Text Mining the History of Economic Thought

Roger Cuscó

Abstract

This project analyses text from some of the greatest economists of all time to provide insight on their similarities and interests. Although the set of publications used is relatively small, an extended version of the LDA algorithm taking into account varying topic distributions over authors seems to capture some of the key differences between economists. For instance, modern economists seem to have a stronger focus on policy issues.

Data

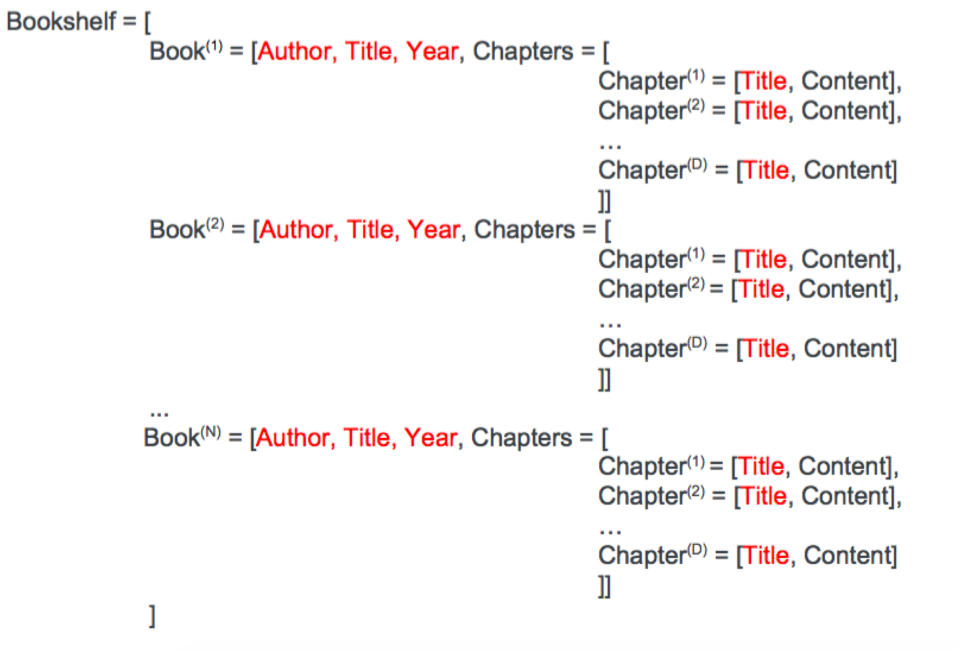
The data used in this project has been scraped from the website econlib.org. It comprises a list of books written by classical economists and political economists from the eighteenth and nineteenth centuries. Each book is divided into chapters labelled with metadata on the author, year and titles.

The data has a hierarchical structure defined as follows (metadata is coloured in red):

The authors selected are Adam Smith, Karl Marx, Ludwig von Mises, Frédéric Bastiat and David Ricardo. In addition, articles from some modern authors such as Gregory Mankiw, Paul Krugman and Milton Friedman are used in the analysis of similarities. However, those articles were collected by hand from public sources (reference), as the volume of data used is small.

The reasons behind this selection are 1) Compare authors that may present notorious differences based on common knowledge of the history of economic thought 2) Compare authors that I personally find amusing.

More details on how the data was collected and stored can be found at the source code.



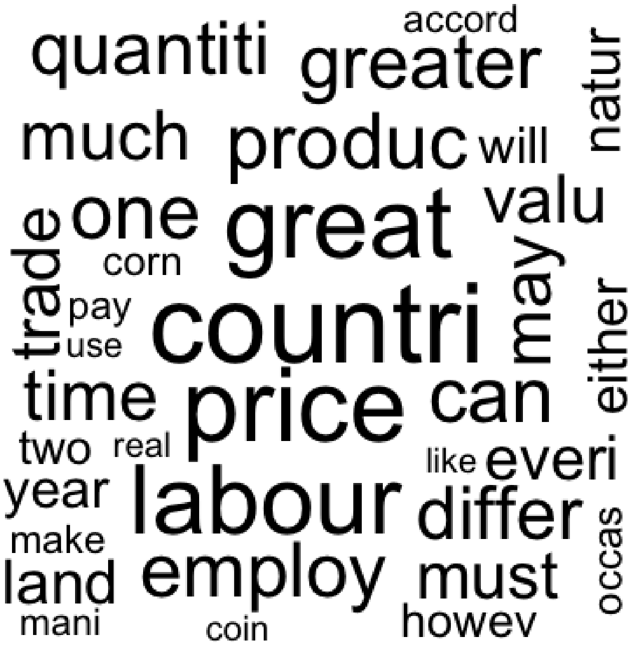
```
Bookshelf = [
          Book^{(1)} = [Author, Title, Year, Chapters = [
                                        Chapter^{(1)} = [Title, Content],
                                        Chapter^{(2)} = [Title, Content],
                                        ...
                                        Chapter^{(D)} = [Title, Content]
                                        ]]
          Book^{(2)} = [Author, Title, Year, Chapters = [
                                        Chapter^{(1)} = [Title, Content],
                                        Chapter^{(2)} = [Title, Content],
                                        ...
                                        Chapter^{(D)} = [Title, Content]
                                        ]]
          ...
          Book^{(N)} = [Author, Title, Year, Chapters = [
                                        Chapter^{(1)} = [Title, Content],
                                        Chapter^{(2)} = [Title, Content],
                                        ...
                                        Chapter^{(D)} = [Title, Content]
                                        ]]
          ]
```

Figure 1: Schema of the dataset

As an initial motivation for our analysis we can use wordclouds to already try to distinguish the stress that each author puts on each topic by the mere frequency of words use.

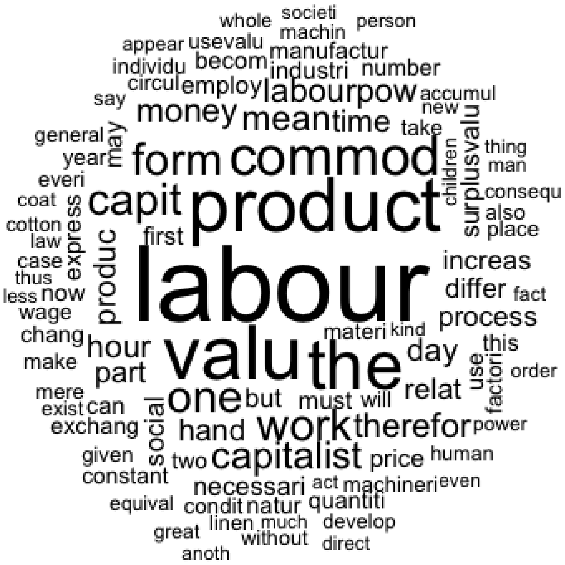
However, not much is revealed just by plotting the most common words of each author. We can see a few similarities between Adam Smith and Karl Marx, as both have labour as one of the most used words.

- Adam smith's wordcloud:



We could say that in Marx's wordcloud shows relatively bigger size for stems such as 'work', 'capitalist', 'commod', while Smith's could shows relatively bigger size for 'price', 'trade', 'quantity' or 'employ'.

- Karl Marx's wordcloud:



However, such an exercise does not tell us much about the topics each economist was more interested in. Or whether their treatment of a certain topic is similar or not. In the next sections we discuss this and other questions and implement a few algorithms to help us answer them.

Question

Economic debates are often noisy and highly ideologized. As a result, the general public tends to adopt a skeptical stance towards economic advice coming from economists and have a noisy idea of what part of an expert's message is an opinion and what is a contrasted economic fact. The main question behind this analysis is: Can text mining techniques help in identifying patterns and similarities in the economic literature that manual categorization might miss? Can we algorithmically establish similarities between authors that coincide with common knowledge?

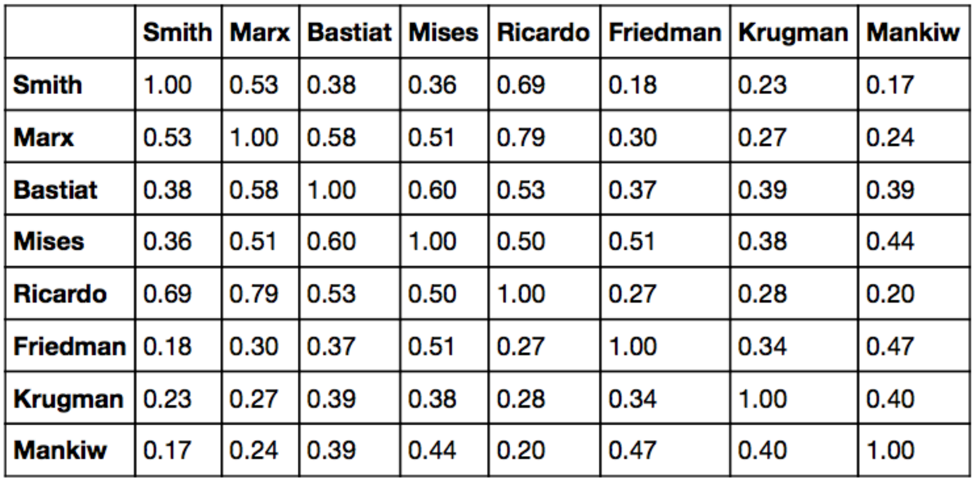
Two interesting follow-up questions would be: Can we predict which economist wrote a certain text from the topics and words used in it? And the perhaps more interesting one: How did economic thinking evolve over time? “An algorithmic history of economic thought” would be a cool title for a study using a much larger dataset of publications spanning over time.

Extracting Content

In order to extract some insight from the data that can help us answer some of the aforementioned questions, we first need to clean it and put it in the right format. In this case, it was not necessary to make any deep changes in the structure of the text or any heavy cleaning process, once the dataset was created as described in the first section.

The cleaning process followed previous to the analysis is simple: 1) tokenization, 2) Removal of puctuation, stopwords (using the list from problem set 2), non-alphanumeric characters and tokens of length less than 4 characters. 3) Stemming, using the Porter stemmer algorithm.

Cosine similarities. A raw measure of similarity between authors is comparing the cosine similarities among each author’s publications. If we compute all the possible comparisons among our 8 economists we obtain the following table:



	Smith	Marx	Bastiat	Mises	Ricardo	Friedman	Krugman	Mankiw
Smith	1.00	0.53	0.38	0.36	0.69	0.18	0.23	0.17
Marx	0.53	1.00	0.58	0.51	0.79	0.30	0.27	0.24
Bastiat	0.38	0.58	1.00	0.60	0.53	0.37	0.39	0.39
Mises	0.36	0.51	0.60	1.00	0.50	0.51	0.38	0.44
Ricardo	0.69	0.79	0.53	0.50	1.00	0.27	0.28	0.20
Friedman	0.18	0.30	0.37	0.51	0.27	1.00	0.34	0.47
Krugman	0.23	0.27	0.39	0.38	0.28	0.34	1.00	0.40
Mankiw	0.17	0.24	0.39	0.44	0.20	0.47	0.40	1.00

Figure 2: Table of cosine similarities

From the table, we see that the most similar authors are Ricardo and Marx, with almost 0.8 cosine similarity. Contrary to what you would expect based on ideological similarity, Smith-Mankiw and Smith-Friedman are the most dissimilar pairs of authors, with just 0.17 and 0.18. In general, eighteenth century authors have a higher cosine similarity between them than between the modern economists.

However, an obvious problem arises when trying to interpret such a table. A high cosine similarity means that authors use a different word composition in their texts, resulting in different vectors of word counts. However, using different words can be the result of having a different approach to an specific topic, but it can also indicate that the authors are talking about different topics altogether. Moreover, it can indicate mere differences in writing style. A more sensible way of comparing authors, that would address this problem, would be to show the same table for each subset of publications sharing a common theme.

LDA. One way to model an author interests would be to use a Latent Dirichlet Allocation (LDA) algorithm that produces a generative model for our text documents. In our case, that amounts to estimating an LDA model that produces a number of topics, defined as generative processes that attach to every word in the corpus a particular propability. In such a context, we would only need to take a look at the posterior probability of each document being generated by a given subset of topics and choose the most probable topics. As we know the author of each document we can then group the distribution of topics over author to have an idea of which topics tend to appear more in a particular author's publications.

This are the topics generated by a regular LDA:

- **Topic 0**: *produc commod profit quantiti increas*
- **Topic 1**: *exchang econom demand credit price*
- **Topic 2**: *gener peopl anoth carri public*
- **Topic 3**: *labour countri greater employ differ*
- **Topic 4**: *polici monetari incom region economi*
- **Topic 5**: *labour commod capit capitalist surplu*
- **Topic 6**: *differ exchang increas consequ system*
- **Topic 7**: *natur foreign interest countri suppos*

- **Topic 8**: *product therefor process social work*
- **Topic 9**: *franc produc peopl becaus principl*

Grouping by author we get:

- **Marx** (top topic: 5)
- **Smith** (top topic: 3)
- **Bastiat** (top topic: 9)
- **Mises** (top topic: 1)
- **Ricardo** (top topic: 0)
- **Friedman** (top topic: 4)
- **Krugman** (top topic: 4)
- **Mankiw** (top topic: 4)

This simple result seems to suggest that modern authors are more interested in topic 4, which could be related to monetary policy or economic policy in general. Mises is more interested in topic 1, which features credit, price and exchange as key words, while both Marx and Smith focus on a topic which features labour as the most probable word. This is not surprising given the importance of the concept of labour in both author's works when defining the notion of value.

However, this grouping over authors has to be done a posteriori, not providing a systematic way of dealing with varying interests. A solution to this problem is to incorporate an extra step in the generative model that incorporates that variation over author. Rosen-Zvi et al. (2014) provide such a model.

Author topic LDA. The Rosen-Zvi Author topic model incorporates the author interests in the classic LDA generative process as depicted in the following graph.

In the traditional LDA, a document's distribution over topics is sampled from a Dirichlet, and then, for each word in a document, we sample a topic. For each word, then we sample from a multinomial distribution specific to the topic sampled for the given word.

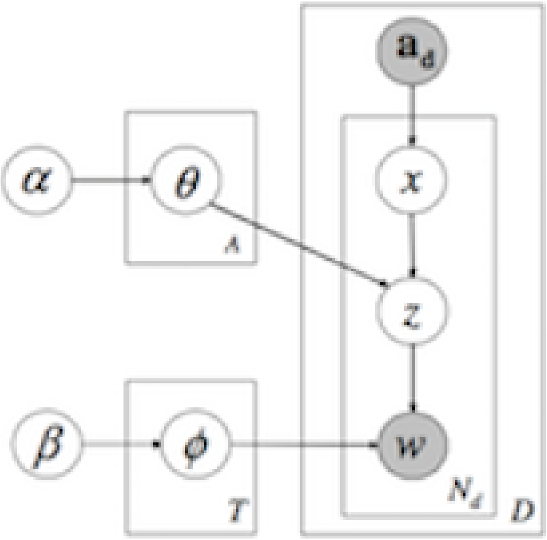


Figure 3: Author Topic model graph

In the author-topic model, words are still sampled from a multinomial specific to each topic, and each topic is still sampled from a Dirichlet distribution. However, the dirichlet distribution from which we draw the model is now specific to each author. To incorporate the possibility of having multiple authors in a given document, they draw one author uniformly for each generated word in the document.

In the graphical model, α and β would be the parameters of the two Dirichlet distributions, over authors and over topics. a_d are the authors of each document. x is the author uniformly drawn from a_d that will produce the word. z is the topic drawn from the dirichlet distribution over authors, and w is the word drawn from the multinomial specific to topic z.

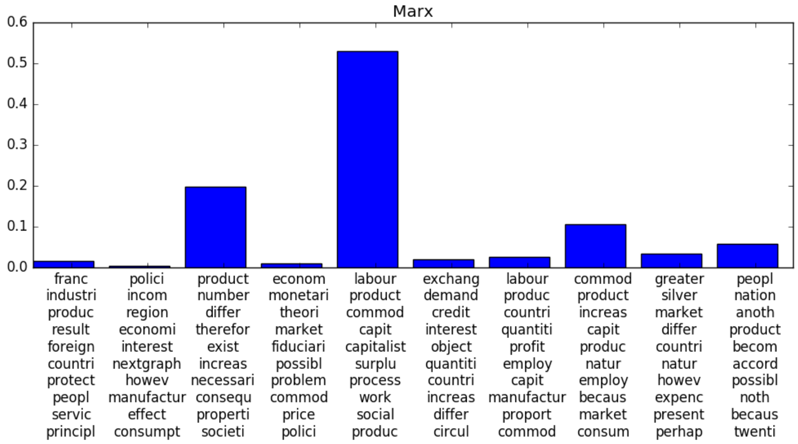
The implementation of this LDA process is relatively straighforward and very similar to the tradictional one. Rosen-Zvi derive a Gibbs sampler that samples jointly pairs of the two latent variables z and x, providing a conditional probability of x and z, and an updating process for the parameters of the Dirichlet distributions analogous to the one obtained in the basic LDA.

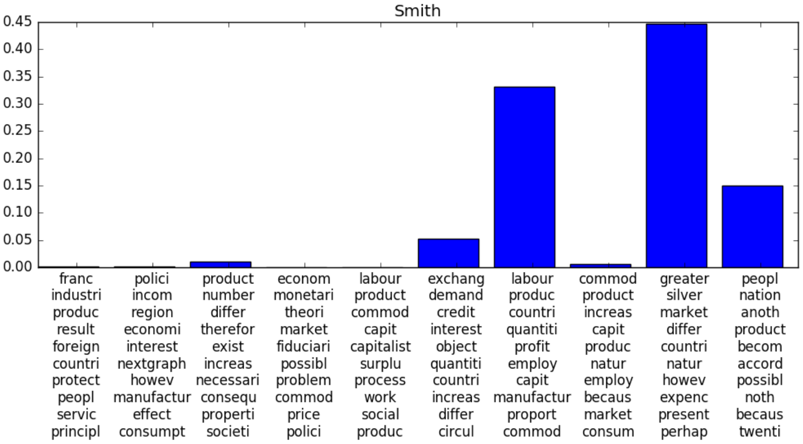
The implementation of Rosen-Zvi author-topic model in python by Dongwoo Kim provides the following results.

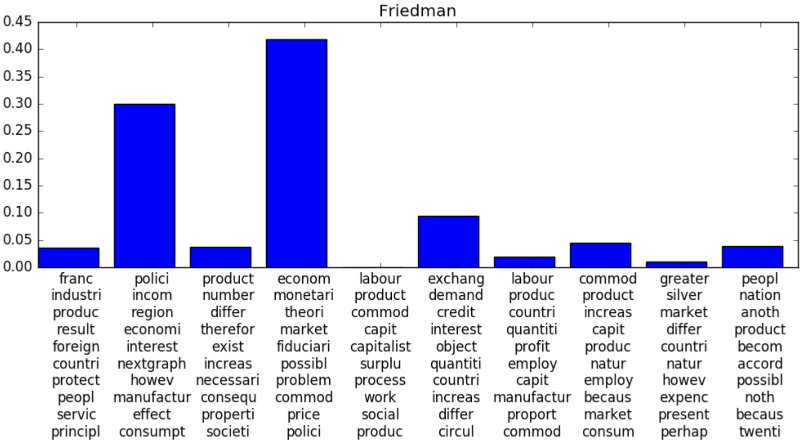
The distribution of words over topic is very similar, as it should be, to the basic LDA:

- **Topic 0**: *franc, industri, produc, result, foreign*
- **Topic 1**: *polici, incom, region, economi, interest*
- **Topic 2**: *product, number, differ, therefor, exist*
- **Topic 3**: *econom, monetari, theori, market, fiduciari*
- **Topic 4**: *labour, product, commod, capit, capitalist*
- **Topic 5**: *exchang, demand, credit, interest, object*
- **Topic 6**: *labour, produc, countri, quantiti, profit*
- **Topic 7**: *commod, product, increas, capit, produc*
- **Topic 8**: *greater, silver, market, differ, countri*
- **Topic 9**: *peopl, nation, anoth, product, becom*

Now, this extended algorithm allows us to extract the distribution of topics by author. As an example, let's examine Smith's, Marx's and Friedman's distribution.







Even with such a rudimental analysis and limited dataset, the results obtained coincide to a certain extent with our popular believe on each author. Smith tends to write on topics related to words like ‘labour’, ‘country’, ‘silver’, ‘people’ or ‘nation’. Marx focuses more on ‘labour’, ‘commodities’, ‘product’, ‘capital’. Friedman leans more towards topics that contain ‘policy’, ‘monetary’, ‘market’, ‘theory’. An easy follow-up exercise that could be done with this results is to use the categorization produced by the author-topic model to predict the author of an unlabelled piece of text.

Conclusion

Our opinions on certain economic thinkers are influenced by our prejudice. This analysis, or an extended one using a similar approach and a larger dataset, can help in identifying similarities between authors and their interests. A more thorough exercise could dig into the positivity and negativity of an author's text on a certain topic. Such an exercise would bring us closer to being able to algorithmically identify an author's ideology. One interesting application is the study of how the focus has shifted in the economic literature from one topic to the other. This rudimental analysis seemed to suggest for instance that more attention is placed now on topics related to policy making.

References

ROSEN-ZVI, M., GRIFFITHS, T., STEYVERS, M., AND SMYTH, P. 2004. The author-topic model for authors and documents. In Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, M. Chickering and J. Halpern, Eds. Morgam Kaufmann, San Francisco, CA, 487–494.