Does Python stand a chance in today’s world of data science?

Radim Řehůřek
YES
Fuck, not another phylum!
RaRe Technologies Ltd.
Python vs. rest

- performance?
- deployment?
- logging, debugging?
- workflow, integration?
### SVD

<table>
<thead>
<tr>
<th>Terms</th>
<th>Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>m1</td>
</tr>
<tr>
<td>trees</td>
<td>1</td>
</tr>
<tr>
<td>graph</td>
<td>0</td>
</tr>
<tr>
<td>minoras</td>
<td>0</td>
</tr>
<tr>
<td>survey</td>
<td>0</td>
</tr>
<tr>
<td>time</td>
<td>0</td>
</tr>
<tr>
<td>response</td>
<td>0</td>
</tr>
<tr>
<td>user</td>
<td>0</td>
</tr>
<tr>
<td>computer</td>
<td>0</td>
</tr>
<tr>
<td>system</td>
<td>0</td>
</tr>
<tr>
<td>EPS</td>
<td>0</td>
</tr>
<tr>
<td>interface</td>
<td>0</td>
</tr>
<tr>
<td>human</td>
<td>0</td>
</tr>
</tbody>
</table>

\[
X = USV^T
\]
English Wikipedia

- ~3.5M docs
- ~2G words
- with 100K vocab, ~0.5G matrix non-zeros
  - very sparse
- small-ish, but known & accessible and **out-of-core**
SVD @ EN wikipedia

- time [h]
- peak mem [GB]

library
- gensim (python)
- scikit-learn (python)
- Mahout (java)
- spark mllib (scala)

peak RAM [GB]

0 1 2 3 4

0 0.75 1.5 2.25 3
Spark mllib

- top level Apache project, Scala
- RDDs, Resilient Distributed Datasets
- ~RAM caching + execution engine
- latest Spark 1.3.0 + mllib
- AWS EMR cluster (4x m3.xlarge)
SVD @ mllib
### Details

<table>
<thead>
<tr>
<th>Type:</th>
<th>Bug</th>
</tr>
</thead>
<tbody>
<tr>
<td>Priority:</td>
<td>Major</td>
</tr>
<tr>
<td>Affects Version/s:</td>
<td>1.1.0</td>
</tr>
<tr>
<td>Component/s:</td>
<td>MLlib</td>
</tr>
<tr>
<td>Labels:</td>
<td>None</td>
</tr>
<tr>
<td>Status:</td>
<td>RESOLVED</td>
</tr>
<tr>
<td>Resolution:</td>
<td>Fixed</td>
</tr>
<tr>
<td>Fix Version/s:</td>
<td>1.2.0</td>
</tr>
</tbody>
</table>

### Description


```
14/10/05 20:16:31 INFO DAGScheduler: Failed to run reduce at RDDFunctions.scala:111
org.apache.spark.SparkException: Job aborted due to stage failure: Task 0 in stage 31.0 failed 1 times, most recent failure: Lost task 0.0 in stage 31.0 (TID 611, localhost): java.lang.ArrayIndexOutOfBoundsException: 4878161

org.apache.spark.mllib.linalg.distributed.RowMatrix$.org$apache$spark$mllib$mlinalg$distributed$RowMatrix$$dsp:
org.apache.spark.mllib.linalg.distributed.RowMatrix$$anonfun$3.apply(RowMatrix.scala:114)
```
Mahout SSVD

- the “scikit-learn” of Hadoop, Java
- originally on MapReduce
- now Mahout Samsara @ Spark, Scala
- newest Mahout 0.10.0
+ “local mode” eats up all disk, then fails
Google’s word2vec

unsupervised ML

- *Berlin* is to *Germany* as *Paris* is to …?
- *king - man + woman* = *queen*
- which word doesn’t fit? “*dinner cereal breakfast lunch*”

http://radimrehurek.com/2014/02/word2vec-tutorial/#app
Word2vec @ Wikipedia

Library

<table>
<thead>
<tr>
<th>Library</th>
<th>Time [h]</th>
<th>Peak Mem [GB]</th>
</tr>
</thead>
<tbody>
<tr>
<td>original (C)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>gensim (python)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>spark mllib (scala)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>deeplearning4j (java)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
C vs. NumPy vs. optimized

<table>
<thead>
<tr>
<th>optimization</th>
<th>words per second</th>
<th>speed-up</th>
</tr>
</thead>
<tbody>
<tr>
<td>NumPy baseline</td>
<td>1.4k</td>
<td>1.0x</td>
</tr>
<tr>
<td>original C word2vec</td>
<td>29.0k</td>
<td>20.7x</td>
</tr>
<tr>
<td>Cython</td>
<td>33.3k</td>
<td>23.8x</td>
</tr>
<tr>
<td>Cython + BLAS</td>
<td>89.8k</td>
<td>64.1x</td>
</tr>
<tr>
<td>Cython + sigmoid table</td>
<td>34.7k</td>
<td>24.8x</td>
</tr>
<tr>
<td>Cython + BLAS + sigmoid table</td>
<td>101.8k</td>
<td>72.7x</td>
</tr>
</tbody>
</table>

(+pure Python: **120x slower** than baseline)
Single machine parallelization

<table>
<thead>
<tr>
<th>implementation</th>
<th># worker threads (speed/peak RAM/accuracy)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C word2vec</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1: 22.6k / 252MB / 27.4%</td>
</tr>
<tr>
<td></td>
<td>2: 42.94k / 252MB / 26.4%</td>
</tr>
<tr>
<td></td>
<td>3: 62.04k / 252MB / 26.8%</td>
</tr>
<tr>
<td></td>
<td>4: 72.44k / 252MB / 27.2%</td>
</tr>
<tr>
<td>gensim word2vec</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1: 109.5k / 591MB / 27.5%</td>
</tr>
<tr>
<td></td>
<td>2: 191.6k / 596MB / 27.1%</td>
</tr>
<tr>
<td></td>
<td>3: 263k / 592MB / 27.3%</td>
</tr>
<tr>
<td></td>
<td>4: 311.7k / 601MB / 28.2%</td>
</tr>
</tbody>
</table>

C (1/2/4 workers): 1.0x / 1.9x / 3.2x
 gensim: 1.0x / 1.75x / 2.85x
streaming (Python generator) for input

>>> for sentence in BrownCorpus('/Users/kofola/nltk_data/corpora/brown'):
...    print sentence
['the/at', 'fulton/np', 'county/nn', 'grand/jj', 'jury/nn', 'said/vb', 'fri
['the/at', 'jury/nn', 'further/rb', 'said/vb', 'in/in', 'term-end/nn', 'pre
['the/at', 'september-october/np', 'term/nn', 'jury/nn', 'had/hv', 'been/be

+ amazing Python ecosystem on either end!

>>> model.most_similar(positive=['woman', 'king'], negative=['man'], topn=1)
[('queen', 0.50882536)]
>>> model.doesnt_match("breakfast cereal dinner lunch".split())
'cereal'
>>> model.similarity('woman', 'man')
0.73723527
INFO BlockManagerInfo: Added rdd_3_0 on disk on ip-172-31-41-18.ec2.internal:45074 (size: 10.6 GB)

WARN TaskSetManager: Lost task 0.0 in stage 0.0 (TID 0, ip-172-31-41-18.ec2.internal): java.lang.illegalArgumentError: size exceeds Integer.MAX_VALUE

at sun.nio.ch.FileChannelImpl.map(FileChannelImpl.java:829)
at org.apache.spark.storage.DiskStore.getBytes(DiskStore.scala:124)
at org.apache.spark.storage.DiskStore.getBytes(DiskStore.scala:133)
at org.apache.spark.storage.BlockManager.doGetLocal(BlockManager.scala:516)
at org.apache.spark.storage.BlockManager.getLocal(BlockManager.scala:431)
at org.apache.spark.storage.BlockManager.get(BlockManager.scala:617)
at org.apache.spark.CacheManager.putInBlockManager(CacheManager.scala:155)
at org.apache.spark.CacheManager.getOrCompute(CacheManager.scala:79)
at org.apache.spark.rdd.RDD.iterator(RDD.scala:242)
at org.apache.spark.rdd.MapPartitionsRDD.compute(MapPartitionsRDD.scala:35)
at org.apache.spark.rdd.RDD.computeOrReadCheckpoint(RDD.scala:277)
at org.apache.spark.rdd.RDD.iterator(RDD.scala:244)
at org.apache.spark.rdd.MapPartitionsRDD.compute(MapPartitionsRDD.scala:35)
at org.apache.spark.rdd.RDD.computeOrReadCheckpoint(RDD.scala:277)
at org.apache.spark.rdd.RDD.iterator(RDD.scala:244)
at org.apache.spark.rdd.MapPartitionsRDD.compute(MapPartitionsRDD.scala:35)
at org.apache.spark.rdd.RDD.computeOrReadCheckpoint(RDD.scala:277)
at org.apache.spark.rdd.RDD.iterator(RDD.scala:244)
at org.apache.spark.rdd.MapPartitionsRDD.compute(MapPartitionsRDD.scala:35)
at org.apache.spark.rdd.RDD.computeOrReadCheckpoint(RDD.scala:277)
at org.apache.spark.rdd.RDD.iterator(RDD.scala:244)
at org.apache.spark.scheduler.ShuffleMapTask.runTask(ShuffleMapTask.scala:68)
When the vocabulary size is large, Word2Vec may yield "OutOfMemoryError: Requested array size exceeds VM limit"

Environment:
Use Word2Vec to process a corpus(sized 3.5G) with one partition.
The corpus contains about 300 million words and its vocabulary size is about 10 million.

Description
Exception in thread "Driver" java.lang.reflect.InvocationTargetException
at sun.reflect.NativeMethodAccessorImpl.invoke0(Native Method)
at sun.reflect.NativeMethodAccessorImpl.invoke(NativeMethodAccessorImpl.java:57)
at sun.reflect.DelegatingMethodAccessorImpl.invoke(DelegatingMethodAccessorImpl.java:43)
at java.lang.reflect.Method.invoke(Method.java:606)
scaling down for Spark

word2vec @ EN wikipedia

<table>
<thead>
<tr>
<th>library</th>
<th>time [h]</th>
<th>peak mem [GB]</th>
</tr>
</thead>
<tbody>
<tr>
<td>original (C)</td>
<td>4</td>
<td>8</td>
</tr>
<tr>
<td>gensim (python)</td>
<td>1</td>
<td>12</td>
</tr>
<tr>
<td>spark mllib (scala)</td>
<td>12</td>
<td>60</td>
</tr>
<tr>
<td>deeplearning4j (java)</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

=> if scaling linearly, Spark needs a cluster of ~12 machines to break even (vs. pySpark)
Deeplearning4j
David Przybilla, Idio Ltd.
https://github.com/idio/wiki2vec

Word2Vec tools:

- Gensim
- DeepLearning4j: Feb 2014, Gets stuck in infinite loops on a big corpus
- Spark's word2vec: Feb 2014, number of dimensions * vocabulary size certain value otherwise an exception is thrown. issue
ANN libs @ Wikipedia
Tools

“Do one thing and do it well.”

Doug McIlroy

"Every program attempts to expand until it can read mail. Those programs which cannot so expand are replaced by ones which can."

Zawinski’s law of software development
APIs & Abstractions

How Standards Proliferate:

(See: A/C chargers, character encodings, instant messaging, etc.)

Situation: There are 14 competing standards.

14?! Ridiculous! We need to develop one universal standard that covers everyone’s use cases. Yeah!

Soon:

Situation: There are 15 competing standards.
Configuration, setup, deployment

... the real work!

Python 2 vs Python 3
Exception in thread main java.lang.NoClassDefFoundError: org/apache/hadoop/hbase/HBaseConfiguration
  at java.lang.Class.forName0(Native Method)
  at java.lang.Class.forName(Class.java:340)
  at org.apache.hadoop.util.RunJar.main(RunJar.java:149)

java.lang.Exception: java.lang.OutOfMemoryError: Java heap space
 at org.apache.hadoop.mapred.LocalJobRunner.runTasks(LocalJobRunner.java:462)
 at org.apache.hadoop.mapred.LocalJobRunner.run(LocalJobRunner.java:522)
Complex pipelines

- Python: Luigi (~Spotify), Pinball (~Pinterest)
- Java: Apache Oozie, Azkaban...
Logging

- UI, job trackers
- tracebacks, continuous
- configurable
- human readable
Navigating the tool landscape

Let it go; if it’s meant to be, it will come back.
Take “progress” easy

mat kelcey @mat_kelcey · May 20
pretty much every paper i’ve ever read....

<table>
<thead>
<tr>
<th>method</th>
<th>score</th>
</tr>
</thead>
<tbody>
<tr>
<td>previous approach</td>
<td>good</td>
</tr>
<tr>
<td>our approach</td>
<td>almost as good</td>
</tr>
<tr>
<td>our approach + last minute hack</td>
<td>slightly better than good!</td>
</tr>
</tbody>
</table>
Summary

Python’s greatest differentiating factors:
● +experienced full stack engineers
● +pragmatic, mature tools
● +HPC & scientific “baggage”
● -meh deployment, orchestration, packaging
● -not as much enterprise “baggage”
Oh, sorry!
Radim Řehůřek
http://rare-technologies.com
(formerly radimrehurek.com)
@radimrehurek

Blog (mostly tech): http://radimrehurek.com/blog/