**TFM**

**Evolution of Science Topics**

**in Spain**

**Alberto Corbella Cruz**

**INDEX:**

[1. Introduction 3](#_Toc25701615)

[2. Data Description 4](#_Toc25701616)

[3. Methodology 5](#_Toc25701617)

[4. Summary 8](#_Toc25701618)

[5. Conclusions 9](#_Toc25701619)

6. Visualization Data Sets ……………………………………………………………………. 10

[7. Front End – User Manual. 11](#_Toc25701620)

1. Introduction

Science publishing is increasing heavily during the last years with more and more publishers in the market. One of the main reasons contributing to this growth is the entry in the market of the Open Access publishers, making science open for the society.

With this study I will try to identify if this global growth is also happening in our country and what science fields are the most investigated in our country, driving this growth.

With a similar study, and for example extending it to a greater area ( i.e Europe or other continents ) and adding additional info ( i.e impact of this articles, in terms of views/ downloads) , can help in future to identify trends in the different areas / topics this would allow publishers to enhance their services / resources in the different areas, by creating new journals or redefining them, acquiring more editors , etc.

As science is dynamic, publishers sometimes react afterwards some topic is growing or showing a descendant trend, so this type of analyses can help, as mentioned before, to anticipate this trends and act accordingly.

2 Data Description

To build my data set I have used an API provided by Dimensions (Digital Science) . This API contains information of papers published all over the world, and related info like authors, authors affiliations, funders, etc

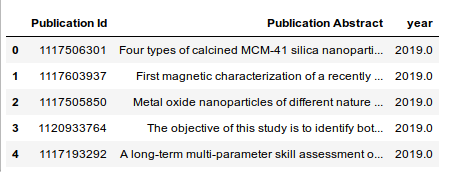
The idea was to get papers published in Spain from 2009, so I have taken all papers coming from the Top 25 Institutions in Spain (this cover for more than 50% of the total papers published in Spain in this period).

Using the language indicated to interact with the API (DSL) we get the list of publications Ids and the year of publication.

After this, as full abstract content is not provided and we can not get it using DSL queries, I created a web scraper that, by using the publication Ids, will iterate over the API interface and retrieve from the html the abstract full text.

After completing this process for all publications ( this is a very long process, taking several hours / even days ), I applied some cleansing in the abstracts obtained, by removing html tags coming with the scraper and keeping only abstracts written in English ( this last step reduced my dataset from around 335 k to 319,821 papers ).

So final data structure is a simple dataset of 3 columns:

* Publication Id
* Publication Abstract (containing the full abstract text)
* Publication year of the paper
* 

3 Methodology

*3.1 Algorithms*

We are trying to identify the topics behind the full set of abstracts. Hence, for this purpose we are using Topic modelling algorithms.

Before starting with these models, we need to carry out a carefully pre process of the abstracts text, by removing stopwords, special characters etc. Also I decided to keep only NOUNs as some studies demonstrated most efficient results keeping only this part of speech (<https://www.aclweb.org/anthology/U15-1013.pdf>)

I have tried 2 Topic Modelling algorithms, Latent Dirichlet Allocation (LDA) and Latent Semantic Indexing (LSI), and compare their performance with their coherence scores.

LSI learns latent topics by performing matrix decomposition on the word – document matrix. It has proven to be a lot faster but less accurate than LDA, giving also topics more difficult to interpret.

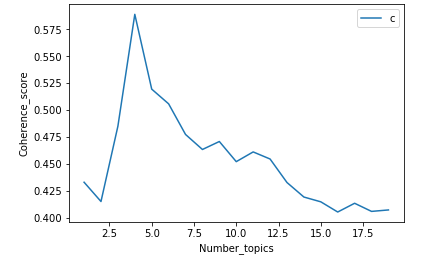
Hence, we have used LDA, which learns the representation of a fixed number of topics given by us and based on this number learn the topic distribution for each document (abstract in our case) in a collection of documents.

*3.2 Evaluation Metrics*

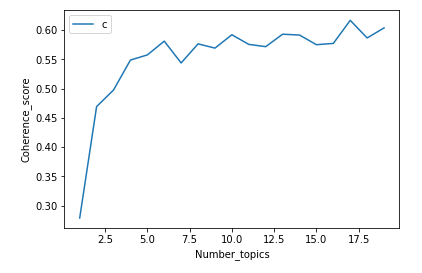
In order to decide between the 2 models, aside from the topic outputs, we have used Coherence Model. Topic Coherence measures score a single topic by measuring the degree of semantic similarity between high scoring words in the topic. These measurements help distinguish between topics that are semantically interpretable topics and topics that are artefacts of statistical inference.

As explained, our LDA and LSI models require a fixed number of topics that we need to decide. Hence by defining a function I have run the models within a range of 1-20 topics and based on the coherence score, I got to choose LDA as it showed more robust results and decided to go for 10 topics for easier interpretability.

LSI Coherence Model Scores

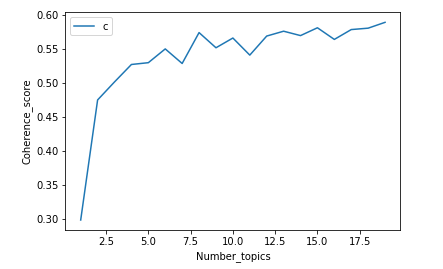


LDA Coherence Model Scores



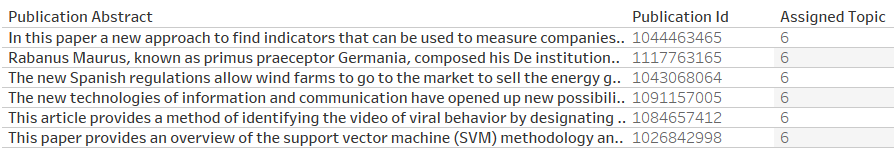
One interesting challenge I have found during the creation of my model was the result I had from lemmatizing the words from the abstracts. While this is a normal process when carrying out topic modelling, the scores I got from the evaluation graph of my model were slightly worse with the words lemmatize.

LDA Coherence Scores – Applying Lemmatization

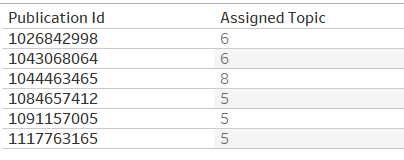


This could be observed specially after assigning the topics given by each of the models. Evaluating the abstracts, I observed unrelated abstracts classified under the same topic with lemmatize LDA, whereas the no lemmatized LDA gave a better and more coherent topic assignation. Below an example with a set of publication Ids

LDA – Lemmatize (abstract with a very different content assigned to same topic)

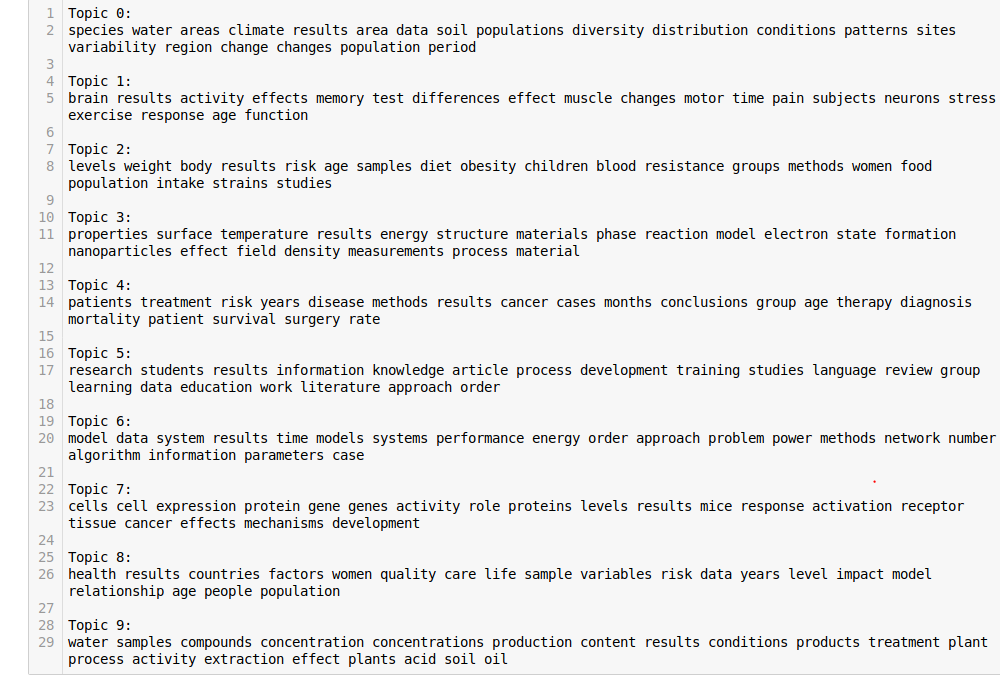


LDA – No lemmatize



4 Summary

We inferred the 10 topics to each of the outputs resulted from our LDA model, based on the distribution of words, giving them a science field given the meaning of the words. Top 20 words for each topic are:



Hence, we can identify the following science fields:

* Topic 0: Ecology & Environmental Sciences
* Topic 1: Neuroscience
* Topic 2: Nutrition and Physiology
* Topic 3: Physics & Chemistry
* Topic 4: Medical Care
* Topic 5: Psychology, Education & Other Social Sciences
* Topic 6: Stats, Computer Science & Engineering
* Topic 7: Cell & Genes Studies
* Topic 8: Sustainability
* Topic 9: Biology

5 Conclusions

Scientific publishing shows a strong growth in our country in this period, with all the scientific topics growing in number of publications. (We need to bear in mind that data is up to November 2019, so not the full 2019 year )

Based on the topics we have identified we see that Physics & Chemistry has been the field with the higher number of publication in our country in this time period, however is the field growing less in the period, even being surpassed by the topic containing publications related to Stats , Computer Science & Engineering. Machine Learning and Data Science increasing demand in these last years might have been one of the reasons of this change. (*Maybe physicians switching to this field due to the high demand on Machine Learning and Data Science?*)

Topic showing the highest growth is sustainability, this is generally defined as the capacity of the biosphere and human civilization to coexist. This socio-ecological field covers different areas of studies all related under this same context (sustainable development, environmental economics, global poverty measures, etc). This field was the 9th with more papers back in 2009, now being the 6th and growing at a faster pace.

From the publisher perspective this is an opportunity also to adapt to this field growth and enhance their resources in terms of journals scope and editors and react on time to this growing field of research.

6 Visualization Data sets

I have used 2 different data sets for the visualization.

* First dataset is the one resulted of the original dataset (containing Publication Id, Publication Abstract and Published year) with the topic assigned from our LDA model. From this dataset I have removed the abstract text as we don’t need it for the visualization and created a lighter csv to feed the visualization
* Second dataset contains 3 columns:

1st column contains the 10 topics from our model.

2nd column contains all the 20 top words for each topic.

3rd column contains weights of each word in each topic.

We will use Tableau to create our visualization and, by having 2 datasets with a common field ( in this case Assigned Topic) , will allow us to create a relationship between them in Tableau and create a dynamic dashboard with action filter ( Tableau way of filtering across graphs based on user selections on the graphs ).

7 User Manual for the Front End

Front end visualization has been built using Tableau.

CONTENT: Visualization contain 3 main parts.

* Big numbers. In here user can see the total number of publications in the entire period for each topic, and the growth from 2009 publications to the papers published to this field up to November 2019.
* Timeline Chart. Here user can see evolution of the topics across the years of the study. Hovering over the lines, a tooltip will come up including the name of the field, year and number of publications. Chart included a couple of annotations to stand out a couple of insights that we get from this chart.
* Word cloud. This visualization shows the top 20 word per each topic identified, with their size assigned by the weight in the distribution of this word in the topic. As in the line chart, hovering over the word, a tooltip will come up with the weight of that word in its topic the topic the word belongs to.

Underneath the title there is a brief description of what the study goes about.

Also, on the top right corner, I have added an info icon including basic information about the source data and methodology used for this analysis.

FUNCTIONALITY:

This is an important part of the dashboard. From the top big numbers, I have added a Tableau action filter, this allows a dynamic dashboard, so when the user clicks on any of the big numbers, the 2 graphs below ( Timeline and Word Cloud ) will be filtered and will show only values related to that topic.