exchange of pieces between different venues, the mixing of different tools for authoring or performing the piece, and ultimately the preservation of the work in a sustainable form that is independent of the technology used to create it.

The goal of SpatDIF is to simplify and enhance the methods of working with spatial sound content. SpatDIF proposes a simple, minimal, and extensible format as well as best-practice implementations for storing and transmitting spatial sound scene descriptions. It encourages portability and the exchange of compositions between venues with different surround-sound infrastructures. SpatDIF also fosters collaboration between artists such as composers, musicians, sound installation artists, and sound designers, as well as researchers in the fields of acoustics, musicology, sound engineering, and virtual reality.

SpatDIF strives to be human-readable, easily understood and unambiguous, platform- and implementation-independent, extensible, and free of license restrictions.

SpatDIF's applications are not limited to the sound-scene concept alone. With both its ability to communicate time-independent metadata and its extensibility with further types of data descriptors, the format is open to other related fields such as sound synthesis, compositional algorithms, and abstract spatial geometry.

SpatDIF is developed as a collaborative effort and has evolved over a number of years. The online community and all related information can be found at www.spatdif.org.

Computer Music Journal, 37:1, pp. 11–22, Spring 2013
doi:10.1162/COMJ_a_00167

Selección de Materiales y Métodos

- pdf → datos divididos sequencialmente por página

- ¡Reto mayor!
- tm: leer pdf en R
- en base a estructura de la revista, reconstruir texto en una columna
- identificar 1ra y última línea de sección métodos
- extraer texto entre ambas
- ejm

Nils Peters,* Trond Lossius,† and Jan C. Schacher**
*Center For New Music and Audio Technologies
University of California, Berkeley
1750 Arch Street
Berkeley, California 94720, USA
nils@icsi.berkeley.edu
†BEK, Bergen Center for Electronic Arts
C. Sundtsgt. 55
5004 Bergen, Norway
trond.lossius@bek.no
**Institute for Computer Music and Sound Technology
Zurich University of the Arts
Baslerstrasse 30
8048 Zurich, Switzerland
jan.schacher@zhdk.ch

The Spatial Sound Description Interchange Format: Principles, Specification, and Examples

Abstract: SpatDIF, the Spatial Sound Description Interchange Format, is an ongoing collaborative effort offering a semantic and syntactic specification for storing and transmitting spatial audio scene descriptions. The SpatDIF core is a lightweight minimal solution providing the most essential set of descriptors for spatial sound scenes. Additional descriptors are introduced as extensions, expanding the namespace and scope with respect to authoring, scene description, rendering, and reproduction of spatial sound. A general overview presents the principles informing the specification, as well as the structure and the terminology of the SpatDIF syntax. Two use cases exemplify SpatDIF's potential for pre-composed pieces as well as interactive installations, and several prototype implementations that have been developed show its real-life utility.

Introduction

SpatDIF, the Spatial Sound Description Interchange Format, presents a structured approach for working with spatial sound information, one that addresses the different tasks involved in creating and performing spatial sound.

A major problem when working on spatial sound is that the methods and the resulting works are often tied to a specific system or infrastructure—for example, with regards to the software used for composition and reproduction or the available speaker arrangement and the characteristics of the physical venue. The lack of flexibility this produces impedes the exchange of pieces between different venues, the mixing of different tools for authoring or performing the piece, and ultimately the preservation of the work in a sustainable form that is independent of the technology used to create it.

The goal of SpatDIF is to simplify and enhance the methods of working with spatial sound content. SpatDIF proposes a simple, minimal, and extensible format as well as best-practice implementations for storing and transmitting spatial sound scene descriptions. It encourages portability and the exchange of compositions between venues with different surround-sound infrastructures. SpatDIF also fosters collaboration between artists such as composers, musicians, sound installation artists, and sound designers, as well as researchers in the fields of acoustics, musicology, sound engineering, and virtual reality.

SpatDIF strives to be human-readable, easily understood and unambiguous, platform- and implementation-independent, extensible, and free of license restrictions.

SpatDIF's applications are not limited to the sound-scene concept alone. With both its ability to communicate time-independent metadata and its extensibility with further types of data descriptors, the format is open to other related fields such as sound synthesis, compositional algorithms, and abstract spatial geometry.

SpatDIF is developed as a collaborative effort and has evolved over a number of years. The online community and all related information can be found at www.spatdif.org.

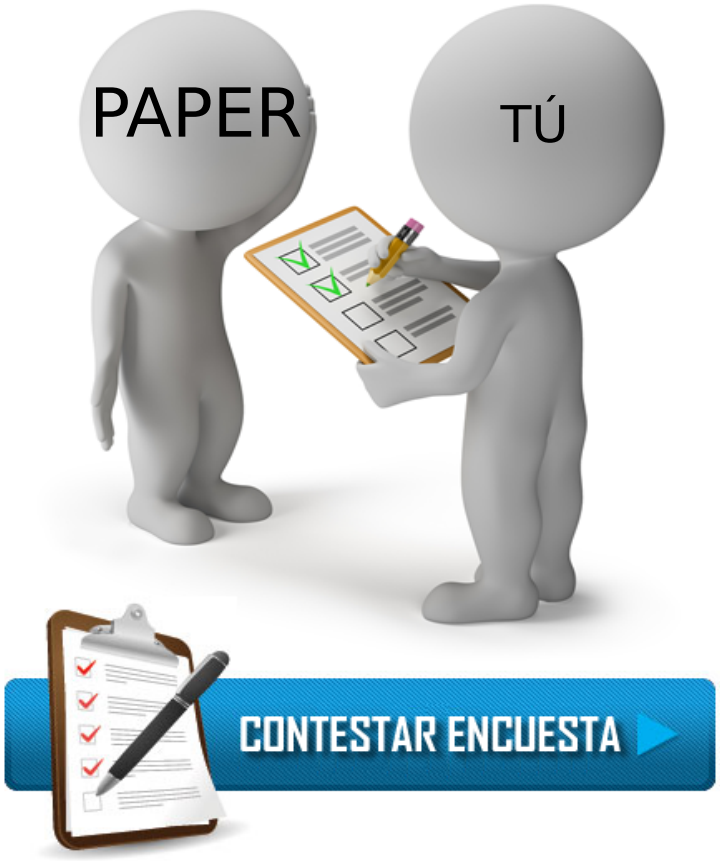
Computer Music Journal, 37:1, pp. 11–22, Spring 2013
doi:10.1162/COMJ_a_00167

Análisis de datos

Análisis de datos

Imaginemos que cada artículo es una persona...

Preguntas abiertas y cerradas



Abiertas

La respuesta es libre:

¿Por qué te gusta el teatro?

¿Qué te motiva a viajar?

¿Qué opinas del presidente?

Cerradas

La respuesta es específica.

¿Escuchas la radio?

a) Si b) No

¿Ves televisión?

a) Si b) No

Conveniencia de las preguntas abiertas o cerradas

Cerradas.

Ventajas

- Son fáciles de codificar.
- Facilidad de análisis.

Desventajas

- Limitan las respuestas de la muestra.

Abiertas

Ventajas

- Son particularmente útiles cuando no tenemos información sobre posibles repuestas de las personas.

Desventajas

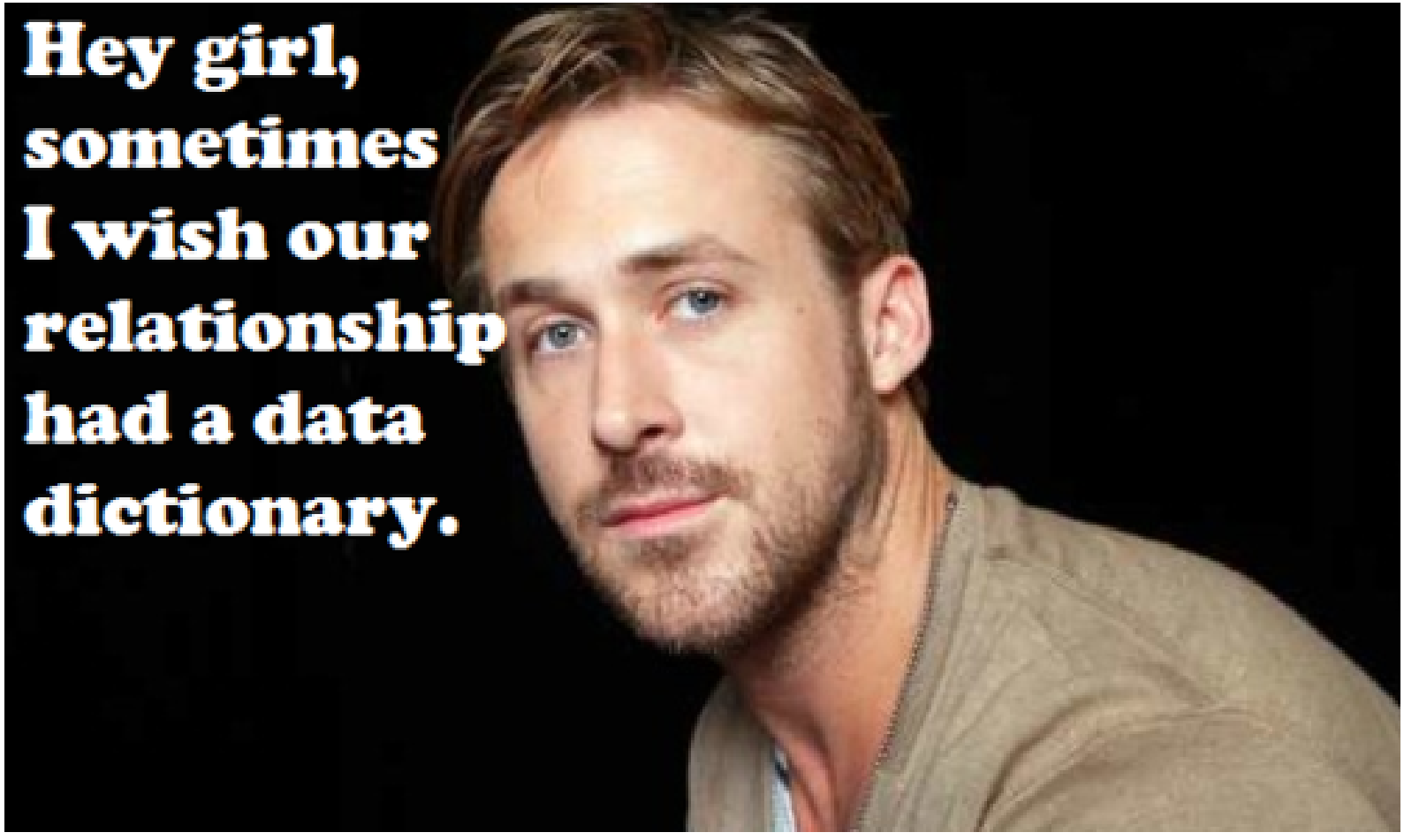
- Son más difíciles de codificar, clasificar y analizar.

Análisis de datos: preguntas cerradas

Desarrollo de diccionarios

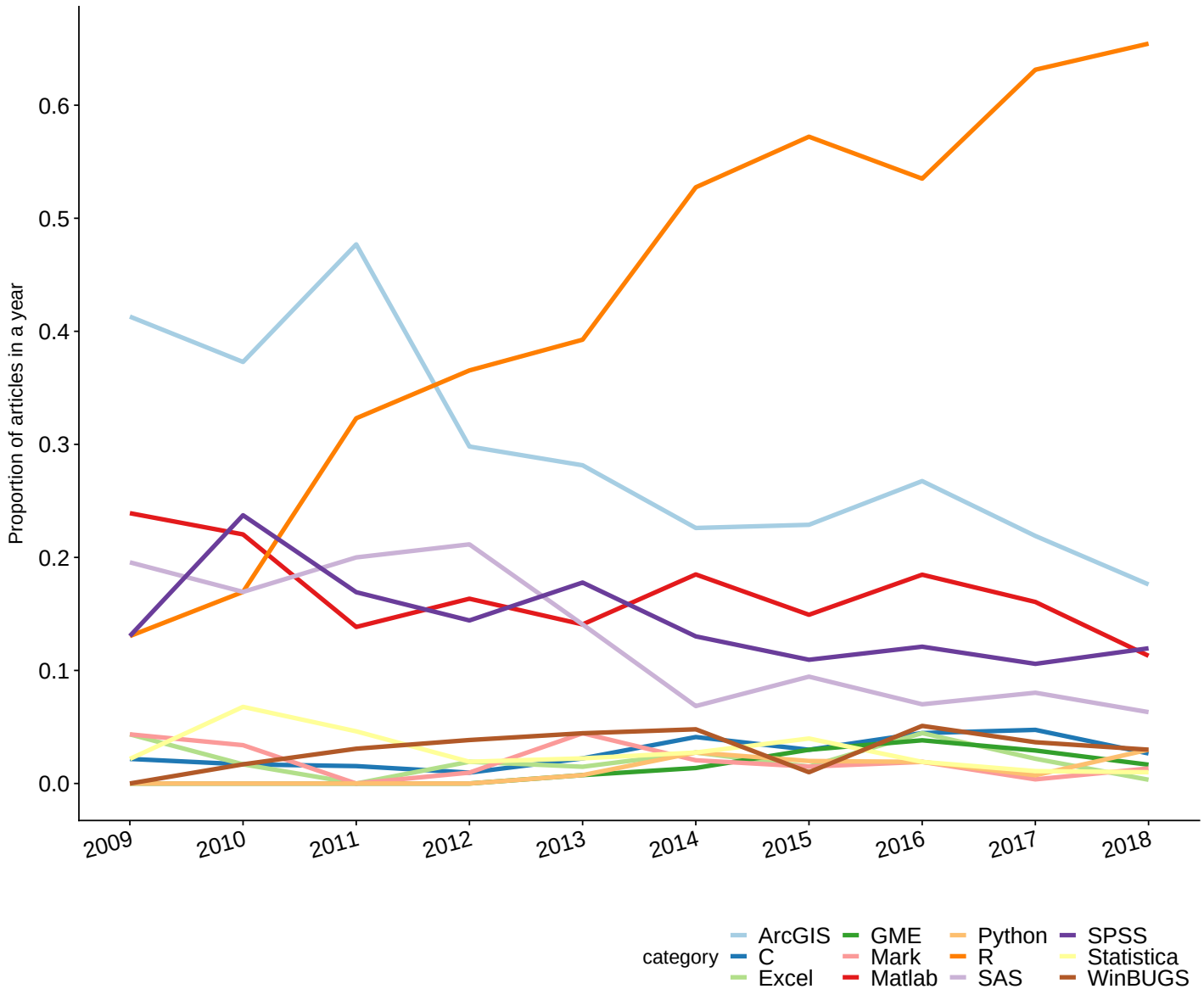
- Ejm: diccionario de software: ¿Qué software utilizas? (encuestado: paper)
- Categorías: R, Matlab, Python, etc.
- Palabras de búsqueda
- Reglas para las palabras
- Procesar la búsqueda
- Control de calidad

Análisis de datos: preguntas cerradas



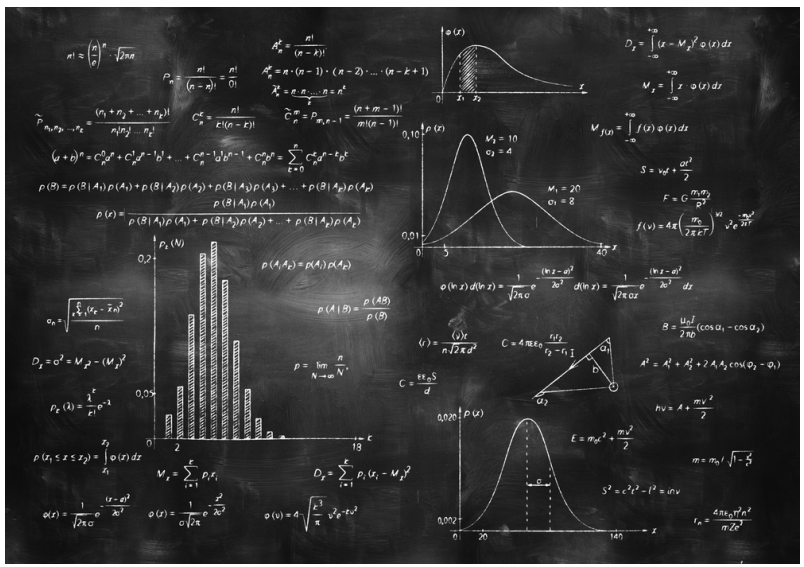
Análisis de datos: preguntas cerradas

Algunos resultados:



Análisis de datos: preguntas cerradas

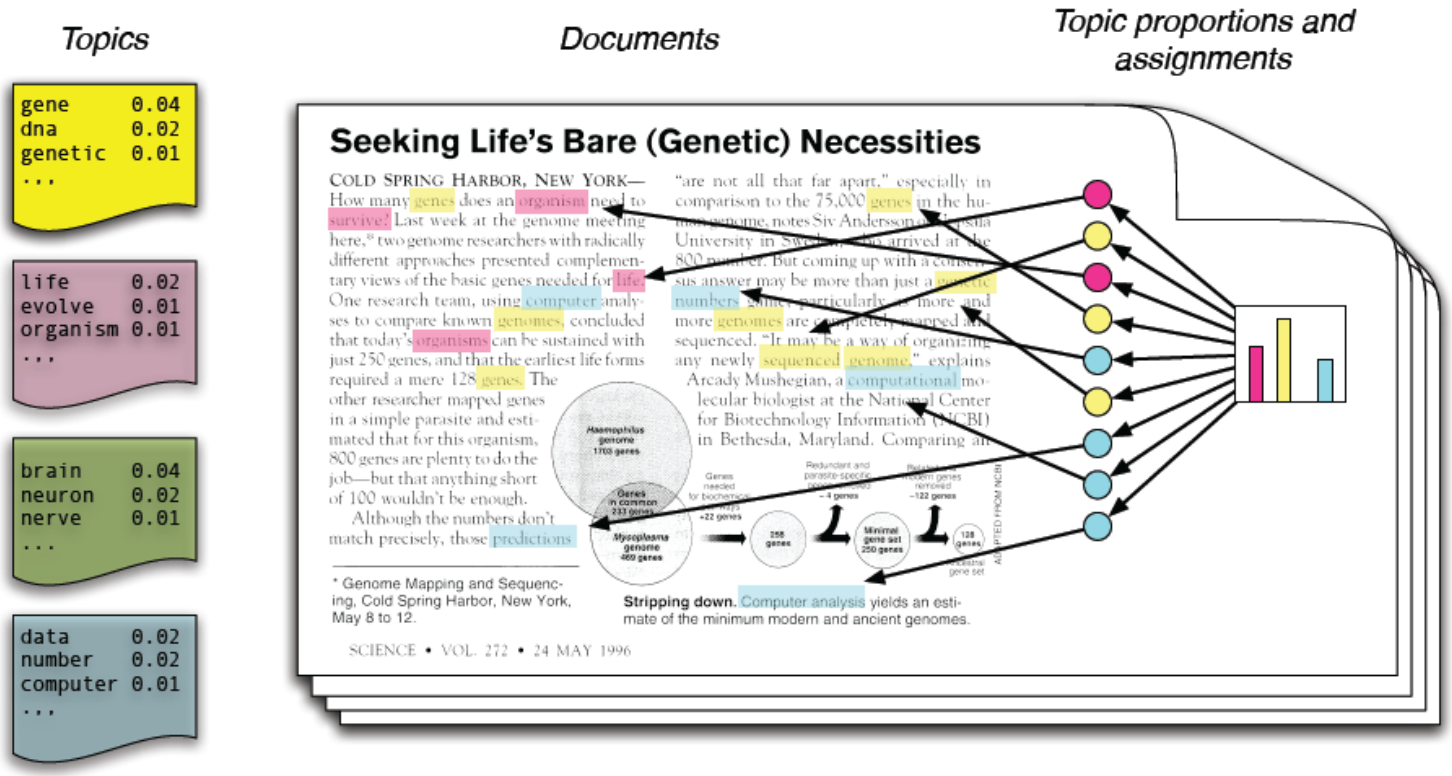
Métodos estadísticos:



Method	Perc.
general	62,2 %
Movement	20,5 %
spatial	7,8 %
time series	7,2 %
social	2,3 %
other	0,7 %
spatiotemporal	0,4 %

Análisis de datos: preguntas abiertas

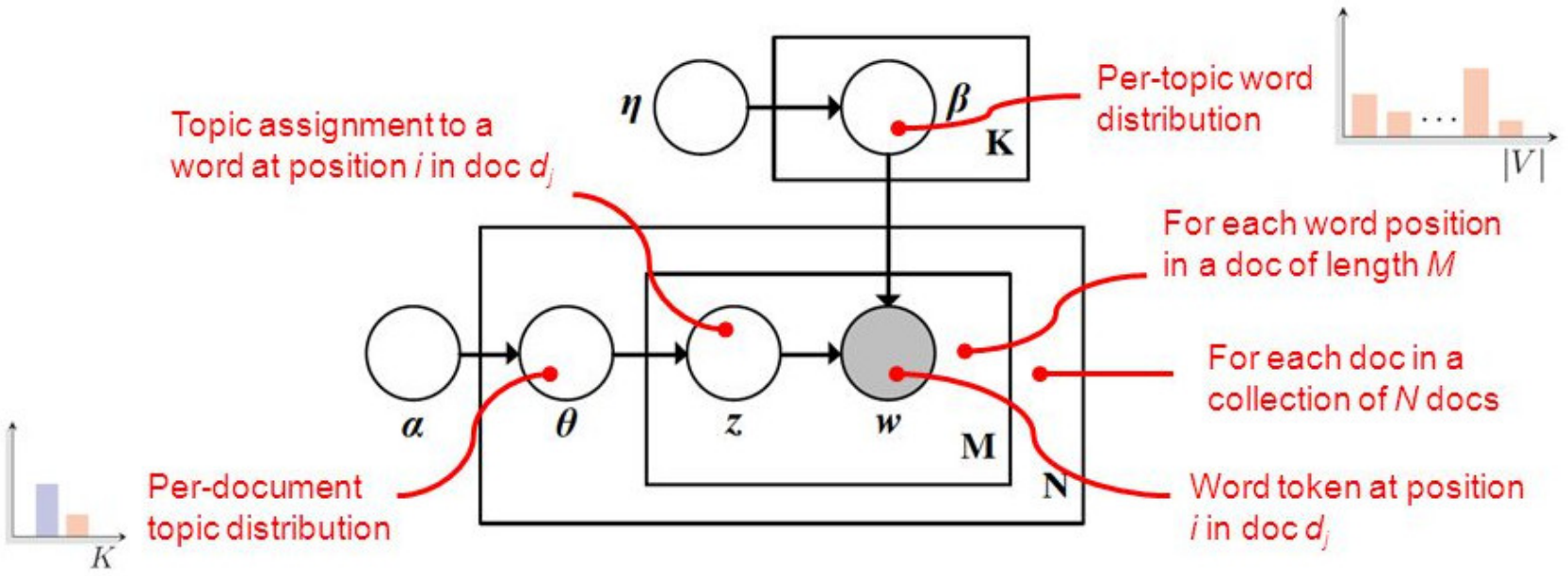
Latent Dirichlet Allocation (LDA)



- Cada **tema** es modelado como una distribución de palabras
- Cada **documento** es una mezcla de temas
- Cada **palabra** sale de uno de esos temas

Análisis de datos: preguntas abiertas

Latent Dirichlet Allocation (LDA)



[Moens and Vulic, Tutorial @WSDM 2014]

$$p(\beta, \theta, z, \omega \mid \alpha, \eta) =$$
$$\prod_{i=1}^{K} \left\{ p(\beta_i \mid \eta) \prod_{d=1}^{N} \left[p(\theta_d \mid \alpha) \left(\prod_{n=1}^{N} p(z_d \mid \theta_d) p(\omega_{d,n} \mid \beta_{1:K}, z_{d,n}) \right) \right] \right\}$$

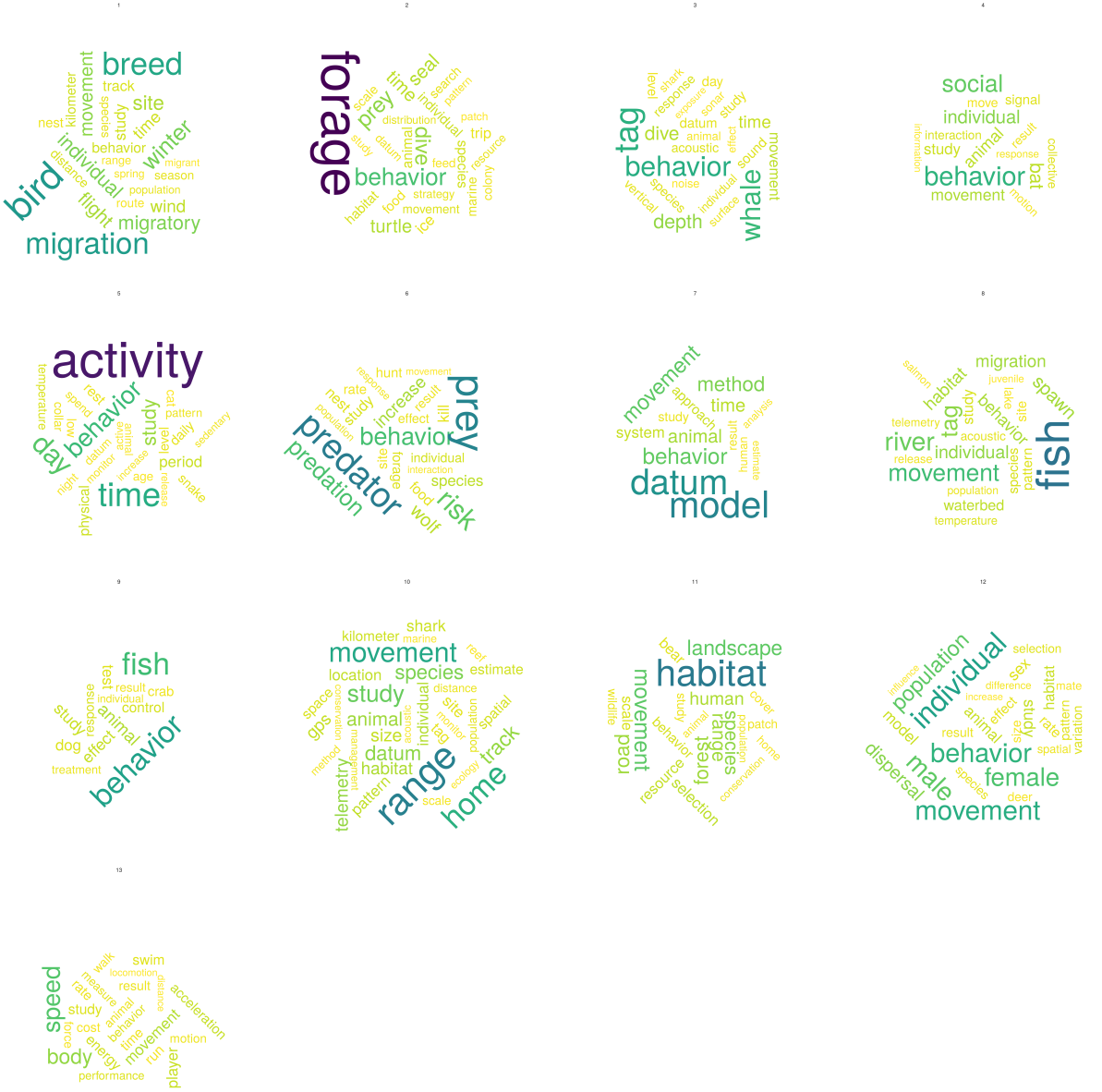
Análisis de datos: preguntas abiertas

Latent Dirichlet Allocation (LDA)

- Estimación de parámetros: Gibbs, EM and variants
- Elegir el número de temas:
  - Perplexity: Perplejidad del modelo sobre los datos, medida por la log-verosimilitud en datos de prueba.
  - Ojo: no necesariamente lleva a temas interpretables
  - Elegir el número mínimo de temas interpretables (ejm)
  - Control de calidad

Análisis de datos: preguntas abiertas

Latent Dirichlet Allocation (LDA)



rocio.joo@ufl.edu

@rocio_joo