Scalable LDA based Twitter Semantic Search

Semantic Technologies

https://github.com/semantic-group1/twitter-lda

Umang Patel
ujp2001

Sarthak Dash
sd2828

Ilambharathi Kanniah
ik2342

Motivation

- Twitter Data - realtime, news-breaking
- Semantic Search through Topic Modelling - for going beyond simple indexing based search.

- Scalability : Need to be able to scale to Web-sized dataset.
- Solution : Use Map-Reduce paradigm based on Hadoop !

Project

Scalable

Amazon S3

Hadoop based

Amazon EMR (MapReduce)

LDA based

Collapsed Variational Bayes

Apache Mahout

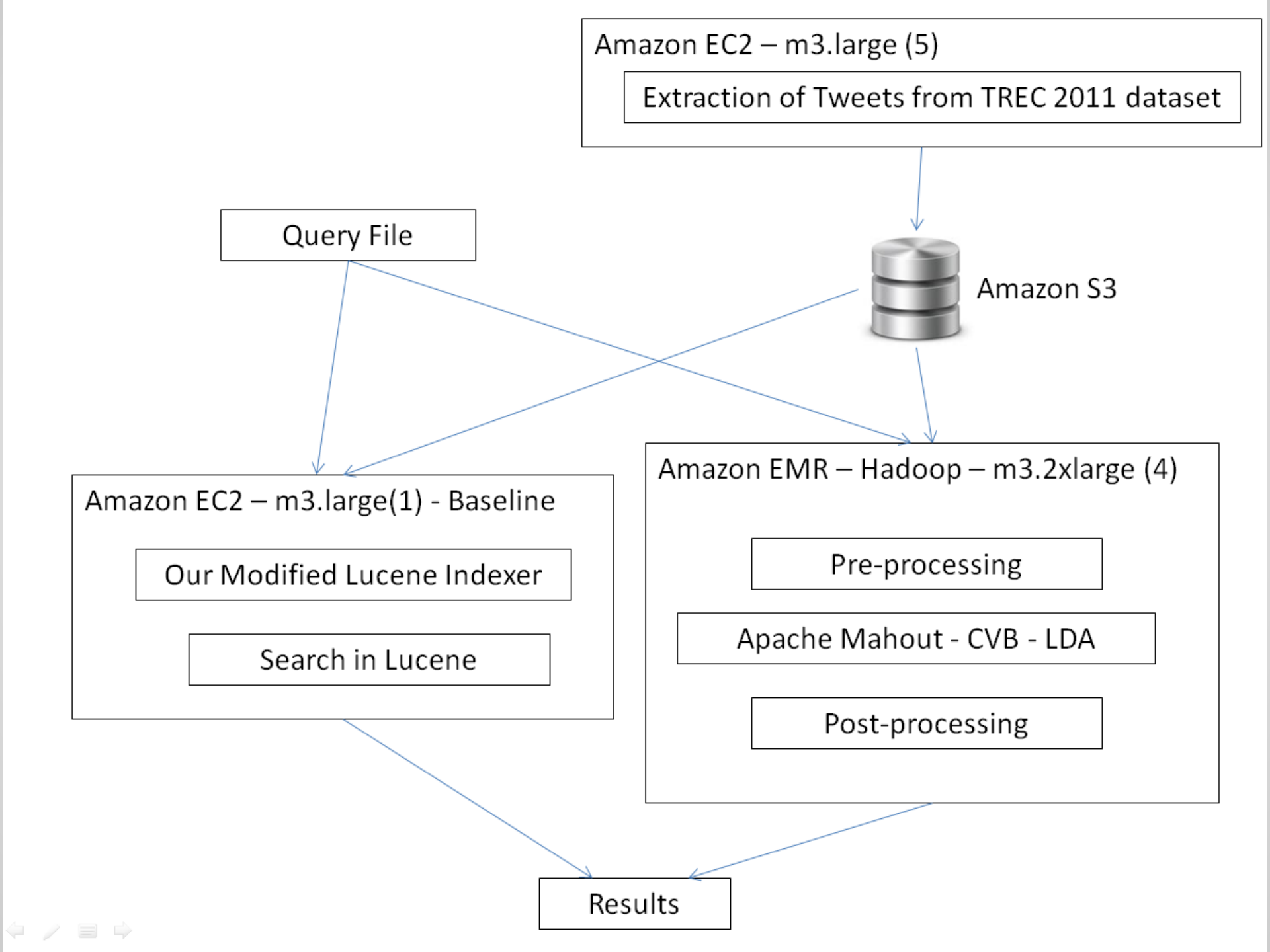
Twitter Semantic Search

Twitter Data

TREC Adhoc Task

Dataset and Evaluation

- TREC Microblog Track 2011 (same for 2012)
- Real-time Ad-hoc task
  - Returns the most recent and most relevant tweets to the query given.
- Evaluation - P@30
- Use the official evaluation scripts and gold data
  - EvalJig.py
  - mb-12eval.py
- Final P@30 for 60 test queries.

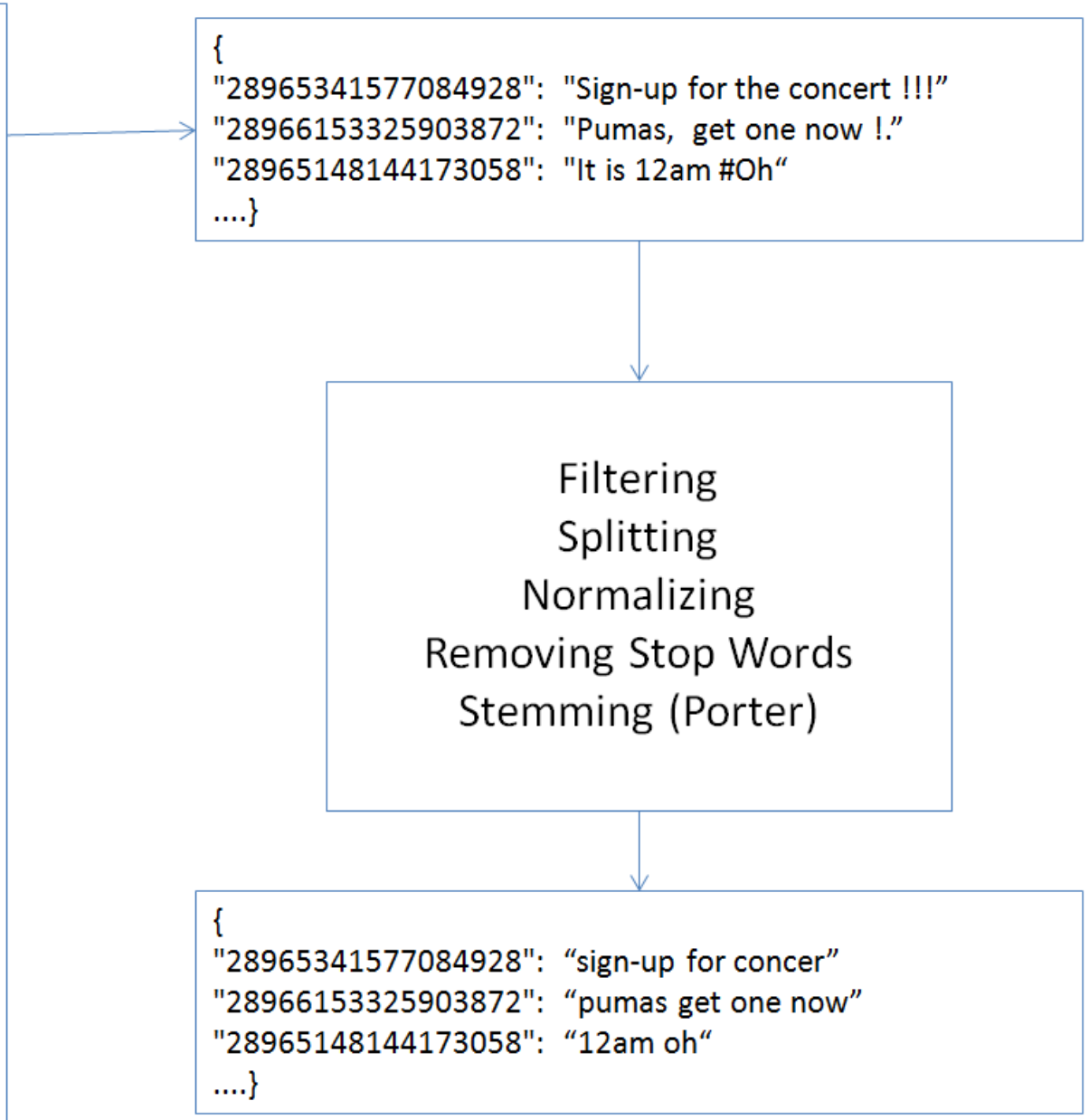

Amazon EC2 – m3.large (5)
Extraction of Tweets from TREC 2011 dataset
Query File
Amazon S3
Amazon EC2 – m3.large(1) - Baseline
Our Modified Lucene Indexer
Search in Lucene
Amazon EMR – Hadoop – m3.2xlarge (4)
Pre-processing
Apache Mahout - CVB - LDA
Post-processing
Results

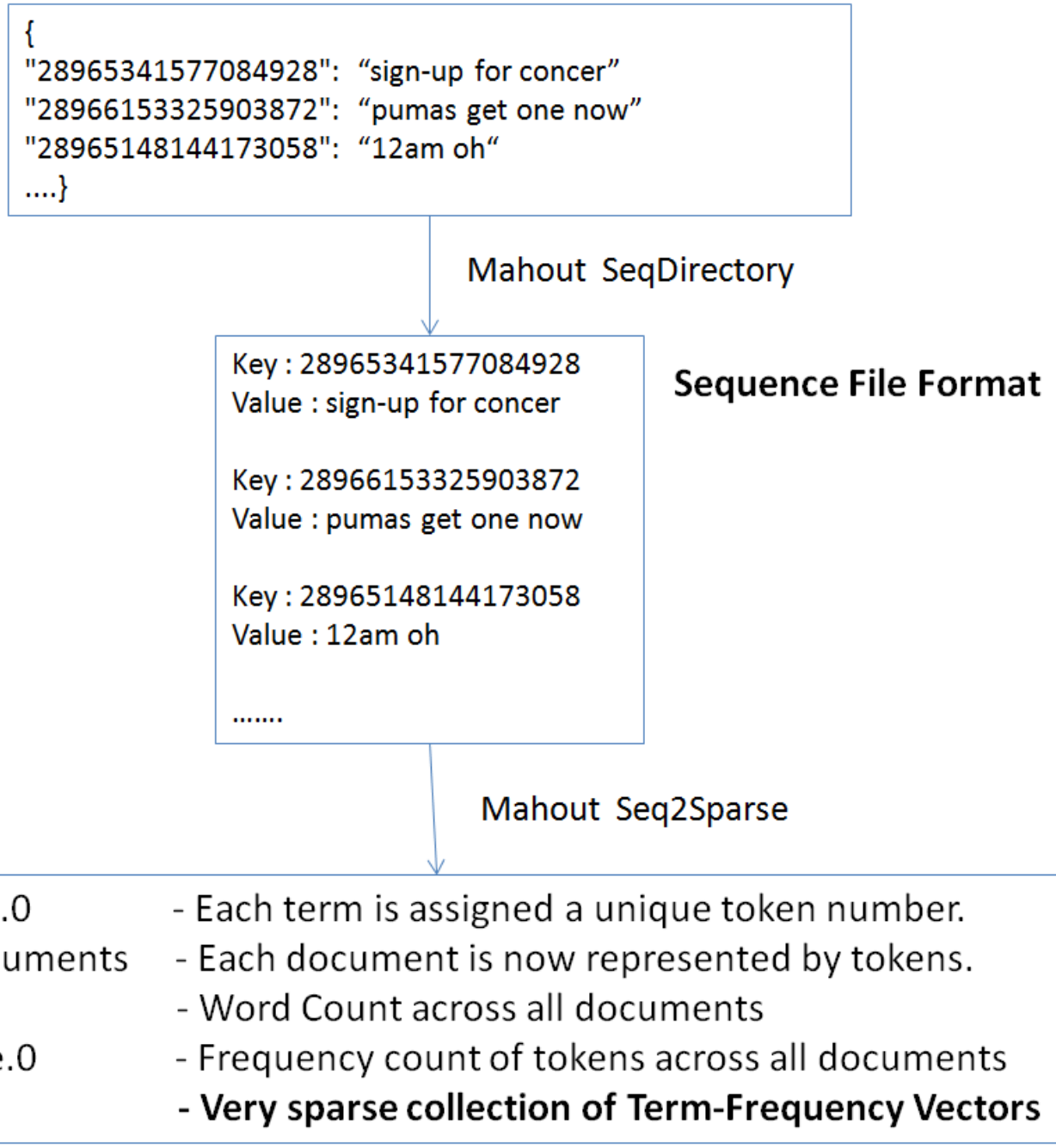
{"28965341577084928":
{"onlytext": "Sign-up for the concert !!!",
"user_id": "wibens46",
"text": "Sign-up for the concert !!!",
"hashtags": "",
"hyperlinks": "",
"file": "20110123-000",
"ptext": "Sign up",
"filenumber": "000",
"id": 28965341577084928,
"filedate": "20110123"},

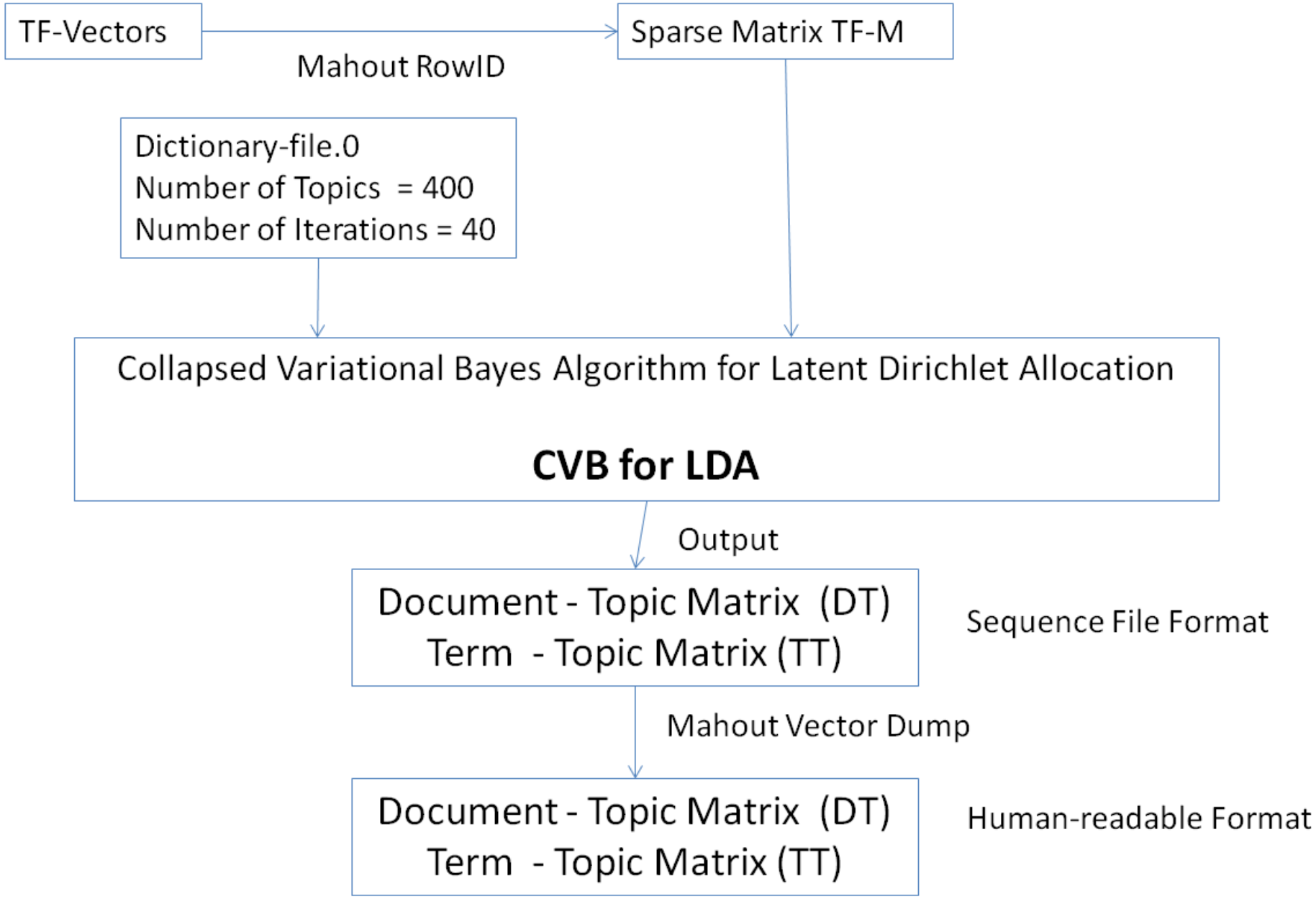
"28966153325903872":
{"onlytext": " Pumas, get one now !.",.",
"user_id": "cuetopo",
"text": Pumas, get one now !.", ",
"hashtags": "",
"hyperlinks": "",
"file": "20110123-000",
"ptext": "Puma cachun cachun",
"filenumber": "000",
"id": 28966153325903872,
"filedate": "20110123"},

"28965148144173058":
{"onlytext": "It is 12am ",
"user_id": "proc",
"text": "It is 12am #Oh",
"hashtags": "Oh",
"hyperlinks": "",
"file": "20110123-000",
"ptext": "It is 12am",
"filenumber": "000",
"id": 28965148144173058,
"filedate": "20110123"},

....}




{
"28965341577084928": "sign-up for concer"
"28966153325903872": "pumas get one now"
"28965148144173058": "12am oh"
....}
Mahout SeqDirectory
Key : 28965341577084928
Value : sign-up for concer
Key : 28966153325903872
Value : pumas get one now
Key : 28965148144173058
Value : 12am oh
.....
Sequence File Format
Mahout Seq2Sparse
Dictionary-file.0 - Each term is assigned a unique token number.
Tokenized Documents - Each document is now represented by tokens.
Word-Count - Word Count across all documents
Frequency-file.0 - Frequency count of tokens across all documents
TF-Vectors - Very sparse collection of Term-Frequency Vectors


TF-Vectors
Mahout RowID
Sparse Matrix TF-M
Dictionary-file.0
Number of Topics = 400
Number of Iterations = 40
Collapsed Variational Bayes Algorithm for Latent Dirichlet Allocation
CVB for LDA
Output
Document - Topic Matrix (DT)
Term - Topic Matrix (TT)
Sequence File Format
Mahout Vector Dump
Document - Topic Matrix (DT)
Term - Topic Matrix (TT)
Human-readable Format

Scoring using LDA

- Each query is converted into a query vector by taking the average of the term-topic vectors (output from LDA) for each of the term in the query.

- We compute Cosine Distance Similarity (CDS) between the query vector and Tweets-Topic vectors (output from LDA) in parallel and output top 1000 tweets ranked by CDS scores.

Interesting Stats !

- 11 million tweets in the TREC 2011 dataset
- Final Term-Topic Matrix - 5.8GB
- Running time
  - Preprocessing - Done when tweets were collected.
  - CVB for LDA - 4 hours / iteration. (24 map tasks, 12 reduce tasks, 3 core/task nodes + 1 master)
  - Post-processing - 3 hours.
  - Query Time - A few minutes (assuming models are loaded in memory already, parallelized CDS scoring)
- Java Heap Size - 20GB (on each node)

Implementation Issues

- We had to run 2 EMR clusters(4 machines each) of m3.2xlarge machines for one complete pass of LDA.
- Reason: Huge incompatibility between Apache Mahout and Hadoop. Ate up most of our development time. Examples !

- Setting up Hadoop environment variables correctly.

Services | Edit | Sarthak Dash | Oregon | Help

Elastic MapReduce | Cluster List | EMR Help

Create cluster | View details | Clone | Terminate

Filter: All clusters | Filter clusters ... | 9 clusters (all loaded)

Name	ID	Status	Creation time (UTC-4)	Elapsed time	Normalized instance hours
Watson_Cluster_20140505_seven	j-ZCTEVA16079I	Terminated User request	2014-05-08 04:46	3 days, 23 hours	3072
Watson_Cluster_20140505_six	j-1WD11F23YS97F	Terminated User request	2014-05-06 15:50	5 days, 12 hours	7980
Watson_Cluster_20140505_five	j-3O3L2K3O0INJ7	Terminated User request	2014-05-06 14:45	1 hour, 6 minutes	60
Watson_Cluster_20140505_four	j-2PAZISJRKZCK2	Terminated with errors Internal error	2014-05-06 05:32	3 days, 19 hours	2304
Watson_Cluster_20140505_three	j-1O5UUZFPCYK73	Terminated User request	2014-05-05 04:37	1 day, 1 hour	832
Watson_Cluster_20140505_two	j-21I8KXJ7TKL1M	Terminated User request	2014-05-05 04:27	7 hours, 16 minutes	256
Watson_Cluster_20140505	j-1GD8NS6UUUN7S	Terminated User request	2014-05-05 03:58	20 hours	672
Watson_Cluster	j-2Z30SYPMQ74Z0	Terminated User request	2014-05-04 00:00	1 day, 3 hours	84
My cluster	j-ULA5RJ598ZCE	Terminated User request	2014-04-23 17:05	17 minutes	3

 Privacy Policy | Terms of Use

Feedback

Results (after Hadoop hard work) ...

Method	P@30
Official Baseline	7.06
Our Lucene Baseline	**9.10**
CVB for LDA	**8.71**

We did not expect P@30 for LDA to be so low. But...

But, here are some issues we believe might be the reason for above result

- Twitter Data
  - 140 characters, brevity
  - Smileys, short-forms, URLs
  - Some tweets are not grammatical
  - Hashtags

Off-the-shelf LDA doesn't work well with Tweets (Zhao, Wayne Xin, et al. "Comparing twitter and traditional media using topic models." *Advances in Information Retrieval*. Springer Berlin Heidelberg, 2011. 338-349.)

- Not all non-English tweets were removed (didn't use an API but language detection)

- Tweet-Topic matrix approximation considered. Explanations !

Other Approaches tried ...

- Moved away from Apache Mahout and tried out a different version of scalable LDA called as Mr.LDA.

- It solved the third issue mentioned in the previous slide, but had ran 10x more slowly than Apache Mahout implementation.

- So, we didn’t proceed with it further.

Where to from here … ?

- Trying this idea out on Twitter dataset, but after removal of all non-english tweets.
- Trying the same scalable LDA based approach, but on actual documents (bigger than 140 characters).
- Tried out the same on Wikipedia dump (smaller size); LDA models worked pretty well.

Conclusions

- We built a web-scalable LDA model using MapReduce paradigm.
- In order to get good semantic search results, extremely careful tuning is required (at Preprocessing step for Tweets)

Thanks. Questions ?