

Bayer META (Message Exchange Text Analytics) Project Final Presentation

Maggie Lu, Siddharth Menon, Yiling
Zhong, Olivia Zheng

December 12, 2018



Who We Are



Maggie Lu
Project Manager

MISM 16



Siddharth Menon
System Administrator

MISM 16



Olivia Zheng
Project Coordinator

MISM 16



Yiling Zhong
Finance Manager

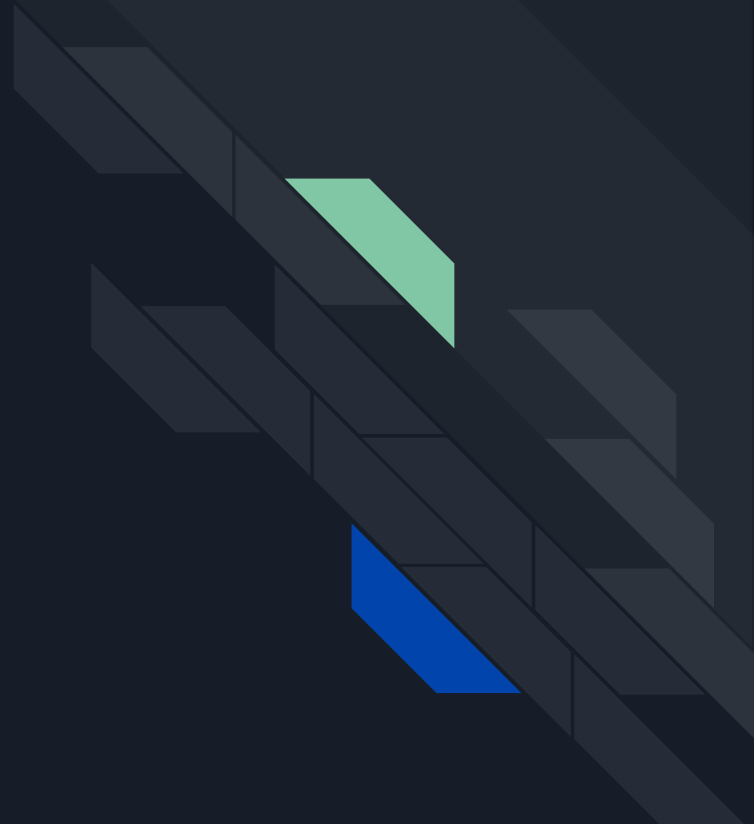
MISM 16



Agenda

- ❑ Project Overview
- ❑ Analytical Methodology
- ❑ Web Application Implementation
 - ❑ Django Implementation
 - ❑ Angular Implementation
- ❑ Demo
- ❑ Summary

Project Overview





Project Overview

❑ Problem Statement

The problem of	Siloed communication among different teams and departments
Affects	All who work on innovative projects
The impact of which is	Lack of collaboration resulting in unachieved potential in productivity
A successful solution would be	A tool that can synergize information and model topics from email correspondence, connect relevant parties with useful information and provide the opportunities to collaborate.



Project Overview

❑ Product Position Statement

For	Bayer scientists, technicians, as well as business professionals
Who	Work on multiple projects across multiple functional departments
Our solution	Would connect previously isolated parties with keywords from their email correspondence, intranet knowledge base, and academic publications as well as provide relevant internet resources
That	Promote effective and efficient collaboration



Needs and Features

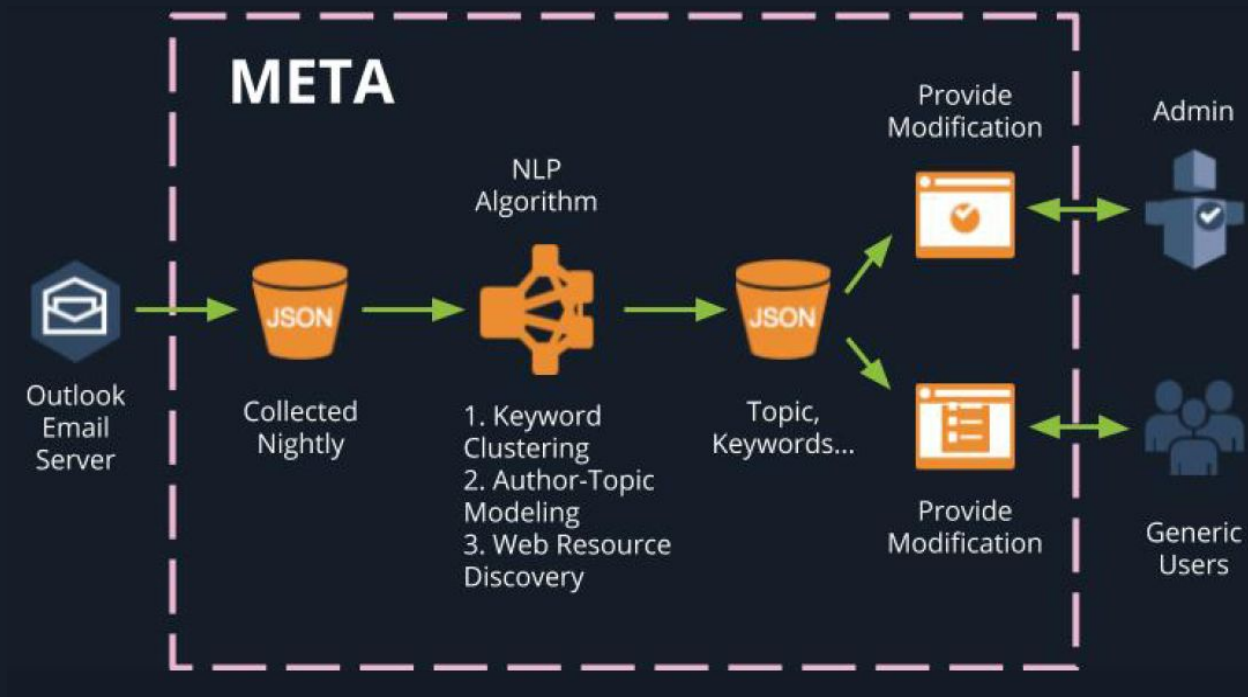
- ❑ Top Priority Needs
 - ❑ Keyword Identification
 - ❑ User Collaboration Identification
 - ❑ Web Resource Discovery
- ❑ Top Priority Features
 - ❑ Web Application Framework
 - ❑ Front-end Visualization



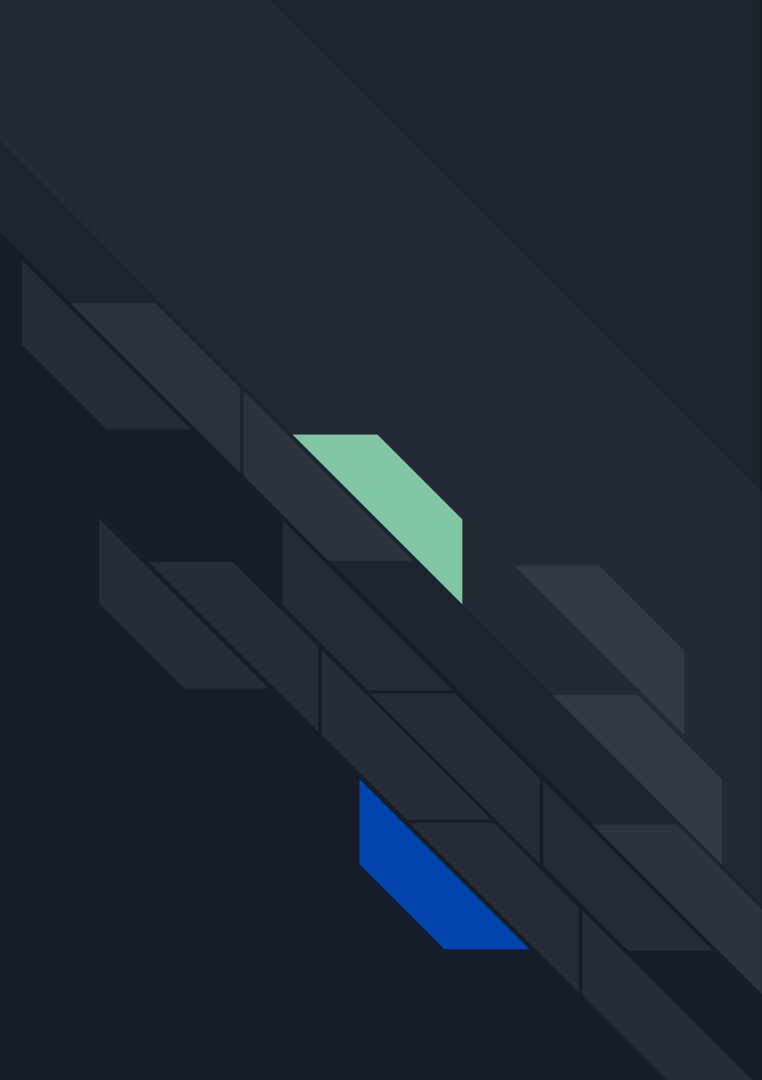
Project Timeline, Scope and Deliverables

- ❑ POC(Proof of Concept): late-August 2018 to mid-October 2018
 - ❑ Define data sources
 - ❑ Research Natural Language Processing Algorithms
 - ❑ Research and demo the application pipeline
- ❑ Development and Implementation: late-October 2018 to mid-December 2018
 - ❑ Develop and construct the application pipeline
 - ❑ Implement working NLP algorithms
 - ❑ Configure and debug code
 - ❑ Front-End Architecture

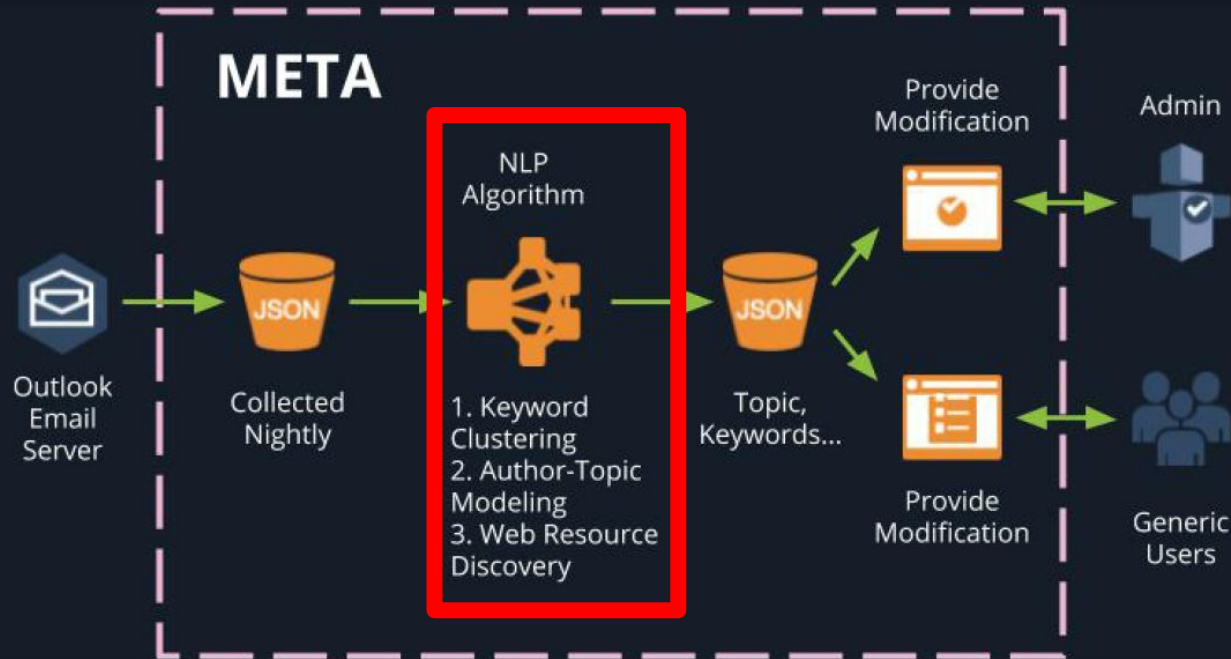
High-Level System Diagram



Analytical Methodology



Analytical Methodology: Natural Language Processing (NLP)





Analytical Methodology: Natural Language Processing (NLP)




- ❑ Data Collection and Formatting
- ❑ Latent Dirichlet Allocation, TF-IDF and Mallet
- ❑ Web Resource Discovery

Data Collection and Cleaning





My Emails

Inbox

Outbox

Drafts

Junk Mail

From: [Bijan Bina](#)

To: [Mohammad Mehdi Amin](#)

Subject: **Data on biosurfactant assisted removal of TNT**

Contamination of environment, especially soil, is in great concern and can cause health problems. Thus, remediation of these pollutants through environmentally friendly methods should be considered. The aim of this data was bioremediation of TNT from contaminated soil. Two plastic pans were used as bioreactor. In each pan, 3/202/kg of soil was used. Concentration of TNT in contaminated soil was 1000/202fmg/kg. Rhamnolipid in concentration of 6002mg/l was added to intended pan. Sampling was done in each two weeks. In order to

My Emails



bayer.outlook.com

Inbox

Outbox

Drafts

Junk Mail

From: [Bijan Bina](#)

To: [Mohammad Mehdi Amin](#)

Subject: **Data on biosurfactant assisted removal of TNT**

Contamination of environment, especially soil, is in great concern and can cause health problems. Thus, remediation of these pollutants through environmentally friendly methods should be considered. The aim of this data was bioremediation of TNT from contaminated soil. Two plastic pans were used as bioreactor. In each pan, 3/202/kg of soil was used. Concentration of TNT in contaminated soil was 1000/202fmg/kg. Rhamnolipid in concentration of 6002mg/l was added to intended pan. Sampling was done in each two weeks. In order to



NLP Deep Dive: Topic Modeling

- ❑ **Latent Dirichlet Allocation (LDA)** is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar.
- ❑ **Term Frequency–Inverse Document Frequency (TFIDF)**, is a numerical statistic that is intended to reflect how important a word is to a document in a collection or corpus.
- ❑ **MALLET** topic model package includes an extremely fast and highly scalable implementation of Gibbs sampling, efficient methods for document-topic hyperparameter optimization, and tools for inferring topics for new documents given trained models.



NLP Deep Dive: Performance Metrics

❑ Coherence Score

The state-of-the-art in terms of topic coherence are the intrinsic measure UMass and the extrinsic measure UCI, both based on the same high level idea. Both measure compute the sum:

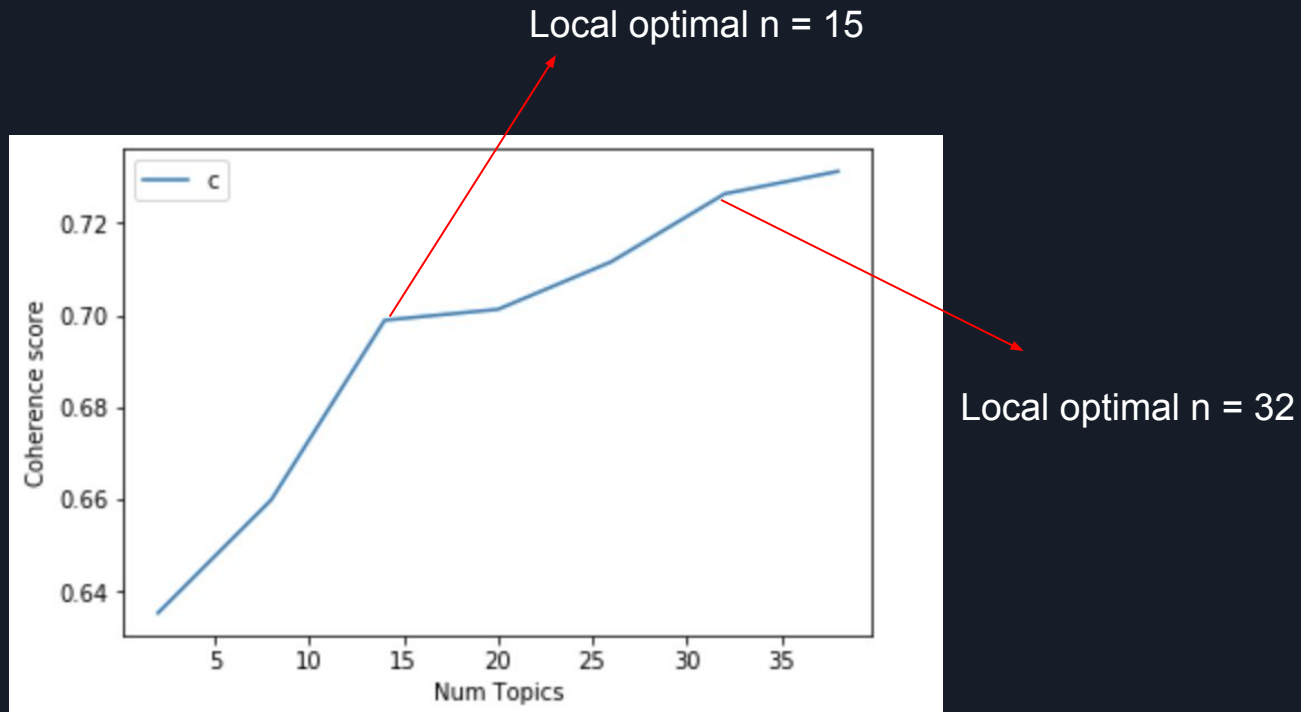
$$\text{Coherence} = \sum_{i < j} \text{score}(w_i, w_j)$$

of pairwise scores on the words w_1, \dots, w_n used to describe the topic, usually the top n words by frequency $p(w|k)$.

❑ Dynamic Model Selection based on input dataset

Do a **10 fold cross validation** to compute the average coherence score!

NLP Deep Dive-Choose **K** topics





Resource Discovery

Tested a few options: Google search, Scholarly, Bing API...

Winner: Bing API v7!

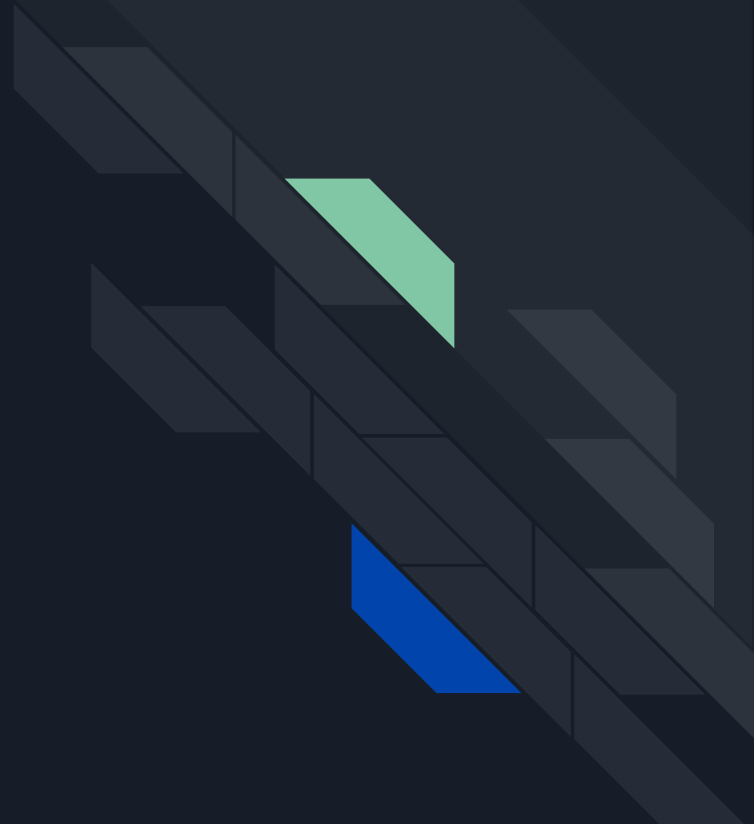
- Up to 3,000 free searches per month
- Multiple keywords search
- The ability to filter the results by number and date range
- Potential image and video searches ability



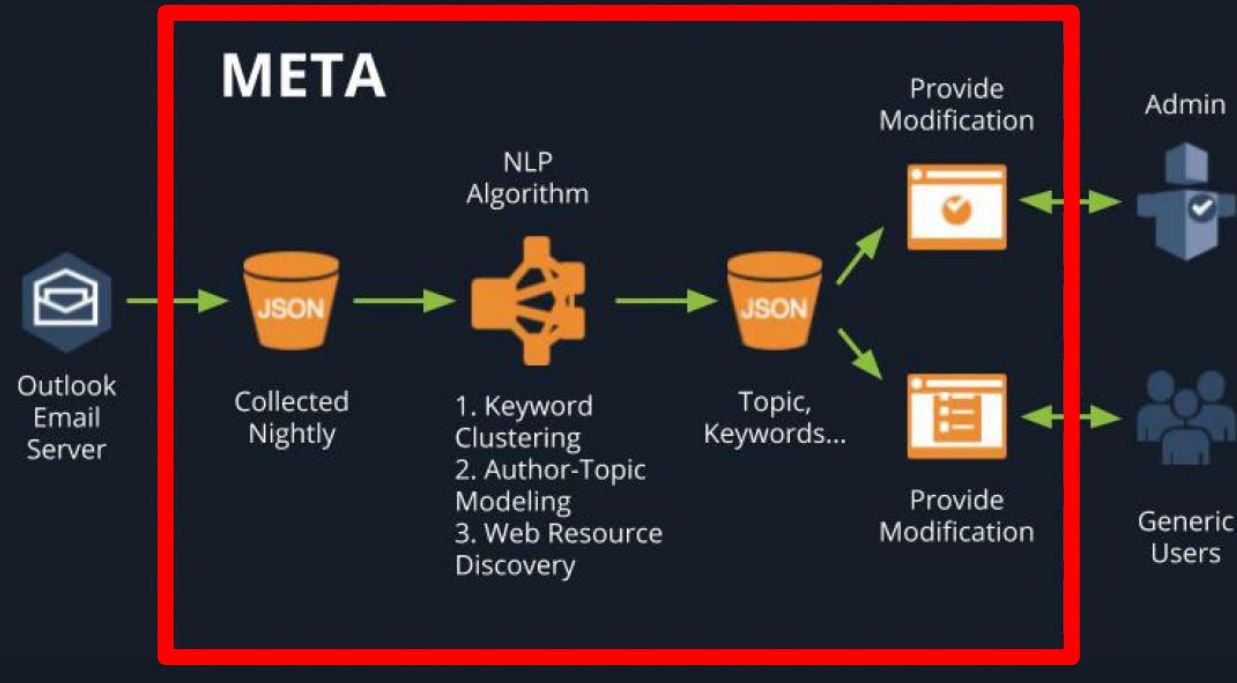
Output: 3 articles that are related to the topic

```
{"url": "https://www.sciencedirect.com/science/article/pii/S1046592818302006", "title":  
"Expression, purification and characterization of the full..."}
```

Web Application Implementation



Web App Implementation





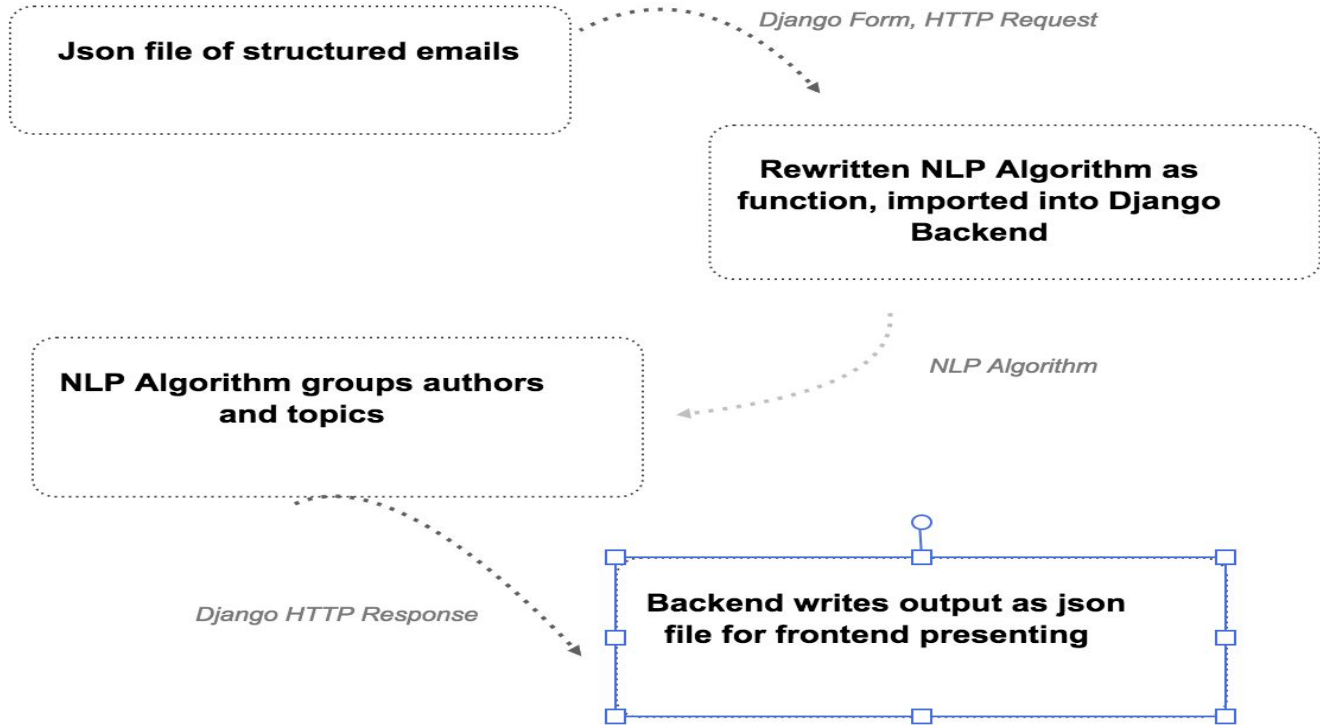
Web Framework Selection

We eventually choose Django for our web application framework

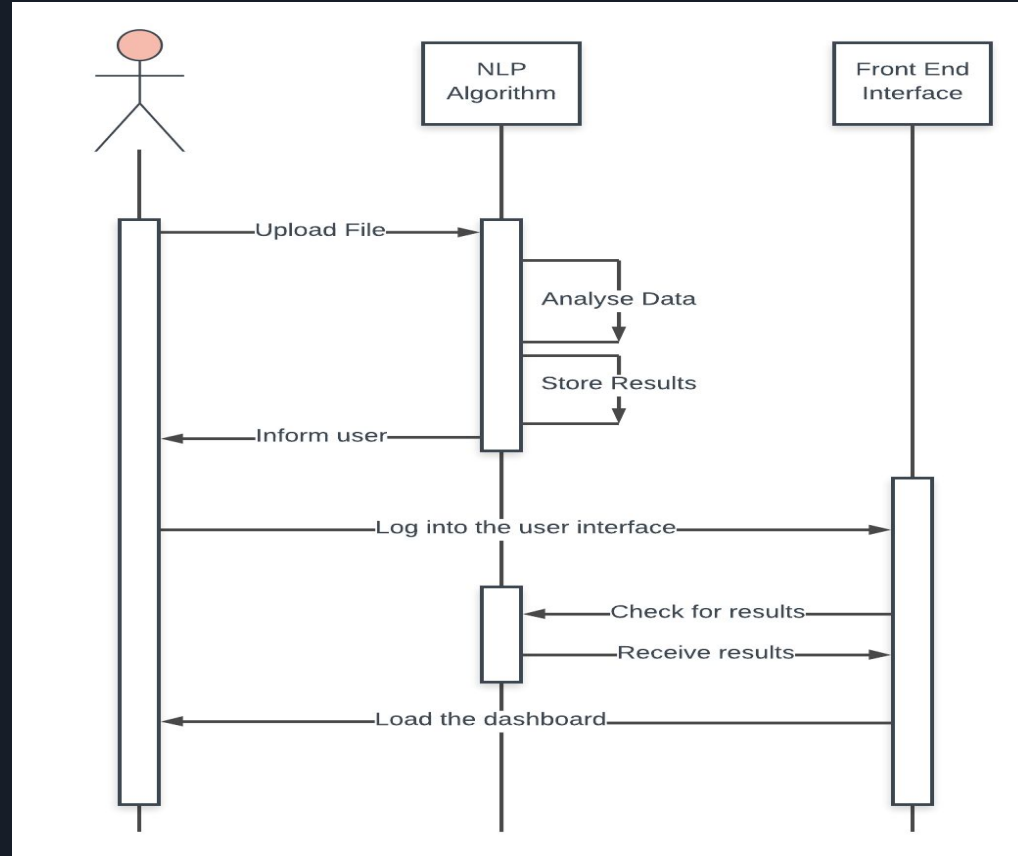
- ❑ Python integrates seamlessly with our nlp algorithm
- ❑ Database support, including models and forms
- ❑ Full-featured MVC framework with built in ORM
- ❑ Prior level of experience



Django Implementation



Web Application Sequence Diagram





Front-end Implementation

- ❑ Initially we created our dashboard using jQuery
- ❑ Due to security concerns and other stability issues we decided to upgrade
- ❑ We looked at 3 different JavaScript frameworks:
Angular, React, and Vue



Frameworks compared

Angular vs React vs Vue

- ❑ Angular and React more popular and well-known and are supported by big companies
- ❑ Both have high “Like percentage”
- ❑ In satisfaction, React leads followed by Vue and then Angular
- ❑ React and Angular offer long term support
- ❑ Angular and Vue are easier to be picked up by developers (less JavaScript required)
- ❑ Angular is a full framework; offers a lot more bundled up within the framework
- ❑ Angular is more consistent due to the use of TypeScript
- ❑ React breaks a lot of established best practices

Our winner: Angular 4



Angular Overview

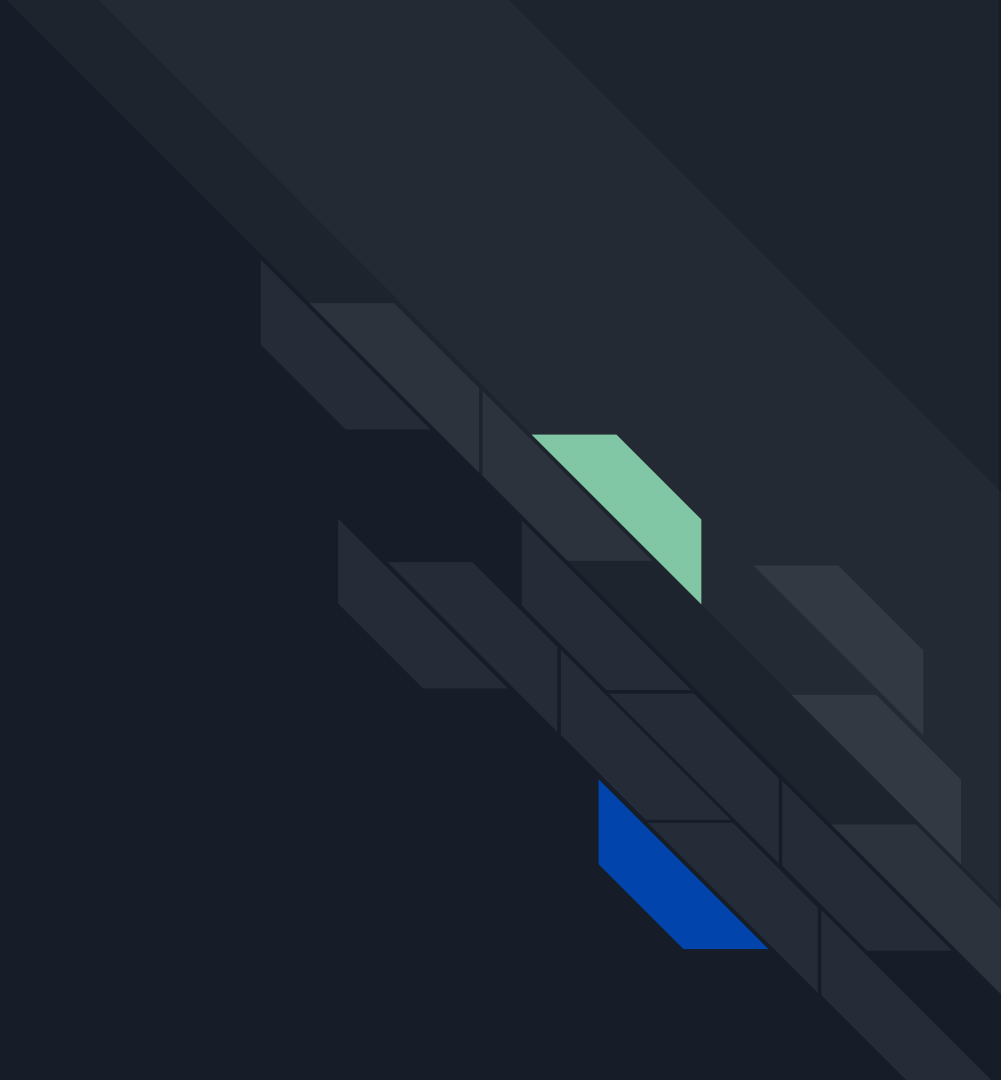
- ❑ Angular is a TypeScript-based JavaScript framework
- ❑ Developed and maintained by Google
- ❑ Initially introduced as AngularJS in 2010; upgraded to Angular in 2016 as version 2
- ❑ Newest version is Angular 7, released in October 2018



Architecture Overview

- ❑ Modules:
 - ❑ Building blocks of code
- ❑ Components:
 - ❑ Contains application data and logic
 - ❑ Associated with an HTML template (the view)
- ❑ Services and Dependency Injection:
 - ❑ Used to share data and logic across components

Demo





Moving Forward...

- ❑ Analytical Backend

 - ❑ Parameter Definition

 - ❑ Dynamic Learning

- ❑ Application Integration

 - ❑ AWS Cloud & Outlook Email server Integration

 - ❑ Active Directory for Tiered Access



Special Thanks To...

- ❑ Our Client -- Bayer
 - ❑ Mr. Jim Koob
 - ❑ Mr. Michael Kremliovsky
- ❑ Our advisor -- Sohel Sarwar

Thank You!

