Embedding Methods on SCOTUS Cases

By:
Ethan Koch,
University of North Carolina at Chapel Hill
Overview

- SCOTUS data set
- Word2Vec
- Doc2Vec
- Node2Vec
- Stochastic Block Model
- SCOTUS Clustering Results
- SCOTUS Classification Results
- Phase Transition Results
Summary

- 27,885 cases (nodes)
- 234,209 citations (edges)
- Majority opinion text for each case (node)

Issue Areas

- Hand-coded issue area (14 in total) for each SCOTUS case

Examples:
- Criminal Procedure
- First Amendment
- Privacy
Distributional Hypothesis

Words that appear in similar contexts are similar in meaning
Word2Vec

“Happy families are all alike; every unhappy family is unhappy in its own way” -- Anna Karenina, Leo Tolstoy

Skip - Gram (SG)

(families, are),
(families, Happy),
(unhappy, family)

Continuous Bag of Words (CBOW)

[(Happy, are), families],
[(all, every), alike],
[(every, family), unhappy]
Doc2Vec

“Happy families are all alike; every unhappy family is unhappy in its own way” -- Anna Karenina, Leo Tolstoy

Each clause = document, with ID = sentence index

Distributed Bag of Words (PV-DBOW)

Distributed Memory (PV-DM)

(0, families),
(1, unhappy),
(1, family)
Node2Vec

Random Walks

(Nashville, Birmingham, Atlanta),
(Charlotte, Atlanta, Charlotte),
(Atlanta, Birmingham, Nashville)
Stochastic Block Model

\[
\begin{array}{cc}
0.8 & 0.8 \\
0.8 & 0.8 \\
\end{array}
\], \ n = 4

\[
\begin{array}{cc}
0.01 & 0.9 \\
0.9 & 0.01 \\
\end{array}
\], \ n = 6
Doc2Vec Similarity Scores

Time vs. Similarity for ID: 2539855

- Case of Interest
- Median Yearly Similarity
- Cosine Similarity
Doc2Vec Similarity Scores

Time vs. Similarity for ID: 2645639

- Case of Interest
- Median Yearly Similarity
- Cosine Similarity

Similarity (between 0 and 1)

Time (in years)
KMeans Clustering as N → 500

KMeans Clustering Results

- V-Measure
- Homogeneity
- Normalized Mutual Information
- Completeness
- Mean of Scores

Score

Number of Clusters
Agglomerative Clustering as $N \rightarrow 500$
Issue Areas

KMeans Clustering vs. Issue Areas

<table>
<thead>
<tr>
<th>Metric</th>
<th>Doc2Vec Score</th>
<th>Node2Vec Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>V-Measure</td>
<td>0.676</td>
<td></td>
</tr>
<tr>
<td>Homogeneity</td>
<td>0.591</td>
<td></td>
</tr>
<tr>
<td>Normalized Mutual Info</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scoring Metric</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Completeness</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Both Concatenated</td>
<td>0.701</td>
<td></td>
</tr>
</tbody>
</table>
Phase Transition: Fixed Walk Length

Explore Phase Change: Walk Length = 50

Scores

Epsilon (c_out/c_in)

- median agreement scores
- Proportion of Successful Iterations
- Undetectable Threshold
- Unrecoverable Threshold
- Empirical Threshold
- raw agreement scores
Phase Transition: Increase Walk Length
Future Work

◎ Work with law expert to understand jump at 2006
◎ Try to understand what information is held by both Node2Vec & Doc2Vec embeddings
◎ See if results hold in other settings, e.g. academic articles
◎ Further investigate phase transition, especially for other parameters