APPROACH DOCUMENT FOR INSURANCE DOMAIN ONTOLOGY

12.08.18

Ver1.0

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*¹, *Wikipedia*³, *ConceptNet*², *Investopedia*⁴, *Google*⁵ to find out the relationship between concepts/terms. Subsequently, those terms were also added and considered during ontology creation.
- 3. Use of **Protégé**⁶ to create the Ontology.
- 4. Use of **WebVOWL1.1.2** ⁷to view the Ontology

Snapshot of Created Ontology:

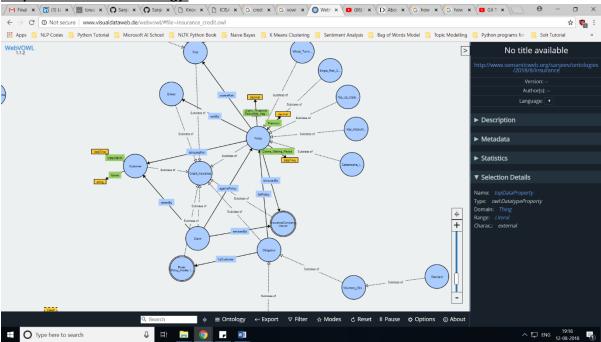


Fig: Snapshot of created ontology

References:

- 1. http://dbpedia.org/page/Reinsurance
- 2. http://conceptnet.io/
- 3. https://en.wikipedia.org/wiki/Outline of finance#Insurance
- 4. https://www.investopedia.com/terms/r/reinsurance.asp
- 5. https://www.google.com/
- 6. https://protege.stanford.edu/
- 7. http://vowl.visualdataweb.org/webvowl.html