Automatic Highlighter of Lengthy Legal Documents
Yanshu Hong, Tian Zhao

Motivation
- Legal Documents are lengthy.
- Reading through each sentence is painful.
- Can we automatically highlight some sentences that require our special attention?

Example: ITUNES user agreement
By using the Services, you acknowledge and agree that the Application Provider is not responsible for examining or evaluating the content, accuracy, completeness, functionality, validity, copyright compliance, legality, decency, or any other aspect of such Third Party Materials or any Other Party. The Application Provider does not warrant or endorse and does not assume and shall not have any liability or responsibility to You for any Third Party Materials or any Other Party. The Application Provider is not responsible for any content, errors or omissions in any Third Party Materials or any Other Party. Third Party Materials are solely to enable you to access them. You agree to indemnify and hold the Application Provider harmless from and against any claim or demand made by any third party or any other party due to or arising from your use of the Application Provider’s Services.

Preprocessing: LDA Model
How to remove common topics words from a given text?
- Build a database of similar documents. Find the common topic words across the database.

Latent Dirichlet Allocation (LDA):
- A word $w$ is the basic unit.
- A document is a sequence of $N$ words.
- A corpus contains $M$ documents.

Generative Process:
- Choose $N \sim$ Poisson($\xi$).
- Choose $\theta \sim$ Dir($\alpha$).
- For each $N$ words, choose a topic $z_n \sim$ Multinomial($\theta$).
- For each $N$ words, choose a word $w_n$ from $p(w_n | z_n, \beta)$.

Example Topic:
Trademark + Services + May + Use + Application + Content + Will + Terms

Feature Extraction: Word2Vec
How to do arithmetic on words and sentences?
- Each word needs to be represented as a vector.
- Each sentence is assumed to be sum of containing word vectors.

Example of a Word Vector:
apple = [0.68, 0.67, 0.66, ...]

Example of Word Similarity:
P(apple, desktop) = 0.684
P(apple, laptop) = 0.604
P(apple, hardware) = 0.633

Clustering: Agglomerative Clustering
How to find non-standard sentences that need our attention?
- Intuition: Cluster sentences with similar meanings; find those farthest from the centroid.
- Challenge: K-means gives unstable results with randomized initial centers.
- Solution: Agglomerative Clustering.

Anomaly Detection Using LOF Model
How to find non-standard sentences that need our attention?
- Intuition: Sentences located with the fewest neighbors are more likely to be anomalies.
- Solution: Find sentences with the largest Local Outlier Factor.

Intuition of LOF:
Result (Plotted first 3D after t-SNE):
LOF compares the local densities of a point with its neighbors.
A has a lower density, and is more likely to be an anomaly.