Thirty Years of IIRA's Rural Research Reports: A Thematic Analysis

Adee Athiyaman[1]

Research Brief, Short Paper
ISSN 2687-8844
Vol. 1, No. 1 (2019 April 11)

Abstract
This paper employs a machine-learning algorithm to analyze the words of all the Rural Research Reports published since 1990 to discover the themes that run through them. The results are then built into an interactive computer application through which one can explore and examine the publications (Rural Research Reports) related to the themes.
Keywords: Research topics, Probabilistic topic model, LDA, IIRA.

Introduction

The IIRA hosts a website that offers visitors a wealth of information including data and research related to rural economic development. Imagine that a visitor to the website wants to search for Rural Research Reports (RRRs) related to a specific theme, for example, 'business creation'; she also wants the flexibility to "zoom in" and "zoom out" to find broader or specific themes, and assess how those themes changed through times or how they are connected to each other. Right now, to gain this kind of information one needs to spend many hours to read and study all of the approximately 200 RRRs to discover and annotate the documents with thematic information.

How could we automate this tedious process and provide readers of RRRs with the kind of browsing experience described above? In the following pages we implement a machine-learning algorithm that analyzes the words of all the RRRs published since 1990 to discover the themes that run through them[2].

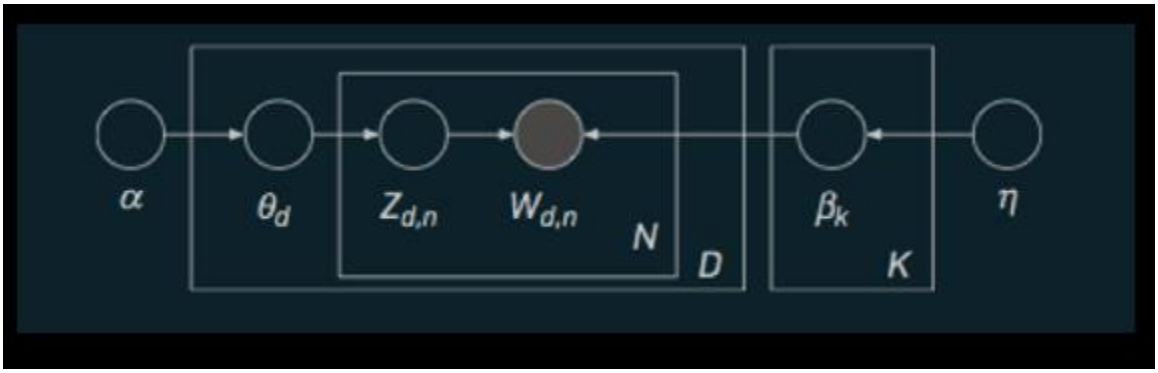
Methodology

The basic idea behind the machine learning algorithm is that documents exhibit multiple topics and each document exhibits the topics in different proportion. In addition, each word in each document is drawn from one of the topics. The computational problem is to use the documents to infer the topic structure. Figure 1 breaks the computational problem down into its components.

---

[1] Professor, Illinois Institute for Rural Affairs.
[2] Latent Dirichlet allocation is the algorithm used in this exercise, see Blei, D., Ng, A., Jordan, M. Latent Dirichlet allocation. *J. Mach. Learn. Res*. 3 (January 2003), 993-1022.

Figure 1: Topic Mining of RRRs: Computational Steps



Θ_d is the distribution of topics for document *d*, a real vector of length *k*. We model this as ~Dir(α) = $\frac{\Gamma(\sum \alpha_i)}{\Pi\Gamma(\alpha_i)} \prod_i^D \theta_i^{\alpha i-1}$; β_k is the distribution of words for topic *k*, a real vector of length *v* which is ~Dir(η) = $\frac{\Gamma(\sum \eta_i)}{\Pi\Gamma(\eta_i)} \prod_i^K \beta_i^{\alpha i-1}$; $Z_{d,n}$ is the topic for n^{th} word in document *d*, and $W_{d,n}$, is the n^{th} word of the d^{th} document.

The joint distribution, p(*W, Z,* θ*,* β |α*,* η) =
$\prod_k p(\beta_k|\eta) \prod_d [p(\theta_d|\alpha) \prod_n p(Z_{d,n}|\theta_d) p(W_{d,n}|Z_{d,n}, \beta)]$

The conditional distribution of the topic structure given the RRRs is the posterior distribution:

$$p(\beta_k, \theta_d, \theta Z_d|W_d) = \frac{\mathrm{p}(\beta_k, \theta_d, Z_d\, W_d)}{\mathrm{P}(W_d)}$$

Since we cannot compute the evidence, the denominator $p(W_d)$, we use the variational-EM algorithm provided by the Gensim library to compute the conditional distribution.

Results

Figure 2 highlights a sample of topics discovered from the RRRs. The topic 'tokens'[3] include nouns, adjectives, and adverbs; all 'stop words[4]' were removed from the corpora. A sample of four topics are shown. Each topic is listed with its most frequent words. Each word's position along the horizontal axis denotes its specificity to the documents. For example, "need" in the first topic is more specific than "housing".

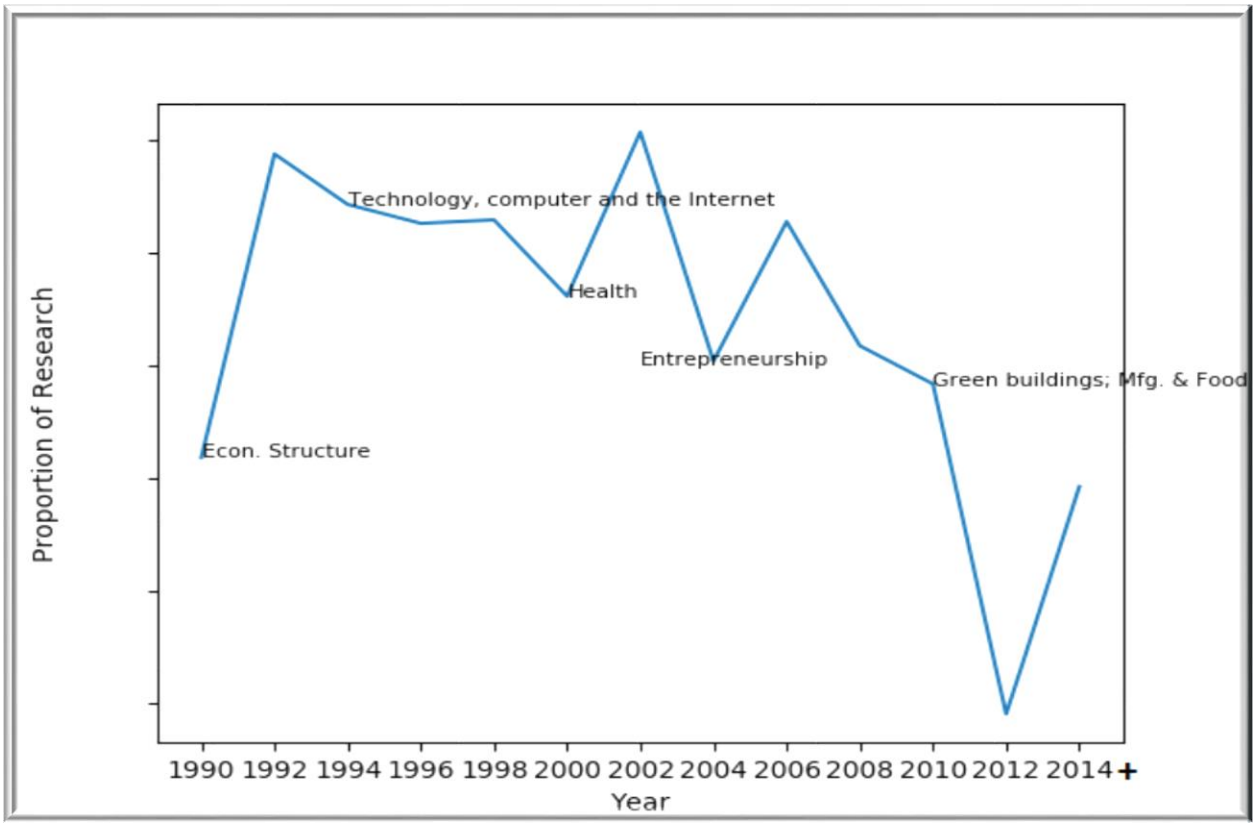
[3] In lexical analysis, tokenization is the process of categorizing a text corpora into meaningful elements such as nouns and adjectives. In all, 600,000 tokens were processed for this topic-modeling exercise.
[4] Stop words are commonly used words such as "the", "and", "or", etc.

Figure 2: Sample RRR Topics

housing	organization	entrepreneur	technology
health	student	government	small
resident	project	public	agency
city	support	available	report
small	work	research	need
information	provide	foundation	
year	job		
need			
care			

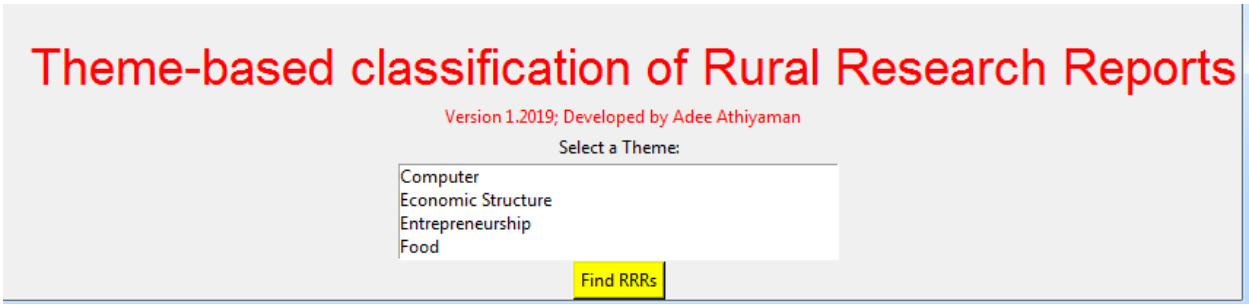
Figure 3 shows the dynamic, time-series analysis of topics. During the introductory years of the RRRs, the early 1990's, research focus was on the economic structure of rural regions; salient topics include: population and employment, township revenue, and tourism facilities. In the late 90's, research on education technology and computer and the Internet were dominant. Health and entrepreneurship including brownfield investments were the topics of research during 2000-2010. Since 2010, research has focused on green consumerism, manufacturing, and food products.

Figure 3: Dynamic Topic Model

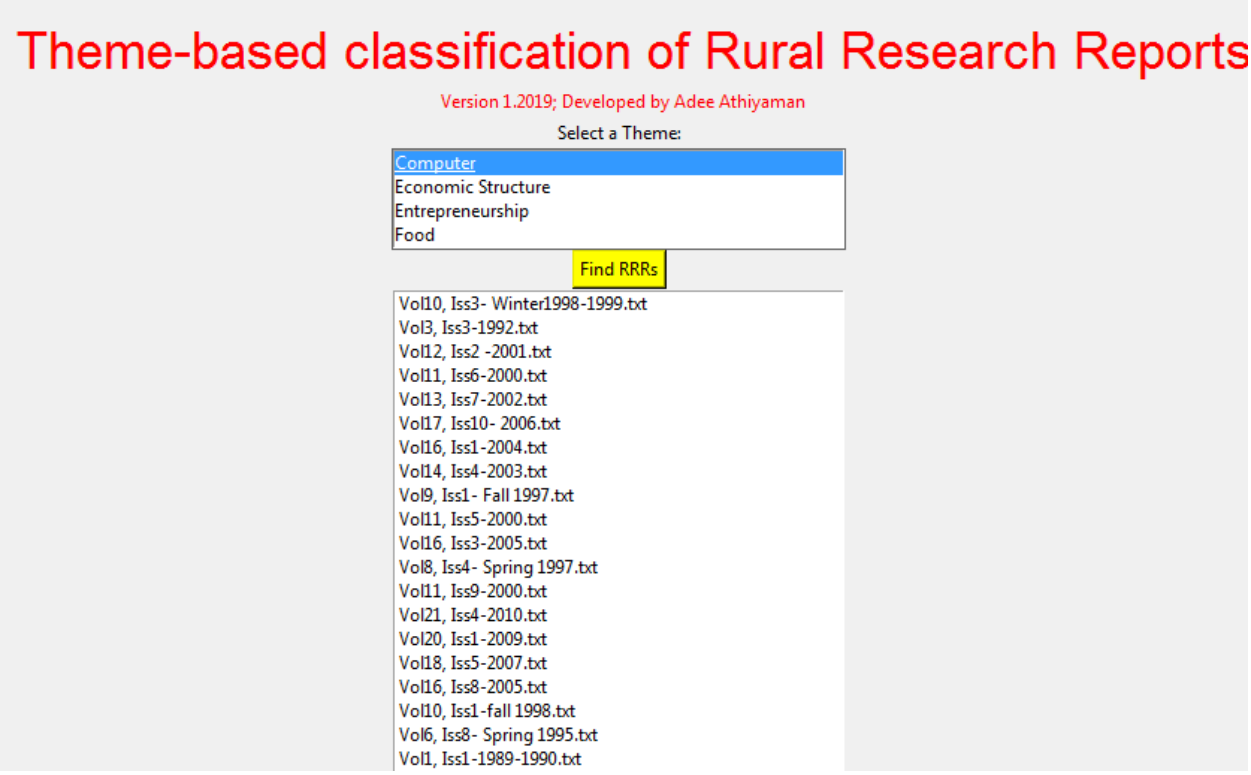


Theme-Search Software, Theme.Exe

Readers interested in thematic analysis of RRR are encouraged to download the software Theme.Exe. The software is provided as a zipped file which can be run on any Windows operating system. Clicking on "Theme.Exe" will open the screen:



Selecting a theme and clicking on the "Find RRRs" button will generate a list of references related to the theme. For example, searching RRRs related to the theme "Computer" will generate the following:



Summary

This paper highlights a new methodology for interacting with the RRR corpora, topic modeling. The interactive computer application could help readers identify the most interesting RRRs related to the themes such as technology and computers. Future research will attempt to summarize and understand the growing digitized archive of information on rural economic development.