

# Academia's Influence on Policymakers

Paul Laliberte'

Amit Khandelwal (Columbia Business School)

David Atkin (MIT)

CBS - Summer Research

August 1, 2017

# Statement

- **Statement:** To inspect whether there exists a disconnect between economic policy proposed by academic and non-academic institutions.
- **How:** Sentimental analysis of published articles.

- **Methodology:**

- ① **Data Collection and Cleaning:**

- Collecting data through web scraping.
    - Cleaning for punctuation and translating foreign text.

- ② **Machine Learning Methods:**

- Stochastic Gradient Descent.
    - Support Vector Classifier.
    - Evaluation Metrics.

- ③ **Natural Language Processing:**

- Term Frequency - Inverse Document Frequency (tf-idf).
    - Latent Dirichlet Allocation (LDA) Topic Models.

- ④ **Domain Expertise in Economics:**

- Amit and David.

# Data Collection and Cleaning

- **Web scraping:**

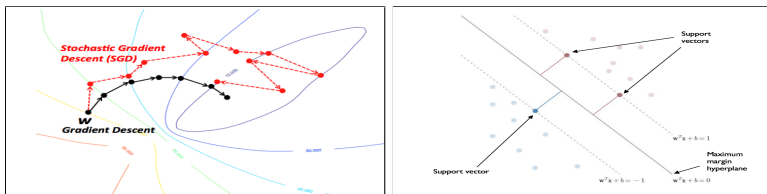
- Final dataset consisted of 2.5 million articles from academic journals and institutional working papers.

- **Cleaning and Translation:**

- Removing punctuation.
- Translate all articles into a common language (English).

# Machine Learning

- Stochastic Gradient Descent (SGD) and Support Vector Classifier (SVC):

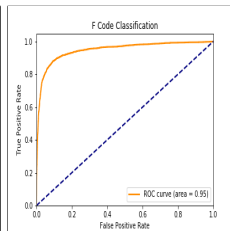
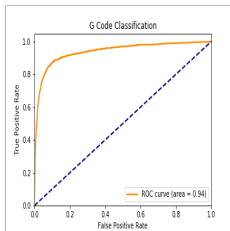
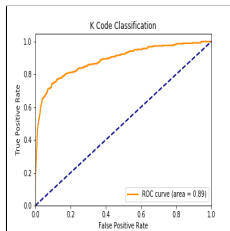


Source: Hong, K. (n.d.): n. pag. Web. 24 July 2017. <<https://bogotobogo.com>>.

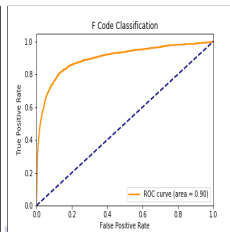
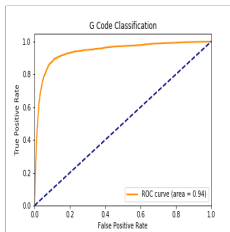
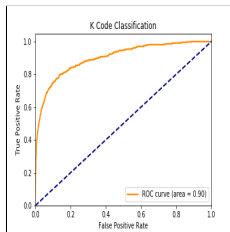
- SGD was used to train the SVC quicker.
- SVC used to predict JEL Codes (D, E, F, G, H, I, J, K, L, and O).
  - Overview: { articles **with** JEL Codes }  $\xrightarrow{\text{train SVC}}$  { article to predict }  $\xrightarrow{\text{Predict JEL Code}}$  { Code or None }.

# Machine Learning (Continued)

- Evaluation Metric: ROC curve.
  - Working papers:



- Journals:



# Natural Language Processing

- **Tf-Idf:**

- Tf: the number of times a term  $t$  appears in a document (article).
- Idf:  $\log \left( \frac{\text{total number of documents in corpus}}{\text{number of documents with term } t} \right)$ .
- Tf-Idf = Tf  $\times$  Idf.

- **n-grams:**

- Example: New York is the best city.
  - Unigrams: {New, York, is, the, best, city}.
  - Bigrams: {New\_York, York\_is, is\_the, the\_best, best\_city}.
  - Trigrams: {New\_York\_is, York\_is\_the, is\_the\_best, the\_best\_city}.

- **Stemming:** reducing words to their derived stem root.

- Example: {learning, learns, learned}  $\xrightarrow{\text{reduced to}}$  {*learn*}.

# Natural Language Processing (Continued)

- **LDA Topic Model:**

- Work in reverse: {collection of documents}  $\longrightarrow$  {inferred topics}.
- Use of the Dirichlet distribution.
- Simplified probability:

$$P(T|t,d) = \frac{\text{total number of tokens } t \text{ in topic } T}{\text{total number of tokens in } T + \beta} \cdot (\text{total words in } d \text{ that belong in } T + \alpha),$$

where  $\beta$  and  $\alpha$  are non-zero constants.

- Example:

<b>"Genetics"</b>	<b>"Evolution"</b>	<b>"Disease"</b>	<b>"Computers"</b>
human	evolution	disease	computer
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Source: David M. Blei. "Probabilistic Topic Models." Communications of the ACM, v. 55, n.4, April 2012.



# Domain Expertise

- Results still need to be interpreted.
  - What is the actual topic?
  - Are the tokens representing a topic even meaningful?
- Example:

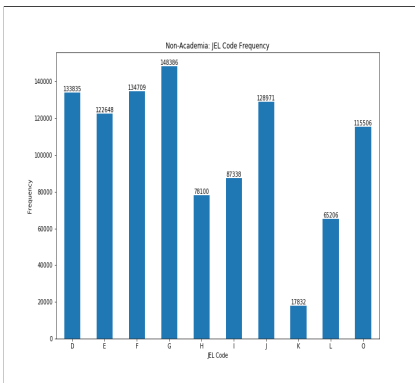
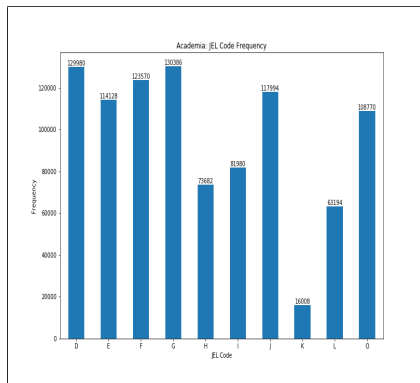
genome	evolutionary	host	models
dna	species	bacteria	information
genetic	organisms	diseases	data
genes	life	resistance	computers
sequence	origin	bacterial	system
gene	biology	new	network
molecular	groups	strains	systems
sequencing	phylogenetic	control	model
map	living	infectious	parallel
information	diversity	malaria	methods
genetics	group	parasite	networks
mapping	new	parasites	software
project	two	united	new
sequences	common	tuberculosis	simulations

Source: David M. Blei. "Probabilistic Topic Models." Communications of the ACM, v. 55, n.4, April 2012.

# Results

- We separated the 2.5 million articles into two groups:
  - ① Written by an individual(s) who spends the majority of their time at an academic institution.
  - ② Those who do not fall into category 1 (FED, Brookings Institute, ect.).
- Results consisted of overall frequencies, frequencies and tf-idf of individual JEL Code, and a topic model.

# JEL Code Frequency



- Not any distinct differences.

# Frequencies and Tf-Idf

Table 1: F Code Frequencies

Academia:	Count	Non-Academia Batch #1: Count	Non-Academia Batch #2: Count:
exchang_rate	5247	exchang_rate	4753
develop_countri	2579	intern_trade	2222
intern_trade	2450	direct_invest	2048
direct_invest	2018	foreign_direct	1950
foreign_direct	1915	foreign_direct_invest	1941
unit_state	1607	develop_countri	1842
free_trade	1603	free_trade	1247
trade_agreement	1450	trade_policy	1170
trade_policy	1346	long_run	1141
real_exchange	1321	unit_state	1136
real_exchange_rate	1314	foreign_exchange	1112
european_union	1184	trade_agreement	1013
trade_liber	1172	real_exchange	993
long_run	1136	real_exchange_rate	987
foreign_exchange	1112	foreign_flow	913
world_trade	1111	european_union	910
econom_growth	963	world_trade	900
trade_flow	918	invest_fdi	882
foreign_trade	892	direct_invest_fdi	877
		exchang_rate	6820
		develop_countri	3478
		intern_trade	3385
		direct_invest	2619
		foreign_direct	2454
		foreign_direct_invest	2434
		free_trade	2110
		trade_agreement	2102
		trade_policy	1909
		real_exchange	1696
		real_exchange_rate	1687
		european_union	1610
		trade_liber	1533
		world_trade	1442
		trade_flow	1400
		foreign_exchange	1392
		long_run	1383
		econom_growth	1227
		latin_america	1173

Table 3: F Code Keyword Frequencies

Academia:	Count	Non-Academia Batch #1: Count	Non-Academia Batch #2: Count:
exchange_rate	11331	exchange_rate	8977
international_trade	3338	international_trade	2933
developing_countries	3277	direct_investment	2909
exchange_rates	3091	exchange_rates	2789
direct_investment	2837	foreign_direct_investment	2662
foreign_direct_investment	2559	developing_countries	2209
free_trade	2502	free_trade	1930
real_exchange_rate	2088	foreign_exchange	1638
current_account	1805	foreign_trade	1402
foreign_exchange	1643	real_exchange_rate	1378
trade_policy	1604	economic_growth	1334
european_union	1536	trade_policy	1320
economic_growth	1500	current_account	1289
trade_agreements	1466	european_union	1227
foreign_trade	1302	trade_agreements	938
terms_of_trade	1063	terms_of_trade	915
comparative_advantage	907	trade_balance	811
economic_integration	843	purchasing_power	784
trade_balance	721	purchasing_power_parity	723
foreign_investment	680	foreign_investment	674
		exchange_rate	14217
		developing_countries	4558
		international_trade	4482
		exchange_rates	3904
		direct_investment	3718
		foreign_direct_investment	3292
		free_trade	3193
		real_exchange_rate	2693
		current_account	2654
		trade_policy	2312
		foreign_exchange	2160
		trade_agreements	2151
		european_union	2042
		economic_growth	1765
		foreign_trade	1545
		terms_of_trade	1454
		economic_integration	1284
		comparative_advantage	1101
		balance_of_payments	847
		foreign_investment	835

Table 2: F Code Tf-Idf

Academia:	Tf-Idf	Non-Academia:	Tf-Idf
trade_wto	0.9303	india_trade	1.0
corpor_govern	0.9081	even_high	1.0
trade_credit	0.8971	eu_access	1.0
human_right	0.8840	semin_contribut	1.0
energ_intens	0.8792	air_servic	0.9633
state_aid	0.8750	human_right	0.9177
currenc_reserv	0.8650	trade_mark	0.9045
humanitarian_emerg	0.8642	gold_exchang	0.9030
knowledg_exchang	0.8507	corpor_govern	0.9018
fair_trade	0.8390	mega_ftas	0.8953
humanitarian_aid	0.8352	foreign_languag	0.8921
price_foreign	0.8228	health_insur	0.8916
foreign_offici	0.8126	trade_secret	0.8879
loss_avers	0.8104	antidump_duti	0.8777
big_mac	0.8087	neo_liber	0.8744
steel_import	0.8062	trade_credit	0.8718
civil_war	0.8049	exit_cost	0.8716
australia_china	0.8045	softwood_lumber	0.8706
payment_balanc	0.8035	energ_subsid	0.8641
domin_posit	0.8020	state_aid	0.8629

# Economic Topic Model

- [http://nbviewer.jupyter.org/github/PaulLaliberte/jupyter\\_notebooks/blob/master/topic\\_model.ipynb](http://nbviewer.jupyter.org/github/PaulLaliberte/jupyter_notebooks/blob/master/topic_model.ipynb)

# Next Steps

- **word2vec**

- Take a sentence and create a vector of word embeddings.
- Example: {england has kings and queens}  $\Rightarrow$  {0.2, 0.0, 0.4, 0.0, 0.4}.

- word2vec embeddings can be used for:

- Similarity of sentences.
- Finding odd-word-out.
- Advanced classification: `model.most_similar('man', 'queen') = "king."`
- Underlying probability of sentences within a topic.