Text Mining Studio ETDs

Hannah Brinkley
Robert Rielage
Jason Vaanjilnorov
Christian Dy

CS 4624: Multimedia, Hypertext, & Information Access
Professor Edward A. Fox
Virginia Tech, Blacksburg, VA 24061

05/05/2020

Outline

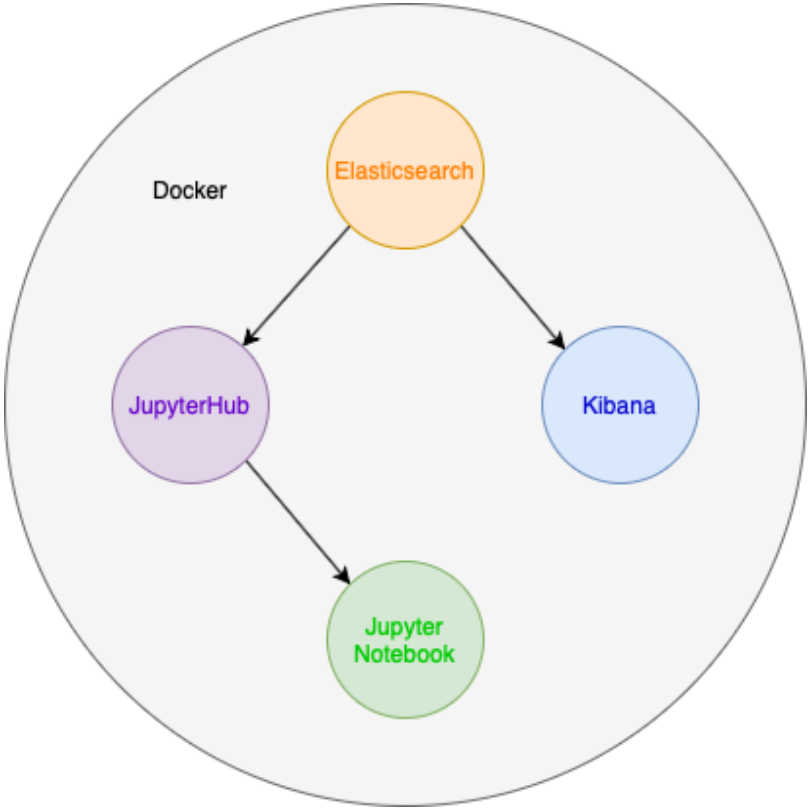
- Project Review
- Project Design
- Deliverables
- Project Demonstration
- Challenges
- Timeline/ Workflow
- Extensibility
- Acknowledgements
- References
- Questions

Project Review

- ETDs: theses and dissertations that have been converted into electronic documents
  - Ideal to used for data analytics, text mining, etc.
- Building a centralized tool for analysis of these ETDs
- Provide user-friendly interface for researchers to perform analysis

Project Design

- ElasticSearch data library backend
- Jupyter Notebook IDE frontend
  - To query and analyze ElasticSearch data
- JupyterHub acts as the launch point for the software system
- Docker containerizes all parts of the software
  - Easy deployment across systems
- Kibana provides visualization of the data within ElasticSearch



Deliverables

1. The full-text from corpus of ETDs, indexed in ElasticSearch
2. A working JupyterHub environment that will allow users to query and work with the data stored in ElasticSearch in a Python Jupyter Notebook
3. Documentation, installation scripts, and Docker containers for project components
    - Indexed in a GitHub Repository

Project Demonstration

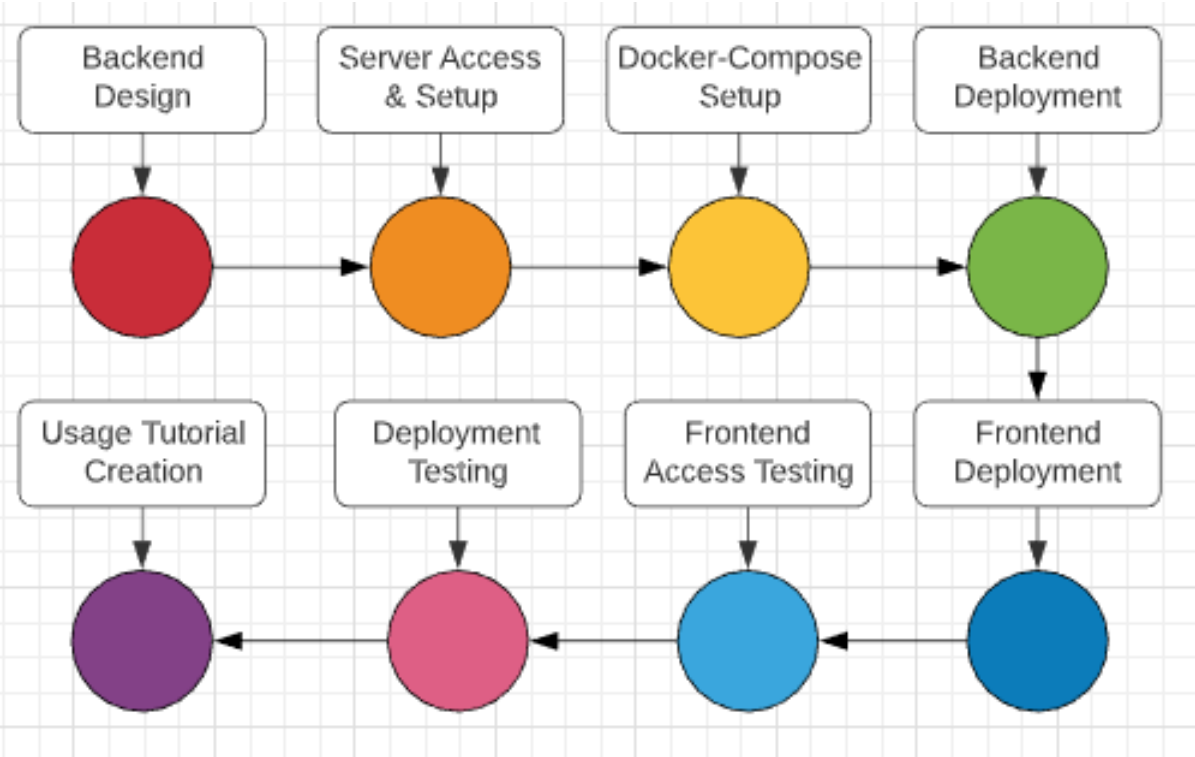
Please watch the Demonstration video
(TextDataMiningStudioETDAnalysisToolDemonstration.mp4)

This video can also be found at the following URL:
https://www.youtube.com/watch?v=90Q8FFhjfkI

Challenges

- COVID-19 communication complexities
- Anaconda auto-update issues
  - Reduced ease of deployment
- Failure to accomplish document ingestion in meaningful quantities
- Overlap with existing efforts

Workflow Timeline



Extensibility

- Better Kibana implementation
- Complete document ingestion
- More sophisticated user tutorials

Acknowledgements

- Client: Bill Ingram, Asst. Dean and IT Director, University Libraries
  - waingram@vt.edu
  - IR server access and navigation
  - Docker Compose tutorial
- Dr. Edward A. Fox
  - Providing graduate student contact information
  - Providing access to VT git group with access to Kibana
- Funding
  - IMLS LG-37-19-0078-19

References

- 2020Project-TextDataMiningStudioETDs
  - https://canvas.vt.edu/courses/104585/pages/2020project-textdataminingstudioetds
- Docker JupyterHub Package
  - https://hub.docker.com/r/jupyterhub/jupyterhub/
- Collection Management of Electronic Theses and Dissertations (CME) CS5604 Fall 2019
  - http://hdl.handle.net/10919/96484
- British Libraries Humanities Workbench Elasticsearch Repository
  - https://github.com/BL-Labs/hwb_elasticsearch

Questions?