Assignment 4: Text Mining

Predict 453

Section 55

Spring Quarter

School of Continuing Studies

Northwestern University

Daniel Prusinski

In compliance for Master of Science Predictive Analytics

Bachelor of Science Business Marketing

Data Analyst

US Bank

220 S 6th St

Minneapolis, MN

Daniel Prusinski
4/20/2013

Text Mining

According to IBM, it is estimated that 80% of an organizations data is in the form of text. At the micro level the text offers valuable information, but at the macro level it offers little value in its form. Text analytics seeks to add value to text data at the macro level through linguistic analysis and Natural Language Processing (NLP), which mines the text for information beyond the original intention of the text. This extracted information is then grouped into categories for macro level analysis that adds value from a predictive modeling standpoint.

The definition of text mining is "the process of analyzing collections of textual materials in order to capture key concepts and themes and uncover hidden relationships and trends without requiring that you know the precise words or terms that authors have used to express those concepts (IBM)." Bruce Ratner's paradigm on data analysis is the following and has been heavily supported in the Northwestern University (NU) Master of Science in Predictive Analytics (MSPA) program:

Problem/Objective: Identify the overall desired outcome.

Data: Identify the data that will be studied.

Analysis: Utilize objective statistical techniques to analyze the data.

Model: From the initial analysis, utilize a model that conforms to the data.

Results/Interpretation: Once the model has been validated and iterations complete, make recommendations based on the model.

IBM's paradigm consists of four steps.

1. Identify text on which analysis will be conducted. This is rather straightforward, and for this assignment I have connected several international development RSS feeds of which to extract text.
2. Apply specific text mining algorithms to the text. This is the most complicated step and will require further explanation.
3. Analyze the results and categorize key concepts that will be used for modeling.
4. Apply numerical predictive modeling techniques to the numerical values associated with the key categorical text information.

The requirements for this assignment will be fulfilled through step three. Before diving into the steps outlined above, further understanding is needed on step two. Specifically, the pulling of information is executed through the following steps:

1. The data is converted to a standard format.
2. Candidate terms are identified through the use of templates, libraries, and resources. It is in the libraries that relationships between words are identified and applied for specific text. This includes part-of-speech-code establishing nouns, verbs, adjectives, etc. A candidate term is a word or group of words that are used to "identify key concepts in the text (IBM)".
3. Equivalent classes, similar phrases, are identified and extracted.

4. Types are assigned through labels such as higher level concepts, positive and negative words, places, etc.

International development is a broad discipline that covers many different sectors including emergency relief, disaster recovery, and country building. The RSS feeds that I am mining falls within the international development sector. Four different RSS feeds were imported utilizing the Web Feed node in SPSS. Each RSS Feed covers the following topics on their own websites:

International Relief & Development – This feed is broad in its nature and covers current developments.

Canadian International Development Agency – This feed highlights the different projects and programs Canada is involved with throughout the world.

Center for Global Development – This feed gives briefs on the latest updates focusing on capacity building around the world.

Journal of International Relations and Development – This feed has more of an international relations focus and is more philosophical in nature.

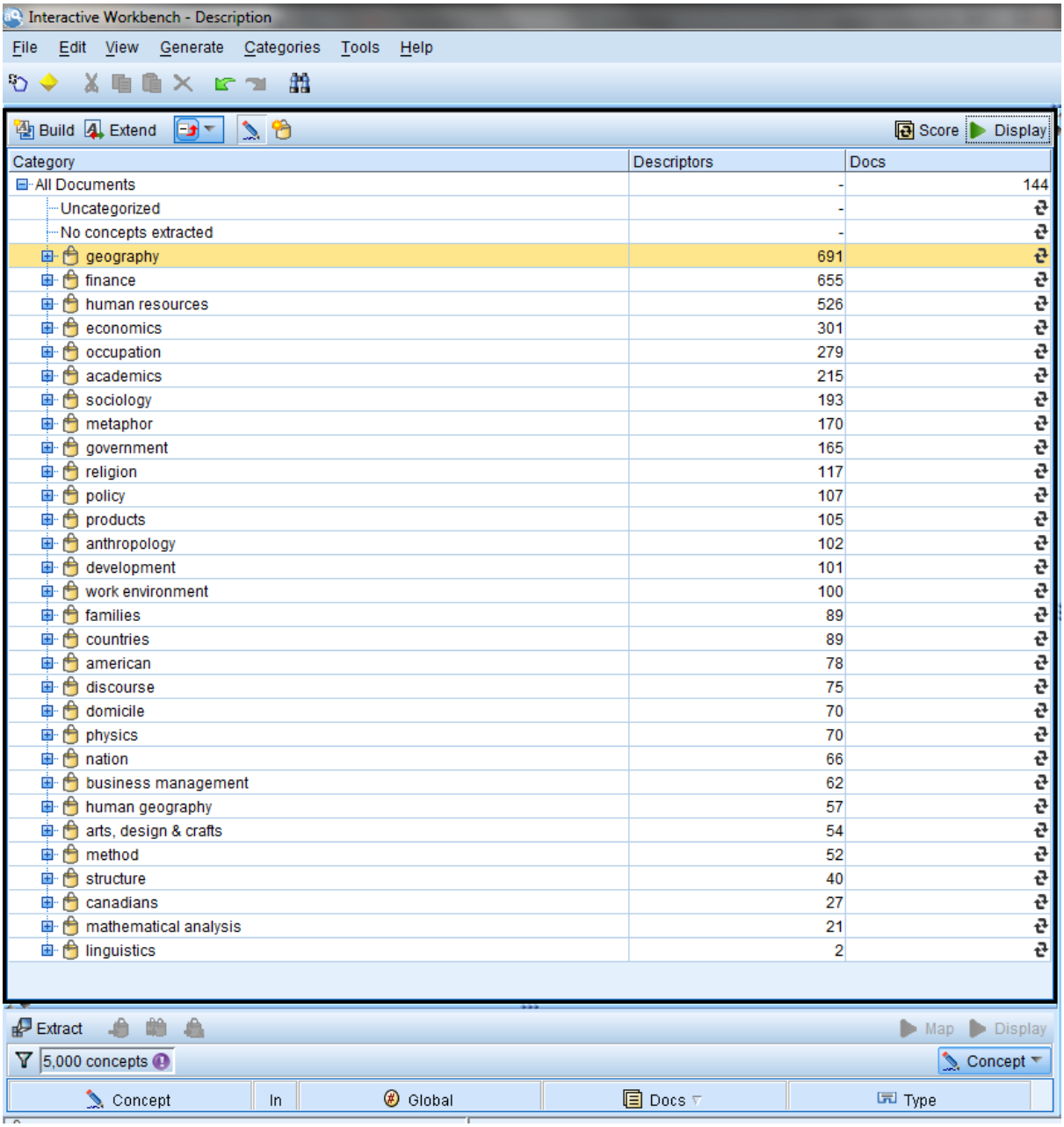
All of these feeds have international development in common but also have subtle differences that text mining will uniquely bring together.

While this is the first exercise I have executed in SPSS Modeler, useful information was gleaned and ideas for inter-organization collaboration was discovered from my elementary mining. After establishing the RSS Feeds in SPSS, I deployed the Text Mining node and utilized the 'Build Interactively' selection to analyze the text in a category format. Five thousand concepts were analyzed

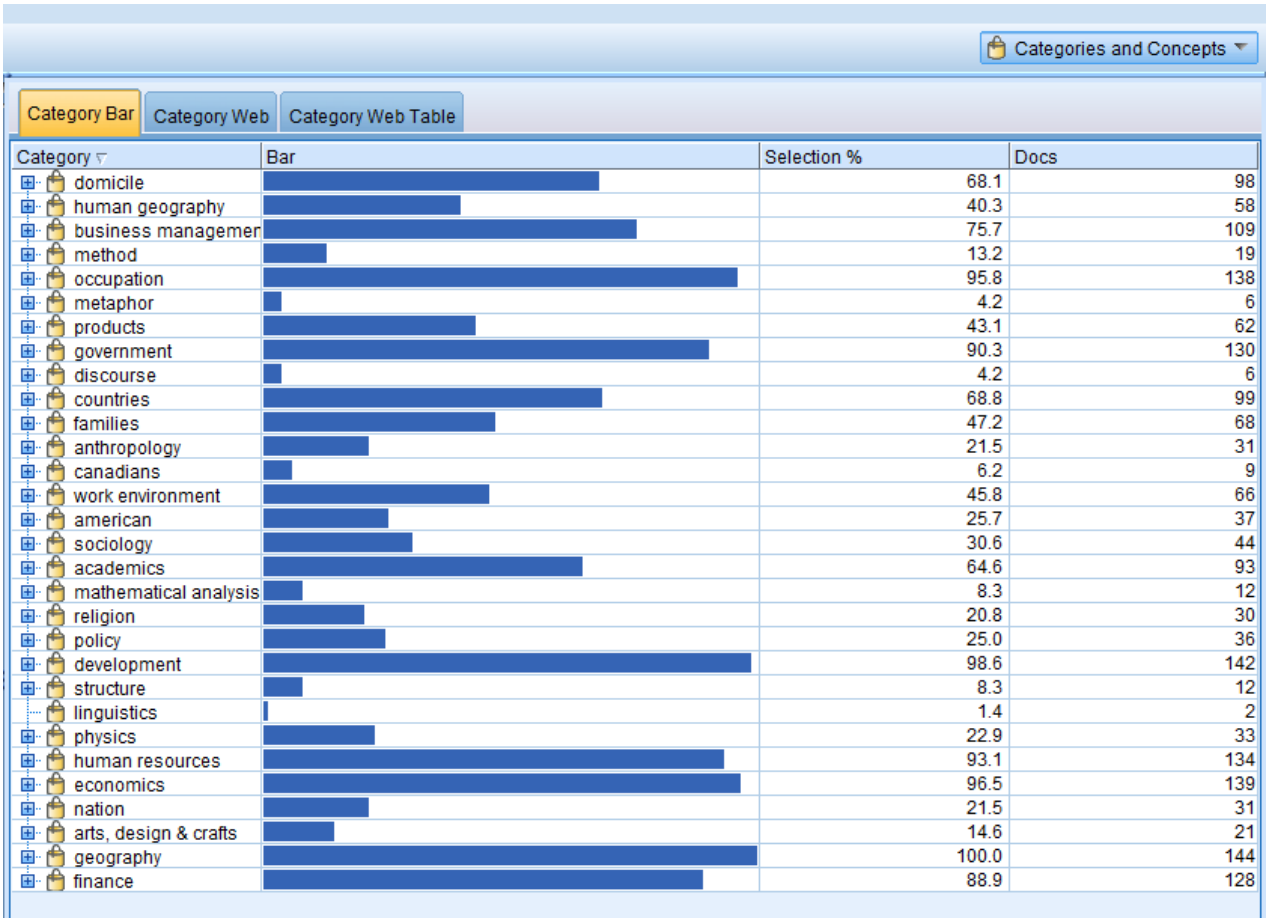
from 144 different documents. I selected a Standard English dictionary from which to analyze the information. Thirty different categories divided all the concepts. The category 'geography' was found in 131 documents and encompassed the whole world. Little information was gleaned from this category except that the feeds are focused on world events, literally. A sub-click on the category 'geography' yielded better information. All the feeds mention America and 74% mention Europe, but Africa and Asia are mentioned in less than 50% of the feeds. I would have expected to see the reverse of this trend based on the feeds that are being monitored. Surprisingly, the category 'economics' is covered in 96% of the documents. This leads me to believe that economics is a holistic theme across all avenues of international development. From this nugget of information, knowledge of economics would be very helpful for someone looking for an occupation within international development. Not surprising, the term government is ubiquitous, but 'law' is the major subset, which leads me to believe that law and international development are intertwined. Occupation is another category that is in every document, but white collar worker is found in 79% of the documents. Domicile is found in almost 70% of the documents, which leads me to believe that shelter is a major trend right now in international development. The category 'family' is found in 47% of the documents and is a rather myopic category within such a large topical study. This leads me to believe that families are a major focus for international development and resources are devoted to this topic.

Overall, I am amazed at the capability of this software. I have just scratched the surface when it comes to understanding the capabilities of this software, but I feel like I have an analytical foundation on what topics are hot right now in international development. Moving forward I would like to become more comfortable in the dictionaries for SPSS and how that affects the analysis.

Appendix 1



Category	Descriptors	Docs
All Documents	-	144
Uncategorized	-	
No concepts extracted	-	
geography	691	
finance	655	
human resources	526	
economics	301	
occupation	279	
academics	215	
sociology	193	
metaphor	170	
government	165	
religion	117	
policy	107	
products	105	
anthropology	102	
development	101	
work environment	100	
families	89	
countries	89	
american	78	
discourse	75	
domicile	70	
physics	70	
nation	66	
business management	62	
human geography	57	
arts, design & crafts	54	
method	52	
structure	40	
canadians	27	
mathematical analysis	21	
linguistics	2	



Category	Bar	Selection %	Docs
domicile		68.1	98
human geography		40.3	58
business managemer		75.7	109
method		13.2	19
occupation		95.8	138
metaphor		4.2	6
products		43.1	62
government		90.3	130
discourse		4.2	6
countries		68.8	99
families		47.2	68
anthropology		21.5	31
canadians		6.2	9
work environment		45.8	66
american		25.7	37
sociology		30.6	44
academics		64.6	93
mathematical analysis		8.3	12
religion		20.8	30
policy		25.0	36
development		98.6	142
structure		8.3	12
linguistics		1.4	2
physics		22.9	33
human resources		93.1	134
economics		96.5	139
nation		21.5	31
arts, design & crafts		14.6	21
geography		100.0	144
finance		88.9	128