

KNOWLEDGE DISCOVERY AND MANAGEMENT

SUMMARIZATION

INSTRUCTOR:

Dr. Yugyung Lee

TEAM 8:

VILAS MAMIDYALA

VIKESH PADARTHI

DINESH KUMAR BANDAM

BHUMIREDDY RANJITHA REDDY

FIRST INCREMENT REPORT - SUMMARIZATION

1. Motivation:

We know that the whole world is awaiting to hear the result of US election which are going to be released by the end of this year. Everyone would like to see how these elections are going to be held. One has an anxiety that who is going to win and what actually the people opinion is and who has more probability to win. These questions stimulate our work towards collecting data about politics which clears all our skeptic things about elections. Since many of the things related to students and their future who have more excitement and worry to get to know the result. Our main motivation behind this project is to analyze the data present in social media like twitter and plot some graphs which shows about which candidate is more famous in social media, the probability of who will be getting elected.

Objective:

Main objective of this project is to use NLP, machine learning knowledge to predict the outcome of election result. Using these we can summarize the result of various blogs, news, and editorial matters in newspapers which are related to elections. We will first plot some graphs based on the twitter data which we have collected. And we want to analyze various text data present in the World Wide Web like Wikipedia and summarize these papers.

Expected outcomes:

By performing these operations using NLP, Machine Learning we want to predict the outcome of the US elections and various views about US elections by the people around the world. The output will be ontology graphs which are developed by analyzing the data sets which are related to US elections.

2. Domain:

Data Set: Twitter Data, provided data sets by Lee.

Technologies: Java, Scala.

Topic: US Politics

IDE : IntelliJ

3. Data Collection:

Twitter data using JAVA and Linux.

4. Task and Features:

- Collected Twitter data using Java code.
- Link for the source code is:

https://github.com/vilasmamidyala/KDM_SM16_SM/tree/master/Source/twit

- NLP processing has been applied to the sample input collected above .

https://github.com/vilasmamidyala/KDM_SM16_SM/blob/master/Sampleoutputs/Nlp%20Output.txt
https://github.com/vilasmamidyala/KDM_SM16_SM/blob/master/Sampleoutputs/Simplecorenlputput.txt

- Word count has been applied to the given same input :

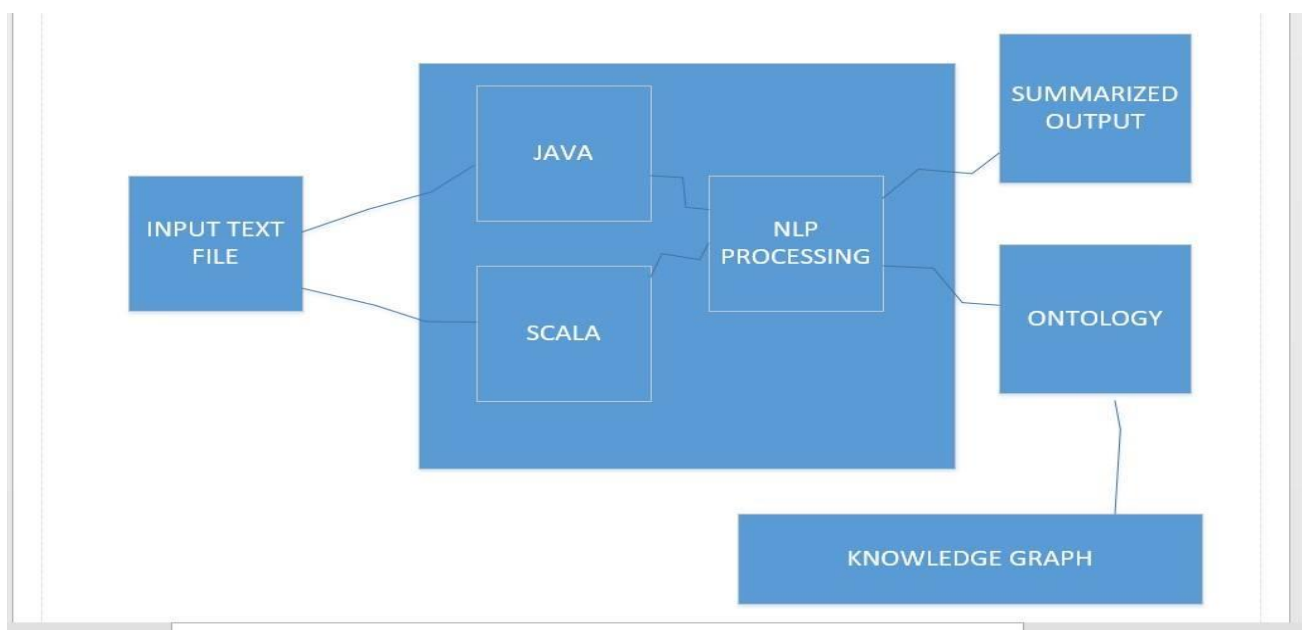
https://github.com/vilasmamidyala/KDM_SM16_SM/blob/master/Sampleoutputs/wordcount_output.txt

- Information Extraction/Retrieval technologies :

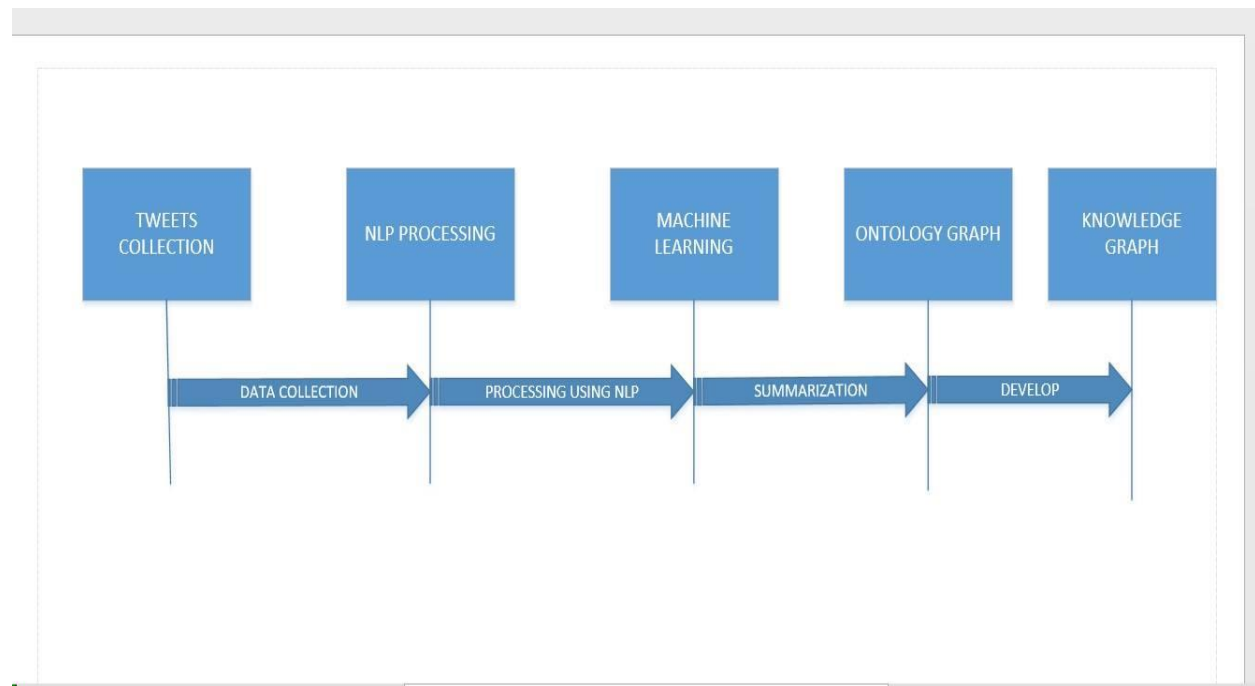
https://github.com/vilasmamidyala/KDM_SM16_SM/blob/master/Sampleoutputs/wordcount_TFID.txt

5. Implementation specification:

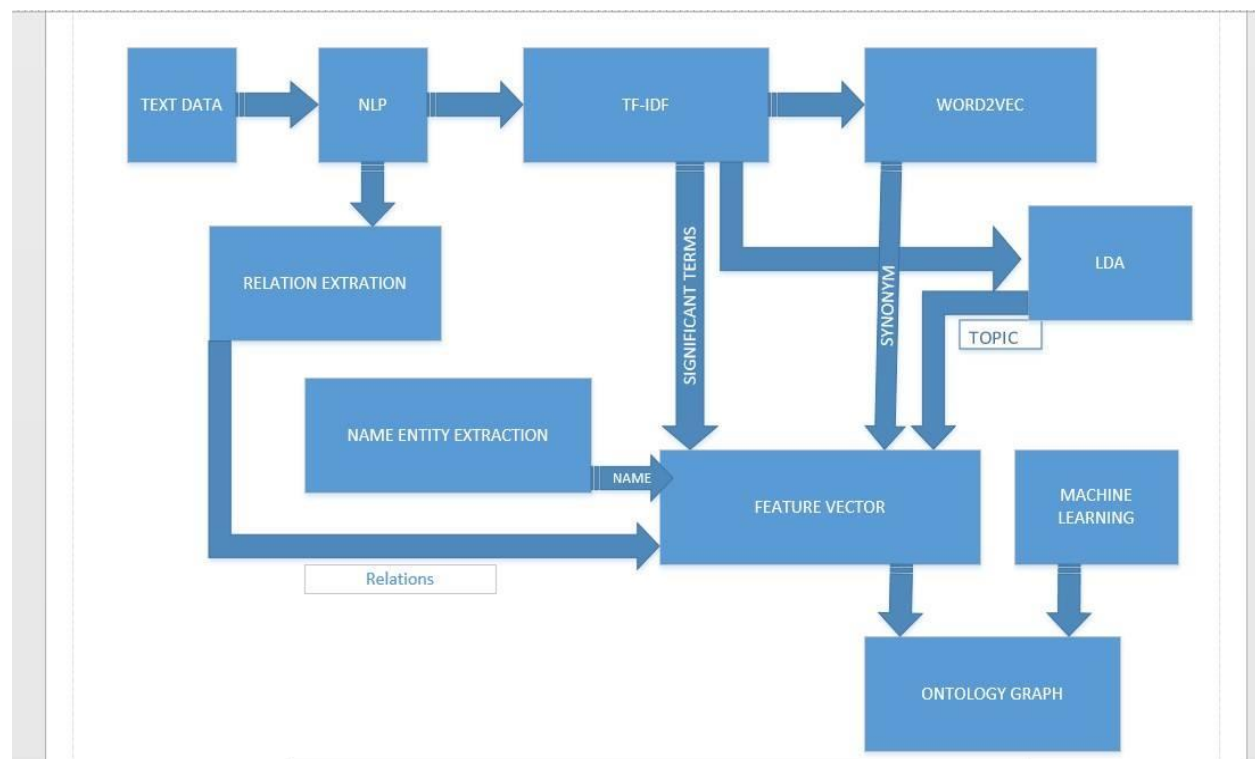
Software Architecture :



SEQUENCE DIAGRAM :



WORKFLOW DIAGRAM :



Existing Services Used:

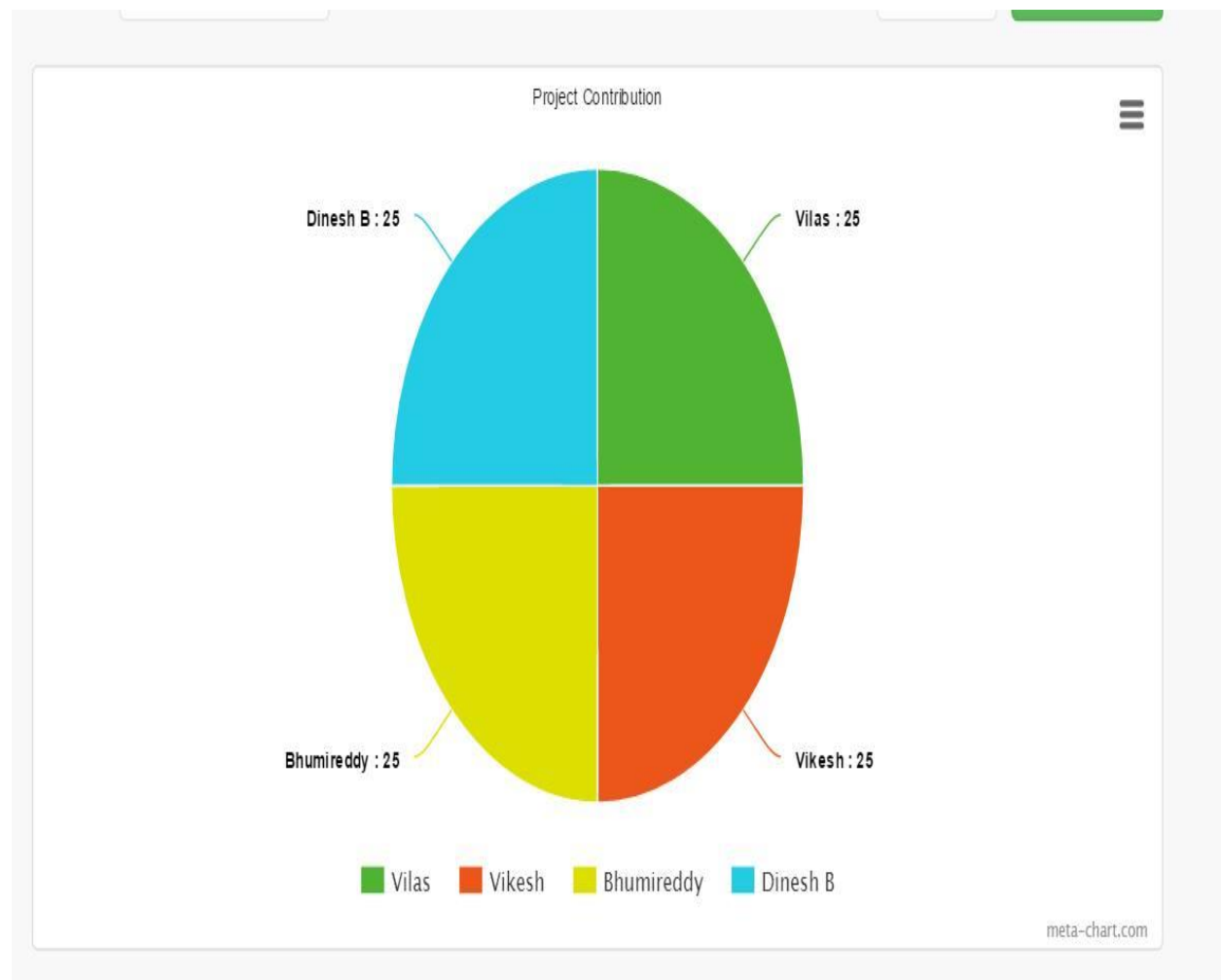
- Implemented word count program using Scala.
- Implemented NLP program.
- Implemented TF-IDF.

New Services:

Tweet collection using Java Code.

6. Project Management:

Contribution of Each member:



Zenhub and Github Screen shots:

vilasmamidyala / KDM_SM16_SM

Unwatch 3 Star 0 Fork 1

Code Issues 3 Pull requests 0 Boards Burndown Wiki Pulse Graphs Settings

Filters is:issue is:closed Labels Milestones New issue

Clear current search query, filters, and sorts

	Author	Labels	Milestones	Assignee	Sort
<input type="checkbox"/> task for wordcount ① #6 opened 10 minutes ago by vilasmamidyala wordcount for the gi...					1
<input type="checkbox"/> Try Information Extraction/Retrieval technologies #4 opened an hour ago by vilasmamidyala Try Information Extra...					1
<input type="checkbox"/> Try NLP processing #3 opened an hour ago by vilasmamidyala Try nlp					1
<input type="checkbox"/> documentation-part1 #2 opened an hour ago by vilasmamidyala task2					1

ProTip! Type **g** or **p** on any issue or pull request to go back to the pull request listing page.

COMP-SCI 5560 (SUMMER 2016) – KNOWLEDGE DISCOVERY AND MANAGEMENT

TEAM 8 : VILAS MAMIDYALA (18) VIKESH PADARTHI (27) DINESH KUMAR BANDAM (2) BHUMIREDDY RANJITHA REDDY (4)

[vilasmamidyala](#) / [KDM_SM16_SM](#) ⌵

[Unwatch](#) 3 [Star](#) 0 [Fork](#) 1

[Code](#) [Issues 3](#) [Pull requests 0](#) [Boards](#) [Burndown](#) [Wiki](#) [Pulse](#) [Graphs](#) [Settings](#)

[Labels](#) [Milestones](#) [New milestone](#)

7 Open

0 Closed

Sort

task 1

Due by June 24, 2016

Last updated about 1 hour ago

Documentation part1

0% complete

0 open

0 closed

[Edit](#) [Close](#) [Delete](#)

task2

Due by June 24, 2016

Last updated 30 minutes ago

create document for class and sequence diagrams

50% complete

1 open

1 closed

[Edit](#) [Close](#) [Delete](#)

Try nlp

Due by June 24, 2016

Last updated 26 minutes ago

do nlp process

100% complete

0 open

1 closed

[Edit](#) [Close](#) [Delete](#)

Try Information Extraction/Retrieval technologies

Due by June 24, 2016

Last updated 27 minutes ago

Try Information Extraction/Retrieval technologies

100% complete

0 open

1 closed

[Edit](#) [Close](#) [Delete](#)

task 3

Due by June 24, 2016

Last updated 37 minutes ago

Project management a. Contribution of each member b. Include ZenH... (more)

0% complete

1 open

0 closed

[Edit](#) [Close](#) [Delete](#)

wordcount for the given sample text

Due by June 24, 2016

Last updated 7 minutes ago

calcuaitte wordcount

100% complete

0 open

1 closed

[Edit](#) [Close](#) [Delete](#)

COMP-SCI 5560 (SUMMER 2016) – KNOWLEDGE DISCOVERY AND MANAGEMENT

TEAM 8 : VILAS MAMIDYALA (18) VIKESH PADARTHI (27) DINESH KUMAR BANDAM (2) BHUMIREDDY RANJITHA REDDY (4)

CS5560 - 2016 Sum x CS5560-Project1.pdf x Boards - vilasmamidi x Facebook x

← → ↻ 🏠 GitHub, Inc. [US] https://github.com/vilasmamidyala/KDM_SM16_SM/issues/7#boards?repos=60785252

Apps Google Gmail Welcome, Vilas Pathway GitHub Instructor Led On mailbox University of Mis K-ROO | UMKC S Bank of America myhr SA: Home KDM-sheet Other bookmarks

This repository Search Pull requests Issues Gist + ToDo

Unwatch 3 Star 0 Fork 1

<> Code Issues 3 Pull requests 0 Boards Burndown Wiki Pulse Graphs Settings

Repos (1/1) show one Labels Milestones Assignees 0 1 2

Search (/) New issue

New Issues 0	Icebox 0	Backlog 1	In Progress 2	Review/QA 0	Done 0	Closed 4
		<div>KDM_SM16_SM #5 Document part 3</div>	<div>KDM_SM16_SM #1 architecture diagram and sequence diagram</div> <div>KDM_SM16_SM #7 collect twitter data</div>			<div>KDM_SM16_SM #6 task for wordcount</div> <div>KDM_SM16_SM #3 Try NLP processing</div> <div>KDM_SM16_SM #4 Try Information Extraction/Retrieval technologies</div> <div>KDM_SM16_SM #2 documentation-part1</div>

COMP-SCI 5560 (SUMMER 2016) – KNOWLEDGE DISCOVERY AND MANAGEMENT

TEAM 8 : VILAS MAMIDYALA (18) VIKESH PADARTHI (27) DINESH KUMAR BANDAM (2) BHUMIREDDY RANJITHA REDDY (4)

The screenshot shows a web browser displaying a GitHub repository page for 'vilasmamidyala/KDM_SM16_SM'. The browser's address bar shows the URL 'https://github.com/vilasmamidyala/KDM_SM16_SM/issues#boards?repos=60785252'. The repository page includes a search bar, navigation links for Pull requests, Issues, Gist, and To Do, and repository statistics (Unwatch 3, Star 0, Fork 1). Below the repository header, there are tabs for Code, Issues (5), Pull requests (0), Boards (selected), Burndown, Wiki, Pulse, Graphs, and Settings. The main content area displays a Kanban board with the following columns and issues:

- New Issues (1)**: KDM_SM16_SM #1 architecture diagram and sequence diagram
- Icebox (0)**: No issues listed.
- Backlog (1)**: KDM_SM16_SM #5 Document part 3
- In Progress (2)**: KDM_SM16_SM #2 documentation-part1, KDM_SM16_SM #3 Try NLP processing
- Review/QA (1)**: KDM_SM16_SM #4 Try Information Extraction/Retrieval technologies
- Done (0)**: No issues listed.
- Closed (0)**: No issues listed.

At the bottom right of the board, there is an 'Add a' button.

COMP-SCI 5560 (SUMMER 2016) – KNOWLEDGE DISCOVERY AND MANAGEMENT

TEAM 8 : VILAS MAMIDYALA (18) VIKESH PADARTHI (27) DINESH KUMAR BANDAM (2) BHUMIREDDY RANJITHA REDDY (4)

This screenshot shows a GitHub repository page for 'vilasmamidyala / KDM_SM16_SM'. The repository has 3 Unwatch, 0 Star, and 1 Fork. The main navigation bar includes links for Code, Issues (2), Pull requests (0), Boards, Burndown, Wiki, Pulse, Graphs, and Settings. Below the navigation bar, there are tabs for Repos (1/1), Labels, Milestones, Assignees, and a search bar. The main content area displays a Kanban board with the following columns:

- New Issues**: Empty column.
- Icebox**: Empty column.
- Backlog**: Contains one issue titled 'KDM_SM16_SM #5 Document part3'.
- In Progress**: Contains one issue titled 'KDM_SM16_SM #6 Document part2'.
- Review/QA**: Empty column.
- Done**: Empty column.
- Closed**: Contains six issues:
 - KDM_SM16_SM #1 architecture diagram and sequence diagram
 - KDM_SM16_SM #7 collect twitter data
 - KDM_SM16_SM #6 task for wordcount
 - KDM_SM16_SM #3 Try NLP processing
 - KDM_SM16_SM #4 Try Information Extracion/Retrieval technologies
 - KDM_SM16_SM #2 documentation-part1

An 'Add a' button is visible next to the Closed column.

Feature concerns/Issues:

- 1) For small amount of data given as input for NLP processing and for other code executions. We found that these programs are working well and giving better results. The issue has occurred when we had tried implement NLP operation on large amount of data the programs were not able to run properly.
- 2) We considered taking Twitter data for the first phase. But we want to know whether twitter data can be useful for summarization? Because each tweet will be independent of the other tweets most of the times. This data alone might not help us for summarization. we think we need to take other different sources of data as well. we will try to figure out about what are the other sources that can be included.

Future Work:

In our further increments we would like focus on how to implement NLP operations on a bit of huge amount of data. We would like to do Word2Vec and LDA analysis on our data and then to get the feature vector for the data. We would like to implement Machine learning and ontology to derive final graphs.