HOW NLP AND SPARK CAN ENRICH CUSTOMER DATA

MARK BALKENENDE
MARK BALKENENDE

Director Technical Product Marketing
Talend

@markbalk

#CraftBeerandData
AGENDA

01 UNSTRUCTURED DATA
02 NATURAL LANGUAGE PROCESSING
03 WAYS TO IMPLEMENT
04 STEPS TO TRAIN
UNSTRUCTURED DATA - TEXT
By 2020, natural-language generation and artificial intelligence will be a standard feature of 90% of modern BI platforms.

By 2020, 50% of analytic queries will be generated using search, natural-language processing or voice, or will be autogenerated.

Through 2020, the number of citizen data scientists will grow five times faster than the number of data scientists.
I WORK FOR A COMPANY SELLING ARTIFICIAL INTELLIGENCE

AN EXISTING OPPORTUNITY IN SALESFORCE

EXAMPLE
<table>
<thead>
<tr>
<th>Description</th>
<th>In the Facebook Live, Mark Zuckerberg says he opposes the spread of fear surrounding the potential of artificial intelligence.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&quot;I have pretty strong opinions on this. I am optimistic,&quot; says Zuckerberg. &quot;I think you can build things and the world gets better. But with AI especially, I am really optimistic.&quot;</td>
</tr>
</tbody>
</table>
## NOTES AND DESCRIPTIONS?

<table>
<thead>
<tr>
<th>WHAT</th>
<th>What value may exist in open text fields?</th>
</tr>
</thead>
<tbody>
<tr>
<td>WHERE</td>
<td>Is there value in Notes or Comments?</td>
</tr>
<tr>
<td>HOW</td>
<td>How do we know if there is value?</td>
</tr>
<tr>
<td>PARSE</td>
<td>Can we parse the data for what our company needs?</td>
</tr>
</tbody>
</table>
NATURAL LANGUAGE PROCESSING (NLP)
NATURAL LANGUAGE PROCESSING (NLP)

- Extract useful information
- Finds people names, companies, tools
- Classify discussions
- Entity linking (e.g. persons and organizations linking)

GREAT NLP USE CASES

- Intelligent search
- Sentiment analysis
- GDPR
- Enterprise Information Extraction
- Emotional meaning
- Chatbots
- Machine translation
WAYS TO IMPLEMENT NLP
HOW TO IMPLEMENT NLP

Service
with broad capabilities
Google
Azure
AWS

Target
pre-packaged models
Klevu - eCommerce
Insight Engines - Enterprise
MindMeld – ChatBots

Data training
Spark ML
Standford
TensorFlow
In the Facebook Live, Mark Zuckerberg says he opposes the spread of fear surrounding the potential of artificial intelligence.

"I have pretty strong opinions on this. I am optimistic," says Zuckerberg. "I think you can build things and the world gets better. But with AI especially, I am really optimistic."
TARGET – PREPACKAGED MODELS

What Agolo does

Scan
Ingests public text and internal documents.

Organize
Identifies relationships among sources.

Summarize
Uncovers the most relevant information.

See it in action

Newsfeed from Reuters
Long list of China high-yield issuers due to make May debt payments

Article from the Wall Street Journal
China Bondholders Lose Their Beijing Bailout Confidence

Internal PDF
People’s Bank of China
TRAIN AN NLP MODEL
TRAIN THE MODEL – USE YOUR DATA

Benefits
- Specific Results
- Industry Focused
- Eliminate Noise

Drawback
- Time Consuming
- Costly – Inefficient
- Data / Taxonomy Gaps

Frameworks
- TensorFlow
- Stanford’s CoreNLP
- OpenNMT
HOW TO TRAIN A NLP MODEL?

Job 0: Text preprocessing

Text Sample ➔ tNLPPreprocessing ➔ tNormalize ➔ Texts

List of tokens
List of tokens
... List of tokens

Job 2: Text labelling

Text 1
Text 2
Text 3
Text 4
Text 5

tNLPPreprocessing ➔ tNLPPredict ➔ NLP Model

Cross-validation, precision, recall, F1 - measure

Job 1: Model training

Data

Annotation
tUser
BASIC FLOW TO NLP TRAINING

Natural language processing basic tasks include:

- Text tokenization
- Label / Annotations
- Syntax/Semantics/Discourse
  - Sentence splitting
  - Named entity recognition

*Stanford CoreNLP
LABEL YOUR DATA!

<table>
<thead>
<tr>
<th>Using</th>
<th>Postgres</th>
<th>with Talend</th>
<th>with Best</th>
<th>Hi</th>
<th>Comma</th>
<th>ppm</th>
<th>install</th>
<th>Text-CVS-EXS</th>
<th>does</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.46</td>
<td>0.64</td>
<td>0.6</td>
<td>0.56</td>
<td>0.56</td>
<td>0.5</td>
<td>0.44</td>
<td>0.5</td>
<td>0.51</td>
<td>0.61</td>
</tr>
<tr>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
</tbody>
</table>

Cross-validation, precision, recall, F1 - measure

Job 1: Model training
Grif Thanks for the reply. I did for get to mention all 3 computers are running XP. Two have the Home Ver and 1 the Pro Ver. The one giving me the problem has the Home Ver and is my "main" desktop computer. The weird thing is the other two (a laptop & 2nd desktop) are still networked. The "main" desktop has my printer connected to it, so until I can get it "networked" with the others, I can't print from either of them. Can I assume that since none of them have Vista that I do not need to download and run the program you suggested? Alan

Alan Grif

Thanks for the reply. I did for get to mention all 3 computers are running XP. Two have the Home Ver and 1 the Pro Ver. The one giving me the problem has the Home Ver and is my "main" desktop computer. The weird thing is the other two (a laptop & 2nd desktop) are still networked. The "main" desktop has my printer connected to it, so until I can get it "networked" with the others, I can't print from either of them. Can I assume that since none of them have Vista that I do not need to download and run the program you suggested? Alan

Contact Information

- Name: Mark Zuckerberg
- Account Name: Facebook
CUSTOMER FLOW

Ingest

Process

Insights

Salesforce

Google Storage

Preparation

NLP

Load to BigQuery To SFDC

Google Storage

Spark

Google Dataproc

Google BigQuery

talend
## TAKE AWAYS

<table>
<thead>
<tr>
<th>EXTRACT</th>
<th>Get value with all your data – text, voice, you name it</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHOICE</td>
<td>Pick your service or build your own</td>
</tr>
<tr>
<td>SPARK</td>
<td>Scale your NLP with existing Spark Platforms</td>
</tr>
<tr>
<td>START</td>
<td>Start now with Talend and Spark and NLP</td>
</tr>
</tbody>
</table>