## Methodology

### Textual Preprocessing
- Tokenization, filtering, word segment, etc.
- Conflation, data are randomly shuffled

### Concept Extraction
- Sequence labeling problem, each word $\in \{B, I, O\}$
- Course- and instructor-agnostic features: stylistic, structural, contextual, semantic and dictionary features
- CRF + semi-supervised machine learning framework
- Self-training, to reduce the cost of labeling

### Definition of Node and Edge
- Node weight for **Fundamentality**: term frequency
- Node weight for **Importance**: term frequency and inverted document frequency
- Edge weight: semantic similarity (cosine distance of concept word vectors)

### Learning Path Generation
- For each concept, iteratively choose top $k$ most semantically similar concepts
- Regard the most **fundamental**/important ones as successors
- Consider locational order of appearance in Lecture Notes

## Experiments

### Data Sets
Teaching materials of an interdisciplinary course conducted on Coursera, including lecture notes, PPTs and questions

<table>
<thead>
<tr>
<th>Source</th>
<th># Sentence</th>
<th># Word</th>
<th># Terminology</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lecture Notes</td>
<td>3,036</td>
<td>69,437</td>
<td>402</td>
</tr>
<tr>
<td>PPTs</td>
<td>2,823</td>
<td>22,334</td>
<td>249</td>
</tr>
<tr>
<td>Questions</td>
<td>268</td>
<td>7,138</td>
<td>95</td>
</tr>
</tbody>
</table>

### Baselines to Extract Concepts
- Term Frequency (TF), a statistic baseline
- Bootstrapping (BT), a rule-based iterative algorithm
- Supervised Concept-CRF (SC-CRF), a supervised CRF
- Semi-Supervised Concept-CRF (SSC-CRF), semi-supervised version of SC-CRF

### Results

<table>
<thead>
<tr>
<th>Precision</th>
<th>Recall</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF@500</td>
<td>0.402</td>
<td>0.500</td>
</tr>
<tr>
<td>TF@1000</td>
<td>0.600</td>
<td>0.746</td>
</tr>
<tr>
<td>BT</td>
<td>0.099</td>
<td>0.627</td>
</tr>
<tr>
<td>SC-CRF</td>
<td>0.890</td>
<td>0.842</td>
</tr>
<tr>
<td>SSC-CRF</td>
<td>0.875</td>
<td>0.783</td>
</tr>
</tbody>
</table>

Figure 2: Comparison between SC-CRF and SSC-CRF

### Two demo learning paths:
- (a) **Node** → **Edge** → **Element** → **Set** → **Alternative** → **Vote**
- (b) **PageRank** → **PageRankAlgorithm** → **SmallWorld** → **Balance** → **NashBalance**

## Conclusion
- Technical and pedagogical feasibility of the novel SCM
- A universal two-phase approach to re-organize MOOC materials with a good efficacy
- The SCMs and learning paths generated can be manually revised up to requirements
- To be scaled to more across-domain courses in the future