MICHIGAN STATE UNIVERSITY

Project Plan

Text Classification of Seller Forums Content

The Capstone Experience

Team Amazon

Maxime Goovaerts
Carl Johnson
Luke Pritchett
Benjamin Taylor
Johnny Zheng

Department of Computer Science and Engineering
Michigan State University

Spring 2015

From Students…
…to Professionals

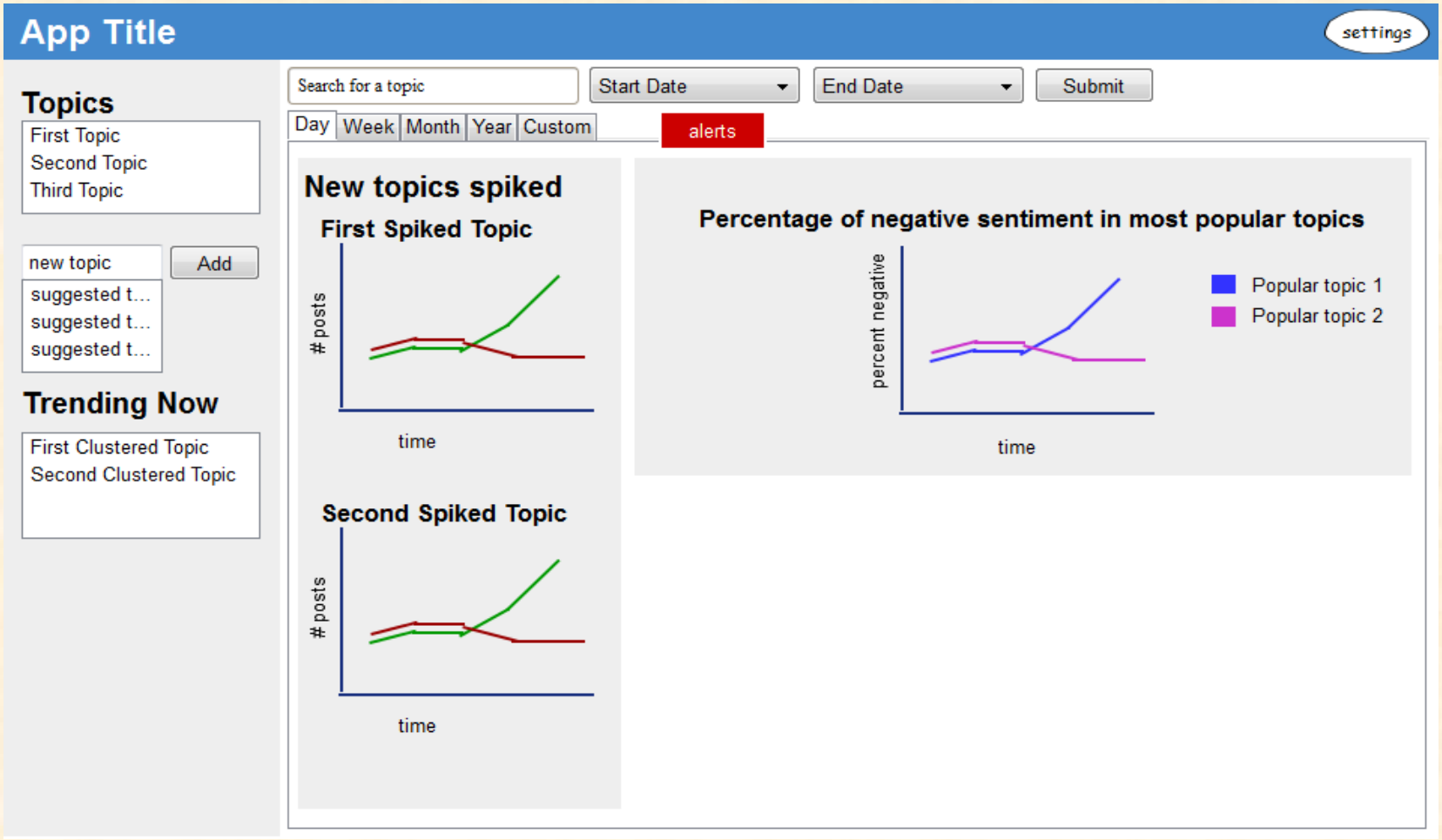
Functional Specifications

- Unlock the value of 3rd party Seller Forums
- Data Organization and Analysis
  - Classification
  - Clustering
  - Sentiment
- Dashboard
  - Graphs- Trending Topics
- Notifications

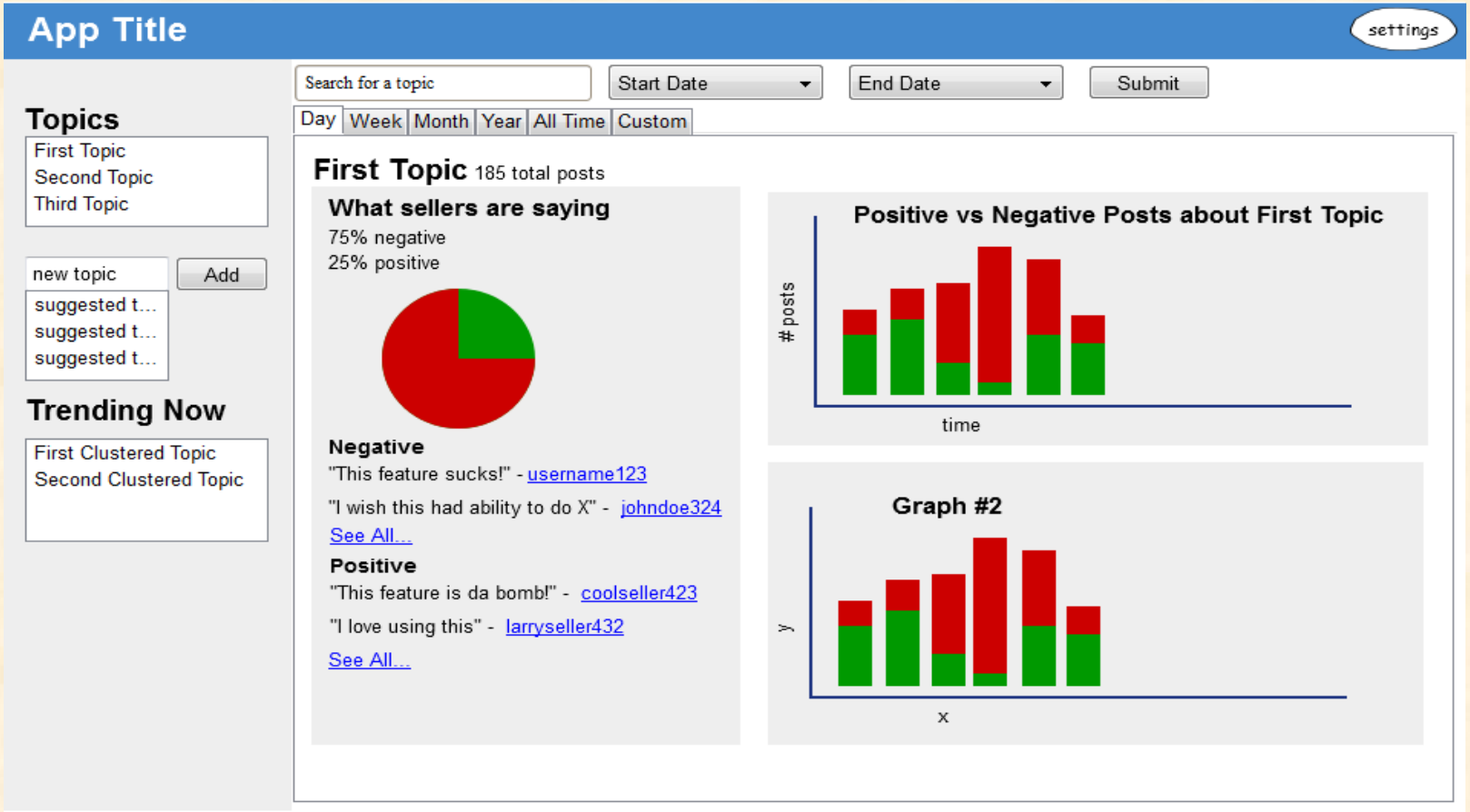
Design Specifications

- Designed for Amazon Employees
  - To visualize the data analysis
- General Analytics
  - The entire seller forums data
- Specific Topic Analytics
  - Each specific topic can be analyzed alone
- Settings Page
  - Add alert instances, add, edit and delete topics

Screen Mockups: *General Analytics*



Screen Mockups: *Specific Topic Analysis*

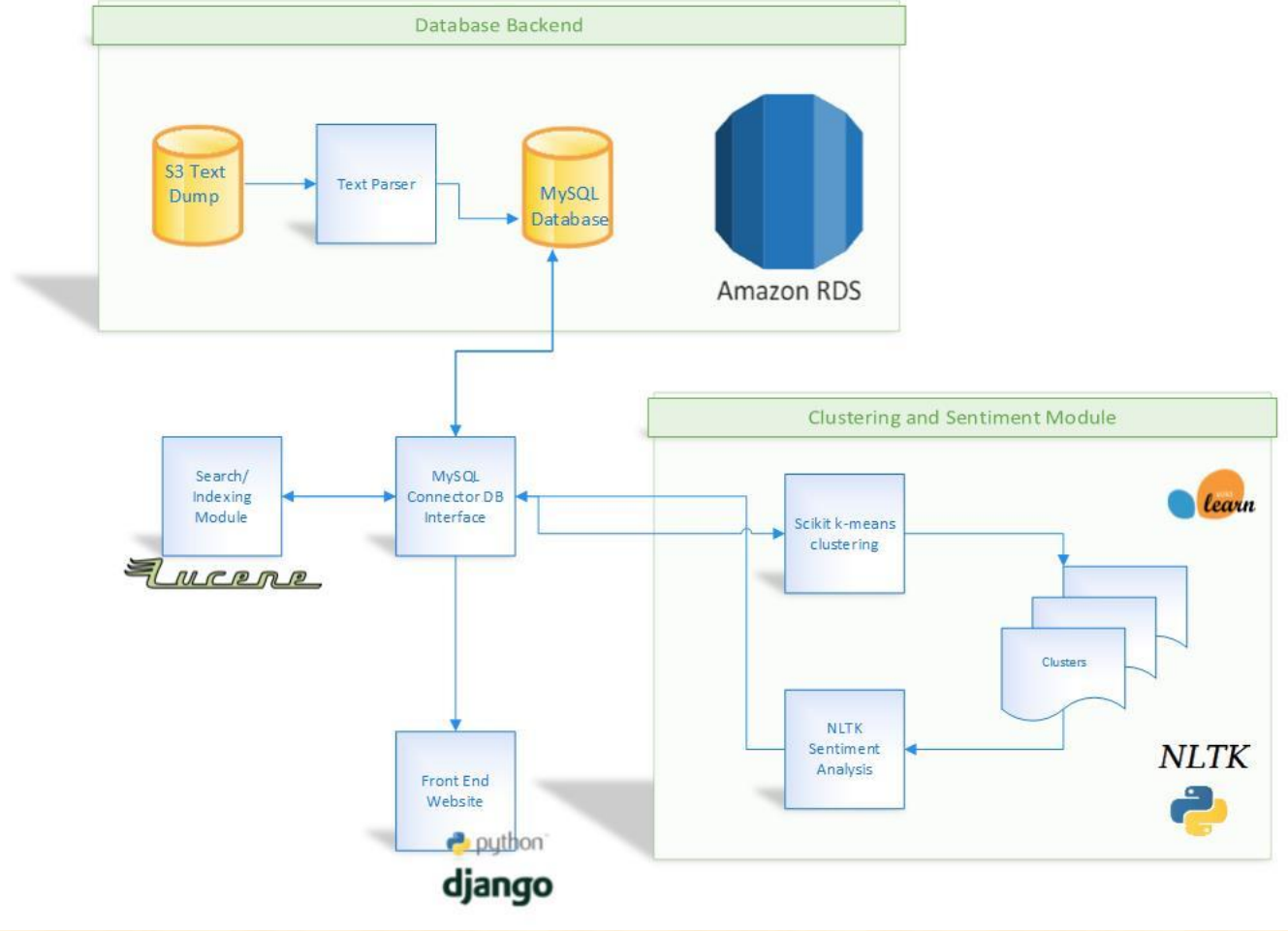


Screen Mockups: *Settings*

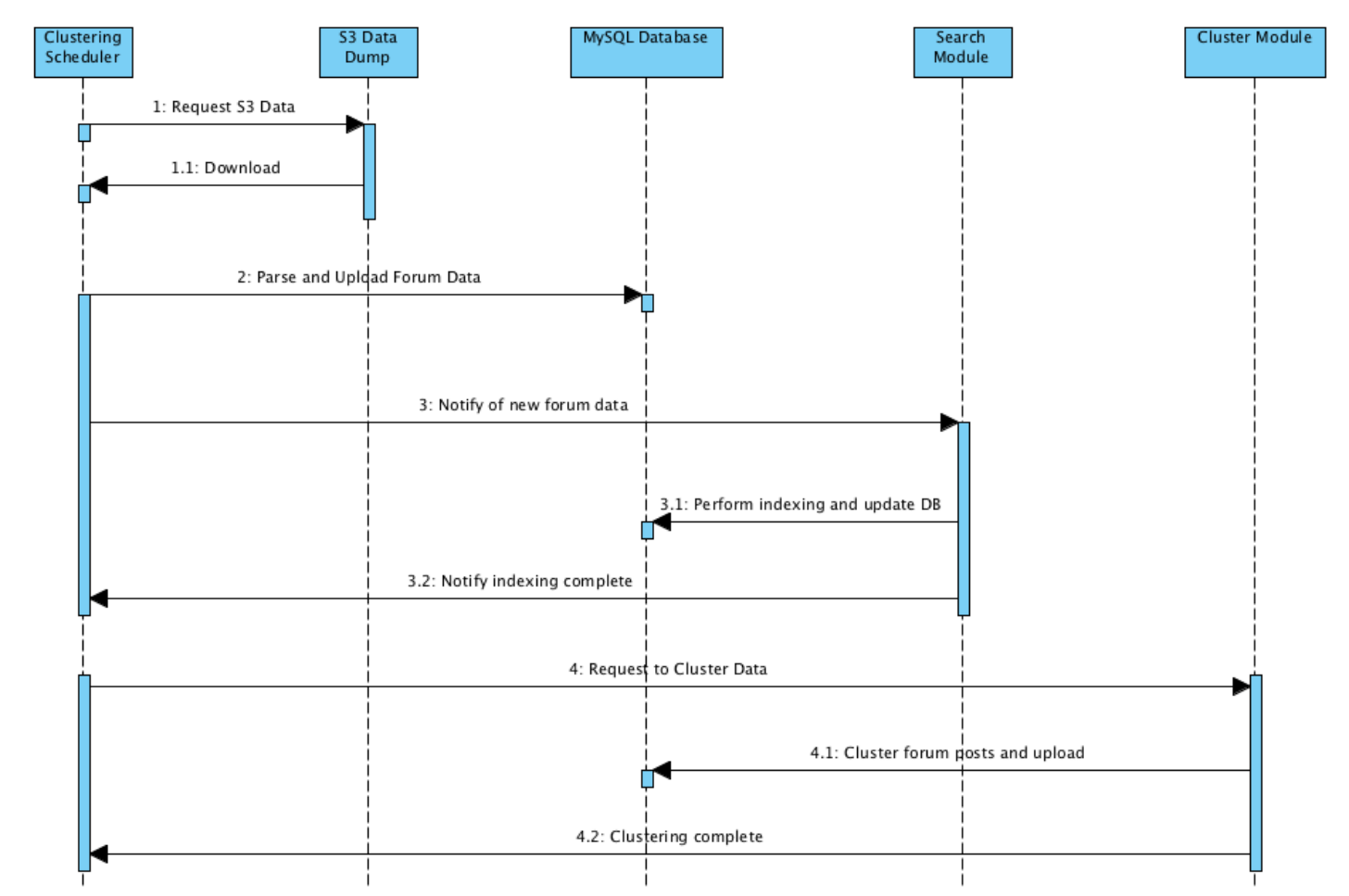
Technical Specifications

- Consume Seller Forums Content in AWS
  - From Text File to SQL DB
- Data Analysis
  - Clustering and Sentiment Module
  - Search Module
- Dashboard
  - Create reports/graphs of forums volume by topics and terms over time
  - Store trending topics

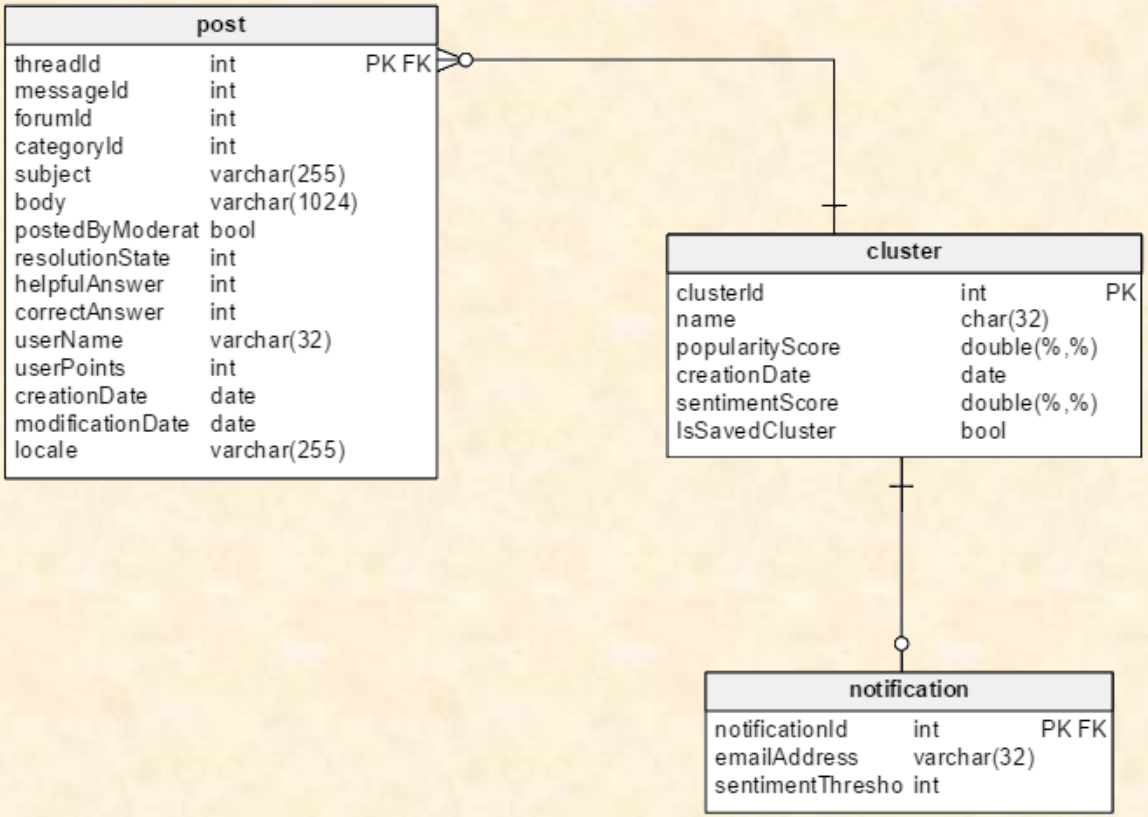
System Architecture



System Arch: *Clustering Schedule*



System Arch: *Database Schema*



System Components

- Hardware Platforms
  - Amazon Web Services
    - AWS RDS Relational Database
    - AWS Elastic Beanstalk Web Hosting

System Components - continued

- Software Platforms / Technologies
  - Python3 Frontend and Backend
  - Django Web Application
  - NLTK + SciKit Natural Language Processing
  - PyLucene Search
  - Python MySQL Connector Database Interface

Testing

- Nose Unit Testing – “Batteries Included”
  - Automatically collects and runs tests
  - Provides coverage information
  - Supports plugins for code profiling
- AWS CodePipeline Continuous Integration
  - Once it launches
  - Backup plan: Jenkins

Risks

- Unfamiliarity with Technologies
  - None of us have used Lucene, SciKit or NLTK
  - Mitigation: Set up test environment and test unfamiliar technologies ASAP

- Feature Creep
  - Ongoing expansion or addition of new features in our project
  - Mitigation: Identify project features and capabilities early

Risks - continued

- Machine Learning
  - Is a tricky topic, can get challenging
  - Mitigation: Use existing libraries rather than developing our own code
- Scalability
  - Our project must be scalable
  - Mitigation: Confirm that the technologies we use are scalable