Job recommendations leveraging Deep Learning using Analytics Zoo on Apache Spark and BigDL
Speakers

Wenjing Zhan
Data Scientist, Talroo

Guoqiong Song
Software Engineer, Intel

Jacob Eisinger
Director of Data, Talroo
Agenda

- Talroo Overview
- Job Search Challenges
- Resume Search Opportunity
- Analytics Zoo and BigDL Overview
- Resume Search Analytics Zoo Solution
- Lessons Learned
Agenda

- Talroo Overview
- Job Search Challenges
- Resume Search Opportunity
- Analytics Zoo and BigDL Overview
- Resume Search Analytics Zoo Solution
- Lessons Learned
Where Talent and Recruiting Intersect

- Data-driven job ad network
- Self-funded growth
- Collaborative and social culture
- Austin, TX headquarters
Talroo reaches active candidates where they are searching.
Talroo turns up what works, and turns down what doesn't, ultimately reducing cost-per-hire.
We power search.
Talroo Monthly Scale

- Billions of Queries
- Billions of Jobs Events
- Tens of Billions Impressions
- Tens of Millions of Interactions
Agenda

- Talroo Overview
- Job Search Challenges
- Resume Search Opportunity
- Analytics Zoo and BigDL Overview
- Resume Search Analytics Zoo Solution
- Lessons Learned
Traditional Information Retrieval Sufferings

Solution challenges: stemming, synonyms, ontologies, sensitivity
Accountant ≠!≠ Accounting
<table>
<thead>
<tr>
<th>Accountant</th>
<th>!=</th>
<th>Accounting</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accountant</td>
<td>=</td>
<td>Accounting</td>
</tr>
</tbody>
</table>
Stemming Sufferings

Accountant ≠!≠ Accounting

Accountant = Accounting

Accountant = Accounting = Account Representative
Registered Nurse

RN
## Synonyms Solution

<table>
<thead>
<tr>
<th>Registered Nurse</th>
<th>(=!) \quad \text{RN}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Registered Nurse</td>
<td>(=) \quad \text{RN} \rightarrow \text{registered nurse}</td>
</tr>
<tr>
<td>Registered Nurse</td>
<td>RN</td>
</tr>
<tr>
<td>------------------</td>
<td>----</td>
</tr>
<tr>
<td>Registered Nurse</td>
<td>=</td>
</tr>
</tbody>
</table>
Dishwasher  !=!=  Back of House
| Dishwasher | =!|= | Back of House |
|------------|-----|----------------|
| Restaurant | =   | Restaurant     |
### ontologies sufferings

<table>
<thead>
<tr>
<th>Dishwasher</th>
<th>=!=</th>
<th>Back of House</th>
</tr>
</thead>
<tbody>
<tr>
<td>Restaurant</td>
<td>=</td>
<td>Restaurant</td>
</tr>
</tbody>
</table>

...
Personalize Results Value

**Job Seekers**
Find the right job faster
Resume

**Employers**
Find the right person
Job description
Agenda

- Talroo Overview
- Job Search Challenges
- Resume Search Opportunity
- Analytics Zoo and BigDL Overview
- Resume Search Analytics Zoo Solution
- Lessons Learned
• **What information** is logged?

• **What do we do** with the logged information?

• **What did we find** from logged information?

• **How** to further tackle the findings?
What Information Do We Log

1. Lead Warehouse Associate
2. Warehouse Workers Needed Immediately ($17-$38/Hour)
3. B2B Outbound Sales VI
4. Telesales Specialist Humana
5. Mail & Package Handler
Replaying the Job Application Process

Application-log
Click-log
Resume, Job, Click
Impression log
Validation
Job1
Job2
Job3
Job4
Job5
Apply
Apply
Real World Example

1. Lead Warehouse Associate
2. Warehouse Workers Needed Immediately ($17-$38/Hour)
3. B2B Outbound Sales VI
4. Telesales Specialist Humana
5. Mail & Package Handler
Real World Example

1. Lead Warehouse Associate
2. Warehouse Workers Needed Immediately ($17-$38/Hour)
3. B2B Outbound Sales VI
4. Telesales Specialist Humana
5. Mail & Package Handler
Bilingual International Sales Executive
09/2007 - 09/2017
Store Manager
03/2001 - 08/2007
International Sales Executive

1. Lead Warehouse Associate
2. Warehouse Workers Needed Immediately ($17-$38/Hour)
3. B2B Outbound Sales VI
4. Telesales Specialist Humana
5. Mail & Package Handler
Real World Example

Bilingual International Sales Executive
09/2007 - 09/2017
Store Manager
03/2001 - 08/2007
International Sales Executive

1. Lead Warehouse Associate
2. Warehouse Workers Needed Immediately ($17-$38/Hour)
3. B2B Outbound Sales VI
4. Telesales Specialist Humana
5. Mail & Package Handler
Investigation Of The Findings
-- First Attempt

Bilingual International Sales Executive
09/2007 - 09/2017
Store Manager
03/2001 - 08/2007
International Sales Executive

3. B2B Outbound Sales VI

Matching Degree

- Industry

- intersect_count

- 0: 54%
- 1: 38%
- 2: 7%
- 3: 0%
• Handle different terminology

• Identify and respect multiple industries

• Recognize other topics
Ideal Resume Search Requirements

- AWS EC2
- Amazon Kinesis
- S3

Diagram showing the integration of AWS EC2, Amazon Kinesis, and S3.
Scoring Before

API Request

Job Pool

job 1
job 5
job 3
job 4
job 6
job 2

Apache Solr

job 1
job 2
job 3
job 4
job 5
job 6
Scoring After - Reranker Solution

API Request

Job Pool
- job 1
- job 2
- job 3
- job 4
- job 5
- job 6

Solution

API

Reranker

Request

Job Pool
- job 1
- job 2
- job 3
- job 4
- job 5
- job 6

Solution

API

Reranker

Request

Job Pool
- job 1
- job 2
- job 3
- job 4
- job 5
- job 6

Solution

API

Reranker

Request

Job Pool
- job 1
- job 2
- job 3
- job 4
- job 5
- job 6

Solution

API

Reranker

Request

Job Pool
- job 1
- job 2
- job 3
- job 4
- job 5
- job 6

Solution

API

Reranker

Request

Job Pool
- job 1
- job 2
- job 3
- job 4
- job 5
- job 6

Solution

API

Reranker

Request

Job Pool
- job 1
- job 2
- job 3
- job 4
- job 5
- job 6

Solution

API

Reranker

Request

Job Pool
- job 1
- job 2
- job 3
- job 4
- job 5
- job 6

Solution

API

Reranker

Request

Job Pool
- job 1
- job 2
- job 3
- job 4
- job 5
- job 6

Solution

API

Reranker

Request

Job Pool
- job 1
- job 2
- job 3
- job 4
- job 5
- job 6

Solution

API

Reranker

Request

Job Pool
- job 1
- job 2
- job 3
- job 4
- job 5
- job 6

Solution

API

Reranker

Request

Job Pool
- job 1
- job 2
- job 3
- job 4
- job 5
- job 6

Solution

API

Reranker

Request

Job Pool
- job 1
- job 2
- job 3
- job 4
- job 5
- job 6

Solution

API

Reranker

Request

Job Pool
- job 1
- job 2
- job 3
- job 4
- job 5
- job 6

Solution

API

Reranker

Request

Job Pool
- job 1
- job 2
- job 3
- job 4
- job 5
- job 6

Solution

API

Reranker

Request

Job Pool
- job 1
- job 2
- job 3
- job 4
- job 5
- job 6

Solution

API

Reranker

Request

Job Pool
- job 1
- job 2
- job 3
- job 4
- job 5
- job 6

Solution

API

Reranker

Request

Job Pool
- job 1
- job 2
- job 3
- job 4
- job 5
- job 6

Solution

API

Reranker

Request

Job Pool
- job 1
- job 2
- job 3
- job 4
- job 5
- job 6

Solution

API

Reranker

Request

Job Pool
- job 1
- job 2
- job 3
- job 4
- job 5
- job 6

Solution

API

Reranker

Request
Evaluation Metric – Gold Standard


CTR = \frac{\text{clicks}}{\text{impressions}}
Evaluation Metric - Quantity

CTR = \frac{\text{clicks}}{\text{impressions}}

Online

Precision = \frac{tp}{tp + fp}

Offline

**Evaluation Metric - Quality**

**Online**

$$CTR = \frac{\text{clicks}}{\text{impressions}}$$

**Offline**

$$\text{Precision} = \frac{\text{tp}}{\text{tp} + \text{fp}}$$

$$\text{MRR} = \frac{1}{Q} \sum_{i=1}^{Q} \frac{1}{\text{rank}_i}$$

---

[Quality](https://en.wikipedia.org/wiki/Evaluation_measures_(information_retrieval))

- **CTR**
  - Clicks / Impressions
- **Precision**
  - True Positives / (True Positives + False Positives)
- **MRR**
  - Mean Reciprocal Rank
Evaluation Metric - Cutoffs

Cutoff:
- 20%
- 40%
- 60%
- 80%
- 100%

Evaluation Metrics

- mrr
- precision
Resume Search Solution

AWS EC2 → Amazon Kinesis → S3 → BigDL

41
Agenda

- Talroo Overview
- Job Search Challenges
- Resume Search Opportunity
- Analytics Zoo and BigDL Overview
- Resume Search Analytics Zoo Solution
- Lessons Learned
BigDL is an open-source distributed deep learning library for Apache Spark* that can run directly on top of existing Spark or Apache Hadoop* clusters.

Ideal for DL Models TRAINING and INFERENCEx

High Performance Deep Learning for Apache Spark* on CPU Infrastructure

No need to deploy costly accelerators, duplicate data, or suffer through scaling headaches!

- Design and Optimized for Intel® Xeon®
- Feature Parity & Model Exchange with TensorFlow*, Caffe*, Keras, Torch*
- Lower TCO and improved ease of use with existing infrastructure
- Deep Learning on Big Data Platform, Enabling Efficient Scale-Out

Powered by Intel® MKL and multi-threaded programming

https://github.com/intel-analytics/BigDL  
https://bigdl-project.github.io/
# Build and Productionize Deep Learning Apps for Big Data at Scale

<table>
<thead>
<tr>
<th>Reference Use Cases</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>• Anomaly detection</td>
<td></td>
</tr>
<tr>
<td>• Sentiment analysis</td>
<td></td>
</tr>
<tr>
<td>• Fraud detection</td>
<td></td>
</tr>
<tr>
<td>• Chatbot, sequence prediction, etc.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Built-In Deep Learning Models</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>• Image classification</td>
<td></td>
</tr>
<tr>
<td>• Object detection</td>
<td></td>
</tr>
<tr>
<td>• Text classification</td>
<td></td>
</tr>
<tr>
<td>• Recommendations</td>
<td></td>
</tr>
<tr>
<td>• Sequence-to-sequence, GAN, etc.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Feature Engineering</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Feature transformations for</td>
<td></td>
</tr>
<tr>
<td>• Image, text, 3D imaging, time series, speech, etc.</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>High-Level Pipeline APIs</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>• Native deep learning support in Spark DataFrames and ML Pipelines</td>
<td></td>
</tr>
<tr>
<td>• Autograd, Keras and transfer learning APIs for model definition</td>
<td></td>
</tr>
<tr>
<td>• Support for model serving/inference pipelines</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Backends</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Spark, BigDL, TensorFlow, etc.</td>
<td></td>
</tr>
</tbody>
</table>

[https://github.com/intel-analytics/analytics-zoo/](https://github.com/intel-analytics/analytics-zoo/)

[https://analytics-zoo.github.io/](https://analytics-zoo.github.io/)
Agenda

- Talroo Overview
- Job Search Challenges
- Resume Search Opportunity
- Analytics Zoo and BigDL Overview
- Resume Search Analytics Zoo Solution
- Lessons Learned
Analytics Zoo solution

- Retrieve document vectors (resume and job) using Glove vectors
- Analytics Zoo recommender model
- End-to-End flow (offline training)
- Evaluation Results
• Words or phrases from the vocabulary are mapped to vectors of real numbers.

• **Global log-bilinear regression model for the unsupervised learning algorithm.**

• Training is performed on aggregated global word-word co-occurrence statistics from a Wikipedia.

• Vector representations showcase meaningful linear substructures of the word vector space.

https://nlp.stanford.edu/projects/glove/
• Neural collaborative filtering, Wide and Deep
• Answer the question using classification methodologies
• Implicit feedback and explicit feedback
• APIs
  • recommendForUser
  • recommendForItem
  • predictUserItemPair

He, 2015
Recommender model

Resume Glove vectors

Job Glove vectors

LogSoftMax

Linear4(2 output)

Linear3(10 output)

Linear2(20 output)

Linear1(40 output)
End to End Flow

Raw data → Feature Extraction → Kmeans

- Cluster 1
- Cluster 2
- Cluster 3

Resume Glove vectors → Linear1(40 output) → Linear2(20 output) → Linear3(10 output) → LogSoftMax → Predict Recommend

Job Glove vectors
Evaluation Results

Precision

MRR
Agenda

- Talroo Overview
- Job Search Challenges
- Resume Search Opportunity
- Analytics Zoo and BigDL Overview
- Resume Search Analytics Zoo Solution
- Lessons Learned
Lessons learned

• Analytics Zoo/BigDL integrates well into existing AWS Databricks Spark ETL and machine learning platform

• Analytics Zoo/BigDL scales with our data and business

• Jobs and resumes can be effectively modeled and processed through embeddings

• Ensembling multiple models and glove embedding feature embedding proved to be very effective for rich content

• More information available at https://analytics-zoo.github.io/
Intel technologies’ features and benefits depend on system configuration and may require enabled hardware, software or service activation. Learn more at intel.com, or from the OEM or retailer.

No computer system can be absolutely secure.

Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase. For more complete information about performance and benchmark results, visit http://www.intel.com/performance.
Questions?

Wenjing Zhan
Data Scientist, Talroo
wenjing@talroo.com

Guoqiong Song
Software Engineer, Intel
guoqiong.song@intel.com

Jacob Eisinger
Director of Data, Talroo
jacob@talroo.com