Assignment 5: Synonym Customization

Predict 453

Section 55

Spring Quarter

School of Continuing Studies

Northwestern University

Janki Vora	&	Daniel Prusinski
Software Engineer		Data Analyst
IBM		US Bank
Dallas/Fort Worth		Minneapolis
TX		MN

In Compliance with Master of Science Predictive Analytics

Bachelor of Science Business Marketing

Daniel Prusinski
4/22/2013

Synonym Customization

IBM's text analytics software is a very powerful tool in the vast field of unstructured text analysis. The software has various dictionaries loaded to aid in the analysis of texts from different sectors. In the standard setting, there is a generic dictionary used for initial analysis. While this software is intuitive, it is necessary to hone in the analysis by creating synonyms. The eleven documents being analyzed all correlate in regard to international development in the Middle East specifically focusing on the Syrian conflict and the specific names of the documents can be found in Appendix 1. Through creating synonyms, the analysis will better reflect the context of the text being analyzed.

Below is the initial text analysis utilizing the general text mining settings for IBM's SPSS. The initial step is to identify concepts where adding

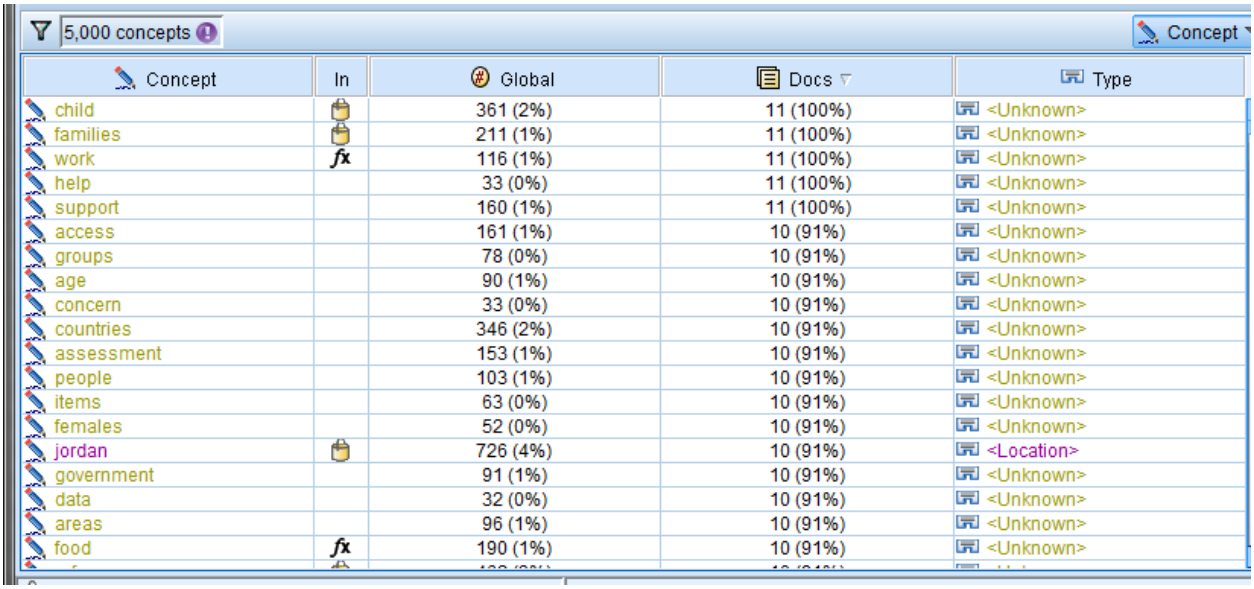
Category	Bar	Selection %	Docs
skills		36.4	4
management		63.6	7
services		81.8	9
careers		63.6	7
programming		27.3	3
occupation		100.0	11
products		100.0	11
biology fields		100.0	11
agriculture		63.6	7
costs		36.4	4
agricultural laborers		18.2	2
health		100.0	11
natural resources		18.2	2
families		100.0	11
refugees		100.0	11
sociology		100.0	11
financial management		36.4	4
food		18.2	2
nature		72.7	8
ecology		100.0	11
mathematical analysis		81.8	9
syrians		63.6	7
human settlements		100.0	11
resources		90.9	10
human resources		100.0	11
economics		100.0	11
house		54.5	6
geography		100.0	11
finance		100.0	11
farming systems		63.6	7

synonyms will aid the extraction process in combining similar terms appropriately. The color coded boxes correspond to terms that are close enough in meaning to be synonyms in the international development sector.

Other concepts that will be used to create synonyms in the concepts field are:

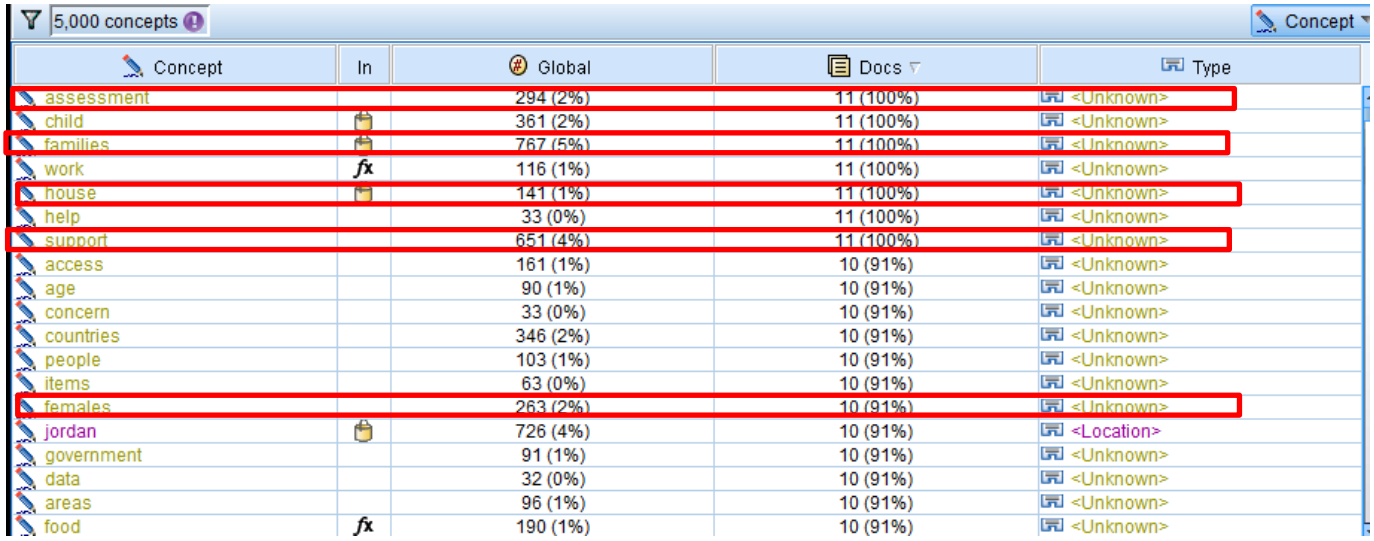
Woman	Community	Assessment	Support	Shelter
Female	Groups	Methodology	Assistance	House
Girls	Families	Monitoring	Humanitarian	Homs
Child Labor	Refugees	Questionnaire	Services	Home
Wife	Relatives	Analysis	Opportunities	Housing
Women	Family	Interviews	Partners	Address
Minor	Refugee	Livelihoods	Agencies	Structure

Before the synonyms were added, the concepts were the following:



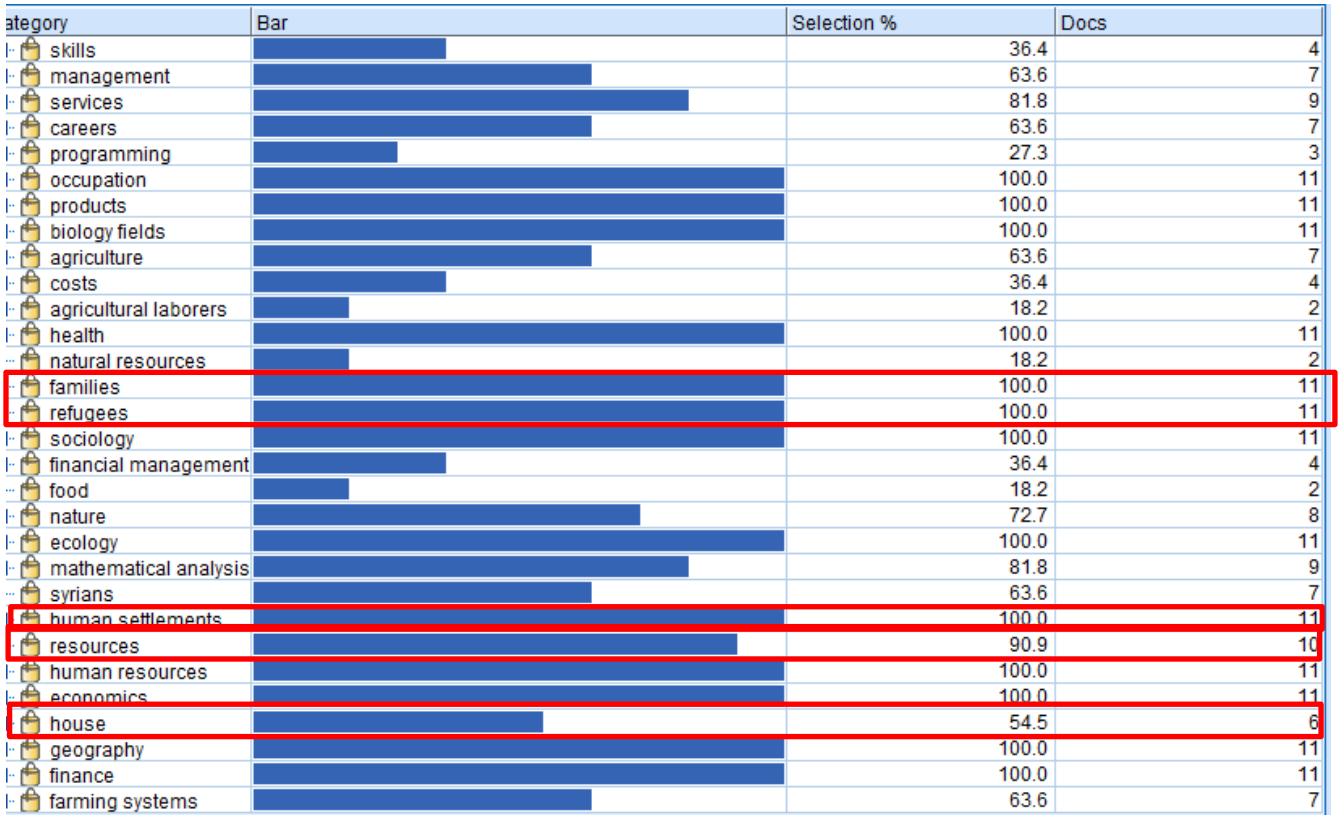
Concept	In	Global	Docs	Type
child		361 (2%)	11 (100%)	<Unknown>
families		211 (1%)	11 (100%)	<Unknown>
work	fx	116 (1%)	11 (100%)	<Unknown>
help		33 (0%)	11 (100%)	<Unknown>
support		160 (1%)	11 (100%)	<Unknown>
access		161 (1%)	10 (91%)	<Unknown>
groups		78 (0%)	10 (91%)	<Unknown>
age		90 (1%)	10 (91%)	<Unknown>
concern		33 (0%)	10 (91%)	<Unknown>
countries		346 (2%)	10 (91%)	<Unknown>
assessment		153 (1%)	10 (91%)	<Unknown>
people		103 (1%)	10 (91%)	<Unknown>
items		63 (0%)	10 (91%)	<Unknown>
females		52 (0%)	10 (91%)	<Unknown>
jordan		726 (4%)	10 (91%)	<Location>
government		91 (1%)	10 (91%)	<Unknown>
data		32 (0%)	10 (91%)	<Unknown>
areas		96 (1%)	10 (91%)	<Unknown>
food	fx	190 (1%)	10 (91%)	<Unknown>

After the synonyms were added, the concepts changed as shown below:



Concept	In	Global	Docs	Type
assessment		294 (2%)	11 (100%)	<Unknown>
child		361 (2%)	11 (100%)	<Unknown>
families		767 (5%)	11 (100%)	<Unknown>
work	fx	116 (1%)	11 (100%)	<Unknown>
house		141 (1%)	11 (100%)	<Unknown>
help		33 (0%)	11 (100%)	<Unknown>
support		651 (4%)	11 (100%)	<Unknown>
access		161 (1%)	10 (91%)	<Unknown>
age		90 (1%)	10 (91%)	<Unknown>
concern		33 (0%)	10 (91%)	<Unknown>
countries		346 (2%)	10 (91%)	<Unknown>
people		103 (1%)	10 (91%)	<Unknown>
items		63 (0%)	10 (91%)	<Unknown>
females		263 (2%)	10 (91%)	<Unknown>
jordan		726 (4%)	10 (91%)	<Location>
government		91 (1%)	10 (91%)	<Unknown>
data		32 (0%)	10 (91%)	<Unknown>
areas		96 (1%)	10 (91%)	<Unknown>
food	fx	190 (1%)	10 (91%)	<Unknown>

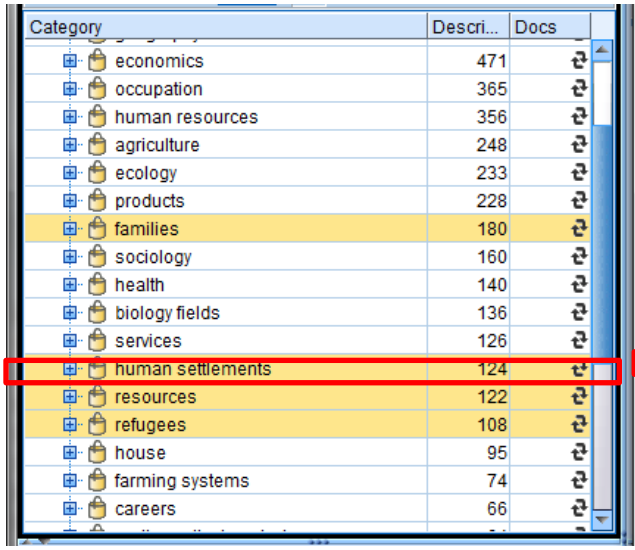
The synonyms that were added greatly increased the occurrence of the corresponding concepts as seen above. The screen shot shows that the linking of the concepts has greatly increased the pervasiveness throughout the documents. At the category level, there is not much difference in regard to added difference from the synonym customization.



ategory	Bar	Selection %	Docs
skills		36.4	4
management		63.6	7
services		81.8	9
careers		63.6	7
programming		27.3	3
occupation		100.0	11
products		100.0	11
biology fields		100.0	11
agriculture		63.6	7
costs		36.4	4
agricultural laborers		18.2	2
health		100.0	11
natural resources		18.2	2
families		100.0	11
refugees		100.0	11
sociology		100.0	11
financial management		36.4	4
food		18.2	2
nature		72.7	8
ecology		100.0	11
mathematical analysis		81.8	9
syrians		63.6	7
human settlements		100.0	11
resources		90.9	10
human resources		100.0	11
economics		100.0	11
house		54.5	6
geography		100.0	11
finance		100.0	11
farming systems		63.6	7

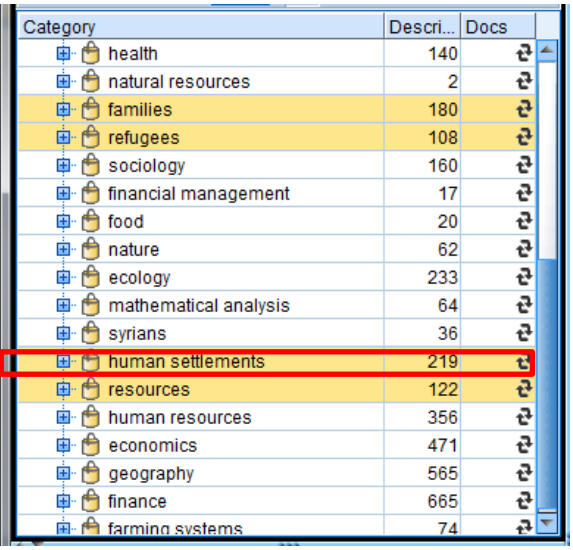
The categories ‘families’, ‘refugees’, and ‘human settlements’ are found in all the documents prior to synonynm customization. These categories encompass the changes I made at the concept level, and there is little demonstration of an affect at the categorical level. The standard dictionay for SPSS took into account the synonynms and grouped them into clusters called categories. Further analysis paring the concept and categorization levels is needed.

Prior to Synonynm Customization

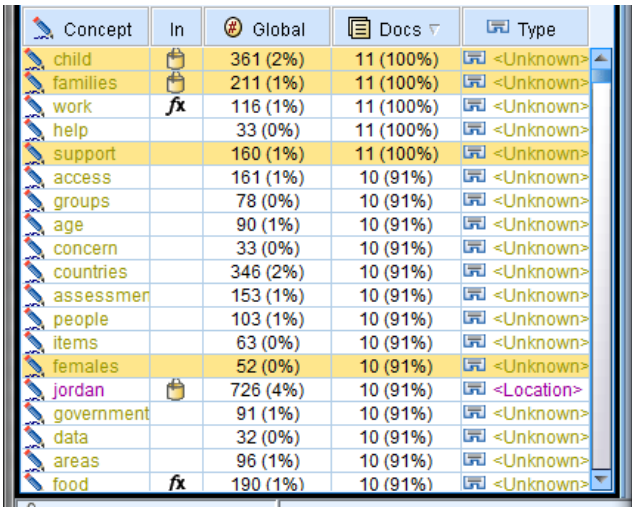


Category	Descri...	Docs
economics	471	
occupation	365	
human resources	356	
agriculture	248	
ecology	233	
products	228	
families	180	
sociology	160	
health	140	
biology fields	136	
services	126	
human settlements	124	
resources	122	
refugees	108	
house	95	
farming systems	74	
careers	66	

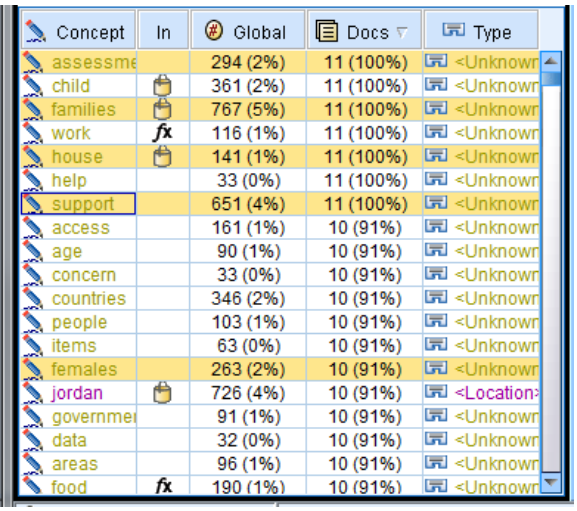
After Synonynm Customization



Category	Descri...	Docs
health	140	
natural resources	2	
families	180	
refugees	108	
sociology	160	
financial management	17	
food	20	
nature	62	
ecology	233	
mathematical analysis	64	
syrians	36	
human settlements	219	
resources	122	
human resources	356	
economics	471	
geography	565	
finance	665	
farming systems	74	



Concept	In	Global	Docs	Type
child		361 (2%)	11 (100%)	<Unknown>
families		211 (1%)	11 (100%)	<Unknown>
work	fx	116 (1%)	11 (100%)	<Unknown>
help		33 (0%)	11 (100%)	<Unknown>
support		160 (1%)	11 (100%)	<Unknown>
access		161 (1%)	10 (91%)	<Unknown>
groups		78 (0%)	10 (91%)	<Unknown>
age		90 (1%)	10 (91%)	<Unknown>
concern		33 (0%)	10 (91%)	<Unknown>
countries		346 (2%)	10 (91%)	<Unknown>
assessmen		153 (1%)	10 (91%)	<Unknown>
people		103 (1%)	10 (91%)	<Unknown>
items		63 (0%)	10 (91%)	<Unknown>
females		52 (0%)	10 (91%)	<Unknown>
jordan		726 (4%)	10 (91%)	<Location>
government		91 (1%)	10 (91%)	<Unknown>
data		32 (0%)	10 (91%)	<Unknown>
areas		96 (1%)	10 (91%)	<Unknown>
food	fx	190 (1%)	10 (91%)	<Unknown>



Concept	In	Global	Docs	Type
assessme		294 (2%)	11 (100%)	<Unknown
child		361 (2%)	11 (100%)	<Unknown
families		767 (5%)	11 (100%)	<Unknown
work	fx	116 (1%)	11 (100%)	<Unknown
house		141 (1%)	11 (100%)	<Unknown
help		33 (0%)	11 (100%)	<Unknown
support		651 (4%)	11 (100%)	<Unknown
access		161 (1%)	10 (91%)	<Unknown
age		90 (1%)	10 (91%)	<Unknown
concern		33 (0%)	10 (91%)	<Unknown
countries		346 (2%)	10 (91%)	<Unknown
people		103 (1%)	10 (91%)	<Unknown
items		63 (0%)	10 (91%)	<Unknown
females		263 (2%)	10 (91%)	<Unknown
jordan		726 (4%)	10 (91%)	<Location>
governme		91 (1%)	10 (91%)	<Unknown
data		32 (0%)	10 (91%)	<Unknown
areas		96 (1%)	10 (91%)	<Unknown
food	fx	190 (1%)	10 (91%)	<Unknown

At the categorization level, 'human settlements' is the only category that changed in regard to an increase in occurrence. The other categories already took into account the synonyms. Analyzing at the concept level better demonstrates the difference the synonym customization made. Three of the greatest changes occurred with 'support', 'females', and 'house'. Occurrence of the concepts occurs at a much greater frequency after the synonym

customization. The screen shots above demonstrate where the most important differences occurred.

All the documents chosen for this assignment were focused on the Middle East and the Syrian conflict. Synonym customization provided a more precise analysis of the text focusing on key international development concepts. While the synonym customization provided a more accurate summary of the text found within the eleven documents, being a novice and prior category grouping muddled the desired results I hoped to accomplish. Perhaps I am not executing a step after the synonym customization, but it appears that little was accomplished beyond increased frequency of defined concepts. An additional factor for this lack of results could be the fact that the documents used for this assignment are rather large. After the synonym customization, one is left with the thought 'now what?' How does one apply this information within the SPSS schema to glean more knowledge beyond the increased frequency of specific terms? Tolerating ambiguity is a necessary characteristic for learning new software, and further analysis is welcomed to further hone in on better utilizing SPSS text mining capabilities.

Appendix 1: Documents Names

1. 5CD49416CE1E16B5852577AB006AB5DF-Full_Report
2. 2012-syrian-refugee-assessment
3. CP&GBVZaatariAssessment(1)
4. FCKupload_file_FAO-Syria-Crisis-Report-en
5. Mission report on Humanitarian situation as a result of confict in Syria external
6. OutreachAnalysisNovember2012
7. Replenishment_2013NeedsAssessment_Report_en
8. Syrian Refugees living in the Community in Jordan - Assessment Report
9. UPP COMPREHENSIVE_ASSESSMENT_SYRIAN_REFUGEES_2012
10. wfp251901
11. CP&GBVZaatariAssessment

☒Include the names of the 10 documents you used at the end of your report.

☒ Using the software tool (IBM SPSS Modeler) read in 10 documents of three to 10 pages in length that relate to your area of interest.

☒Process the text using the standard dictionaries and extraction settings.

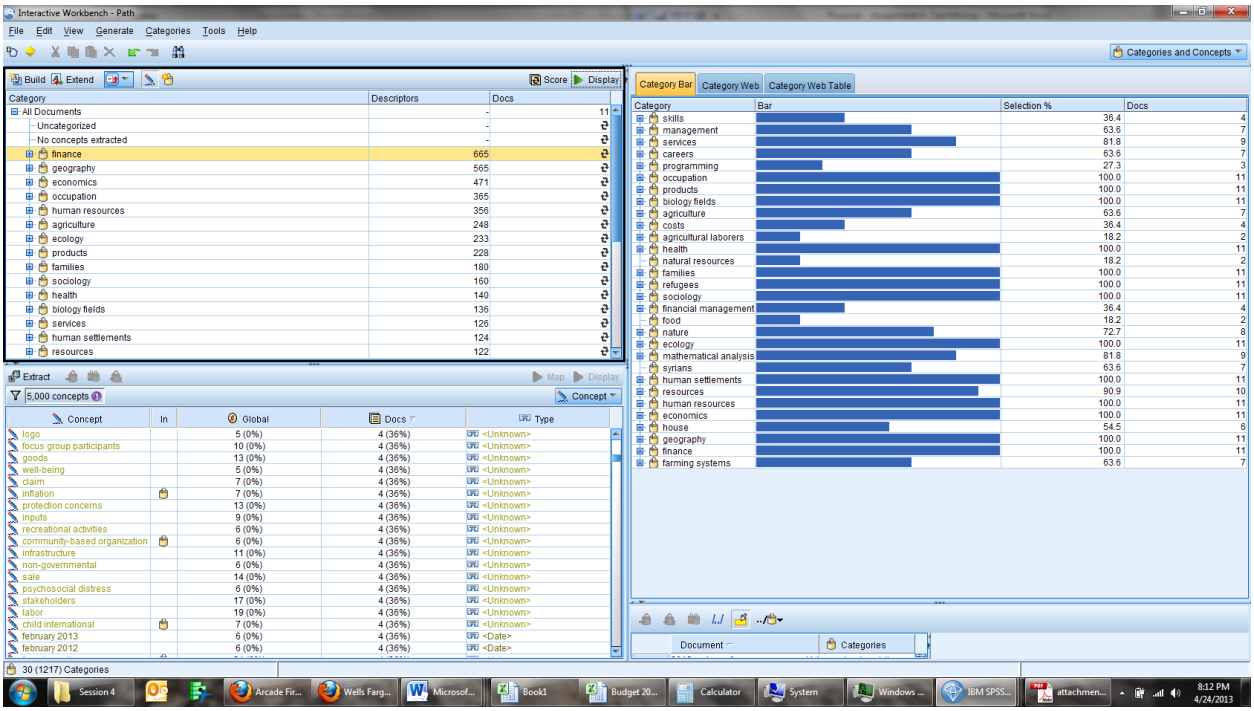
☒Identify content in the text where adding synonyms to the extraction process will combine similar terms appropriately.

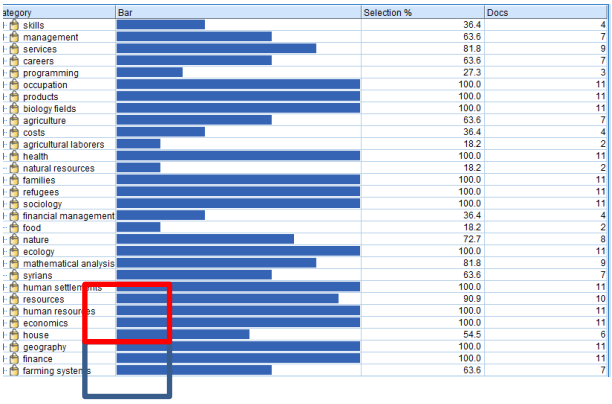
☒Make the necessary changes (at least five) and run the extraction again.

☒Use screen shots of the results to show the differences in the two methods.

☒Describe the impact of the changes you made referring to the results in the screen shots to demonstrate where the most important differences occurred.

☒Explain the value of the changes you made in terms of the improvements they provided with respect to accuracy of interpretation of the original text.


Interactive Workbench - Path
File Edit View Generate Categories Tools Help
Build Extend Score Display
Category Descriptors Docs
All Documents 11
Uncategorized
No concepts extracted
finance 665
geography 565
economics 471
occupation 365
human resources 356
agriculture 248
ecology 233
products 228
families 180
sociology 160
health 140
biology fields 136
services 126
human settlements 124
resources 122
Extract Map Display
5,000 concepts Concept
Concept In Global Docs Type
logo 5 (0%) 4 (36%) <Unknown>
focus group participants 10 (0%) 4 (36%) <Unknown>
goods 13 (0%) 4 (36%) <Unknown>
well-being 5 (0%) 4 (36%) <Unknown>
claim 7 (0%) 4 (36%) <Unknown>
inflation 7 (0%) 4 (36%) <Unknown>
protection concerns 13 (0%) 4 (36%) <Unknown>
inputs 9 (0%) 4 (36%) <Unknown>
recreational activities 6 (0%) 4 (36%) <Unknown>
community-based organization 6 (0%) 4 (36%) <Unknown>
infrastructure 11 (0%) 4 (36%) <Unknown>
non-governmental 6 (0%) 4 (36%) <Unknown>
sale 14 (0%) 4 (36%) <Unknown>
psychosocial distress 6 (0%) 4 (36%) <Unknown>
stakeholders 17 (0%) 4 (36%) <Unknown>
labor 19 (0%) 4 (36%) <Unknown>
child international 7 (0%) 4 (36%) <Unknown>
february 2013 6 (0%) 4 (36%) <Date>
february 2012 6 (0%) 4 (36%) <Date>
30 (1217) Categories
Categories and Concepts
Category Bar Category Web Category Web Table
Category Bar Selection % Docs
skills 36.4 4
management 63.6 7
services 81.8 9
careers 63.6 7
programming 27.3 3
occupation 100.0 11
products 100.0 11
biology fields 100.0 11
agriculture 63.6 7
costs 36.4 4
agricultural laborers 18.2 2
health 100.0 11
natural resources 18.2 2
families 100.0 11
refugees 100.0 11
sociology 100.0 11
financial management 36.4 4
food 18.2 2
nature 72.7 8
ecology 100.0 11
mathematical analysis 81.8 9
syrians 63.6 7
human settlements 100.0 11
resources 90.9 10
human resources 100.0 11
economics 100.0 11
house 54.5 6
geography 100.0 11
finance 100.0 11
farming systems 63.6 7
Document Categories
Session 4
Arcade Fir...
Wells Farg...
Microsof...
Book1
Budget 20...
Calculator
System
Windows ...
IBM SPSS...
attachmen...
8:12 PM
4/24/2013


ategory Bar Selection % Docs
skills 36.4 4
management 63.6 7
services 81.8 9
careers 63.6 7
programming 27.3 3
occupation 100.0 11
products 100.0 11
biology fields 100.0 11
agriculture 63.6 7
costs 36.4 4
agricultural laborers 18.2 2
health 100.0 11
natural resources 18.2 2
families 100.0 11
refugees 100.0 11
sociology 100.0 11
financial management 36.4 4
food 18.2 2
nature 72.7 8
ecology 100.0 11
mathematical analysis 81.8 9
syrians 63.6 7
resources 90.9 10
human resources 100.0 11
economics 100.0 11
house 54.5 6
geography 100.0 11
finance 100.0 11
farming systems 63.6 7