TEXT MINING SCIENTIFIC PAPERS with R

Rocío Joo

UF UNIVERSITY of FLORIDA

MIAMI R-Ladies

rocio.joo@ufl.edu

@rocio_joo

Colaborators:

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

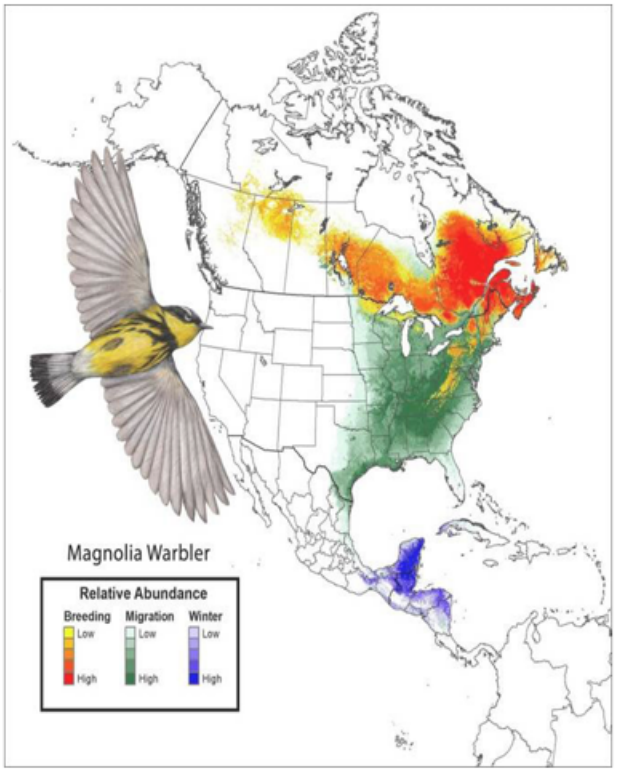
Context: movement ecology

Context: movement ecology

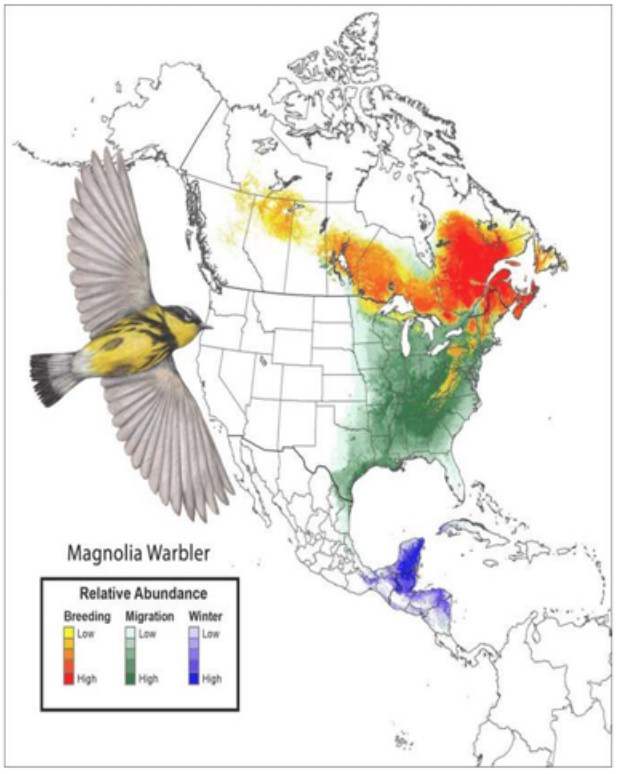
Context: movement ecology

Context: movement ecology

Context: movement ecology

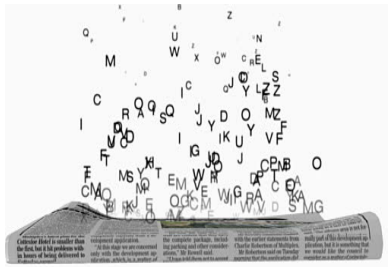


Context: movement ecology

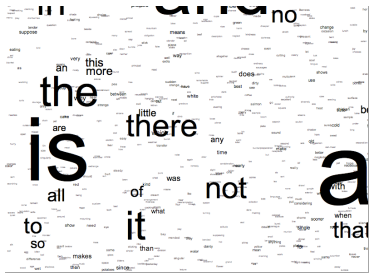


Workflow

Identify
mov-eco papers



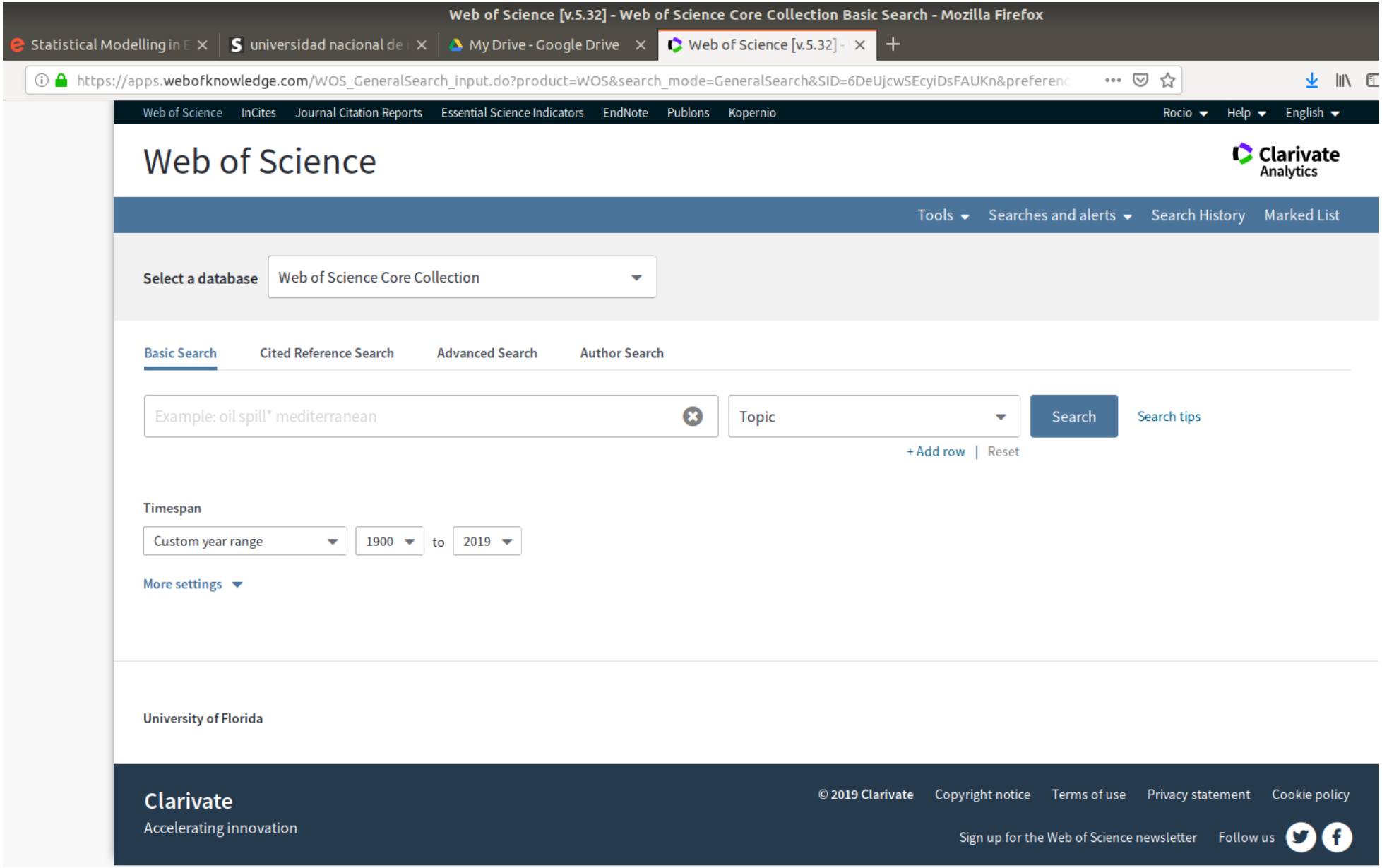
Download &
extract text



Extract key
information

Statistical
analysis

Searching for papers



Final keywords

Movement ecology papers: Papers that studied the voluntary movement of one or more living individuals

Final keywords

Movement ecology papers: Papers that studied the voluntary movement of one or more living individuals

4 groups of keywords were established for the search in WoK:

1. **Behavior**: behavio
2. **Movement**: moveme, 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

Papers had to have words from **at least 3 groups** to be selected.

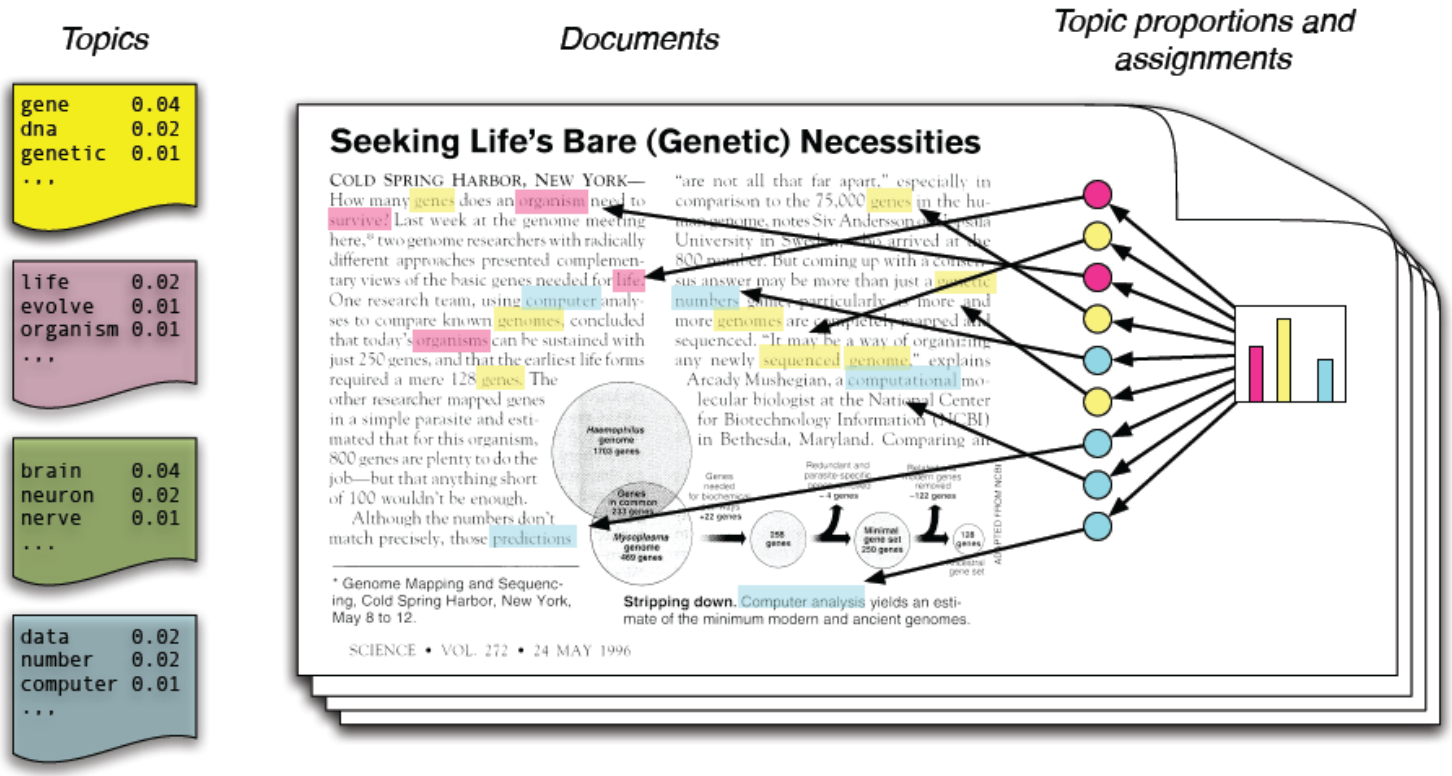
Cleaning query results

- Query results are processed again in R
- Mostly because of this exception:
  - **Unless** they had words from **group 3**, they shouldn't contain:
  - **Missleading words**: cell, DNA, enzyme, strain, neuro, atom, molecule, lymph, cortex, cortic, receptor, patient, prosthese, eye, particle, tectonic, counsel, cognit, market, spine, questionnaire, sedentary, insulin, peristal, muscle, amput, nervous, retinal, psychiatric, disease, virus, remotely-guided, tissue, polymer, mri

Example in R

Topic analysis

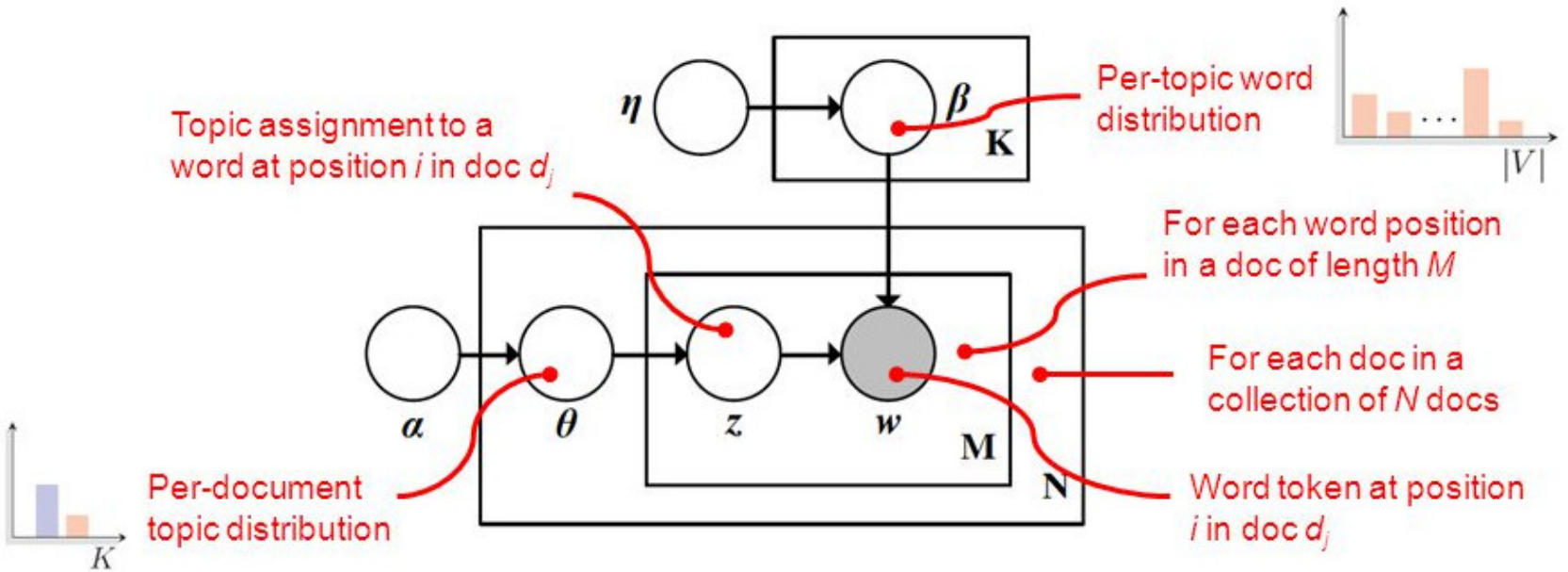
Latent Dirichlet Allocation (LDA)



- Each **topic** is modeled as a distribution of words
- Each **document** is a mixture of topics
- Each **word** comes out of one of these topics

Topic analysis

Latent Dirichlet Allocation (LDA)



[Moens and Vulic, Tutorial @WSDM 2014]

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

Topic analysis

Steps to follow in practice:

- Define what a **document** is: an abstract
- **Filter out** non informative words (e.g. prepositions) and lemmatize
- Prepare input for LDA: list of words per document and their frequency
- Fit **LDA** with fixed number of topics
- **Interpret** topics

Example in R

rocio.joo@ufl.edu

rocio_joo