juxt smart mandate

Introduction to Text Mining & Analytics

IIM Lucknow
January 2017

By

Jaidev Deshpande
Practice Lead – Data Sciences,
Juxt-SmartMandate Analytic Solutions Pvt Ltd.

Get in touch :
jaidev@smartmamdate.com

Contents

Today's Agenda

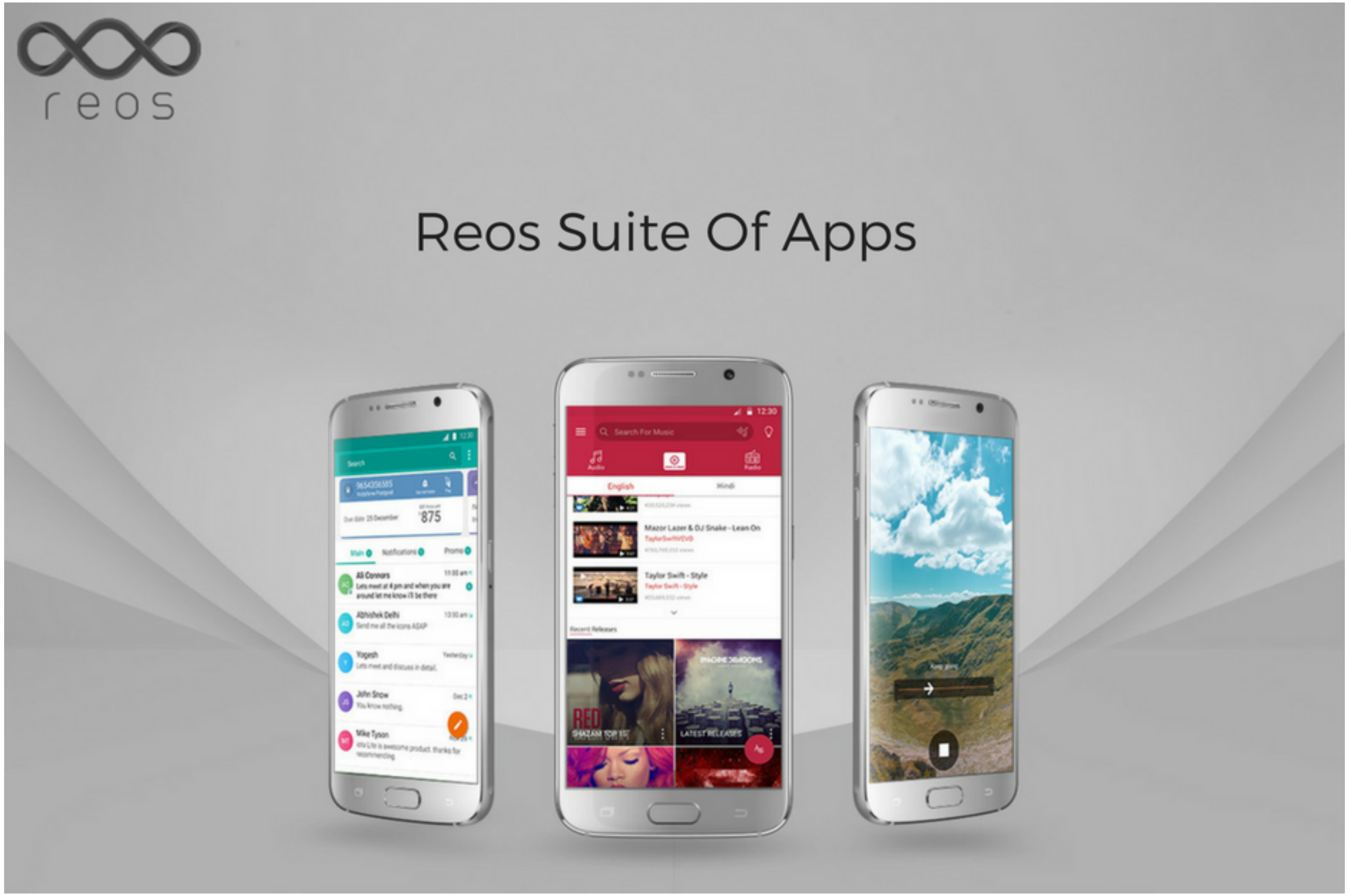
Motivation and Examples (20 minutes)

Basics of Web Scraping with Exercises (30-45 minutes)

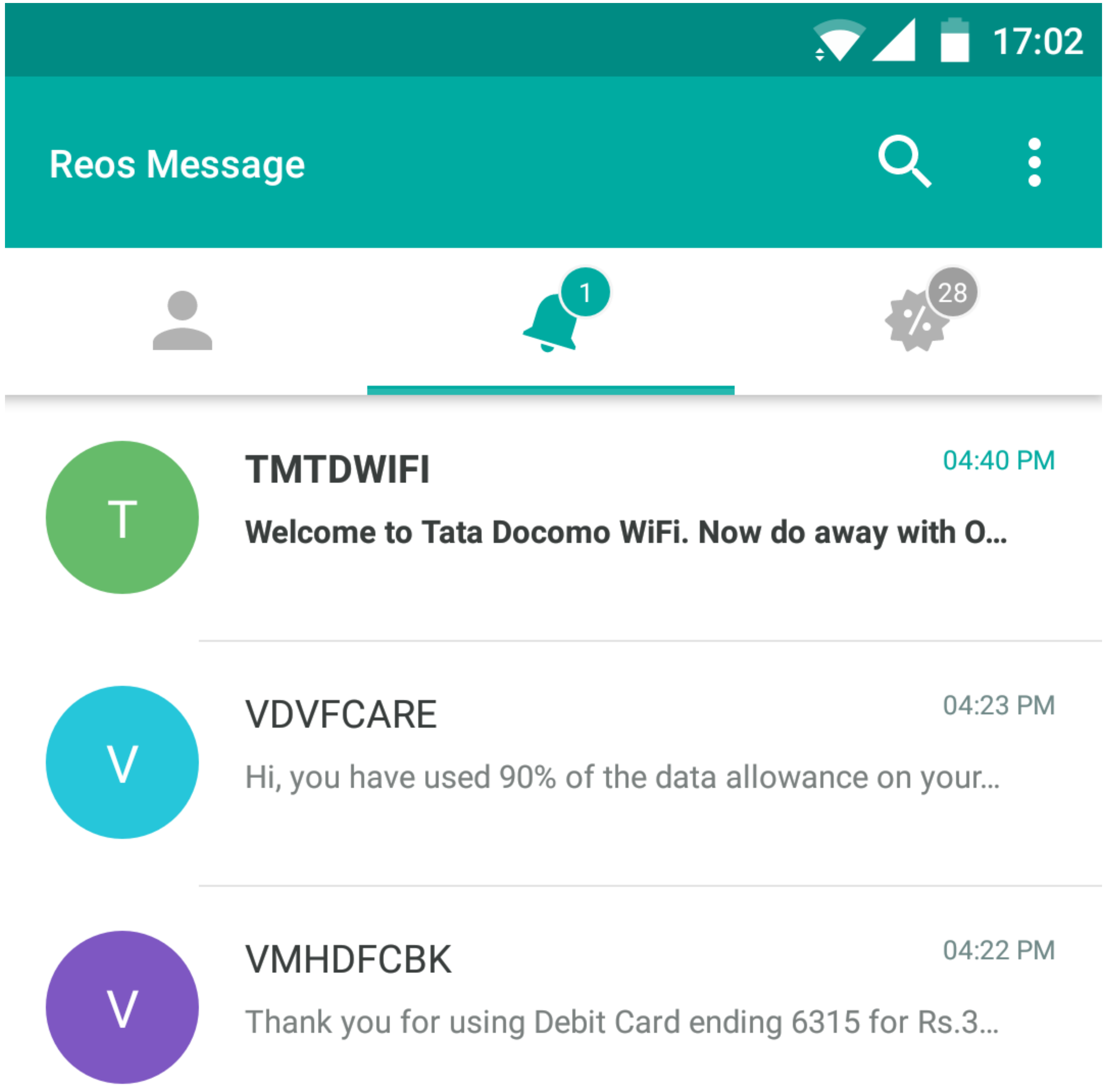
Exploratory Data Analysis & Visualization (30-45 minutes)

Building a model and drawing inferences (30-45 minutes)

Document Classification Example: The ReosMessage App



Example: The ReosMessage App



TMTDWIFI 04:40 PM
Welcome to Tata Docomo WiFi. Now do away with O...

VDVFCARE 04:23 PM
Hi, you have used 90% of the data allowance on your...

VMHDFCBK 04:22 PM
Thank you for using Debit Card ending 6315 for Rs.3...

(Simple) Document Classification

1. Documents are simply an unordered collection of words.

2. The frequency and not the sequence of words determines the characteristics of documents.

3. A “document” can be quantified into a vector – each element is a measure of the frequency of a word.

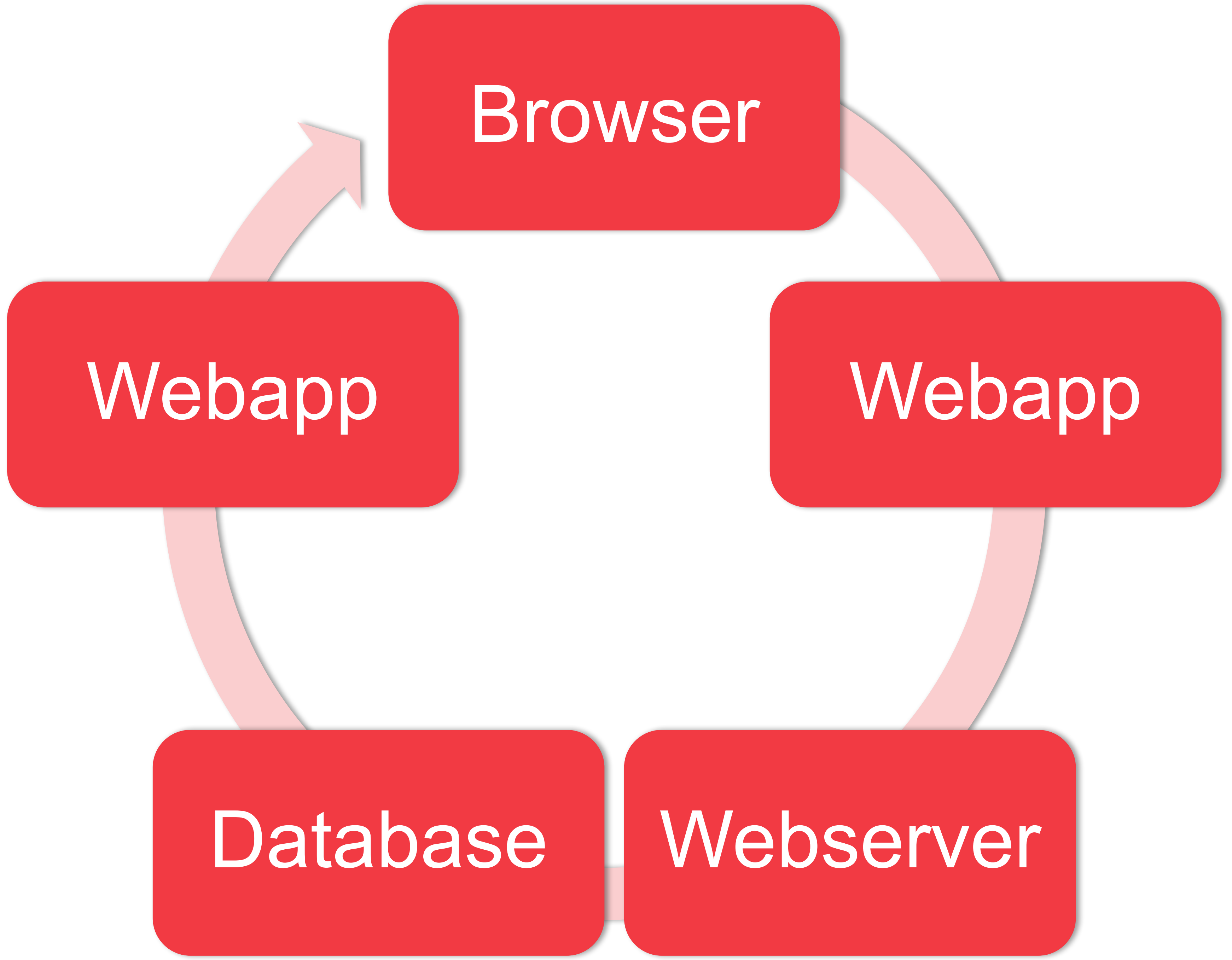
4. A “corpus” is a collection of documents.

5. The “vocabulary” of the corpus is the set of unique words in the whole corpus.

Getting Text from the Web

- Most websites are CRUD applications
- All databases support CRUD
- The WWW uses HTTP to implement CRUD

Operation	HTTP Request
Create	PUT / POST
Read	GET
Update	PUT / POST
Delete	DELETE



Problem: Why can’t you always use a browser?

Working Exercise: Predicting Talk Selection

- Data Sources:
  - The Junction Platform: https://github.com/pythonindia/junction
  - The Funnel Platform: https://github.com/hasgeek/funnel

- Problem:
  - Given historical data, can we predict the probability of a proposal being selected?
  - What can we infer about the talk selection process from what the model has learnt?

Exercises!

Head over to:

http://github.com/jaidevd/iiml_textmining