

CS15-319 / 15-619

Cloud Computing

Recitation 13

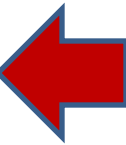
April 15th and April 17th, 2014

Last Week's Project Module

- Provision your own Hadoop cluster
- Write a MapReduce program to construct inverted lists for the Project Gutenberg data
- Run your code from the master instance
- Piazza Highlights
 - Different versions of Hadoop API: Both old and new should be fine as long as your program is consistent

Module to Read

- UNIT 5: Distributed Programming and Analytics Engines for the Cloud
 - Module 16: Introduction to Distributed Programming for the Cloud
 - Module 17: Distributed Analytics Engines for the Cloud: MapReduce
 - Module 18: Distributed Analytics Engines for the Cloud: Pregel
 - Module 19: Distributed Analytics Engines for the Cloud: GraphLab

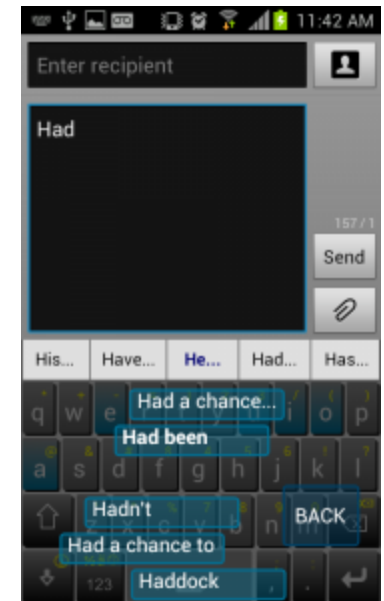


Input Text Predictor

- Suggest words based on letters already typed

wiki	
wikipedia	250,000,000 results
wikipedia encyclopedia	16,300,000 results
wiki answers	24,400,000 results
wikimapia	12,000,000 results
wikihow	1,780,000 results
wikiquote	3,280,000 results
wikispaces	7,800,000 results
wikitavel	2,270,000 results
wikimedia	55,700,000 results
wikipedia dictionary	20,300,000 results
	close

[Advanced Search](#)
[Preferences](#)
[Language Tools](#)



n -gram

- An n -gram is a phrase with n contiguous words

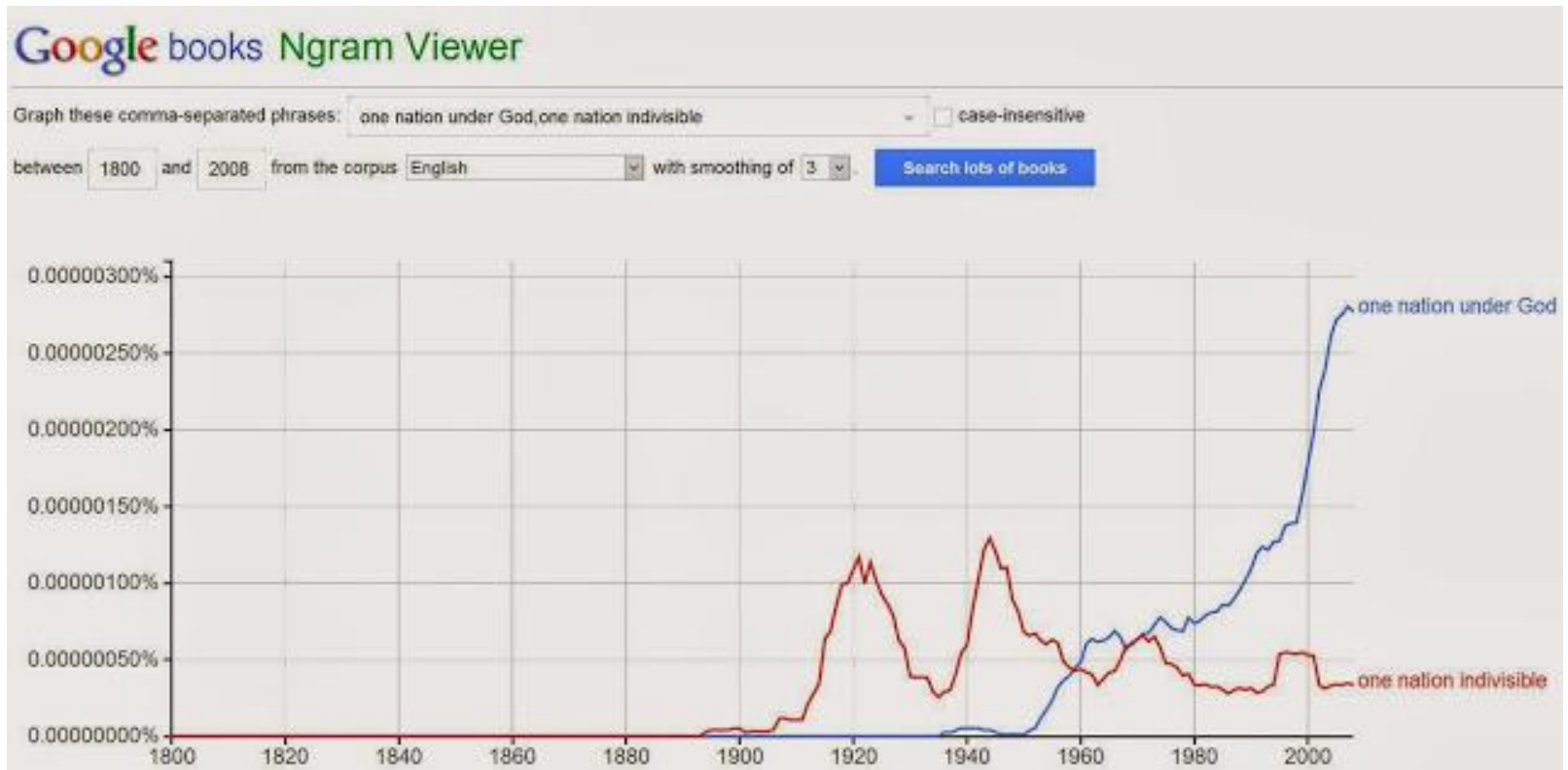
Example Phrase: This is interesting because this is a cloud computing course							
#	1-gram	Count	2-gram	Count	3-gram	Count	
1	this	2	this is	2	this is interesting	1	
2	is	2	is interesting	1	is interesting because	1	
3	interesting	1	interesting because	1	interesting because this	1	
4	because	1	because this	1	because this is	1	
5	a	1	is a	1	this is a	1	
6	cloud	1	a cloud	1	is a cloud	1	
7	computing	1	cloud computing	1	a cloud computing	1	
8	course	1	computing course	1	cloud computing course	1	
#	4-gram	Count	5-gram	Count	6-gram	Count	
1	this is interesting because	1	this is interesting because this	1	this is interesting because this is	1	
2	is interesting because this	1	is interesting because this is	1	is interesting because this is a	1	
3	interesting because this is	1	interesting because this is a	1	interesting because this is a cloud	1	
4	because this is a	1	because this is a cloud	1	because this is a cloud computing	1	
5	this is a cloud	1	this is a cloud computing	1	this is a cloud computing course	1	
6	is a cloud computing	1	is a cloud computing course	1			
7	a cloud computing course	1					
8							

Google-Ngram Viewer



- The result seems logical: the singular “is” becomes the dominant verb after the American Civil War.

Google-Ngram Viewer



- “one nation under God” and “one nation indivisible.”
- “under God” was signed into law by President Eisenhower in 1954.

How to Construct an Input Text Predictor?

1. Given a language corpus

- Project Gutenberg (2.5 GB)
- English Language Wikipedia Articles (30 GB)

2. Construct an n-gram model of the corpus

- An n-gram is a phrase with n contiguous words
- For example a set of 1,2,3,4,5-grams with counts:

• this	1000
• this is	500
• this is a	125
• this is a cloud	60
• this is a cloud computing	20

How to Construct an Input Text Predictor? (Next Week)

3. Build a statistical language model that contains the probability of a word appearing after a phrase
4. Store and index the words and their probabilities to use in an application

This Week's Goal

Construct an n-gram model of the corpus

- An n-gram is a phrase with n contiguous words
- For example a set of 1,2,3,4,5-grams with counts:
 - this 1000
 - this is 500
 - this is a 125
 - this is a cloud 60
 - this is a cloud computing 20

Upcoming Deadlines

- Project 4:

[Project 4](#)

[Input Text Predictor: NGram Generation](#)

NGram Generation

[Checkpoint](#)

11:59PM

04/20/2014



- 15-619 Project:

- Phase 3 (last phase) is due on April 22nd

- Live-test will be announced



15-619 Project

- Live test for phase 2 is completed
- You should have received feedback
- Phase 3 is ongoing!
 - 75% of the total grade
 - **Pick one** between MySQL and HBase
 - 6 queries in total
 - 4 hour live test at the end to determine your performance and **the winning team!**

15-619 Project: Phase 3

- Q4: Text of tweets
 - A tweet may contain multiple lines
- Q5: Find tweets by location and during a particular time range
 - The **text** of tweet contains a given place
 - All possible places come from “place” object in the data set
 - Text of tweet needs pre-processing (see write up)
- Q6: Number of tweets
 - The number of tweets in a **given data set**

15-619 Project:

Rumors and the Truth

- EMR cost is for cluster: **No!**
 - EMR cost is per instance per hour. A cluster of 9 m1.large will consume $\$0.044 * 9 = \0.396
- Budget is only development cost: **No!**
 - \$75 is for the whole phase including live test. Please intelligently plan how to spend
- We can start until this weekend: **No!**
 - The amount of data you will process will be larger than last phase, leading to increased risks for ETL
 - You may need more time to optimize your design: new queries tend to be more difficult to achieve a good score
 - You should be doing ETL now

15-619 Project: How We Test

- We use JMeter
- Multiple threads (up to 50) keep issuing requests to your IP address
- Your responses are compared to the correct responses
- Requests are not ordered: they are generated randomly based on some rules to fully explore your throughput
- For q5 and q6, expect large ranges (such as 100K userids in q6)
- For q4, expect many responses that are large in size

15-619 Project:

What You Should Know

- Set port configuration of ELB as **TCP** 80 -> TCP 80 instead of HTTP 80 -> HTTP 80
- Do not use the **same connection** for every request: significant negative effects
- Table design is not the whole world: find the bottleneck in your system
- Re-test q1 – q3 in phase 3: Something may be different

Check AWS Services Charges

Amazon Web Services

Sign Up My Account / Console English

AWS Products & Solutions AWS Product Information Support

Account

- Account Activity
- AWS Identity and Access Management
- AWS Management Console
- Consolidated Billing
- Reserved Instance Marketplace Setting
- DevPay
- Manage Your Account
- Payment Method
- Personal Information
- Security Credentials
- Usage Reports
- Billing Alerts
- Billing Preferences

Cost Allocation Report

- Manage Cost Allocation Report

Account Activity

W... | Sign Out
Account Number 1732-1996-3889

New Billing Console

The Account Activity page is moving to the AWS Management Console, and it has a fresh new look. You're invited to [Preview the Billing Console](#) and to leave us feedback using the link located at the bottom of the Billing Console page.

You are eligible for the [AWS Free Usage Tier](#). See the [Getting Started Guide AWS Free Usage Tier](#) to learn how to get started with the free usage tier.

Your account is enabled for monitoring estimated charges. [Set your first billing alert](#) to receive an e-mail when charges reach a threshold you define. [Learn More](#)

This Month's Activity as of January 12, 2014

The statement period for this report is January 1 - January 31, 2014. The charges on this page currently show activity through approximately 01/12/2014 21:06 GMT.

Select a different statement:

Summary

Check AWS Services Charges

This Month's Activity as of August 26, 2013

The statement period for this report is August 1 - August 31, 2013. The charges on this page currently show activity through approximately 08/26/2013 19:01 GMT.

Select a different statement:

Summary

AWS Service Charges	\$0.00
Usage Charges and Recurring Fees <small>(More Info)</small>	\$0.00
<small>View estimated charges</small>	
Total new charges for this statement	\$0.00

All charges this month will be paid for by AWS Account 9392-2385-3384.

Details

[Expand All Services](#) | [Collapse All Services](#) | [Printer Friendly Version](#)

AWS Service Charges

Amazon Elastic Compute Cloud	\$0.00	
<small>Download Usage Report ></small>		
EDU_Course_Sakr_CarnegieMellon_Summer2013	Credit	-4.94
		-4.94

US East (Northern Virginia) Region

Amazon EC2 running Linux/UNIX		
\$0.060 per M1 Standard Small (m1.small) Linux/UNIX instance-hour (or partial hour)	6 Hrs	0.36
\$0.006 per Micro Instance (t1.micro) instance-hour (or partial hour) (blended price)*	650 Hrs	4.04
Amazon EC2 EBS		
\$0.080 per GB-month of provisioned storage under monthly free tier (blended price)*	6.785 GB-Mo	0.54
\$0.00 for the first 2 million I/O requests under monthly free tier	116,945 IOs	0.00
Amazon CloudWatch		
\$0.00 per request - first 1,000,000 requests	7 Requests	0.00
		4.94

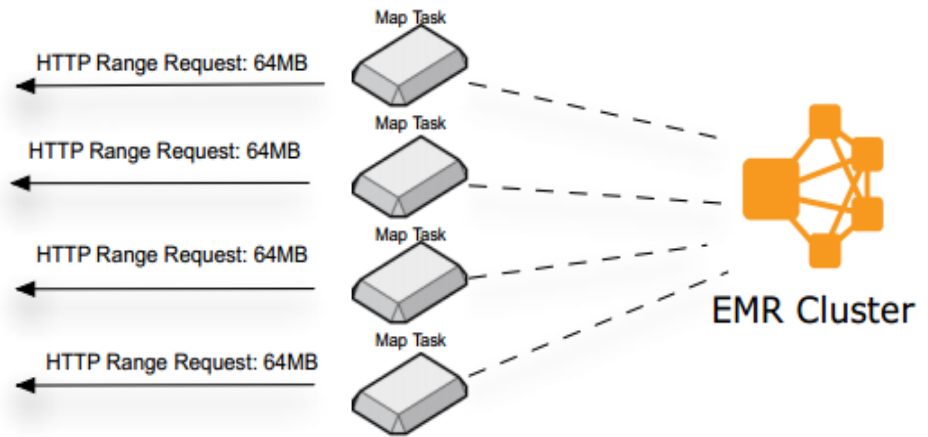
Amazon SimpleDB

Amazon SimpleDB	\$0.00
<small>Download Usage Report ></small>	

Demo Outline

- 1. Hadoop Commands

- Hadoop fs -help
- hadoop fs -put
- hadoop fs -get
- hadoop distcp



- 2. N-Gram Generation

- Google Instant
- Input Text Predictor
- N-Gram Generation

Recommendation

- Use small text to test your code and debug before running the entire big dataset
- Optimize your code to accelerate MapReduce before seeking other optimization methods
- Start Early
- Reference:
 1. http://hadoop.apache.org/docs/r1.0.4/commands_manual.html
 2. http://docs.aws.amazon.com/ElasticMapReduce/latest/DeveloperGuide/UsingEMR_s3distcp.html
 3. Amazon AWS EMR Best Practices (link posted on Piazza)