CS15-319 / 15-619
Cloud Computing

Recitation 13

November 18th and 20th, 2014

Announcements

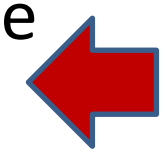
- Encounter a general bug:
  - Post on Piazza
- Encounter a grading bug:
  - Post Privately on Piazza
- Don’t ask if my answer is correct
- Don’t post code on Piazza
- Search before posting
- Post feedback on OLI

Last Week's Project Reflection

- 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

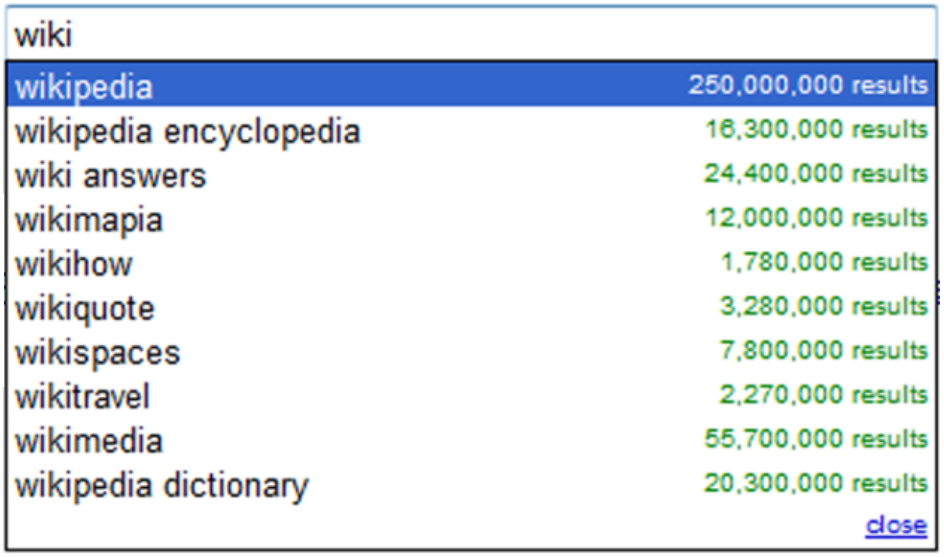


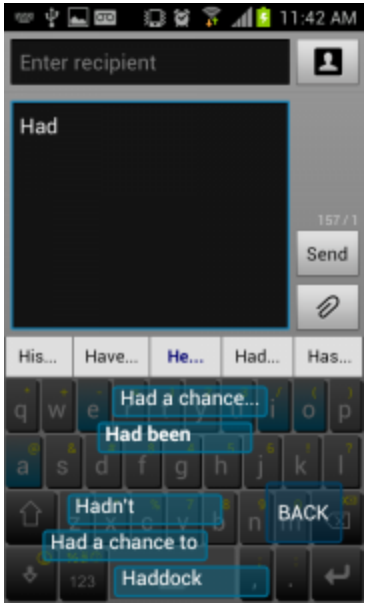
Project 4

- MapReduce
  - Hadoop MapReduce
- Input Text Predictor: NGram Generation
  - NGram Generation
- Input Text Predictor: Language Model and User Interface
  - Language Model Generation

Input Text Predictor

- Suggest words based on letters already typed





n-gram

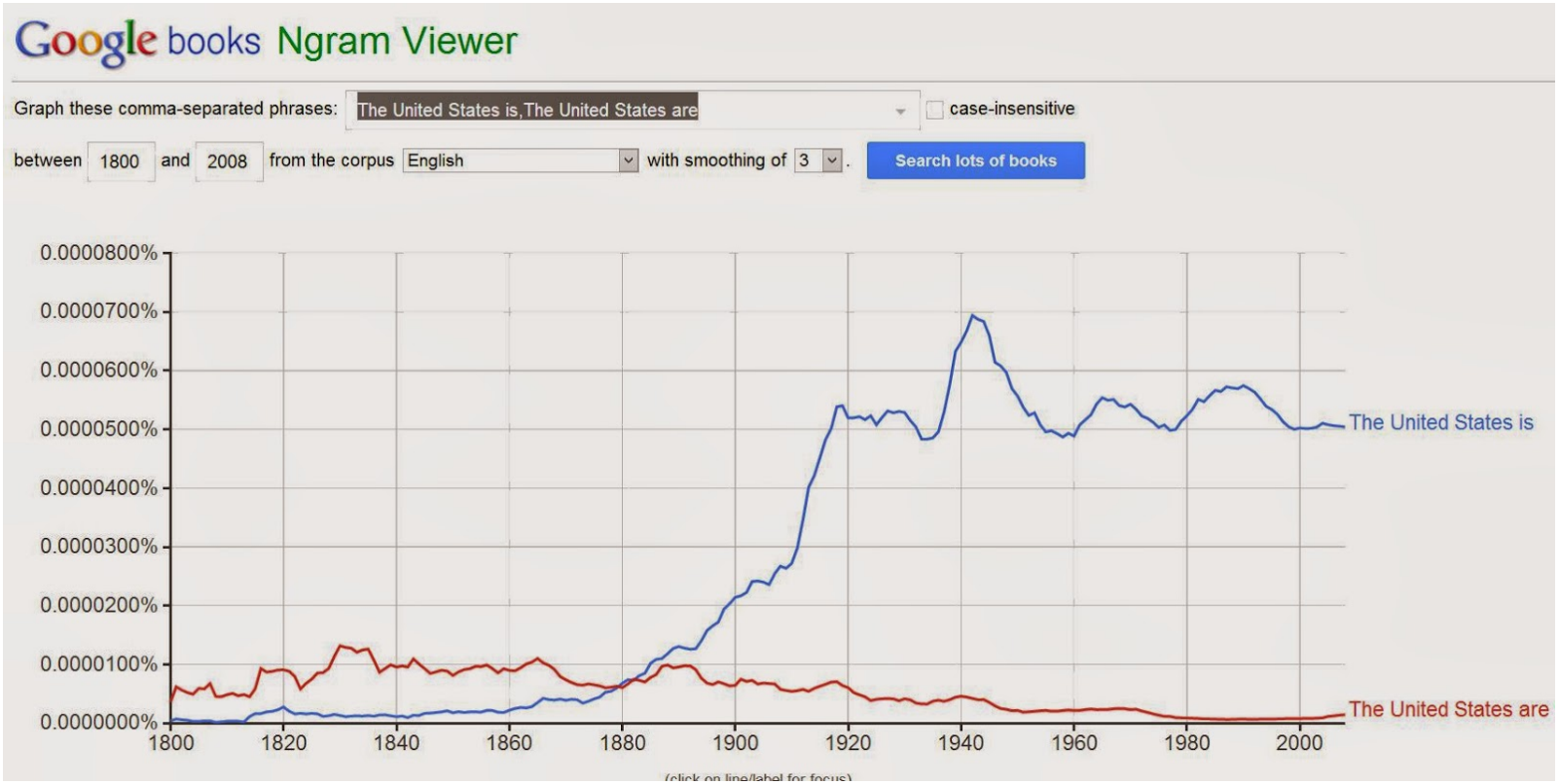
- An *n*-gram is a phrase with *n* contiguous words

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

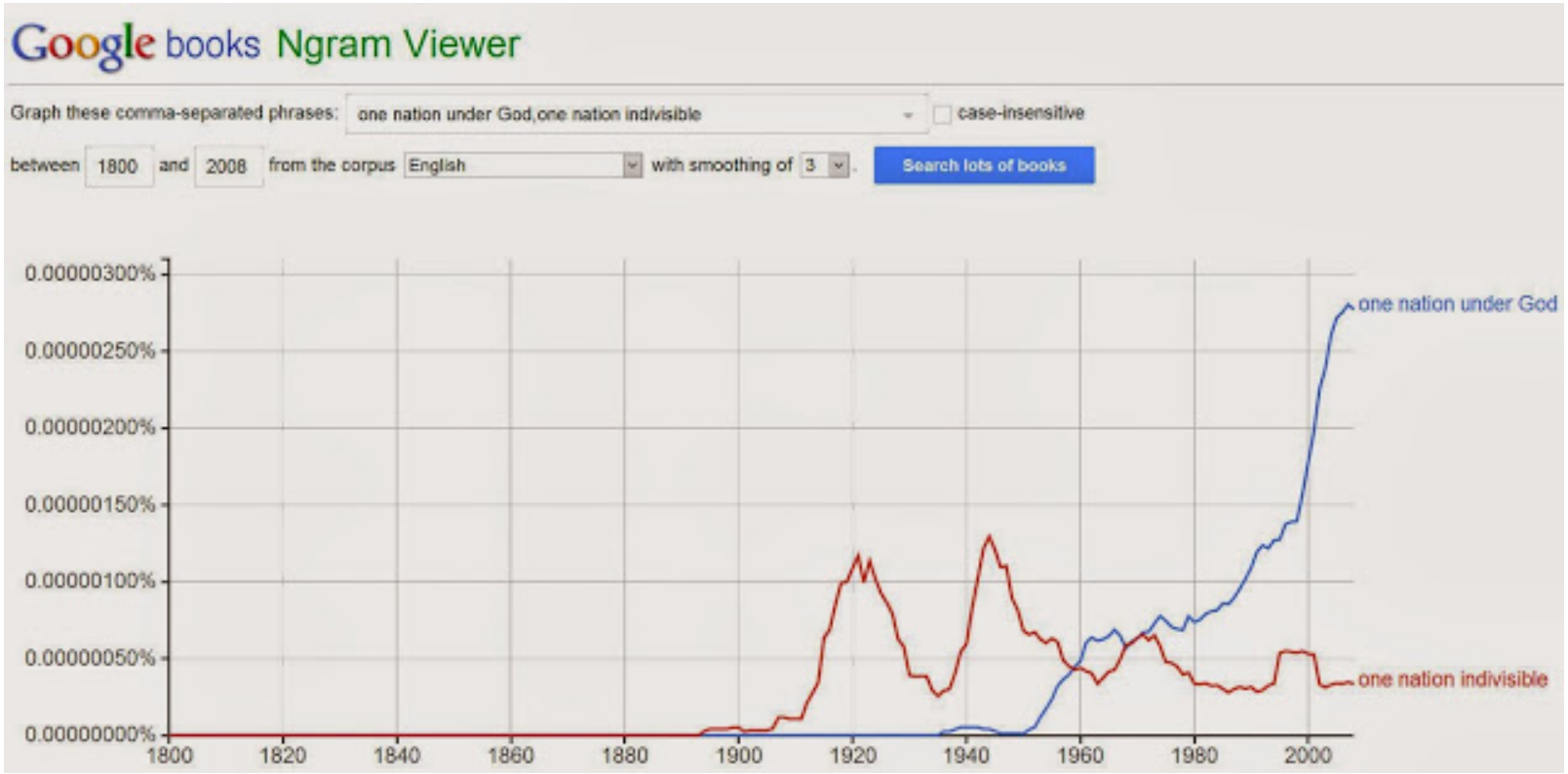
#	4-gram	Count	5-gram	Count	6-gram	Count
1	this is interesting because	1	this is interesting because this	1	this is interesting because this is	1
2	is interesting because this	1	is interesting because this is	1	is interesting because this is a	1
3	interesting because this is	1	interesting because this 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
  - $\Pr(is|this) = \frac{Count(this\ is)}{Count(this)} = \frac{500}{1000} = 0.5$
  - $\Pr(a|this\ is) = \frac{Count(this\ is\ a)}{Count(this\ is)} = \frac{125}{500} = 0.25$
- 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

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

Upcoming Deadlines

- Project 4:

Project 4			
Input Text Predictor: NGram Generation			
	NGram Generation	***Checkpoint***	***11:59PM 11/23/2014***

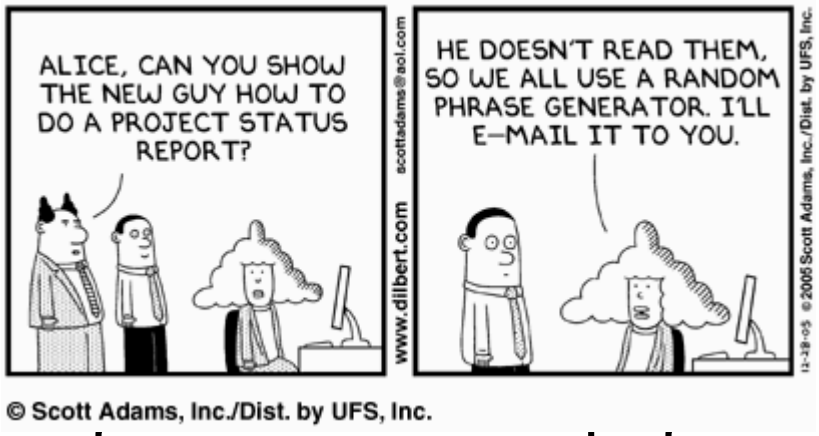
- 15-619 Project:
  - Phase 3 (last phase) is due on November 20th
  - Live-test will begin at 12:00 am, November 21th

TWITTER ANALYTICS: THE 619 PROJECT

Important Dates

- Phase 3 dummy test, Nov. 18
- Phase 3 test submission due 23:00 ET, Nov. 20
- Phase 3 report due 23:59 ET Nov. 21

Phase 3 Report [VERY IMPORTANT]



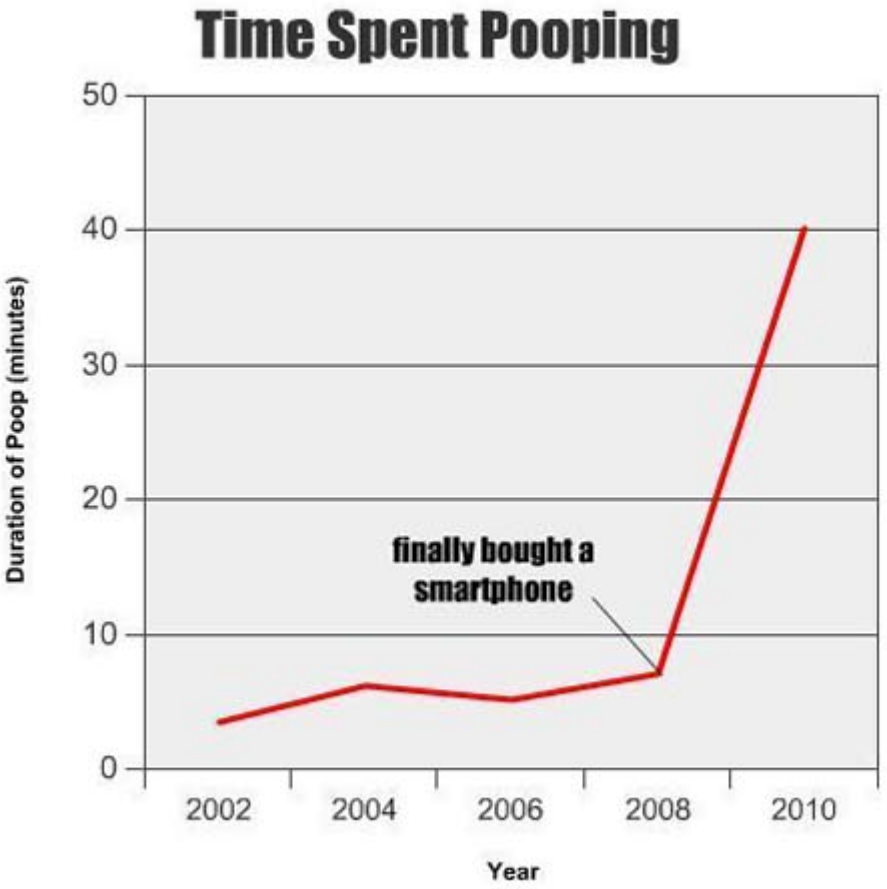
- Start early
- Document your steps
- Identify and isolate the performance impact of each change you make
- Document your ideas and experiments

MAKE A QUANTITATIVE, DATA-DRIVEN REPORT

Good reports contain Tables

Idea	***Expected Result***	***Submission id***	***Observed Result***	***Inference***
Use a Godzilla caching proxy server	***200% improvement in throughput (based on reference[1])***	***10022, 10024***	***No improvement (8k rps)***	***Caching proxy did not help as the front end server we use already caches responses***
Use ZooomZoom Web Server	***1 million rps (based on YouTube video [2])***	***11211-11217***	***25k rps***	***LG is not fast enough***
...	***...***	***...***	***...***	***...***

Good reports contain Charts/Graphs



Illustrate the effectiveness of your optimization technique by tweaking parameters and plotting the delta

Live Test

- 4-hour live test
  - 30 minute warm-up
  - 3 hours Q1-Q6
  - 30 minutes mix-Q1Q2Q3Q4Q5Q6
- $1.6 per hour
  - EC2 instance, EBS, ELB

Timeline (+/- delta) EST

Start Time	***End Time***	***What's Happening***
0000	***0030***	***Warmup***
0030	***0100***	***Q1***
0100	***0130***	***Q2***
0130	***0200***	***Q3***
0200	***0230***	***Q4***
0230	***0300***	***Q5***
0300	***0330***	***Q6***
0330	***0400***	***Mix***

Leaderboard (as of 3 AM on Nov 18)

Q1	***Q2***	***Q3***	***Q4***	***Q5***	***Q6***
apt143	***Transcendence***	***FDU***	***MJMCloud***	***FDU***	***SVM***
cloudreaper	***CumulusNimbus***	***Cyan***	***wanbaoC2***	***cloudlol***	***cloudlol***
FDU	***FDU***	***161Santiago***	***Cyan***	***CMUETC***	***FDU***

Exciting stuff !!!

How to do well in a Live Test?

- Don’t crash. If you do, recover fast!!!
- Make smart trade-offs (focus on your score)
- No benefit of database pre-caching
  - (unless you’re really smart)
- Hopefully you have simulated a Live Test
  - self-warmup
  - sequential 20 minutes

Grading

- The report really matters!!
  - Dense, not long
  - What you tried and what you measured matters
- 60% of the grade of the 619 Project
  - 45% Live Test
  - 15% Report
- But, improve a lot and you can partially cover up for a poor Phase 1 or Phase 2

Suggestions / Improvements

- UI
- Favicon
- Design
- Report comments
- Features
- Bugs
- Feedback

http://bit.ly/1roJsvU

Call for Teaching Assistants

- If you are interested in being evaluated for a TA position for S15 or F15
  - We will be releasing a TA application form soon
  - We will also be releasing the TA interview dates soon
    - Most likely around 11/24 & 11/25
- Being a TA is an excellent learning opportunity
  - Ask one of the TAs now
- You can work on a variety of teams
  - Project development
  - Testing system/scoreboard improvements
  - Cheat checking
  - Budget and tag checking
  - Grading and office hours
  - Etc…

Any questions?