Capstone Project 2: Final Report

Name: Jaime Ortiz

Science and the Language of Invasive Species in the Galapagos Islands

1. Introduction and Problem Statement

Ecologists often find themselves in the need to respond to managers’ problems, it is often the case that scientific research does not necessarily represents the problems set as priority by state managers. In other cases, scientific research reflects the views set by funders. Thus it is important to analyze the trends in ecological research to understand the role of scientists in shaping the conservation management effort, especially in places such as the Galapagos Islands. Considering that many scientific articles are electronically available it is possible to perform large data analysis of scientific trends.

It is widely recognized that invasive species are a serious problem for biodiversity around the world. Invasive species are especially problematic for fragile and unique ecosystems such as the Galapagos Islands. One of the main barriers that conservation managers in the Galapagos face when dealing with invasive species is the miscommunication and interpretation of scientific literature to identify and prioritize management strategies in the field. Using Natural Language Processing (NLP) techniques my objective is to understand the relationships between scientific research and conservation management goals.

2. Analysis

Once the data was cleaned and preprocessed, we applied topic modeling for the statistical analysis of a large collection of unstructured text data (i.e. scientific articles, abstracts, and grey literature) using Latent Dirichlet Allocation (LDA) (Blei et al. 2003) in the R statistics software (R Core Team 2016) using the package “mallet” (Mimno 2013), “dfrtopics” and “dfr-browser” (Goldstone 2016). For instance, this approach has been successfully implemented to distinguish trends over time of specific research interests (Hall et al. 2008; Cohen Priva and Austerweil 2015). The main assumption in Topic Models considers that every word in every document by one of various topics. In this context, topics are mixtures of words Dirichlet-distributed and documents are mixtures of topics Dirichlet-distributed.

Topics in this case can be defined as a distribution of high probability words over a fixed vocabulary (Blei 2012). For instance, the *plant eradication* topic has words related to plant eradication with high probability.

There were various topics that were stable and of good quality, however we will consider only the topic related to introduced species for each dataset, in the case were more than one topic related to introduced species for a dataset the topic with the highest quality was chosen. In addition to this, the topic corresponding to the scientific literature only shows the data beginning from the year 1960. Finally, to study the temporal relationships in our dataset we chose to use a post-hoc calculation based on the weighted contribution of each topic T in each year y using the equation proposed by Cohen Priva and Austerweil (2015).

3. Final Outcomes

The results from our Topic Model using Latent Dirichlet Allocation (LDA) showed that there is a clear evidence for differences in which the topic of introduced species was treated among the different datasets, considering that each dataset represents a key stakeholder in conservation. The importance of the introduced species topic varies greatly along the time scale, reaching its highest peak in the year 2002 for all stakeholders (Fig. 1). This peak coincides with the implementation of the biggest ever conservation project for the Galapagos Islands and was mainly devoted to the control of introduced species, thus deriving in many documents related to management and scientific research.

It is also interesting to observe that the earliest peak for the topic of introduced species was in the early 1980s just after the designation of Galapagos as a UNESCO World Heritage site. However, the importance of this topic declined sharply after the most devastating “El Niño” climatic phenomena ever recorded in the islands, indicating that after that natural disaster great part of the effort from managers and scientists was concentrated in recovering from this natural disaster and understand its effects on the local ecosystems.

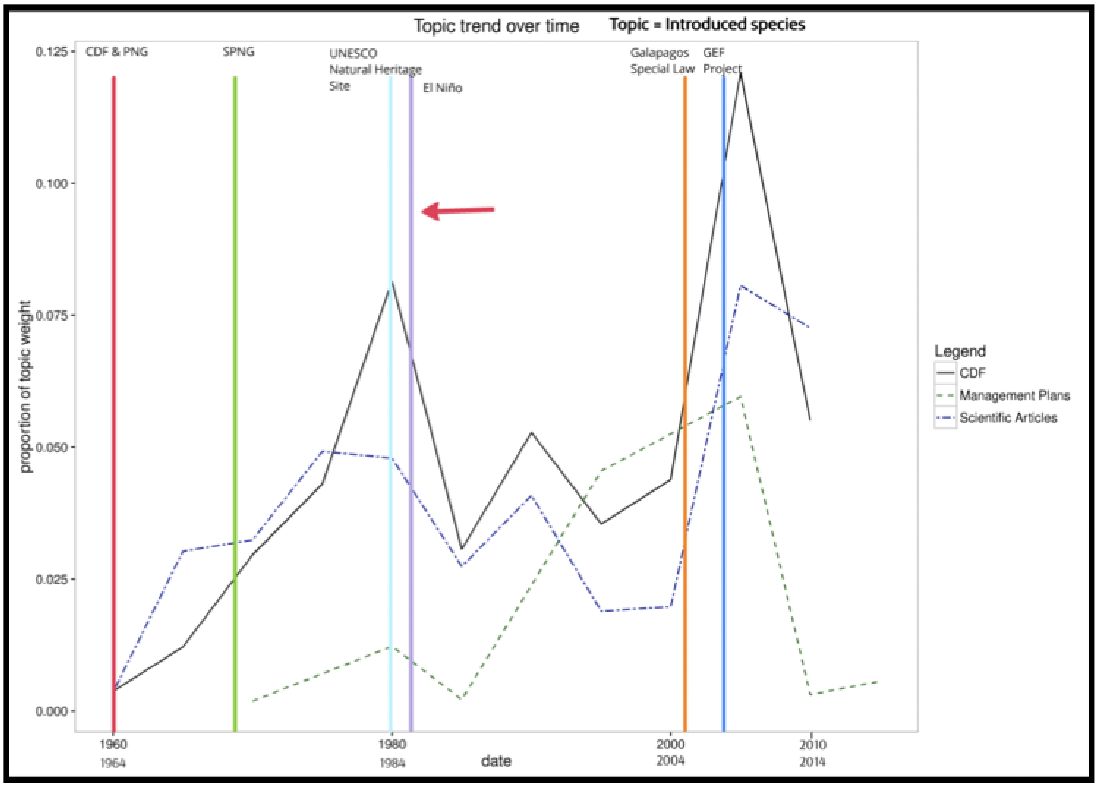

Figure 1. Introduced species topic trend over time from different corpus. CDF: Charles Darwin Foundation, Management Plans from the Galapagos NAtional PArk and Scientific literature. Vertical lines indicate key events for conservation management in the Galapagos.

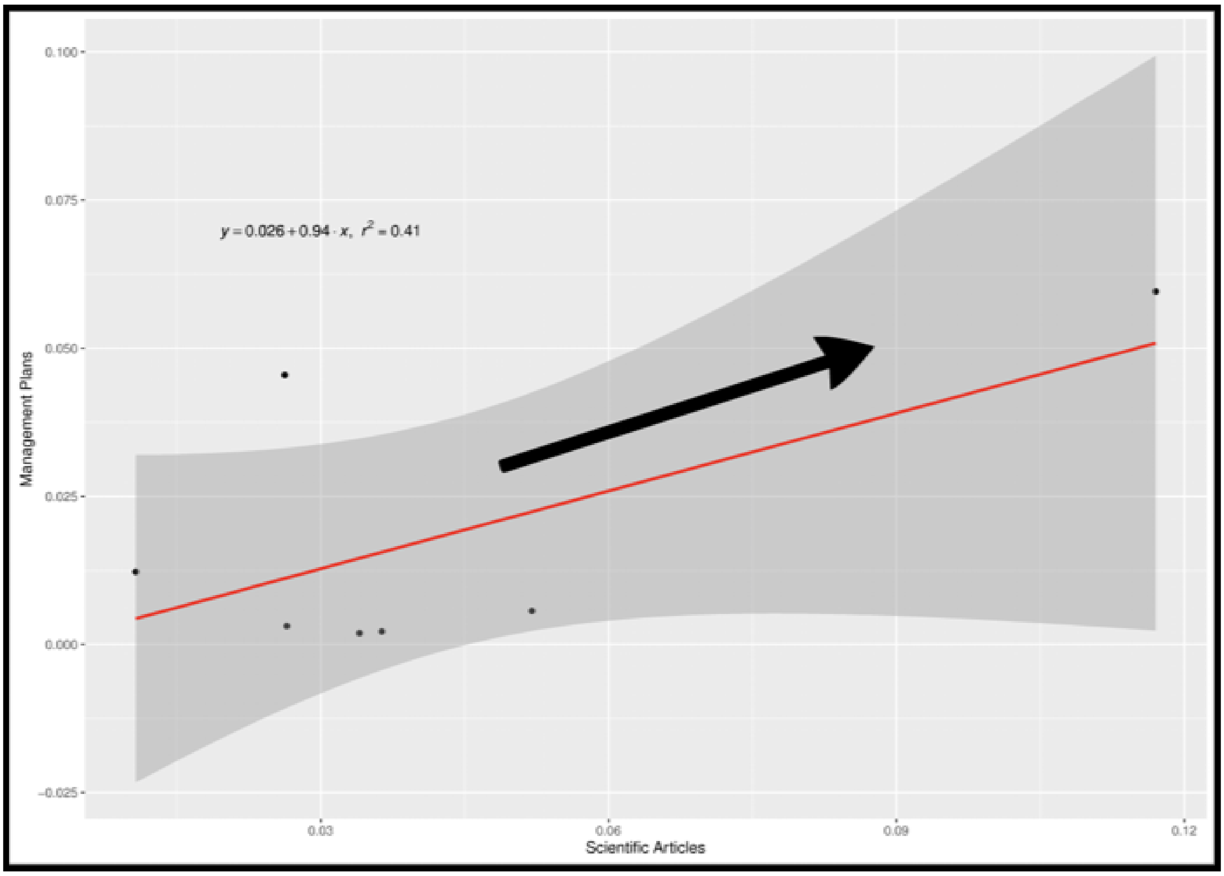

Figure 2. Relationship between scientific publications and management decisions as reflected in management plans published for the conservation of the Galapagos.

Additionally, the linear regression model on our text data suggests that there is a positive correlation between scientific publications and management plans in Galapagos (Fig. 2). This particular results can be considered as an indicative that science has an important role in shaping how managers are making decisions to protect the islands. In this context, the NLP analysis provides critical insights in the importance of written communication among stakeholders in the Galapagos.

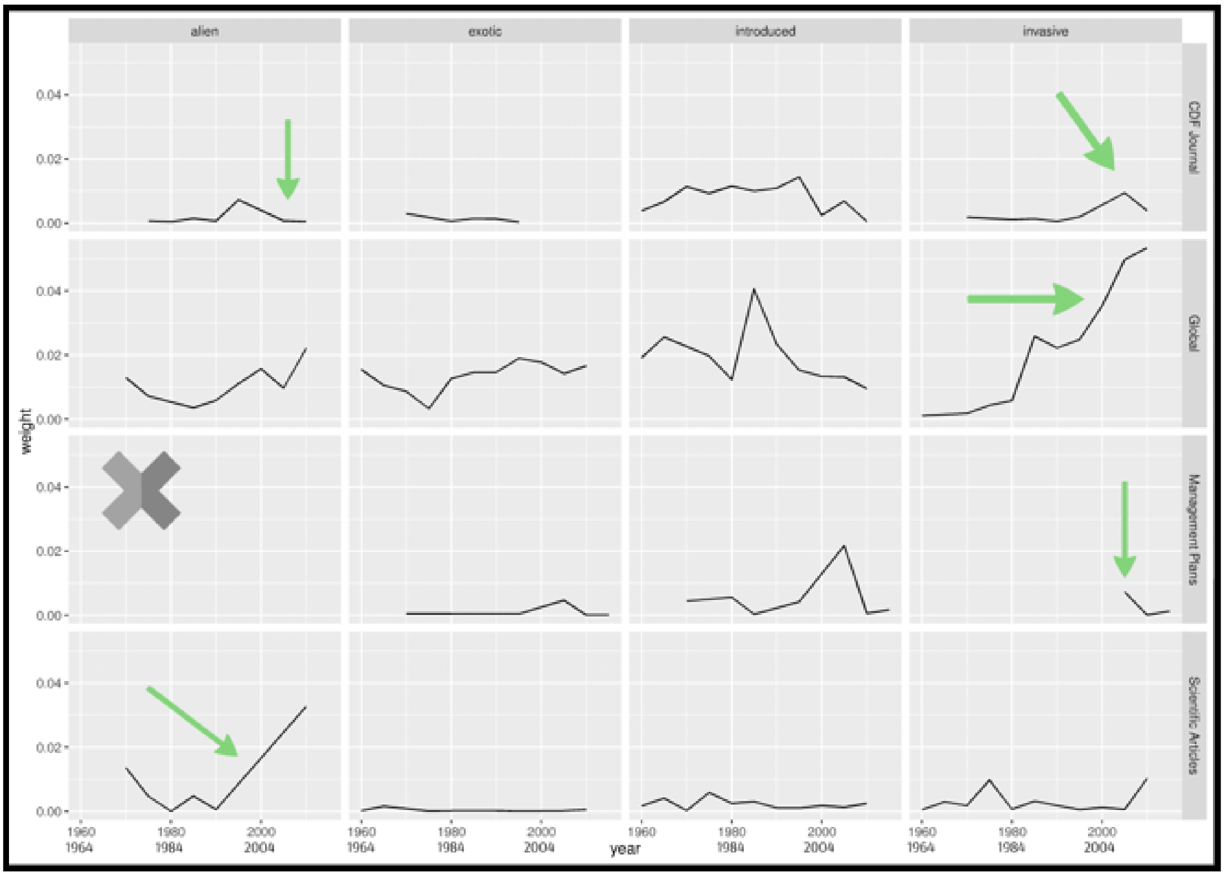

Figure 3. Word usage among different stakeholders across time. A fourth dataset was included in this analysis, "Global" indicating scientific literature related to the theme of introduced species globally and not only in the Galapagos.

NLP results at a finer scale showed that different stakeholders tend to use different words to communicate about introduced species. For instance the word "alien" is not used by managers, but heavily popular among scientists (Fig. 3). On the other hand, the term "invasive" has grown exponentially in popularity among scientists at a global scale, which surprisingly does not match the trends observed for Galapagos in particular.

Finally, the analysis presents strong evidence to suggest that prioritization of introduced species for management and research does not match among stakeholders at any point in time (Fig. 4). This result is highly relevant because indicates that resources for scientific research and management are used inefficiently and resulting in failure to protect the fragile ecosystems of the islands.

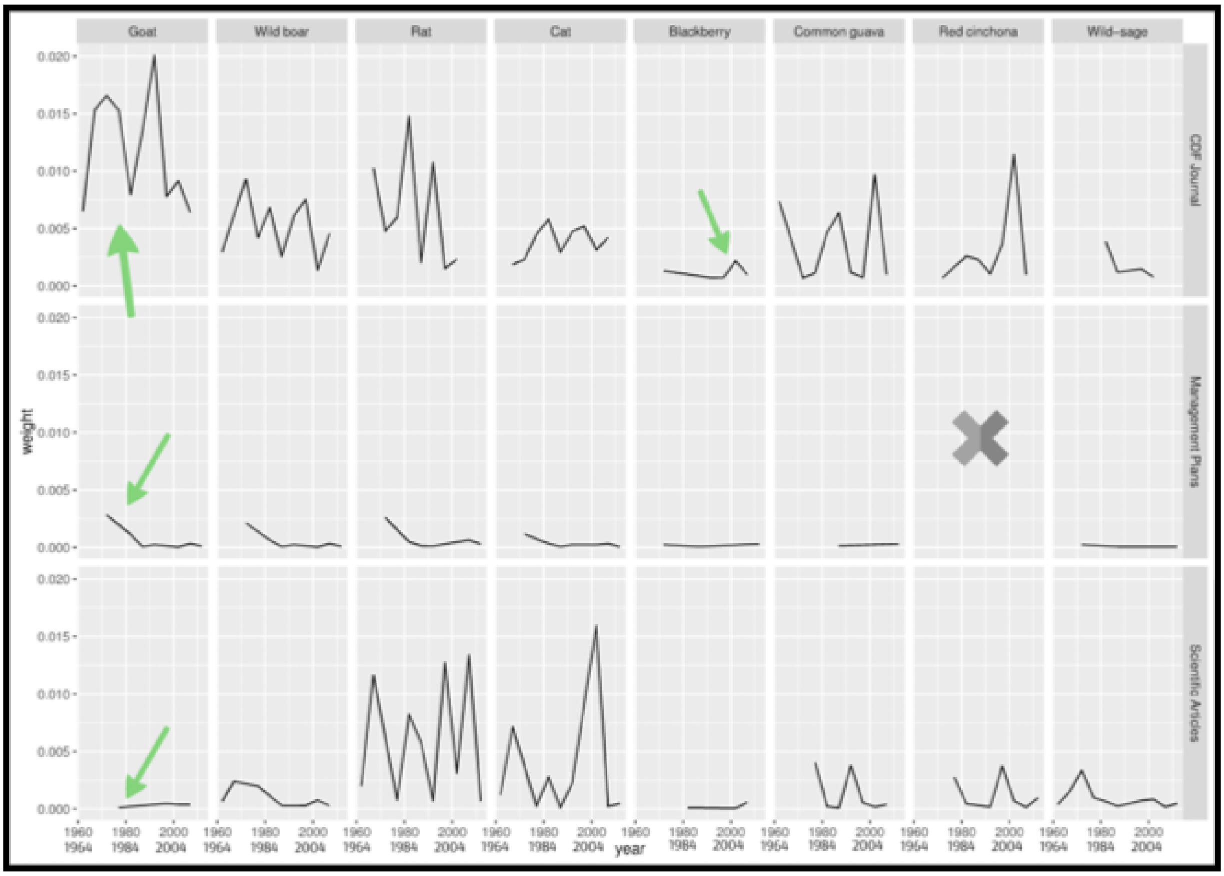

Figure 4. Word usage among different stakeholders across time.

4. Recommendations

The results from this research project showed that NLP techniques can be successfully implemented to discover and quantify differences in language use among two key stakeholders for the conservation of the fragile ecosystems of the Galapagos Islands. It was clearly observed that Topics related to introduced species changed among stakeholders and also throughout time. The model strongly suggests that scientists and environmental managers use different language to communicate about introduced species in the Galapagos. For instance, scientist prioritize different set of introduced species and use different words to assign labels to these introduced species. Based on the model results my recommendation would be that key stakeholders in the Galapagos islands work on a common communication strategy that allow better information flow among users and enables them to make more effective decisions and in turn increase the efficiency on scarce resource allocation for introduced species control programs on the islands.

References:

Blei DM (2012) Probabilistic topic models. Communications of the ACM 55:77. https://doi.org/10.1145/2133806.2133826

Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. Journal of machine Learning research 3:993–1022

Cohen Priva U, Austerweil JL (2015) Analyzing the history of Cognition using Topic Models. Cognition 135:4–9. https://doi.org/10.1016/j.cognition.2014.11.006

Goldstone A (2016) dfrtopics - An R package for exploring topic models of text. In: GitHub. https://github.com/agoldst/dfrtopics. Accessed 13 Oct 2016

Hall D, Jurafsky D, Manning CD (2008) Studying the history of ideas using topic models. In: Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics, pp 363–371

Mimno D (2013) Package “mallet”