Motivation

- Over 70% of small business rely on social media for Consumer-to-Consumer (C2C) business opportunities.
- Sellers post product offers & buyers post their needs.
- These messages get hidden among so many others posts.
- Both buyers & sellers could benefit if such relevant messages can be detected and matched as they get posted.

Social Media Messages

- Buying & selling related tweets for training and testing, size of 2 million
- Linked data/knowledgebase data to label our training set.
- N-Quads(RDF) & JSON data from Amazon and web scrapping
- Amazon dataset includes 3 main product domains such as Headphones, Phones and TV

Problem Statement

How to develop an architecture/framework which will provide real-time C2C matching, using customers’ text-based social media data?

Product Attribute Extraction

- Supervised Random Fields
- Conditional Random Fields
- Multi class logistic regression

Product attributes
- Product brand
- Product group
- Product name
- Selling status
- Product model

Overall System Architecture

- Storm cluster
- CEP Publisher
- CEP Processor
- WSO2 DAS
- Cassandra NoSQL DB

Real-time Extraction Using Distributed Stream Processing

- Worker (JVM)
- Nimbus
- Task
- Executor (Thread)

Cluster architecture
- Parallelism in a single JVM

Conceptual architecture of a trivial Storm topology

Low Latency in-Memory Processing with NoSQL

- Real-time semantic stream
- Query
- Related streams
- Spark (distributed in-memory computing)
- Cassandra

Complex Event Processing Based Product Matching

- CEP is a popular solution to process structured real-time stream
- To match buying & selling messages we used time-based matching as well as content based matching
- Time window & length window are the main 2 caching structure we using

Results

```
<table>
<thead>
<tr>
<th>Module Name</th>
<th>Accuracy</th>
<th>Recall</th>
<th>Precision</th>
<th>F1</th>
<th>Latency (MS)</th>
<th>Parallel Instances</th>
<th>Training Set Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brand NER</td>
<td>0.821</td>
<td>0.873</td>
<td>0.912</td>
<td>0.901</td>
<td>0.333</td>
<td>12</td>
<td>2,03,851</td>
</tr>
<tr>
<td>Product NER</td>
<td>0.84</td>
<td>0.904</td>
<td>0.922</td>
<td>0.913</td>
<td>0.644</td>
<td>10</td>
<td>8,83,101</td>
</tr>
<tr>
<td>Status Classification</td>
<td>0.985</td>
<td>0.974</td>
<td>0.993</td>
<td>0.983</td>
<td>0.533</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Product Group classification</td>
<td>0.948</td>
<td>0.96</td>
<td>0.944</td>
<td>0.952</td>
<td>0.402</td>
<td>10</td>
<td>9,10,951</td>
</tr>
<tr>
<td>In-memory data manipulation</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>3.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CEP based matching</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
```

Accuracy measures of selling status classification

Accuracy measures of product group classification

Latency