Capstone Project 2: Project Milestone 1

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

The scientific text data was obtained from the “Data for Research” service provided by JSTOR (http://dfr.jstor.org) using the following search string: [“Galapagos” OR “Galápagos”]. This first search produced more than ten thousand articles; this result was then refined to full-length journal research articles only and English language resulting in a total of 3,717 articles dated from 1812 to 2015. The grey literature sample was obtained by digitizing the available management plans for the region and the protected areas of the Galapagos Islands, nine in total (1974, 1984, 1988, 1996, 1999, 2005, 2007, 2014, 2015). Another set of grey literature “Galapagos Report” (http://www.galapagos.org/about_galapagos/about-galapagos/library/galapagos-reports/), is composed by a set of articles (in this analysis the publications since 2008 were used) written by different actors in the Galapagos environmental governance (i.e. researchers, protected area managers, urban planners, natural science students, conservationists). The final data set is composed by another scientific laden literature, the CDF journal (1964-2010) (http://www.darwinfoundation.org). This set of data differs from the international scientific articles in two ways: 1) the journal was aimed to a very specific public concerned with the conservation and scientific research in the Galapagos Islands; and 2) the content, at least, in the beginning of the journal included a summary of news deemed relevant for this public.

I pre-processed the data by removing all punctuation, converting all characters to lowercase, removing high frequency words (i.e. stop words) using a specific and manually constructed stop list containing more than 22,000 terms (English and Spanish common words, ascii characters, roman numbers, common names, abbreviations, and other terms that were considered unnecessary for the analysis). Although, this might be a time consuming process the construction of a comprehensive stop-list specifically built for the dataset under analysis is an important requisite to get good results and removing words with a frequency lower than five.

3. Initial findings

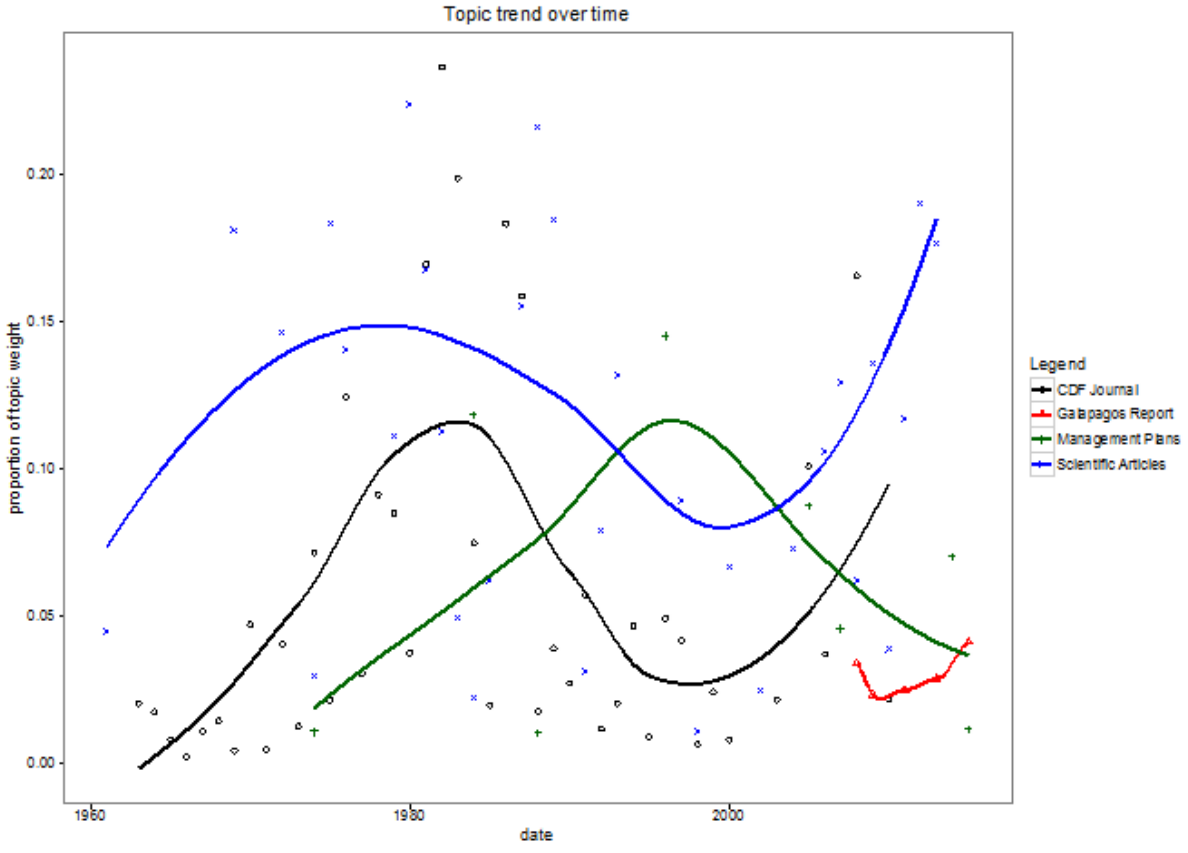

fig 1. Plot showing topic related to introduced species for all sets of documents (corpus), smoothing applied using loess.

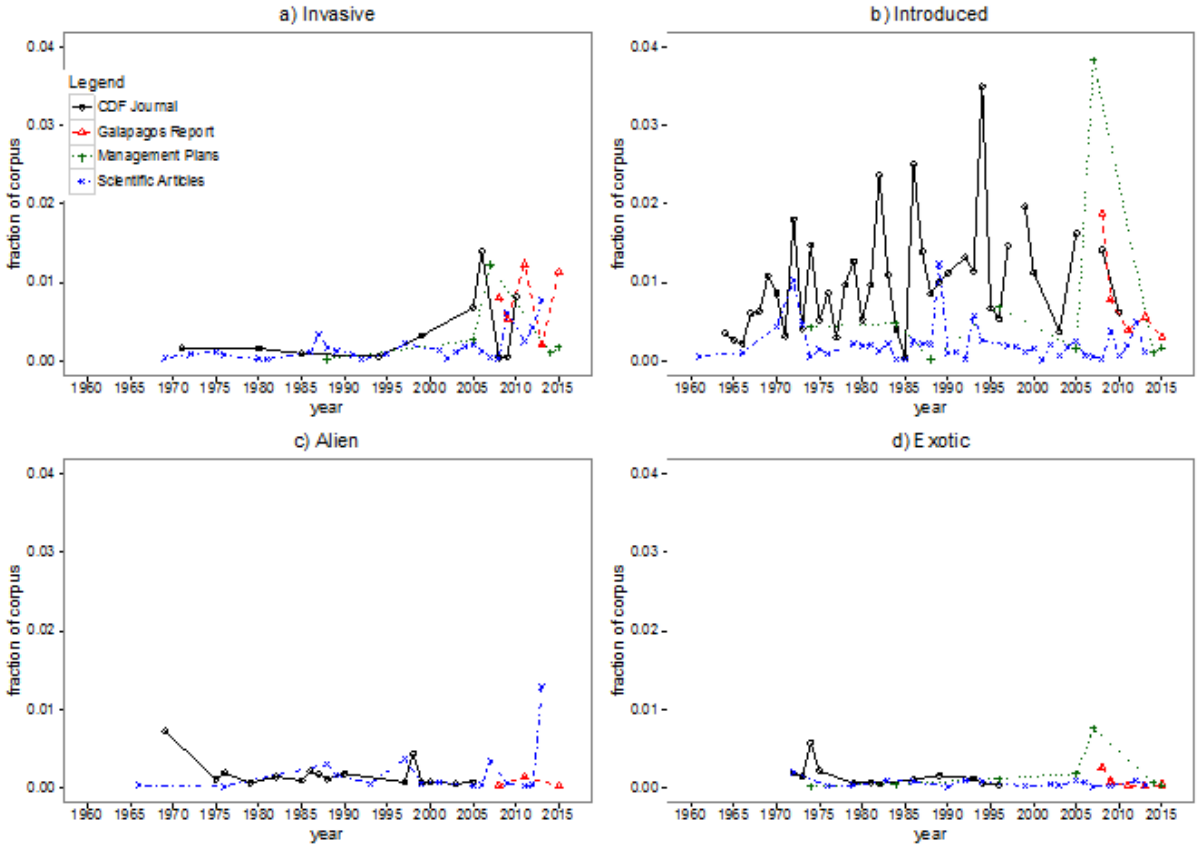

fig 2. Plot showing the temporal distribution of terms related to introduced species for each corpus, grouped by term, raw data. Fraction of corpus indicates the proportion of contribution of that specific word to the whole corpus, this allows for comparison among different set of corpus.

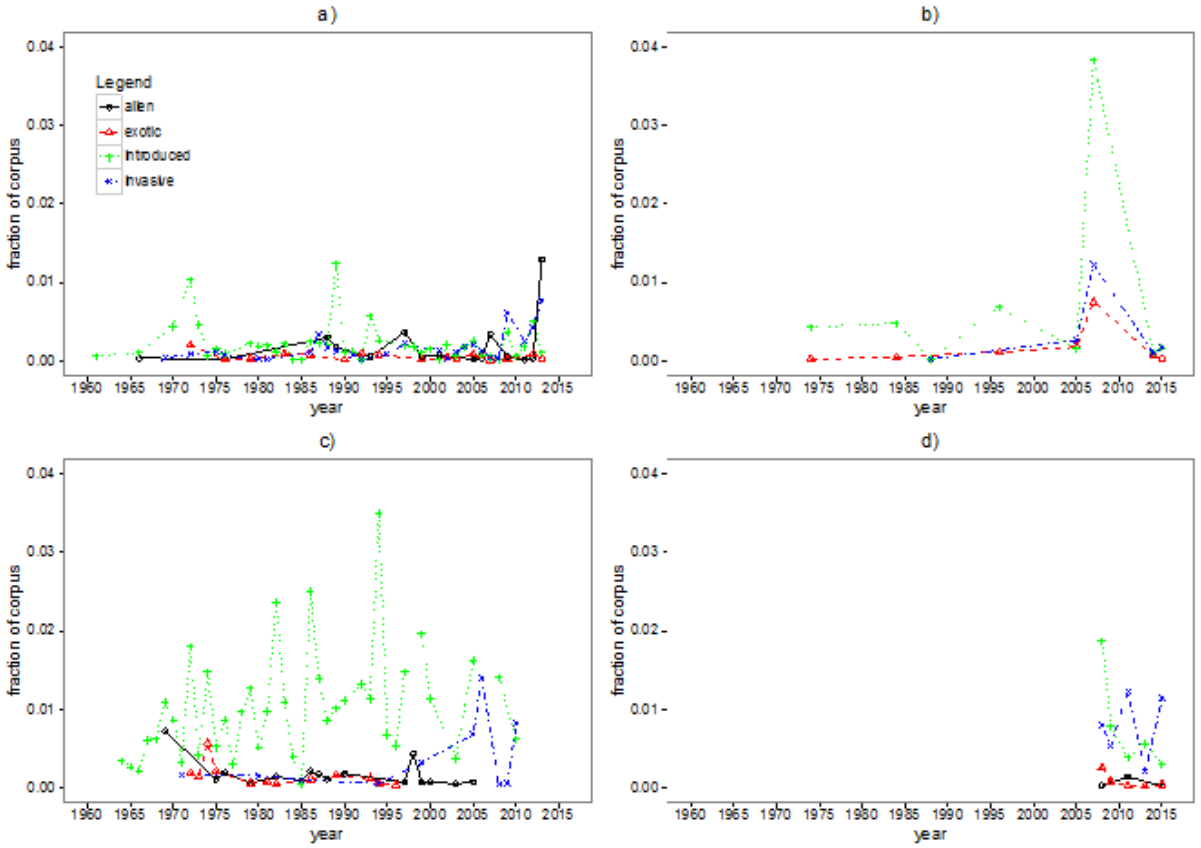

fig 3. Plot showing the temporal distribution of terms related to introduced species for each corpus, grouped by corpus (a. Scientific Articles, b. Management Plans, C. CDF Journal, d. Galapagos Report), raw data. Fraction of corpus indicates the proportion of contribution of that specific word to the whole corpus, this allows for comparison among different set of corpus.