APPROACH DOCUMENT FOR INSURANCE DOMAIN ONTOLOGY

Ver1.0

12.08.18

Hi Marc,

Trust you are doing well. I made an attempt to create the domain ontology by using a semi-automatic approach to reduce the manual effort to some extent.

Below are the approach details:

Automatic Process:

1. Conversion of PDF into Text format using PyPDF2.
2. PDF contains (Double split Columns) so, PyPDF had some issue in identifying sentence boundary.
3. Converted input text has content from German language as well after conversion.
4. Tried to clean data a bit (ex: ”……..”)
5. NLTK Analyzers like: Tokenization & Parts of Speech Tagger) were used.
6. NP Chunker was used on pos tagged data.
7. Output of NP Chunkers was considered as a potential concepts/terms for ontology.
8. NP Chunker output could have been better (if PDF format was slightly better instead of double split columns and it contained data from English Language only).

NOTE: Because of datatype being very unstructured NLP analyzers had some issues during Tokenization, Parts of Speech Tagging and NP Chunking.

Manual Process:

1. Concepts/Terms were picked from the Vocabulary.
2. Use of DBpedia, Wikipedia, ConceptNet, Investopedia, Google to find out the relationship between concepts/terms. Subsequently, those terms were also added and considered during ontology creation.
3. Use of **Protégé** to create the Ontology.
4. Use of **WebVOWL1.1.2** to view the Ontology

Snapshot of Created Ontology:

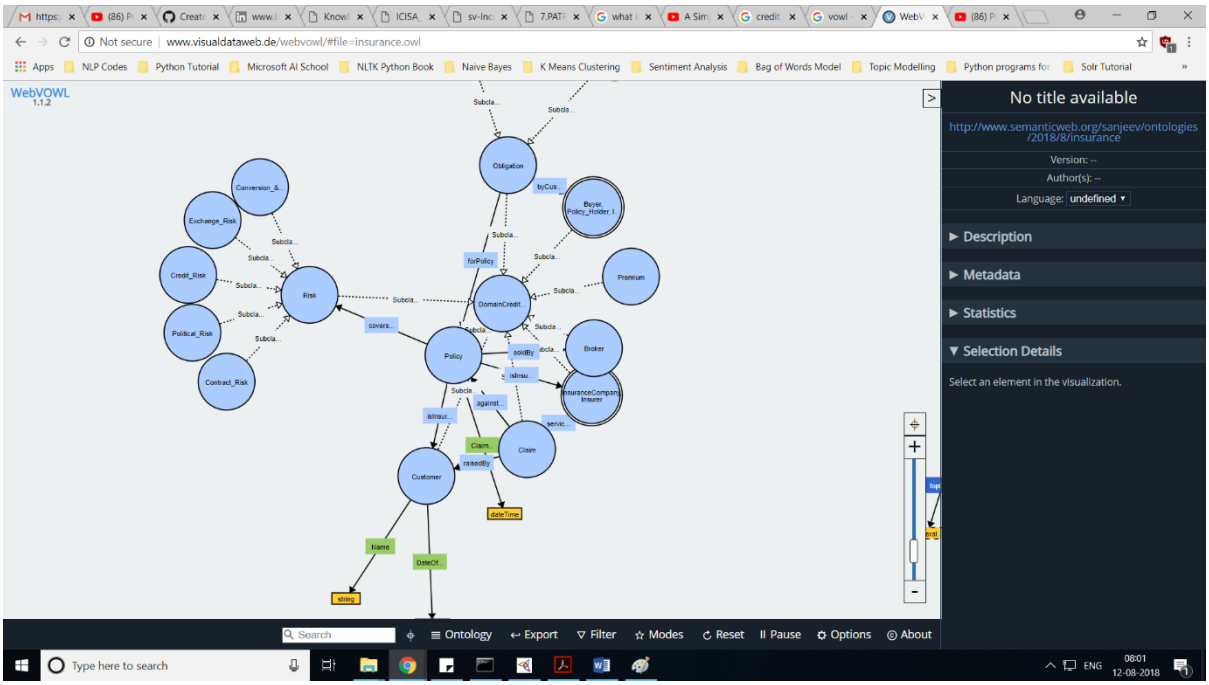

Fig: Snapshot of created ontology

References:

1. http://dbpedia.org/page/Reinsurance
2. http://conceptnet.io/
3. https://en.wikipedia.org/wiki/Outline_of_finance#Insurance
4. http://wordnetweb.princeton.edu/perl/webwn?s=Premium&sub=Search+WordNet&o2=1&o0=1&o8=1&o1=1&o7=1&o5=1&o9=&o6=1&o3=1&o4=1&h=000000
5. https://protege.stanford.edu/
6. http://vowl.visualdataweb.org/webvowl.html