NERank: Ranking Named Entities in Document Collections
Chengyu Wang, Rong Zhang, Xiaofeng He, Aoying Zhou
Institute for Data Science and Engineering
East China Normal University
Shanghai, 200062, China

Introduction
- Named entity ranking is necessary to bring semantics to plain documents.
- Rank order of named entities should be determined by the relative importance considering the document collection.
- NERank is the first attempt to tackle the problem of named entity ranking directly from documents.

NERank Workflow
- Tripartite Graph Construction
  - The part aims to model the semantic relations between entities and documents indirectly by topic modeling. A weighted, tripartite graph is employed to represent < document, topic, entity > relations.
  - Prior Topic Rank Estimation
    - The part is responsible for assigning prior ranks to each topic in the tripartite graph.
  - Random Walk Process
    - This part is designed to propagate prior topic ranks to documents and entities through a random walk process on the tripartite graph.

Tripartite Graph Construction
- Named Entity Recognition and Normalization
  - Given a document collection D, perform NER and NEN to generate the entity set M and map each m ∈ M to the normalized form e ∈ E.
  - Entity-Aware Topic Modeling
    - Model a document in D as the union set of common words and normalized named entities in E.
    - Estimate document-topic distribution Θ and topic-word distribution Φ by Gibbs sampling in LDA.
  - Graph Construction
    - Nodes: documents D, topics T and entities E.
    - Edges: assign weights of < document, topic > and < topic, entity > edges by respective document-topic and topic-word probabilities.

Prior Topic Rank Estimation
- Estimate the prior rank for topic t ∈ T: ρ(t).
- Three quality metrics:
  - Prior probability: the probability that topic t is discussed in D
    \[ p(t) = \frac{1}{|D|} \sum_{d \in D} \theta_{d,t} \]
  - Entity richness: the proportion of entities in words related to topic t
    \[ e(t) = \frac{1}{|E|} \sum_{j=1}^{E} \phi_{j,t} \]
  - Topic specificity: whether the topic is specific about certain aspects or only provides background information
    \[ s(t) = \frac{1}{|Z_{ts}|} \sum_{j=1}^{Z_{ts}} \theta_{j,t} \log \theta_{j,t} \]

Random Walk Process
- Select a topic t ∈ T with probability \( \rho(t) \) as the starting point.
- Make one of the following three transfers iteratively until the system reaches equilibrium (α and β are parameters where \( \alpha > 0, \beta > 0 \) and \( \alpha + \beta < 1 \)):
  - With probability \( \alpha \), the random surfer walks through the path \( t \rightarrow d_j \rightarrow t \). \( d_j \in D \) is selected with probability \( w_j \sum_{d \in D \cap E} \theta_{d,t} \).
  - With probability \( \beta \), the random surfer walks through the path \( t \rightarrow e_j \rightarrow t \). \( e_j \in E \) is selected with probability \( w_t \sum_{e \in E} \phi_{j,e} \).
- Compute the rank of entity \( e \) by random surfer.

Experiments
- Datasets: Newswire collections where each collection is related to a major international event.
- Metrics: Average Precision@K and MAP (with paired t-test).
- Methods: TF-IDF, TextRank, NERank\(_{in}\) (which assigns prior topic ranks uniformly), NERank\(_{avg}\) (which sets \( \alpha = 0 \) in random walk process) and NERank\(_{full}\) (proposed approach).
- Results: NERank\(_{full}\) outperforms all the baselines.

<table>
<thead>
<tr>
<th>Method</th>
<th>AvgP@5</th>
<th>AvgP@10</th>
<th>AvgP@15</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>TF-IDF</td>
<td>0.85</td>
<td>0.79</td>
<td>0.73</td>
<td>0.81</td>
</tr>
<tr>
<td>TextRank</td>
<td>0.87</td>
<td>0.83</td>
<td>0.73</td>
<td>0.83</td>
</tr>
<tr>
<td>NERank(_{in})</td>
<td>0.80</td>
<td>0.75</td>
<td>0.71</td>
<td>0.78</td>
</tr>
<tr>
<td>NERank(_{avg})</td>
<td>0.72</td>
<td>0.61</td>
<td>0.51</td>
<td>0.62</td>
</tr>
<tr>
<td>NERank(_{full})</td>
<td>0.92</td>
<td>0.87</td>
<td>0.79</td>
<td>0.89</td>
</tr>
</tbody>
</table>

Conclusion and Future Work
- NERank is an effective method to rank named entities in documents with little human intervention.
- Future work includes:
  - A general framework for entity ranking from different types of texts (i.e., documents, tweets, etc.).
  - A complete benchmark for evaluating entity ranking.

References