Vector representation with a finance corpus

Student: Ritesh Bachhar
University of Rhode Island
riteshbachhar@uri.edu

Mentor: Murat Aydogdu
Rhode Island College
maydogdu@ric.edu

Date: June 8, 2022
Vector representation with a finance corpus

- NLP & Word embedding
- Glove implementation of words
- Co-occurrence matrix

T: frog
1. frogs
2. toad
3. litoria
4. etc

(https://nlp.stanford.edu/projects/glove/)
Vector representation with a finance corpus

- Impact of domain-specific representation of word vector
- Wikipedia articles (5.8 mil)
- Finance documents - 10-K filings by publicly traded companies
- e.g. Interest (Distinct meaning in general and finance context)
Goals

- Vector representation; Glove
- Small scale test of existing code
- Large scale representation of word vector using two corpora
- Setting up workflow on Unity
### Timeframe

<table>
<thead>
<tr>
<th>Month</th>
<th>Activities</th>
</tr>
</thead>
<tbody>
<tr>
<td>May</td>
<td>Glove; Initial setup</td>
</tr>
<tr>
<td>June</td>
<td>Tokenizer; Testing code; Resource management</td>
</tr>
<tr>
<td>July</td>
<td>Run Glove on large scale dataset; streamline workflow</td>
</tr>
</tbody>
</table>
Vector representation with a finance corpus

• What I hope to learn
  – NLP & vector representation
  – Bash scripting, slurms
  – Large scale computing with HPC
  – Parallel computing
Vector representation with a finance corpus

- Help needed
  - Resource management (partitions, nodes, ntasks, cores etc.)
  - Shared directory
  - Parallelization