Capstone Project 2: Project Milestone 2

Name: Jaime Ortiz

Science and the Language of Invasive Species in the Galapagos Islands

1. Problem Statement

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. Dataset 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, Ng & Jordan 2003) in the R statistics software (R Core Team 2016) using the package “mallet” (Mimno 2013), “dfrtopics” and “dfr-browser” (Goldstone 2016). 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. New Important findings

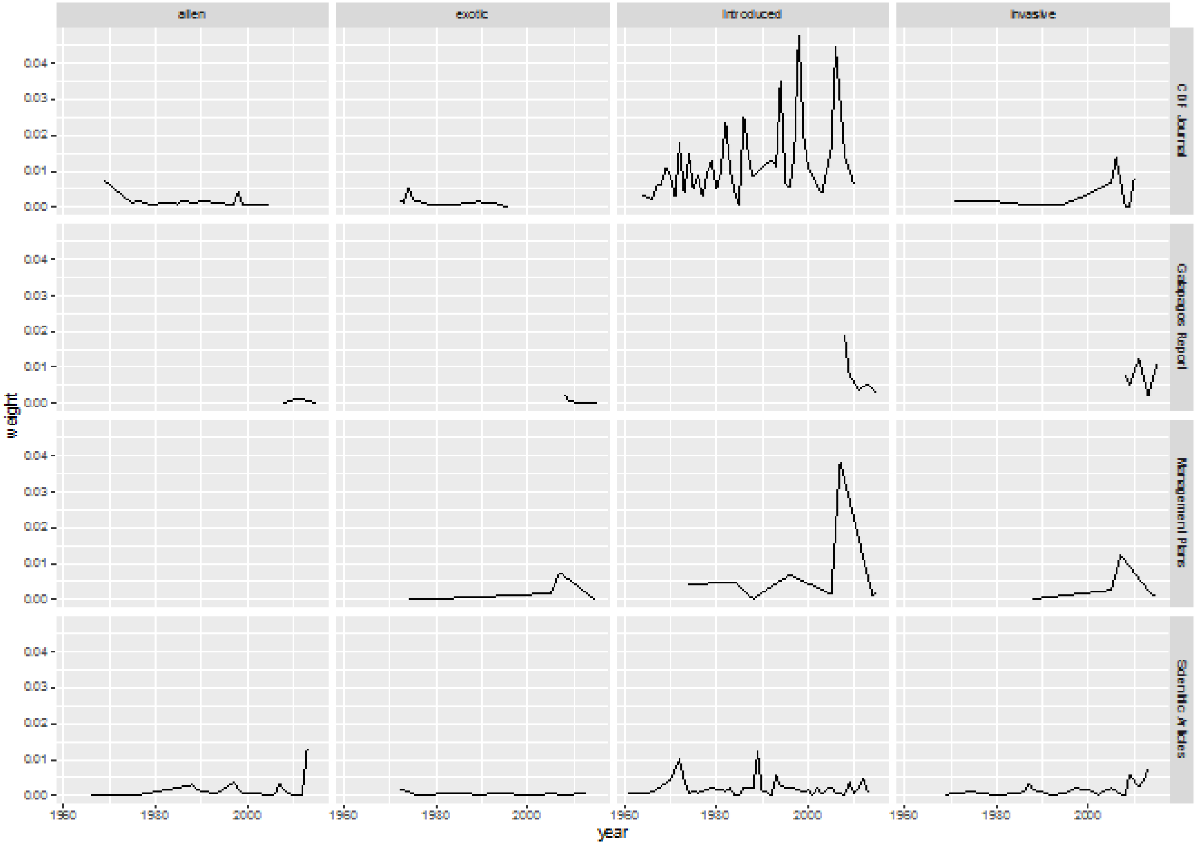

Fig 1. Plot showing the temporal distribution of terms related to introduced species for each corpus, facet grid style, raw data. Weight indicates the proportion of contribution of that specific word to the whole corpus, this allows for comparison among different set of corpus.

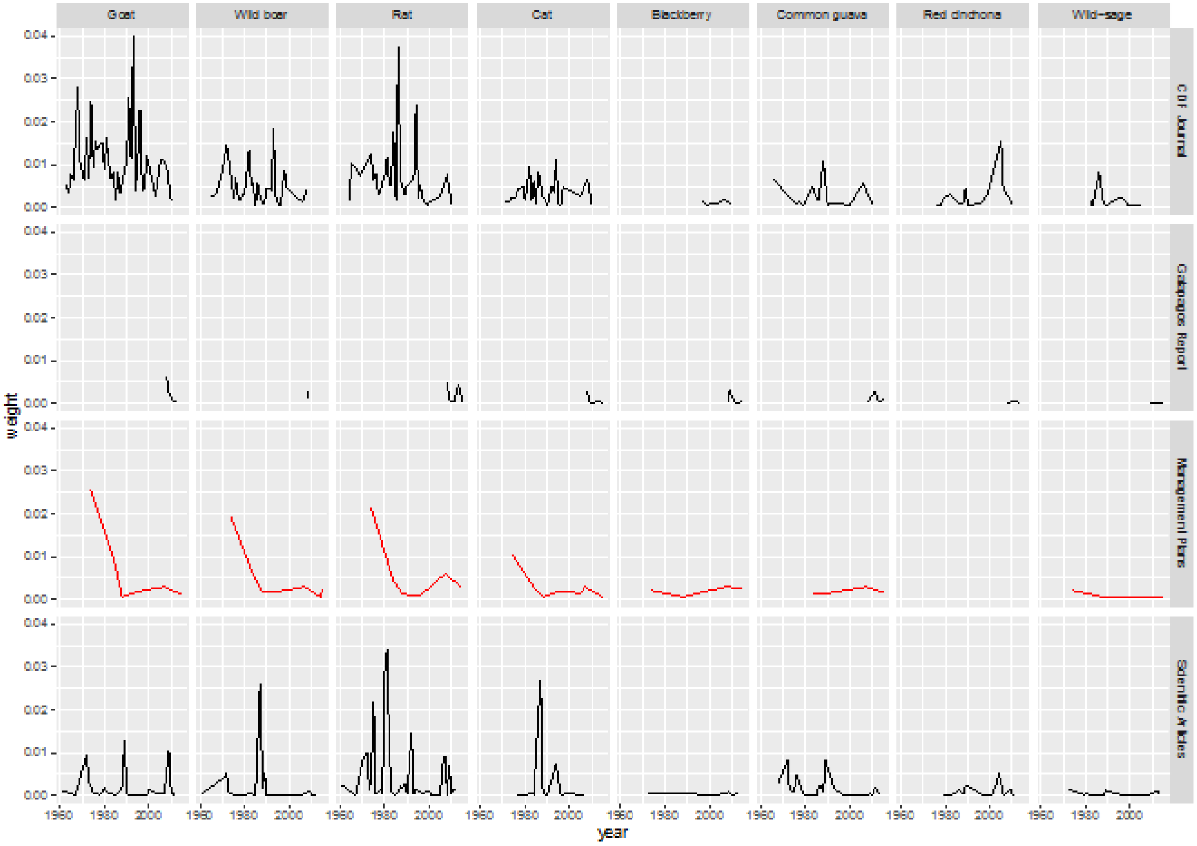

Fig 2. Plot showing the temporal distribution of introduced plants and vertebrate species for each corpus, facet grid style, raw data. * Management plans data has been multiplied by 10 in order to provide better visualization of the pattern in the use of names of introduced plants and vertebrates.

Weight indicates the proportion of contribution of that specific word to the whole corpus, this allows for comparison among different set of corpus.

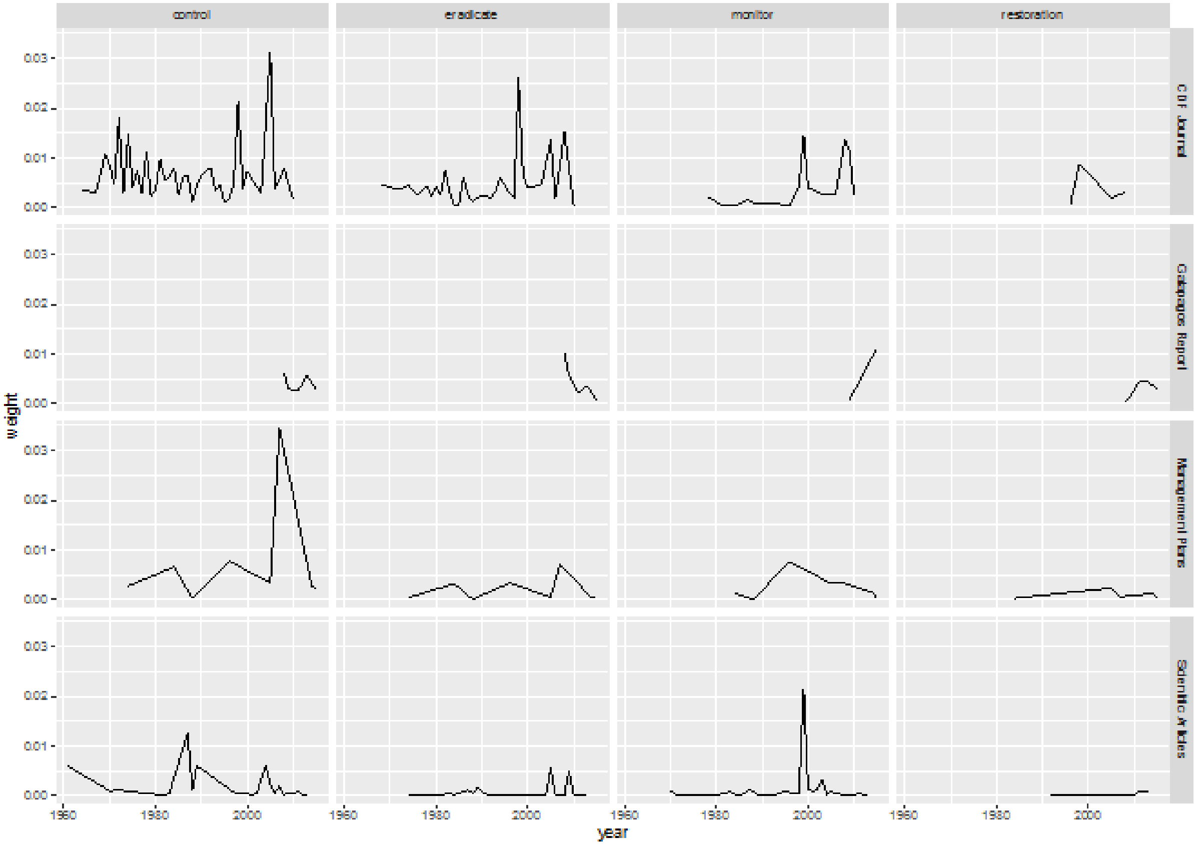

Fig 3. Plot showing the temporal distribution of terms related to actions to manage introduced species for each corpus, facet grid style, raw data.

Weight indicates the proportion of contribution of that specific word to the whole corpus, this allows for comparison among different set of corpus.

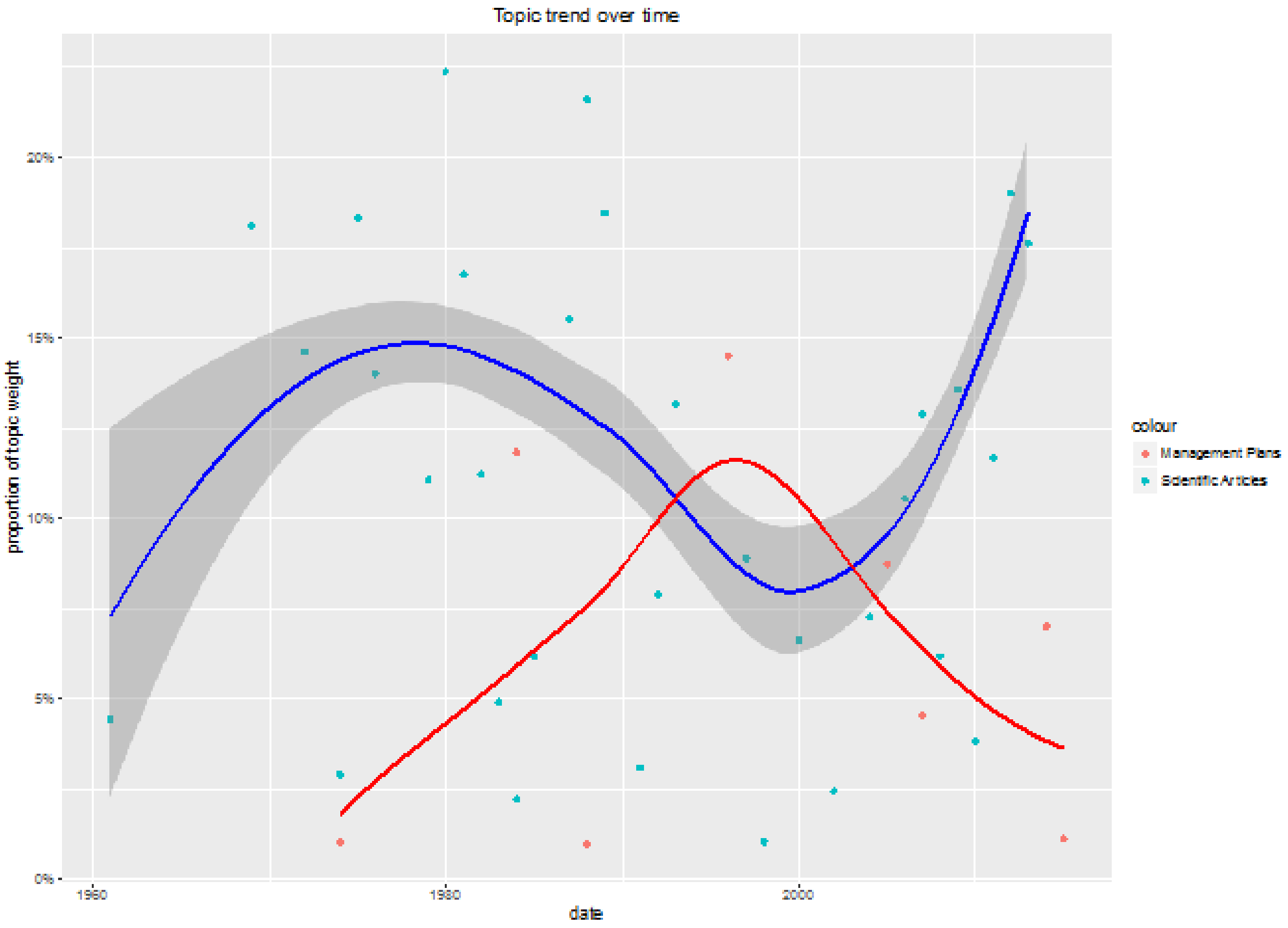

Fig. 4. Trends in the prevalence of the introduced species topic in the Galapagos Islands within management plans (red) and scientific publications (blue). The prevalence of the topic shows a time lapse between both data sets, suggesting a relationship where science writing has an influence on management intention, since the latter appears to follow up but not catching up with the trend set by scientists. The curves are best-fit polynomial functions using the loess smoothing in the ggplot2 R function and shaded area around the science literature curve denotes 95% confidence intervals.

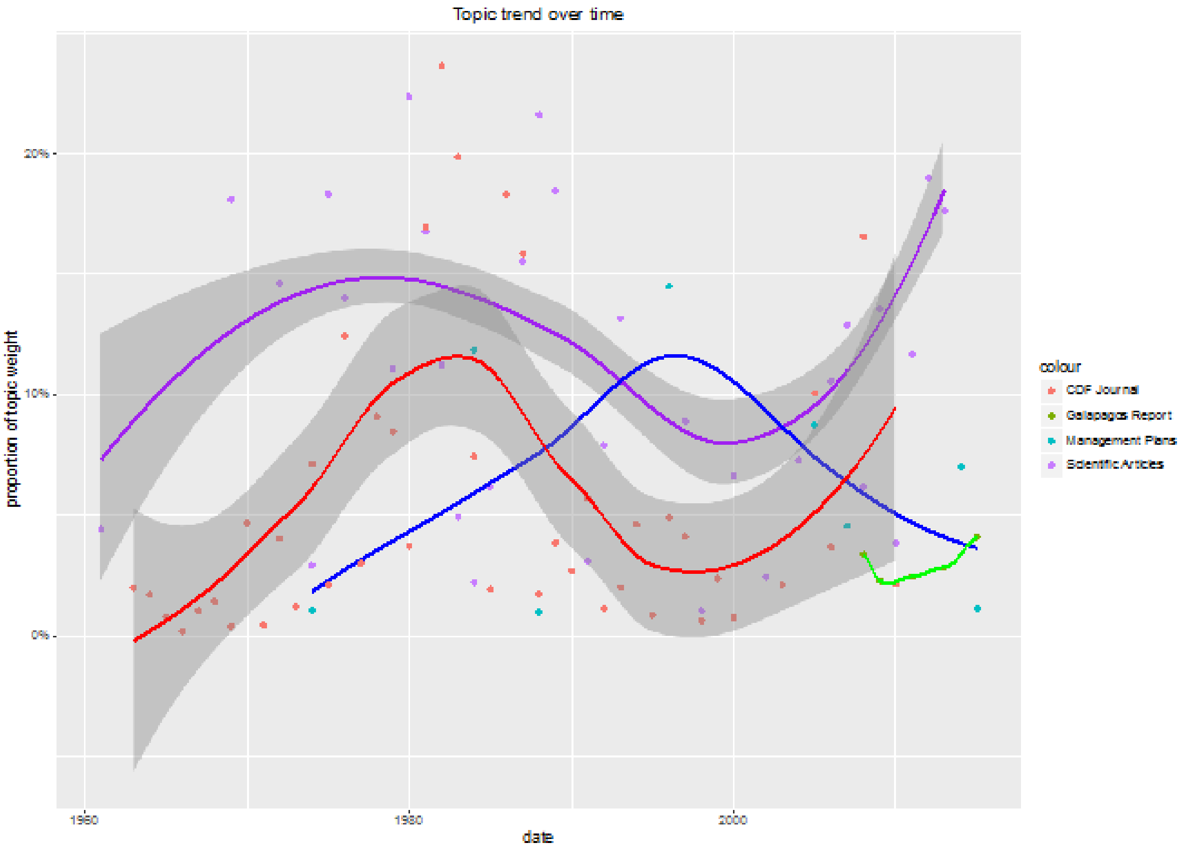

Figure 5. Trends in the prevalence of the introduced species topic in the Galapagos Islands within the Charles Darwin Foundation (CDF) journal (red); Galapagos Report (green); management plans (blue) and scientific publications (purple). The prevalence of the topic shows a time lapse between the more scientific data sets (Scientific articles & CDF journal) and management plans, suggesting a relationship where science writing has an influence on management intention, since the latter appears to follow up but not catching up with the trend set by scientists. The curves are best-fit polynomial functions using the loess smoothing in the ggplot2 R function and shaded area around curves denotes 95% confidence intervals.