Topic Modelling

Preprocessing of Text

History:
Tokenization benefits from link metadata removal.
Sum grams produce good Key Phrases
Gensim Phraser works, but hyper parameter control seems blackboxy and fragile

1. Lower case
2. Sentence/Context/Word/Phrase tokenization
3. PoS Tagging to extract only Noun Phrases as key terms
4. Clean known non-topic or Key Phrase entities e.g. Citations
5. Replacing known ngrams to be one term (Saffron example)
6. N-gram tokenization/ Sumgrams

Vectorize

This choice probably is tied to the actual Topic modeling Algo/solution

1. BoW - Counts and/or 1 hot or TF-IDF
2. Albert Vectors

Topic Modelling

1. Topic Modelling - Try to see if SPARK has a solution already
    a. LDA
    History:
    SPARK LDA
    Hyper param to be tuned - number of topics
        i. Inference possibilities
            1. https://stats.stackexchange.com/questions/209027/understanding-lda-inference
            2. https://towardsdatascience.com/topic-modeling-and-latent-dirichlet -allocation-in-python-9bf156893c24- Last section on the testing model on unseen document
    b. HLDA
    History:
    It produces a lot of topics and noisy Doesn't
    generate any hierarchy of topics
    c. Hierarchical Topic Models and the Nested Chinese Restaurant Process
    d. Topic Modeling: Beyond Bag-of-Words

Evaluation

1. pyLDAvis
2. https://github.com/dice-group/Palmetto
3. Topic Coherence scores etc

Clustering and TOpic Modelling

Representing the Documents

1. BoW - Split a case into a list of words
2. Vector Representations
    a. TF-IDF
    b. Embeddings
        i. Word2Vec - Combine word2vec
        ii. Doc2Vec
        iii. BERT
        iv. Any Out of the Box Document/Sentence/Word Embeddings

Clustering Techniques

1. K-Means
2. EM
3. DBSCAN
3. Mean Shift
4. Guassian Mixture Models
5. Hierarchical Clustering
6. LDA

Picking terms from Clusters

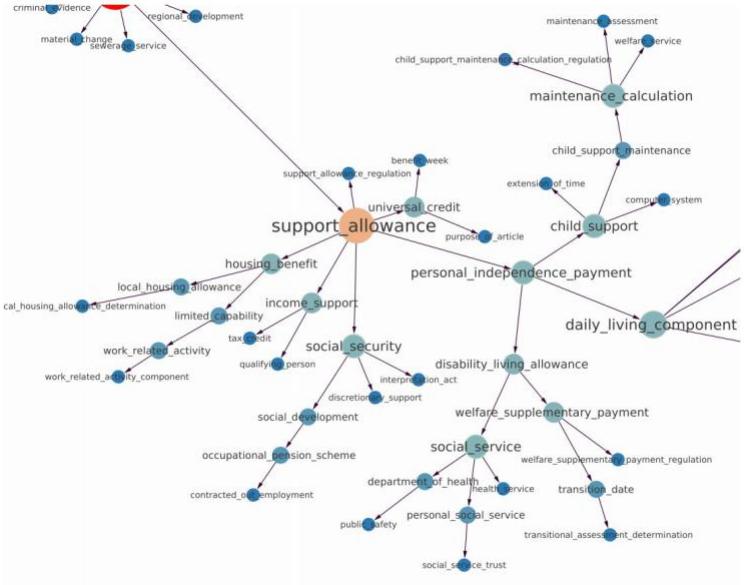
1. Pick popular ngrams from the documents in the cluster
2. Pick distinctive ngrams common to the documents in a cluster
3. Boost terms if found in a dictionary or the original Statute?

Clustering Services

1. Carrot - http://project.carrot2.org/
2. https://radimrehurek.com/gensim/
    a. https://radimrehurek.com/gensim/auto_examples/tutorials/run_lda.html#sp hx-glr-auto-examples-tutorials-run-lda-py
3. **http://brandonrose.org/clustering**
4. https://scikit-learn.org/stable/auto_examples/cluster/plot_cluster_comparison.htm l#sphx-glr-auto-examples-cluster-plot-cluster-comparison-py
5.

Topic Modelling Resources Cluster

Tasks

1. Unsupervised Topic Modelling
2. Clustering + Topic Modelling
3. Flat vs Hierarchical
4. 

5. Granularity?
6. Testing/Evaluation -
    a. Evaluate Topic Models: Latent Dirichlet Allocation (LDA)
    b. http://topicmodels.info/ckling/tmt/part4.pdf
7.

Resource List in Progress

Papers

1. https://github.com/iwangjian/Paper-Reading#topic-modeling
2. http://saffron.insight-centre.org/acl/topic/topic_model/topics/
3. http://ceur-ws.org/Vol-2143/paper7.pdf-> Saffron for hierarchical
4. http://saffron.insight-centre.org/acl/search/?query=clustering
5. https://papers.nips.cc/search/?q=topic+modelling
6. https://papers.nips.cc/paper/4291-improving-topic-coherence-with-regularized-topic-mod els
7. https://www.aclweb.org/anthology/D19-1504.pdf(?)
8. Evaluation - https://www.aclweb.org/anthology/D19-1349.pdf
9. Reinforcement - https://www.aclweb.org/anthology/D19-1350/

10.

Git Repos

1. https://github.com/baidu/Familia
2. https://github.com/lda-project/lda
3. https://github.com/bmabey/pyLDAvis-> Simon probably tried this for hack day
4. https://github.com/xiaohuiyan/BTM- BiTerm Bigrams
5. https://github.com/meereeum/lda2vec-tf
6. https://github.com/larsmaaloee/deep-belief-nets-for-topic-modeling
7. https://github.com/vi3k6i5/GuidedLDA
8. https://github.com/qiang2100/STTM- Short Text Topic Modelling

OOTB Public Code

1. http://mallet.cs.umass.edu/topics.php
2. https://nlp.stanford.edu/software/tmt/tmt-0.4/
3. https://docs.aws.amazon.com/comprehend/latest/dg/topic-modeling.html
4. Google Entity Analysis