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

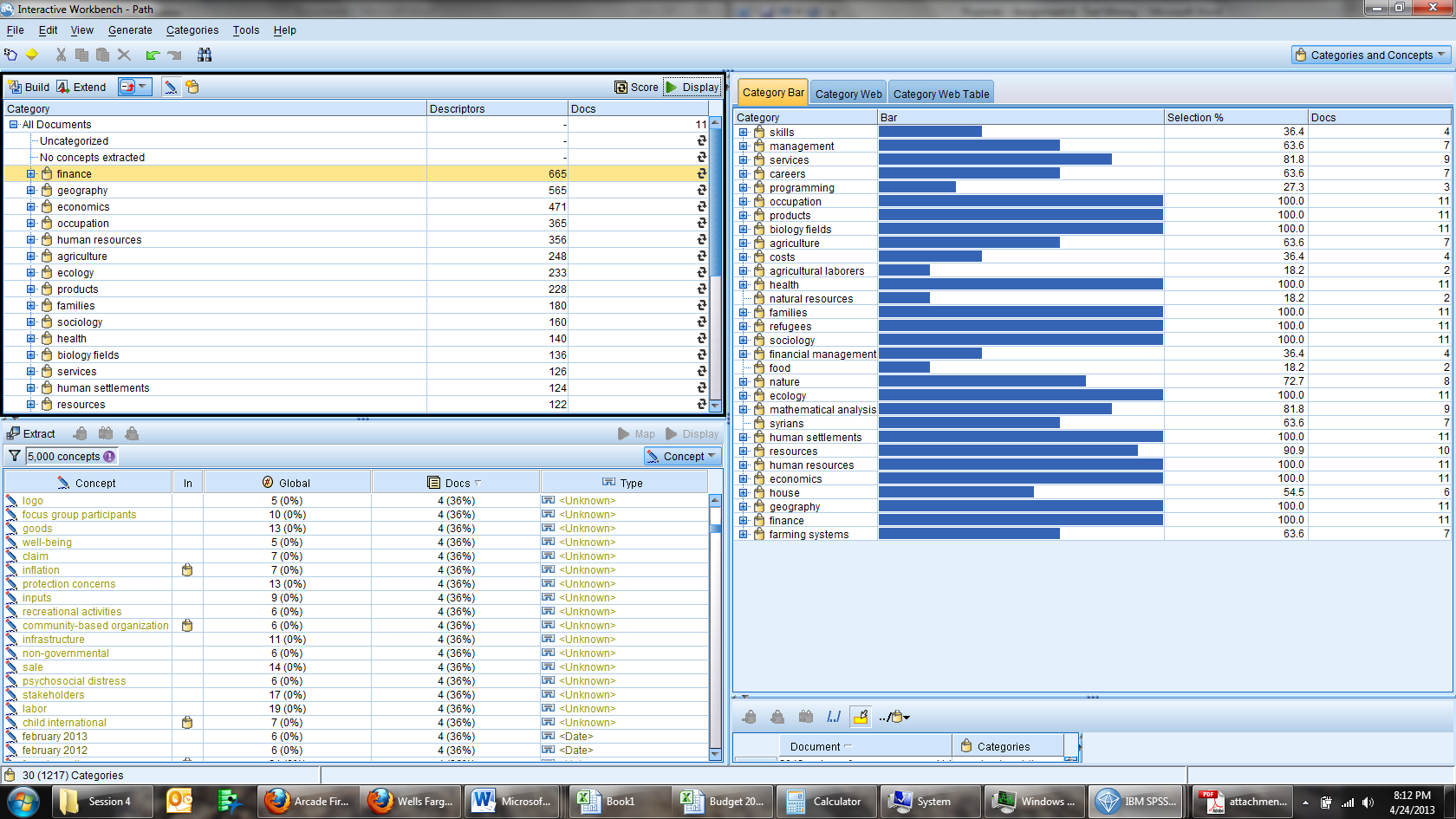
In Compliance with Master of Science Predictive Analytics

Bachelor of Science Business Marketing

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.

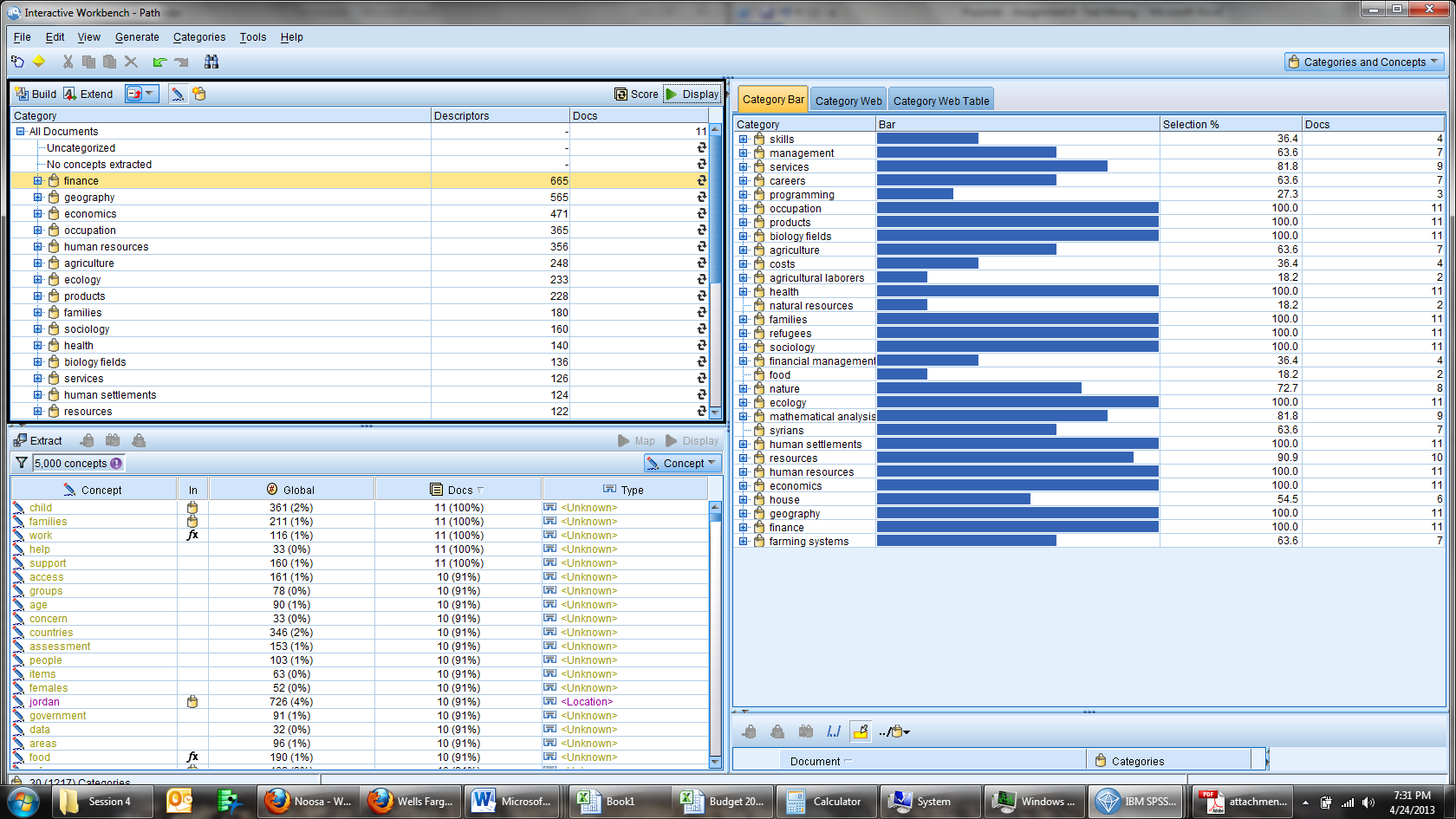
Below is the initial text analysis utilizing the general text mining settings for IBM’s SPSS. The initial step is to identify concepts where adding

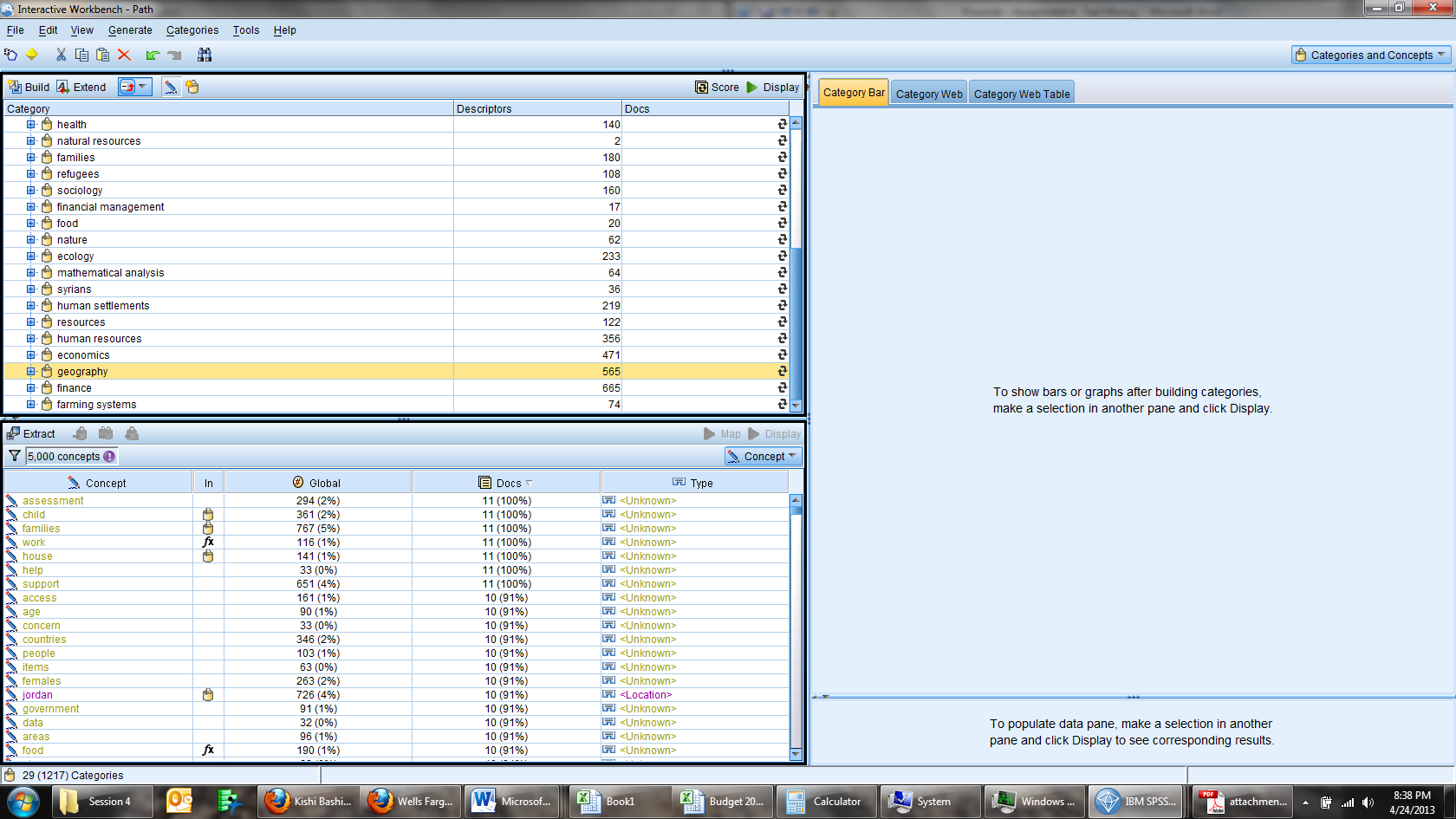
synonyms will aid the extraction process in combining similar terms appropriately. The color coded boxes correspond to terms that are close enough in meaning to be synonyms in the international development sector.

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

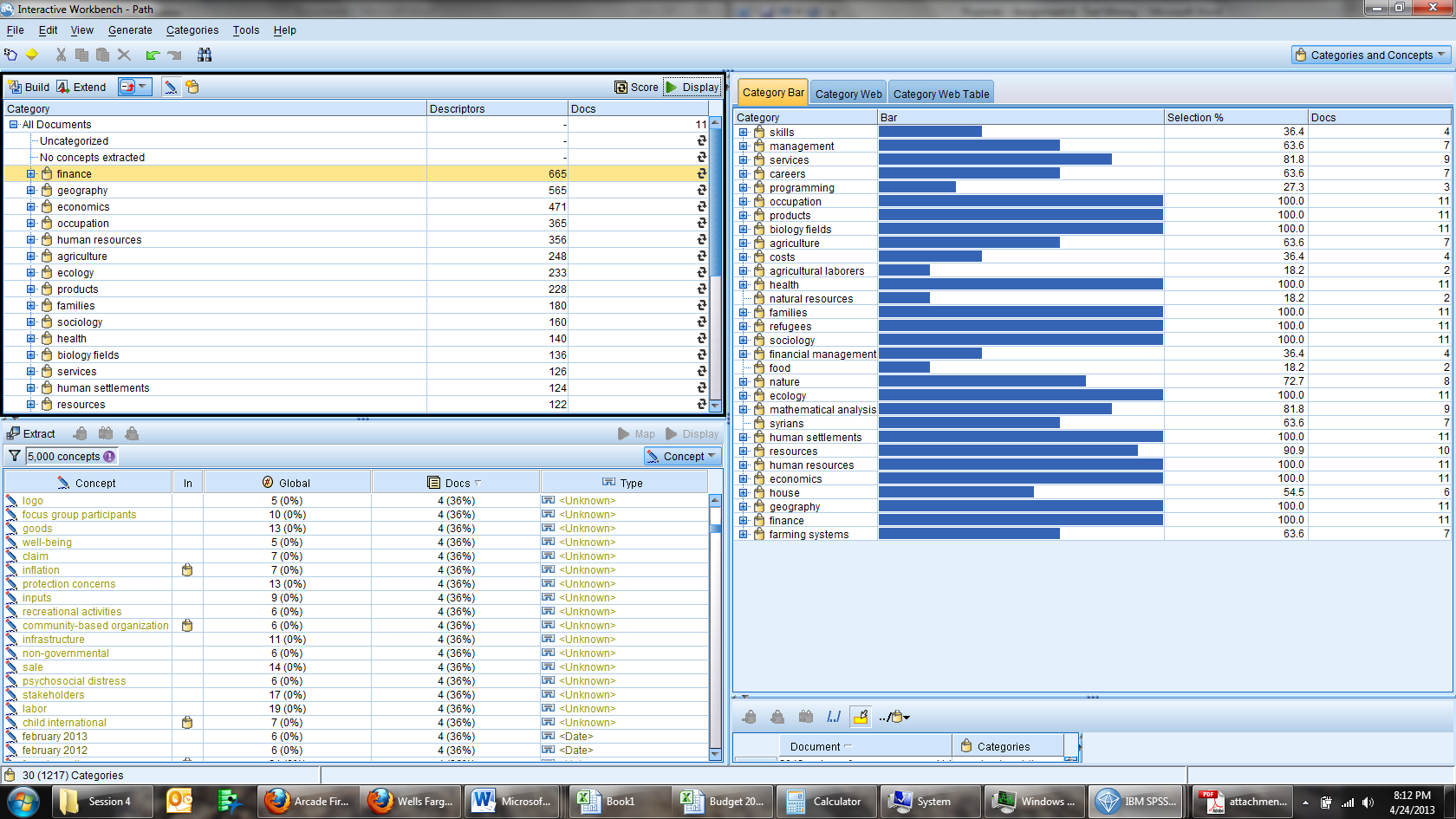
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Woman** | **Community** | **Assessment** | **Support** | **Shelter** |
| Female | Groups | Methodology | Assistance | House |
| Girls | Families | Monitoring | Humanitarian | Homs |
| Child Labor | Refugees | Questionnaire | Services | Home |
| Wife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

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



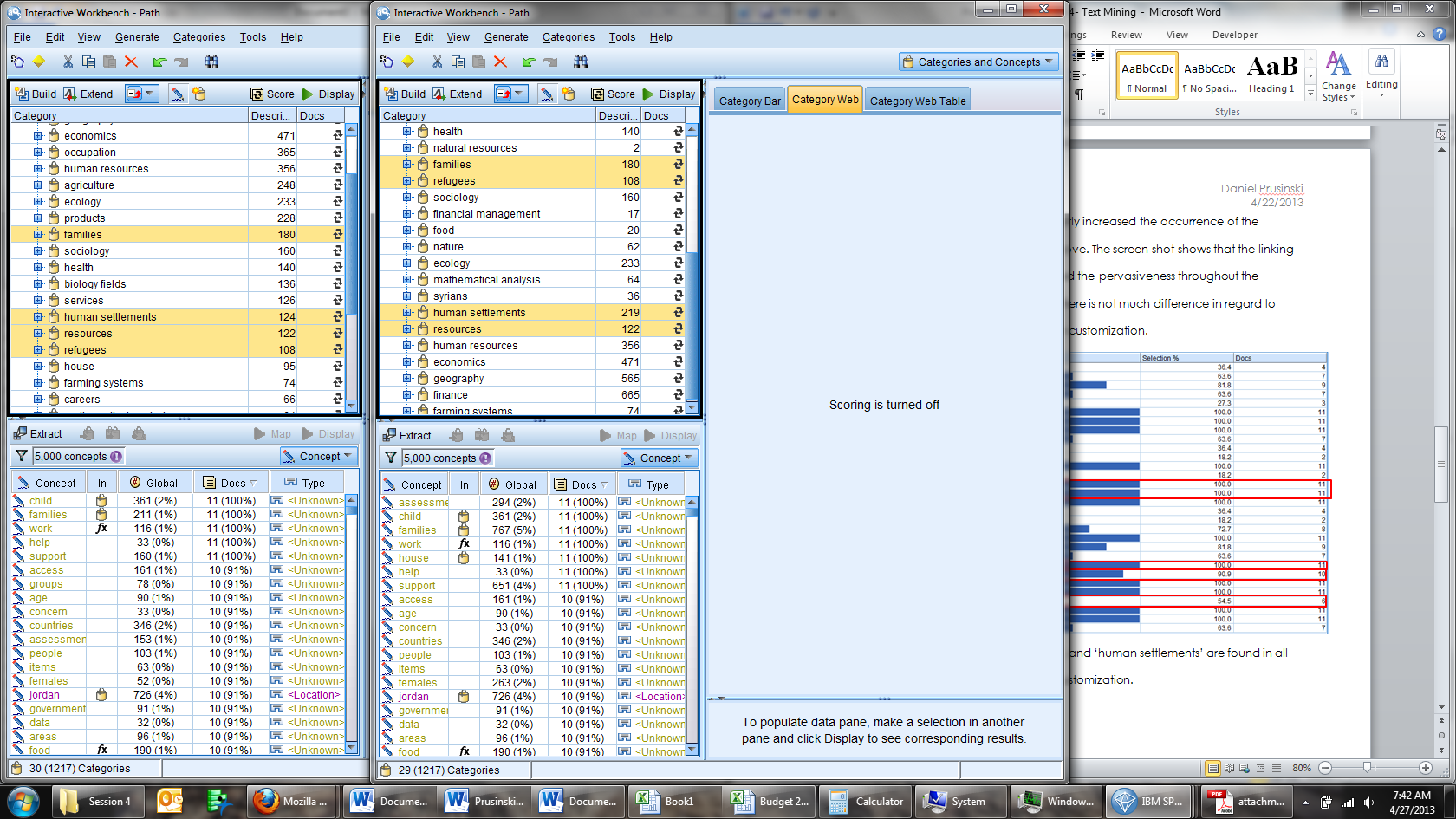
After the synonyms were added, the concepts changed as shown below: 

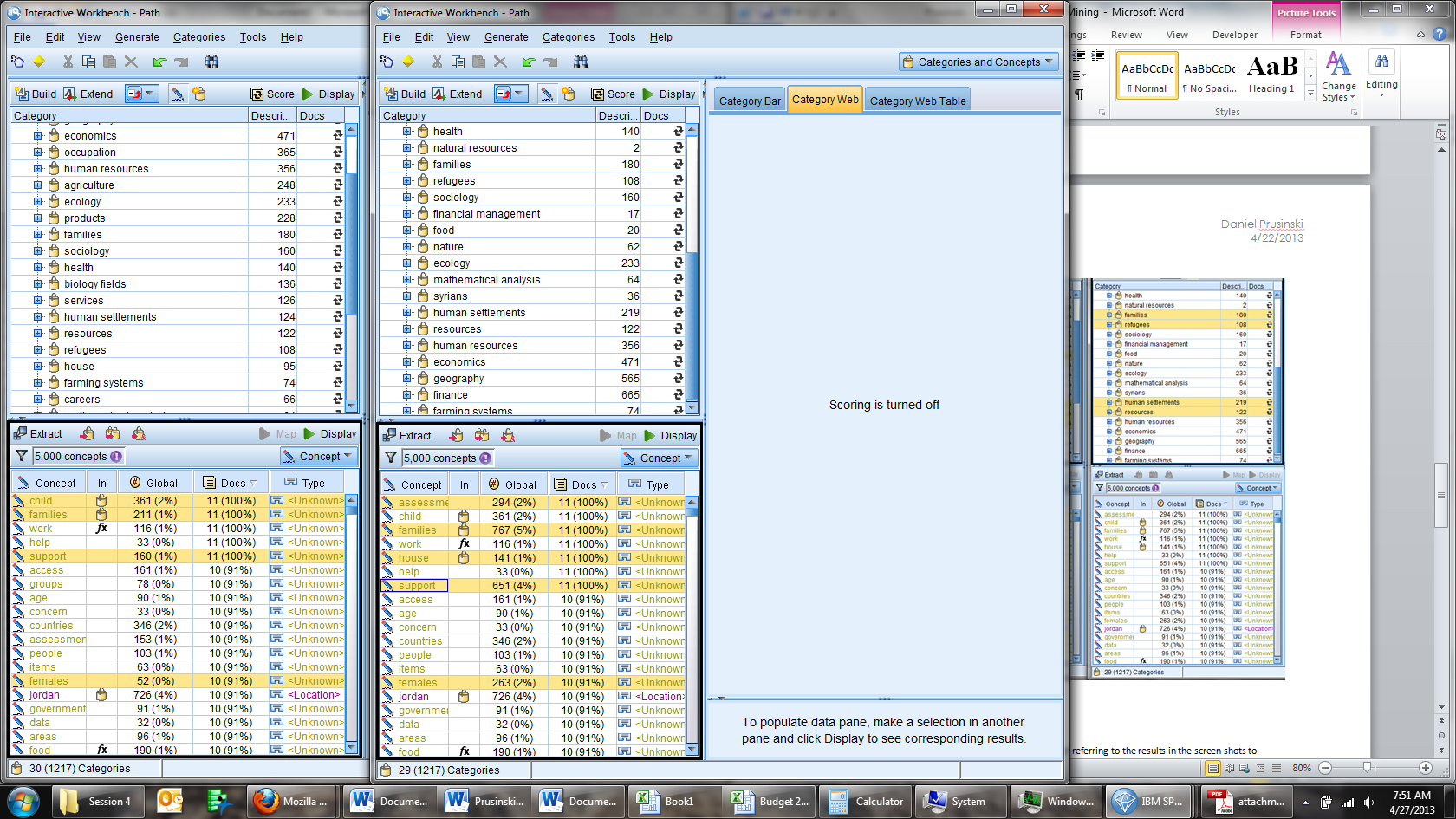
The synonyms that were added greatly increased the occurrence of the corresponding concepts as seen above. The screen shot shows that the linking of the concepts has greatly increased the pervasiveness throughout the documents. At the category level, there is not much difference in regard to added difference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



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 After Synonynm Customization





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.

