Mining Public Datasets
using Apache Zeppelin (incubating), Apache Spark and Juju

by Alexander Bezzubov
NFLabs for AppacheCon ’16 NA
Alexander Bezzubov

Software Engineer at NFLabs, Seoul, South Korea

Co-organizer of SeoulTech Society

Committer and PPMC member of Apache Zeppelin (Incubating)

Graduated Maths at St.Petersburg State University, Russia

github.com/bzz
@seoul_engineer
PUBLIC DATASETS: Number, Size & Growth

Web Crawls
Structured data (RDF, micro-formats, tables)
Hackers News\Reddit\Twitter\StackOverflow\Wikipedia
Reviews (movies, restaurants, beer, wine)
Emails (Enroll, ASF public ML archives)
Census Data (US, UK, UN, Japan, etc)
Transportation (Taxi, Flights, Bicycles)
Climate
Genome
PUBLIC DATASETS: Number, Size & Growth

Web Crawls
Structured data (RDF, micro-formats, tables)
Hackers News\Reddit\Twitter\StackOverflow\Wikipedia
Reviews (movies, restaurants, beer, wine)
Emails (Enroll, ASF public ML archives)
Census Data (US, UK, UN, Japan, etc)
Transportation (Taxi, Flights, Bicycles)
Genome

order of Tbs
PUBLIC DATASETS: Number, Size & Growth

Web Crawls
Structured data (RDF, micro-formats, tables)
Hackers News\Reddit\Twitter\StackOverflow\Wikipedia
Reviews (movies, restaurants, beer, wine)
Emails (Enroll, ASF public ML archives)
Census Data (US, UK, UN, Japan, etc)
Transportation (Taxi, Flights, Bicycles)
Genome

order of Tbs

Yahoo Webscope https://webscope.sandbox.yahoo.com/
Stanford Network Analyser Project http://snap.stanford.edu/data/

Physics Research http://opendata.cern.ch
PUBLIC DATASETS: Number, Size & Growth

Web Crawls
Structured data (RDF, micro-formats, tables)
Hackers News\Reddit\Twitter\StackOverflow\Wikipedia
Reviews (movies, restaurants, beer, wine)
Emails (Enroll, ASF public ML archives)
Census Data (US, UK, UN, Japan, etc)
Transportation (Taxi, Flights, Bicycles)
Genome

Yahoo Webscope https://webscope.sandbox.yahoo.com/
Stanford Network Analyser Project http://snap.stanford.edu/data/
Physics Research http://opendata.cern.ch

order of Tbs

order of Pbs
PUBLIC DATA = OPPORTUNITY
I. Tools

II. Data
TOOL TO PURSUIT THE OPPORTUNITY:
Overview Big Data eco-system
TOOL TO PURSUIT THE OPPORTUNITY:
Overview Big Data eco-system
TOOL TO PURSUIT THE OPPORTUNITY:

Todays choice Zeppelin, Spark, Juju

Apache Spark
Scala, Python, R

Apache Zeppelin
Modern Web GUI, plays nicely with Spark, Flink, Elasticsearch, etc.

Warcbase
Spark library for saved crawl data (WARC)

Juju
Scales, integration with Spark, Zeppelin, AWS, GCE
APACHE ZEPPELIN: Overview
# Zeppelin: Brief history

[Zeppelin Icon](http://zeppelin.incubator.apache.org)

<table>
<thead>
<tr>
<th>Date</th>
<th>Event Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>12.2012</td>
<td>Commercial App using AMP Lab Shark 0.5</td>
</tr>
<tr>
<td>10.2013</td>
<td>Prototype Hive/Shark</td>
</tr>
<tr>
<td>08.2013</td>
<td>NFLabs Internal project Hive/Shark</td>
</tr>
<tr>
<td>12.2014</td>
<td>Enters ASF Incubation</td>
</tr>
<tr>
<td>01.2016</td>
<td>3 major releases</td>
</tr>
<tr>
<td>05.2016</td>
<td>Graduation vote passed</td>
</tr>
</tbody>
</table>
Interactive Visualization
APACHE SPARK

From Berkeley AMP Labs, since 2010
Joined Apache since 2014
1000+ contributors
REPL + Java, Scala, Python, R APIs

http://spark.apache.org
Service modelling at scale

Deployment/configuration automation
  + Integration with Spark, Zeppelin, Ganglia, etc
  + AWS, GCE, Azure, LXC, etc

https://jujucharms.com/
$ apt-get install juju-core juju-quickstart
  # or
$ brew install juju juju-quickstart
$ juju generate-config
#LXC, AWS, GCE, Azure, VMWare, OpenStack

$ juju bootstrap
$ juju quickstart apache-hadoop-spark-zeppelin
$ juju expose spark zeppelin
$ juju add-unit -n4 slave

http://bigdata.juju.solutions/getstarted
http://bigdata.juju.solutions/getstarted

7 node cluster designed to scale out
**APPROACH:** local, small cluster, big cluster

<table>
<thead>
<tr>
<th>1 core</th>
<th>Prototype</th>
<th>Your laptop</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>10s PC</strong></td>
<td>Estimate the cost</td>
<td>AWS spot instances</td>
</tr>
<tr>
<td><strong>1000 instances</strong></td>
<td>Scale out</td>
<td>Deployment automation</td>
</tr>
</tbody>
</table>
I. Tools

II. Data
DATA: GitHub

- 300Gb compressed
- Collaboration google and github engineers
- Events on PR, repo, issues, comments, etc in JSON

http://githubarchive.org
Commit Logs From Last Night

Tried to do something smart but fuck it.

fix shit

fixed the shit

hacky presentation ready kind of shit

http://www.commitlogsfromlastnight.com/
http://zoom.it/kCsU
DATA PRODUCT: Get notified when project goes Open Source
DATA PRODUCT: Exploration
**DATA PRODUCT: Sketch**

We are going to build a Notebook that sends you a digest email:

---

**Opensourced today: 25 Sep 2015**

The newest projects on github from the industry leaders

<table>
<thead>
<tr>
<th>Company</th>
<th>Repository</th>
<th>Language</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>google/primarysources</td>
<td>C++</td>
<td></td>
</tr>
<tr>
<td>Facebook</td>
<td>facebook/device-year-class</td>
<td>Java</td>
<td>A library that analyzes an Android device's specifications and calculates which year the device would be considered &quot;high end&quot;.</td>
</tr>
<tr>
<td>Facebook</td>
<td>facebook/fresco</td>
<td>Java</td>
<td>An Android library for managing images and the memory they use.</td>
</tr>
<tr>
<td>Google Cloud Platform</td>
<td>GoogleCloudPlatform/appengine-multitenancy-python</td>
<td>Python</td>
<td></td>
</tr>
</tbody>
</table>
DATA PRODUCT: pieces (flow-chart)

We are going to build a Notebook that:

• Downloads the latest data from GitHub Archive
• Read & explore the dataset
• Imports, filters the PublicEvent
• Join logs w/ more data from Github API calls
• Shows HTML template, to visualise the list
• Sends email notifications
• Does all above automatically, once a day
DATA PRODUCT: Full impl
I. Tools

II. Data
DATA: Common Crawl

[Image]

https://commoncrawl.org

Nonprofit, by Factual

On AWS S3 in WARC, WAT, formats since 2013, monthly: ~150Tb compressed, 2+bln ulrs
URL Index by Ilya Kreymer of @webrecorder_io
http://index.commoncrawl.org/

February 2016 Index Info Page

title: February 2016 Index

Search a url in this collection: (Wildcards -- Prefix: http://example.com/*  Domain: *.example.com)

wikipedia.org

Show Number Of Pages Only

(See the CDX Server API Reference for more advanced query options.)

Back To All Indexes
$N$-gram Counts and Language Models from the Common Crawl

Christian Buck†, Kenneth Heafield‡, Bas van Ooyen*
†University of Edinburgh, Edinburgh, Scotland
‡Stanford University, Stanford, CA, USA
* Owlin BV, Utrecht, Netherlands
christian.buck@ed.ac.uk, heafield@cs.stanford.edu, bas@owlin.com

Abstract

We contribute $n$-gram counts and language models trained on the Common Crawl corpus, a collection over 9 billion web pages. This release improves upon the Google $n$-gram counts in two key ways: the inclusion of low-count entries and deduplication to reduce boilerplate. By preserving singletons, we were able to use Kneser-Ney smoothing to build large language models. This paper describes how the corpus was processed with emphasis on the problems that arise in working with data at this scale. Our unpruned Kneser-Ney English 5-gram language model, built on 975 billion deduplicated tokens, contains over 500 billion unique $n$-grams. We show gains of 0.5–1.4 BLEU by using large language models to translate into various languages.

Keywords: web corpora, language models, multilingual

1. Introduction

The sheer amount of data in multiple languages makes web-scale corpora attractive for many natural language processing tasks. Of particular importance is language modeling, where web-scale language models have been shown to improve machine translation and automatic speech recognition performance (Brants et al., 2007; Chelba and Schalkwyk, 2013; Guthrie and Hepple, 2010). In this work, we contribute $n$-gram counts and language models trained on

2. Data Preparation

The Common Crawl\textsuperscript{2} is a publicly available crawl of the web. We use the 2012, early 2013, and “winter” 2013 crawls, consisting of 3.8 billion, 2 billion, and 2.3 billion pages, respectively. Because both 2013 crawls are similar in terms of seed addresses and distribution of top-level domains in this work we only distinguish 2012 and 2013 crawls.

The data is made available both as raw HTML and as text
Wikipedia for Schools
schools.wikipedia.org
Welcome to Wikipedia for Schools! This selection of articles from Wikipedia matches the UK National Curriculum and can be used by school children around the...

Wikipedia Review
www.wikipedia-review.com
Putting the wakeup alarmclock to Wikipedia's head since... Oh god, it's been that long?

Wikiwix » Wikipedia
www.wikiwix.com
Search Last visited websites Categories Favorite websites

Wikipedia Toolbar: Home
wikipedia.mozdev.org
Wikipedia Toolbar gives you quick access to usefull commands for the Mediawiki Software. It also provides a backend for other types of Wiki software. It also...

Wikipedia (TheFreeDictionary.com mirror)
www.encyclopedia.farlex.com
Wikipedia is a Web-based, free-content encyclopedia written collaboratively by volunteers and sponsored by the non-profit Wikimedia Foundation. It contains...
Measuring the impact of Google Analytics

Objective: estimate % of pages/domains that use Google Analytics/Facebook

Existing research from 2013
DATA: CommonCrawl - Data Product

Measuring the impact of Google Analytics

Copy to HDFS vs read from S3
Verify using grep

```bash
hadoop jar hadoop-examples.jar grep /grep-data/ \
 /grep-output/ '[Bb]ig [Dd]ata is ([a-zA-Z]{5,})'
```

Verify using grep
DATA: CommonCrawl - Data Product

Feb 2016 Crawl:
- 48Tb compressed
- 100 segments (dir on S3)
- 30,000 files, ~1Gb each
AWS optimisations:
- pick spot instance prices
- pick instance type (net throughput)
- use Juju instead of EMR (2x $$ savings!)

Spark optimisations:
- IO-bound, so increase `spark.executor.cores`
  `spark.executor.memory`
DATA: CommonCrawl - Data Product
Zeppelin Viewer

Community service for sharing example notebooks
http://zeppelinhub.com/viewer
TAKEAWAY

There are plenty of free tools out there

To crunch the data for fun and profit

They are easy (not simple) to learn and generic enough
Questions?

Alexander Bezzubov

@seoul_engineer
github.com/bzz
Thank you

Alexander Bezzubov
NFLabs, Seoul (we are hiring!)