Implementing Latent Dirichlet Allocation on NASA ADS

Internship report submitted to the Industrial Engineering and Operations Research Department of Columbia University in the City of New York

Tarun Ruchandani

August 2016

Advisor: Dr. Michael Kurtz

06/01/2016 – 08/31/2016

Introduction

This summer I had the most excellent opportunity to intern at Harvard-Smithsonian Center for Astrophysics.

The proposed project was to build an LDA space on recent Phys-Astro-Geo literature (nearly a million papers) on the Astrophysics Data System (ADS) database.

Prof. Kurtz’s PhD student, Kriste Krstovski, recently concluded his thesis on using LDA on ADS XML text collection (every refereed Physics, Astronomy, and Geophysics written in the last 10 years) where he used his own software. Since he began, Apache Spark has been released with a state-of-the-art LDA capability.  This will make these kind of investigations much easier. This project is envisioned to implement LDA on Spark. There has been prelim work done on getting the vocab and removing the duds.

Next, this space shall be used to classify the readers.  About 50,000 different people download more than 100 papers/year via ADS, so there should be enough signal. Thus, the potential next step could be using the choices of similar individuals to make recommendations.

If you think you could implement a SPARK instance and bring up the LDA (Latent Dirichlet Allocation) portion quickly enough to get a result before the end of the summer

My PhD student, Kriste Krstovski, has just finished.  Much of his thesis used LDA on the ADS whole XML text collection (every refereed Physics, Astronomy, and Geophysics written in the last 10 years).  He used his own software.  He will move to Colombia to work with Blei at the beginning of the summer.

Since he began SPARK has been released with what appears to be a state-of-the-art LDA capability.  This will make these kind of investigations much easier.

The project I envision (very preliminary) is to build an LDA space on the recent Phys-Astro-Geo literature (nearly a million papers, Kriste has done a lot of the prelim work in getting the vocab and removing the duds).

Next we will use this space to classify the readers.  About 50,000 different people download more than 100 papers/year via ADS, so we should have enough signal.

Lessons Learnt

8/9: Perform Clustering on LDA output

Preface:

As Dr. David Blei

We humans are naturally prone to finding patterns. Of connecting human intuition comprehend the complexities around any topics.

Acknowledgements:

profoundly grateful to my advisor

Dr. Kurtz : pointing me to right papers to read, right people to reach out to, and encouraging me to implement. I am grateful to him for all this, and much more!

For kickstarting Kriste

ADS group:

Alberto, Alex, Jonny, Donna , Carolyn, Zack, Edwin,

Donna Wyatt - NASA

Grateful to Roman whose friendship, and enlightening lunchtime conversation made my time at Harvard CfA so memorable.

Prof. Cunningham

1. Duties and responsibilities

Building Data pipelines

2. Notable Projects

4. Skills and qualifications acquired

Machine Learning knowledge

Programming skills

Motivation to work in academia with astrophysicists.

Look at the next step, and work backwards from there

Working to work. Thinking of architecture.

3. A typical day during this internship

Weekly scientific talks

Daily scrum calls - daily goals

Thursday meetings with Dr. Kurtz and ADS group

Mornings: Read and learn the theoretical concepts.

Watch Blei’s lectures, Jaff’s NLP videos.

Post lunch: build the pipeline.

Writing and documenting all of the learnings.

Parts of Professor Dan Jurafsky’s course on Natural Language Processing

Reading Blei’s papers

Learning Louvain Methods

Meeting with fellow interns on Summer Symposiums on Thursdays

**Learnings:**

Connecting iPython with pyspark

PATH="/usr/bin:/bin:/usr/sbin:/usr/local/bin:/usr/X11/bin"; export PATH;

echo "export PATH=$PATH:/Users/tarunruchandani/Desktop/HarvardSummer2016/spark-1.6.1/bin" >> .profile

echo "export PYSPARK\_DRIVER\_PYTHON=ipython" >> .profile

echo "export PYSPARK\_DRIVER\_PYTHON\_OPTS='notebook' pyspark" >> .profile

cd /Users/tarunruchandani/Desktop/HarvardSummer2016/spark-1.6.1

export SPARK\_HOME="/Users/tarunruchandani/Desktop/HarvardSummer2016/spark-1.6.1"

source .profile

pyspark

Usually not: IPYTHON\_OPTS="notebook" pyspark

~~Scratch all of the above~~

What worked on 8/4

New Terminal

Anaconda open. Just the app, not Jupyter or its terminal.

export SPARK\_HOME="/Users/tarunruchandani/Desktop/HarvardSummer2016/spark-1.6.1"

cd $SPARK\_HOME

IPYTHON\_OPTS="notebook" ./bin/pyspark

WORKS!

./Infomap [options] network\_data dest

cd /Users/tarunruchandani/Desktop/HarvardSummer2016/Infomap

Using the Python example throws error of not being able to import \_infomap module. - Unresolved.

./Infomap JS\_output2.txt mymaps/ -N 10 --tree --bftree

**Learn what’s community detection.**

run demo with Python

cd examples/python

make

python [example-networkx.py](http://example-networkx.py/)

Demo with C++

cd examples/cpp/minimal

make

./example

./Infomap –clu /Users/tarunruchandani/Desktop/HarvardSummer2016/mymaps/JS\_output.txt /Users/tarunruchandani/Desktop/HarvardSummer2016/mymaps –k -–map

Sample Output

**Docker**

5. Contribution to your career plans

I applied for the with the motivation and goal of pursuing a doctoral program at the end of my , and this internship opportunity ha

NYTimes LDA paper

Communicating to non-technical audience

Why I chose an academic internship over an industrial one is to further my education and maximize my learning. Also wrote the code, Got the best of both the worlds.

A potential pursuit of doctoral degree.

Working with ADS group helped understand how large scale systems work.

Everything fell into place.

6. Relationship to your coursework at Columbia

Stat ML

NLP

A mathematical mindset to approaching problems

Future work:

Persistent storage

Build a recommendation engine

7. References:

Blei’s paper

<https://www.cs.princeton.edu/~blei/papers/Blei2012.pdf>

Topik

ADS

NetworkX

Louvain Method