LDA: Extracting Topics from Tweets and Webpages

Sarunya Pumma
Xiaoyang Liu

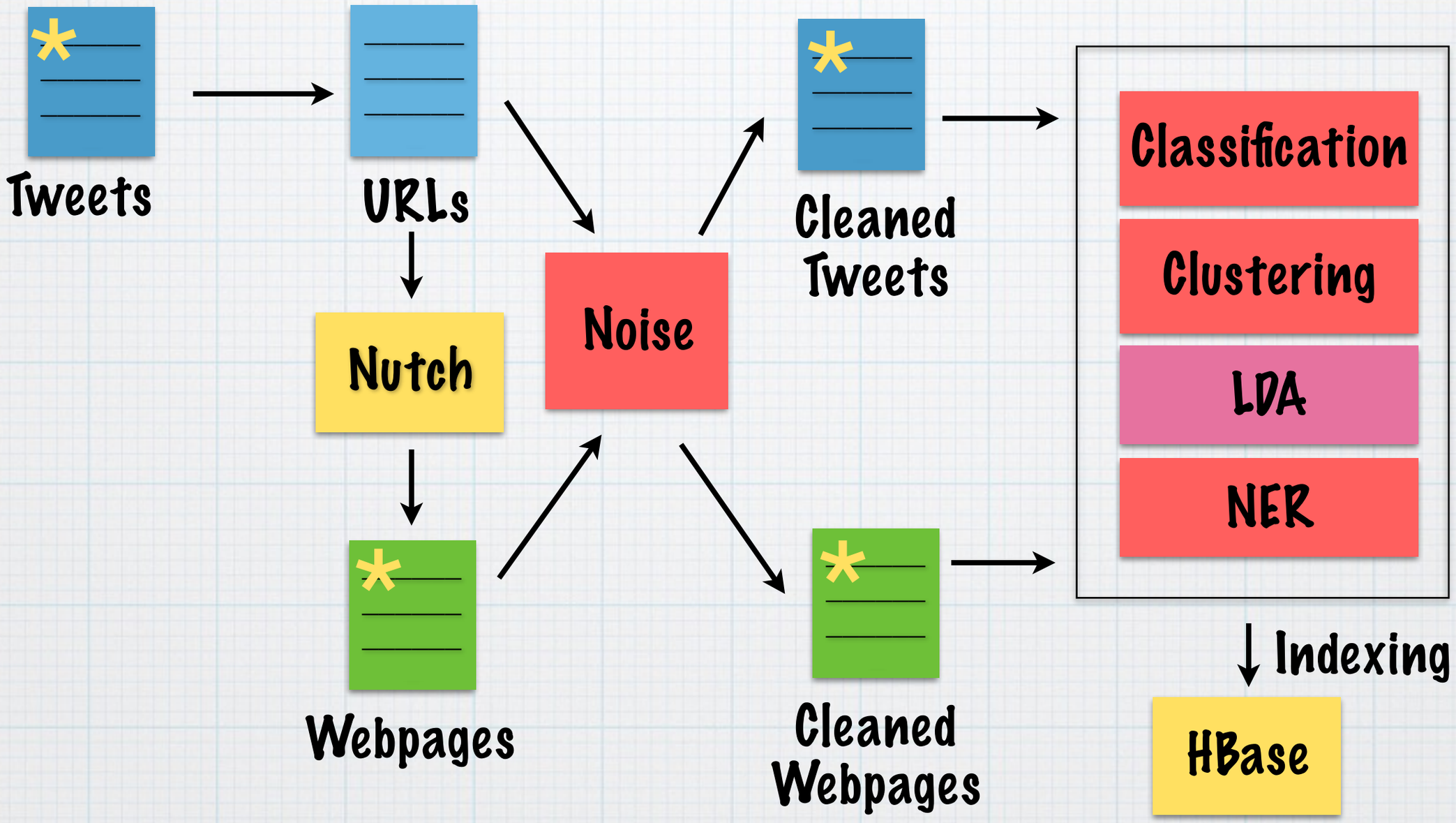
CS5604 Information Retrieval
Instructor: Dr. Edward Fox
April 30, 2015

Outline

* Introduction
  * The IDEAL System
  * LDA
* Design and Implementation
  * Data Preparation
  * Topic Extraction
  * Data Storing
* Evaluation
  * Human Judgement
  * Evaluation Against the Clustering Team
* Conclusion

Introduction

The IDEAL Project



Inter-dependence between teams

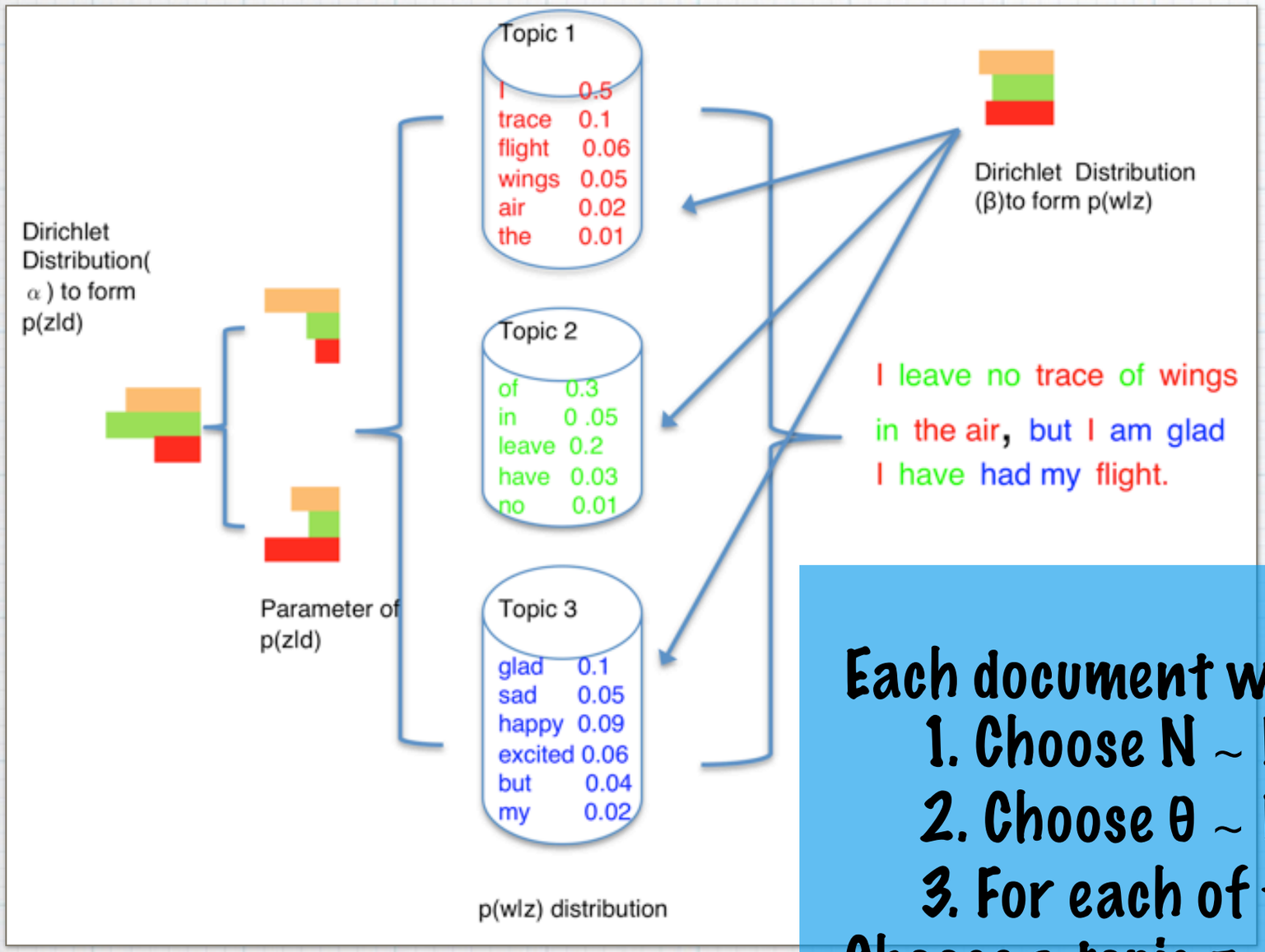
What is LDA?

* Latent Dirichlet Allocation (LDA) is a generative topic model
* Each document is a mixture of topics
* Each word is attributable to one of the document's topics

Why is LDA Useful?

* Extract topics of documents and words associated with topics
* Find semantical similarity between documents
* Unsupervised method: training model is not needed
* Dimensionality reduction: lower-dimensional representation for documents in corpus

How Does LDA Work?



Each document w in a corpus D:

1. Choose N ~ Poisson(ξ).
2. Choose θ ~ Dir(α).
3. For each of the N words wn:

Choose a topic z_n ~ Multinomial(θ).

Choose a word w_n from $p(w_n|z_n,\beta)$, a multinomial probability conditioned on the topic.

How Does LDA Work?

Example:

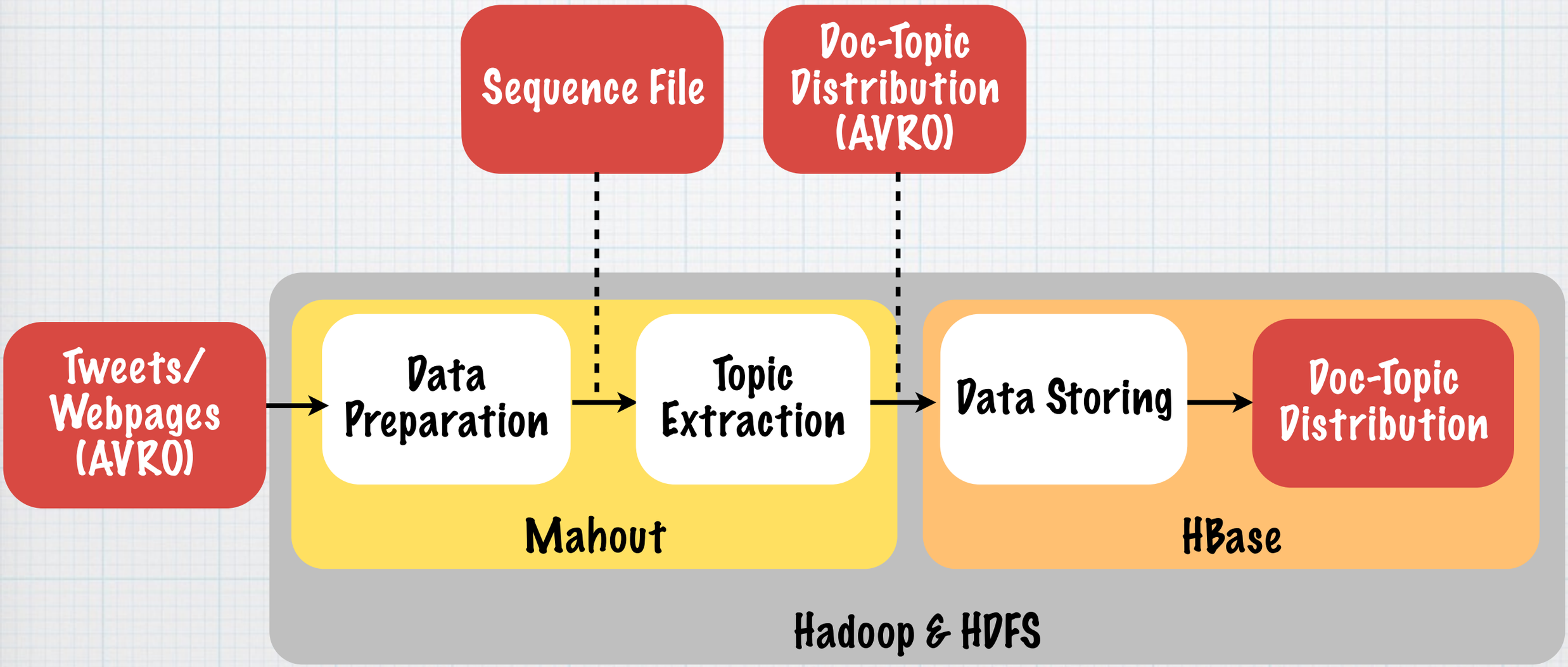
Sentence 1: Monkeys like banana
Sentence 2: Cats like fish
Sentence 3: Banana and apple are popular fruits

sentence 1 contains 50% topic 1 and 50% topic 2
sentence 2 contains 100% topic 2
sentence 3 contains 100% topic 1

Topic 1 (plant): banana apple orange grape ...
Topic 2 (animal): cat fish tiger lion ...

Design & Implementation

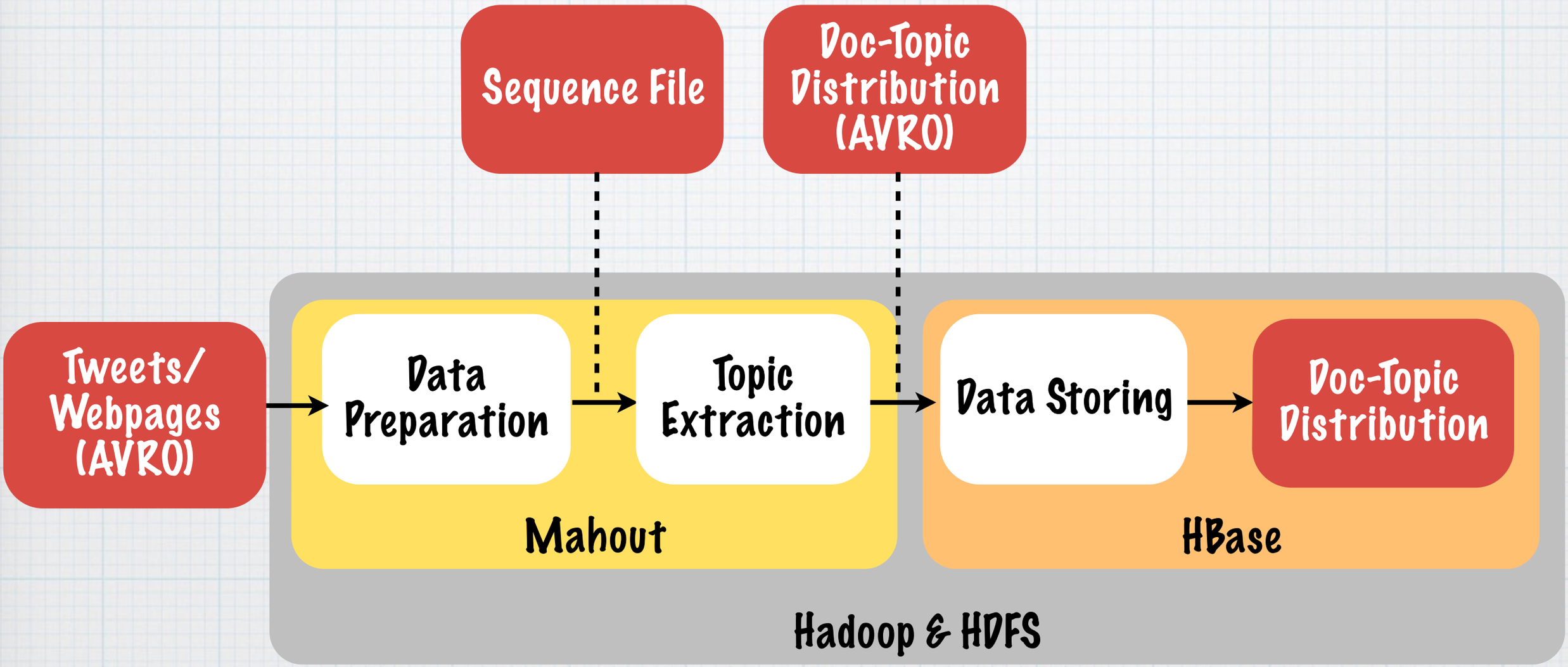
Architectural Design



(1) Data Preparation

* Tweets
  * Use the cleaned data
  * Use the Java program to convert AVRO file to sequence file
* Webpages
  * Crawl the webpages based on the URLs
  * Use Mahout to convert webpages to sequence file

Architectural Design



(2) Topic Extraction

* Implementation
  * Java using Mahout library (CVB - Collapsed Variational Bayesian Inference)
* Steps
  1. Convert the sequence file to a sparse vector based on TF-IDF
  2. Decompose the vector to singular value decomposition vectors (SVD)
  3. Run CVB algorithm on SVD vectors

(2) Topic Extraction

* Output
  * Document-topic distribution
* Structure (N topics)

$\{topic_1:P_1(topic_1), topic_2:P_1(topic_2), \ldots, topic_N:P_1(topic_N)\}$

.
.
.

$\{topic_1:P_M(topic_1), topic_2:P_M(topic_2), \ldots, topic_N:P_M(topic_N)\}$

} M documents

(2) Topic Extraction

How many topics for each collection?

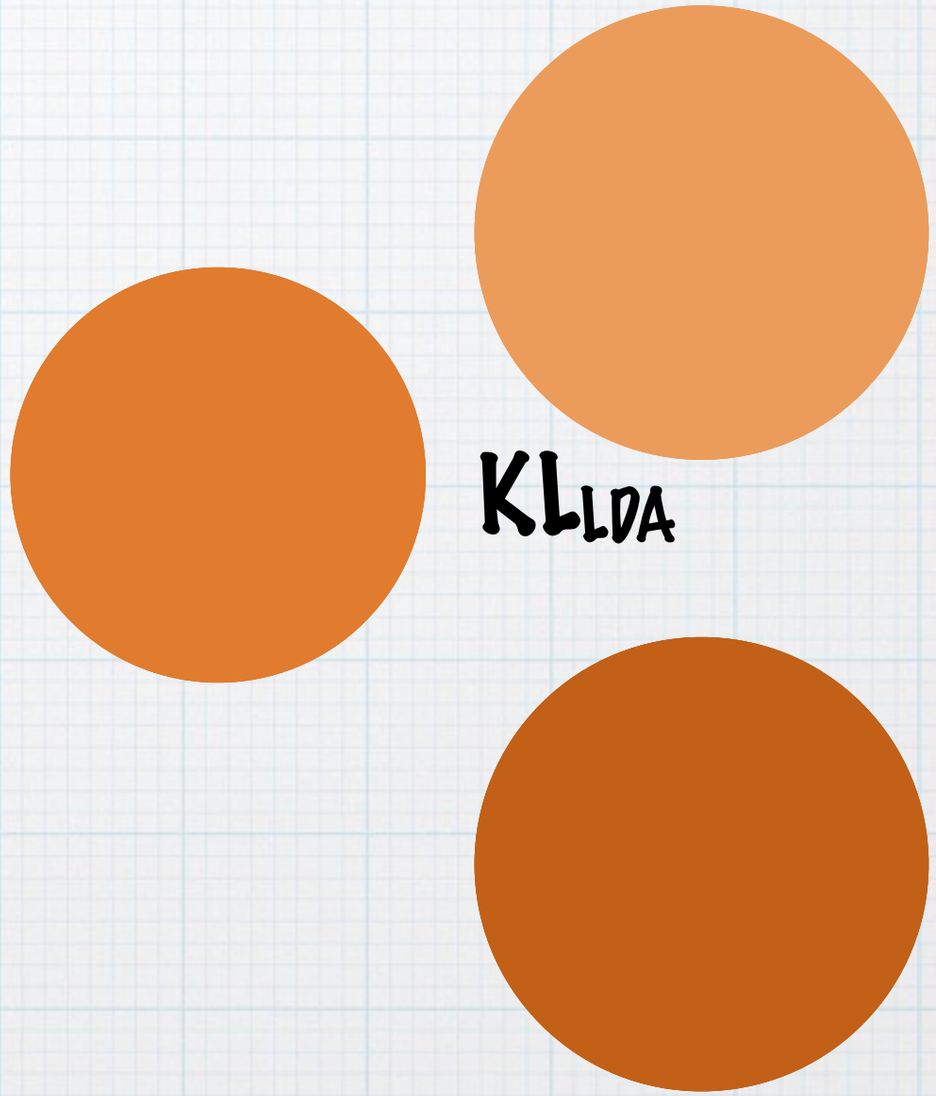
Empirical study!

How to evaluate the quality of LDA?

Kullback Leibler (KL) Divergence

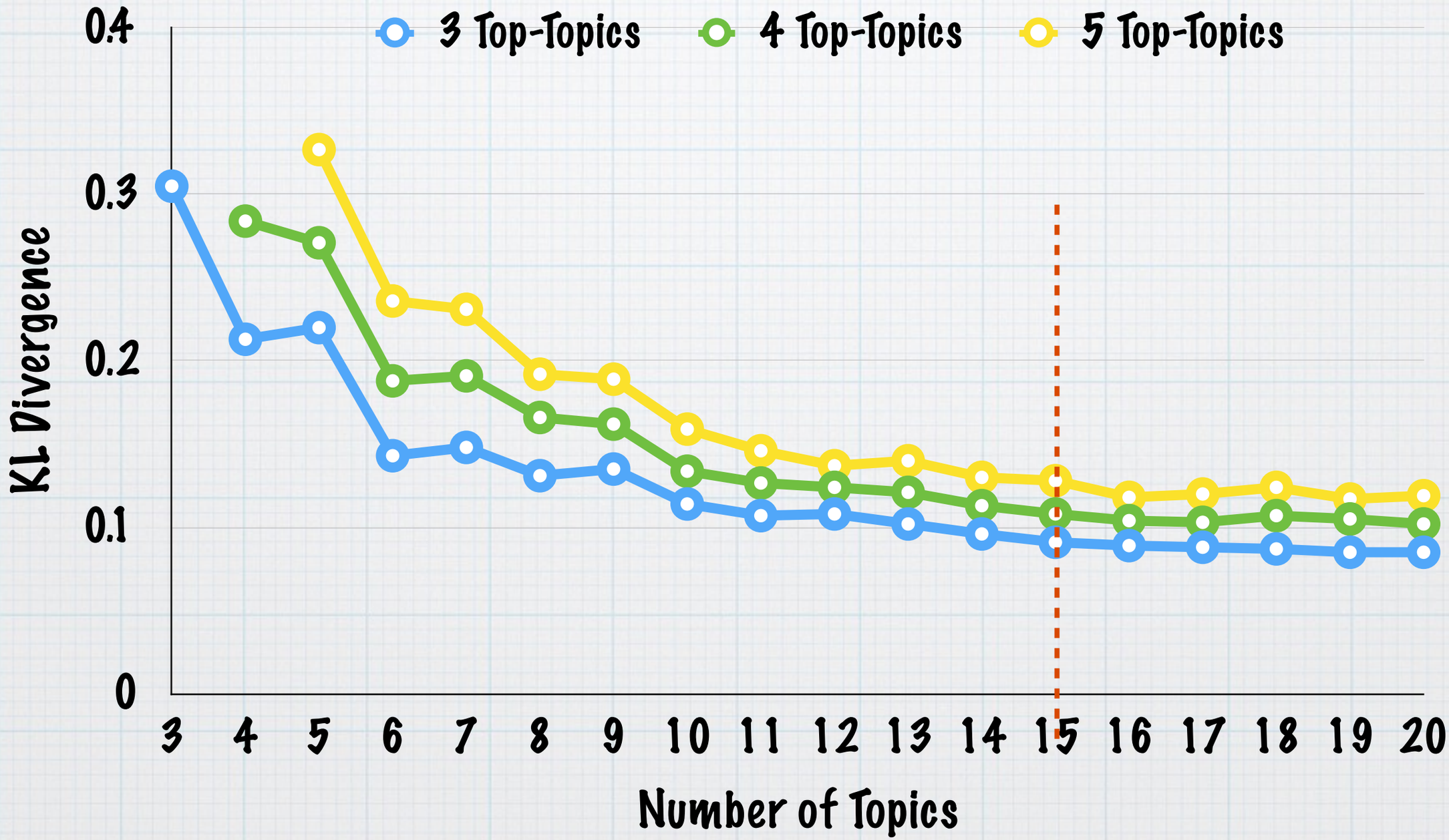
(2) Topic Extraction

* LDA Evaluation
  1. Group the documents based on X top topics
  2. Compute the average KL divergence in the cluster
  3. Compute the average KL of all the clusters



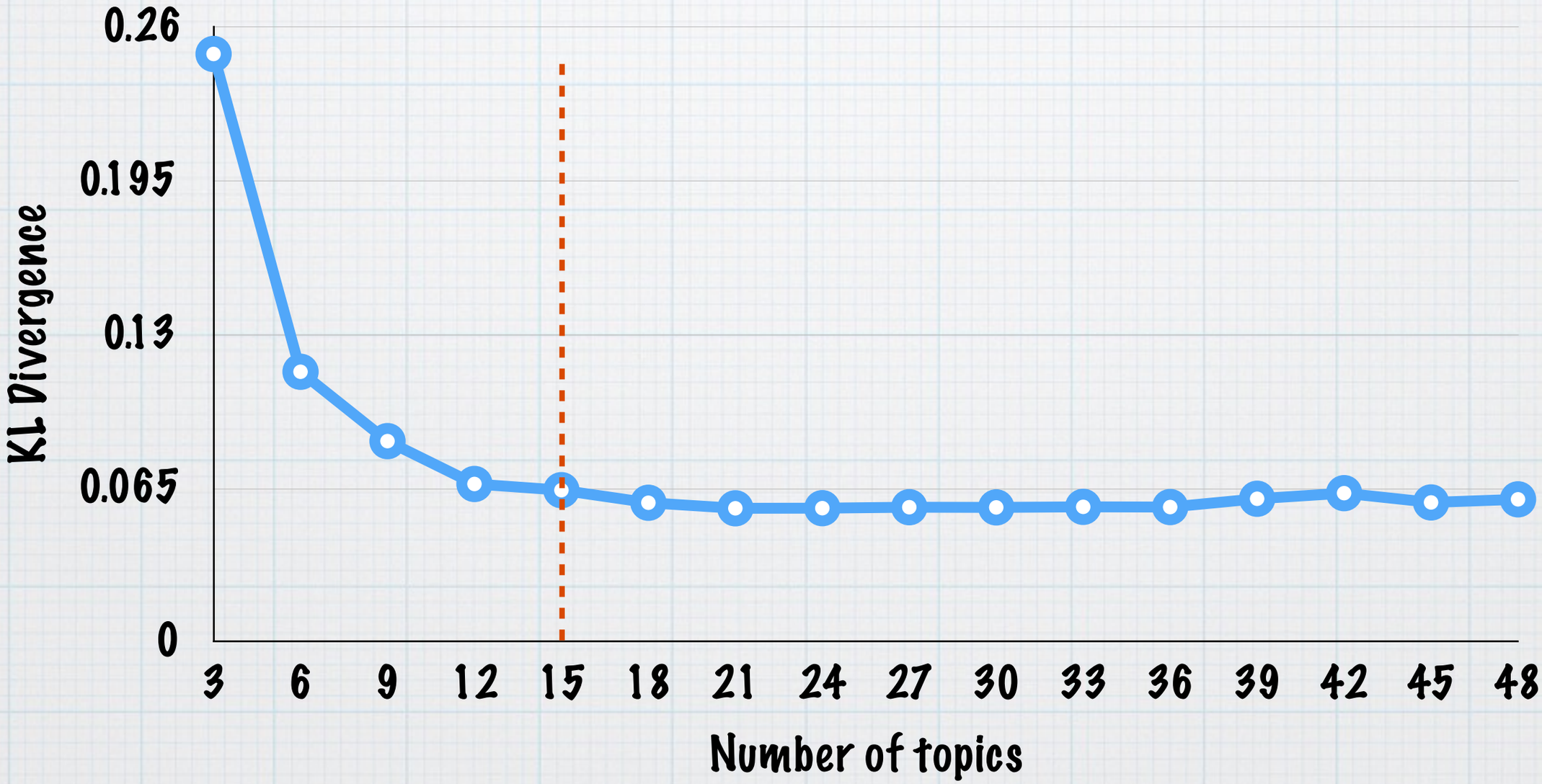
(2) Topic Extraction

* Small collection

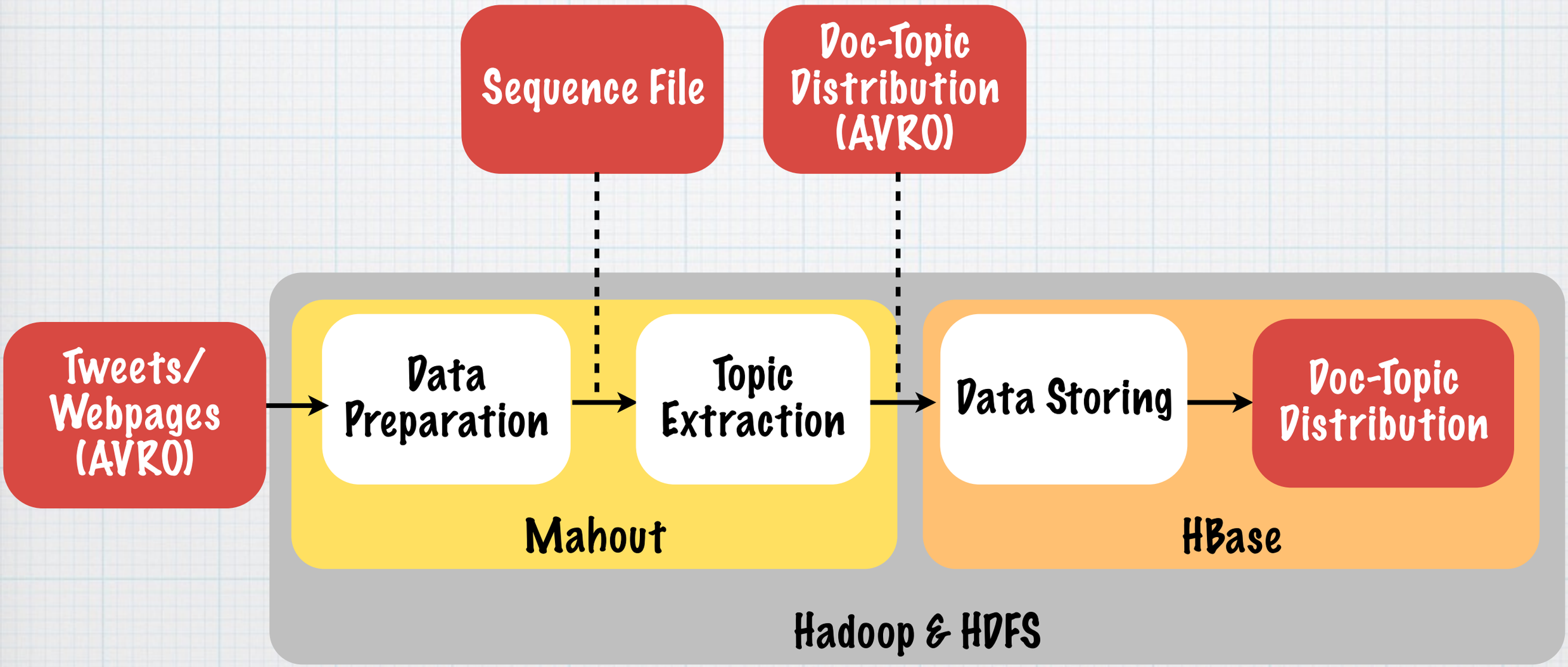


(2) Topic Extraction

* Large collection (Top topics = 3)



Architectural Design



(3) Data Storing

* On-going process
* We have talked to the Hadoop team
* The output will be converted to AVRO file
* It will be loaded into HBase using the script provided by the Hadoop team

Evaluation

Human Judgement

Word intrusion

$$MP_k^m = \sum_s 1(i_{k,s}^m = w_k^m)/S$$

w (m,k) is the intruding word among the words of k th topic generated by mth model.

i(m,k,s) is the intruder selected by subject s on the testing set.

S is the number of subjects

find the word which doesn't belong with the others

{dog, cat, horse, **apple**, pig, cow}

Human Judgement

Topic intrusion

$$TLO_d^m = (\sum_s \log \Theta^m_{d,j^m_{d,*}} - log \log \Theta^m_{d,j^m_{d,s}})/S$$

θ is the possibility of the topic in the documents

Index * means it is the true intruder

index s, j donate to the intruded topic selected by subjects

Finding the topics which don't consistent with the others

Higher the value of TLO and MP, greater the correspondence

Evaluation Against the Clustering Team

* On-going process
* Measurement: Cosine Similarity
* Comparison: T-Test
* Hypotheses
  * H0: $\mu_{LDA} - \mu_{Clustering} > 0$
  * H1: $\mu_{LDA} - \mu_{Clustering} <= 0$

Conclusion

Conclusion

* We have applied LDA on tweets and webpages
* The number of topics for each collection is determined by the empirical study
* Topics quality will be evaluated by human judgement and cross validation with the Hadoop team
* The results from the remaining works will be presented in the final report

:)

-LDA team