Knowledge Discovery and Management

Final Report

Semantic Search

- Movie Plots

Prepared by

Sai Venkatesh Gatiganti (Class ID: 08)
Chaitanya Sai Manne (Class ID: 20)
Sri Chaitanya Patluri (Class ID: 32)
Karthik Reddy Vundela (Class ID: 43)

Introduction:

An application to search the movies based on the plot provided by the user. We have an applications based on the keywords to find details about the keyword. Now, we are developing an application where user gives a story line or plot based details and can fetch the related movie either based in the genre or topic based. It's a simple application to find the movie names based on the plot.

Motivation:

There is an existing Semantic search engine to find the details about the movies. We can find the details by giving a key word about the movie like movie name, cast etc. We are building an application to find the details based on the plot. We can search by giving a plot of a movie and find the movies similar to that plot.

Objective:

To design an application, Semantic Search Engine that provides search results based on the plot given by the user on movies which are obtained from DBpedia and Wikipedia.

Expected Outcome:

To obtain search results based on context besides keywords for better accuracy. For example, if a user searches for a plot in the search engine, the movies with similar plot as entered by the user are shown as the search results.

Project Domain: Movies (Plot Based Semantic Search Engine)

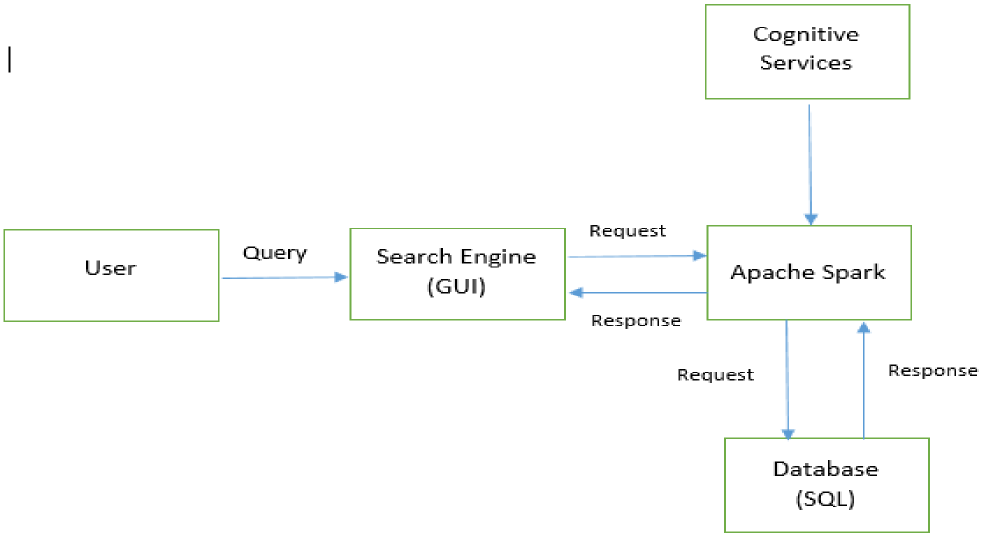
Datasets:

Amazon Review Data collected by Julian McAuley, UCSD -
http://jmcauley.ucsd.edu/data/amazon/links.html
Image-based recommendations on styles and substitutes *J. McAuley, C. Targett, J. Shi, A. van den Hengel* SIGIR*, 2015*
Inferring networks of substitutable and complementary products *J. McAuley, R. Pandey, J. Leskovec, Knowledge Discovery and Data Mining, 2015*

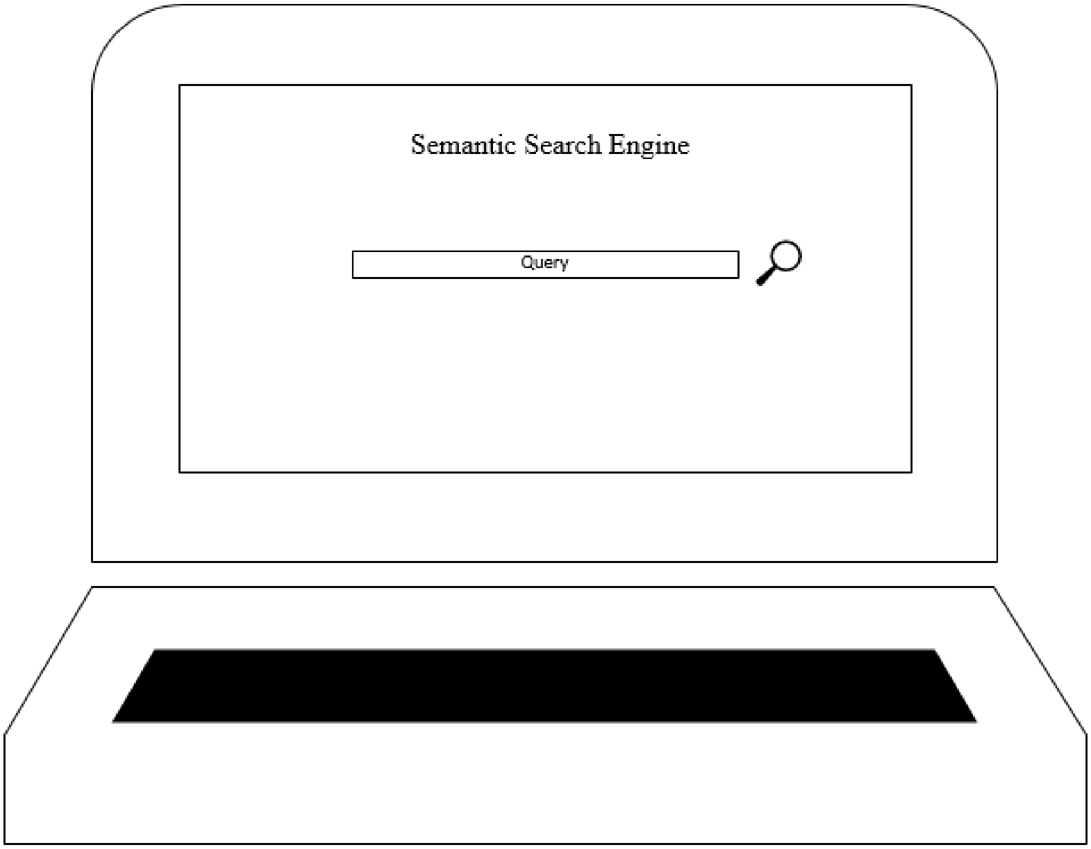
Wikipedia - https://dumps.wikimedia.org/

DBpedia - http://wiki.dbpedia.org/

System Architecture:

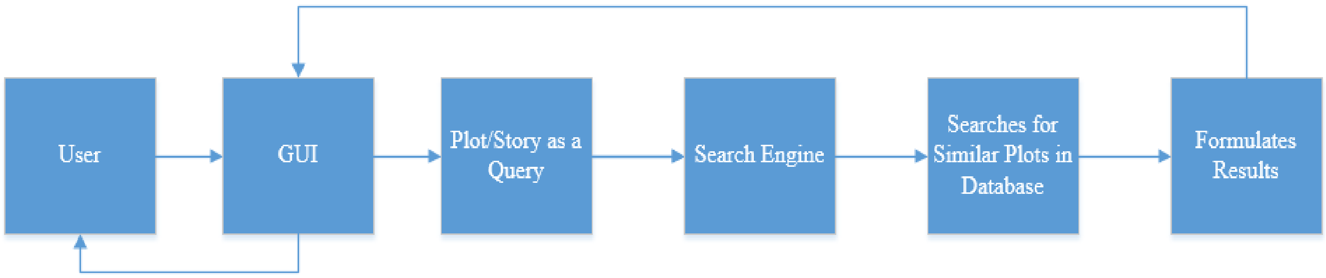


Wireframe of Search Engine:

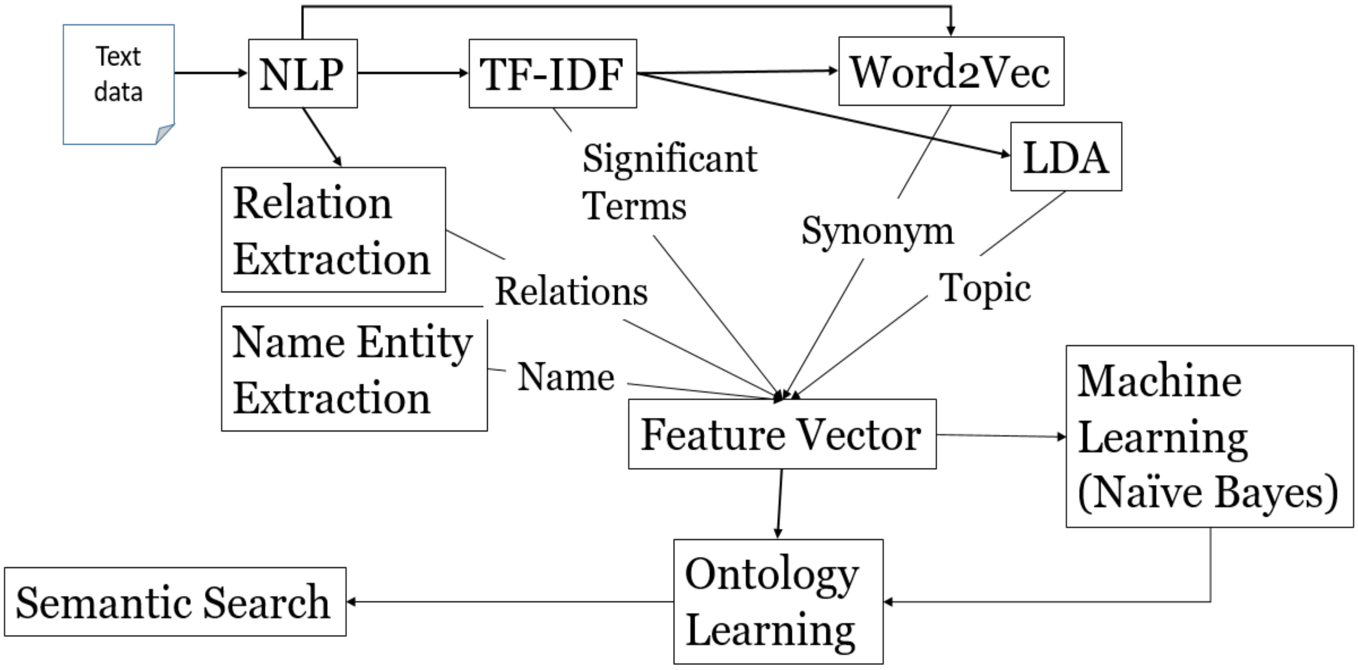


Workflow Diagram:

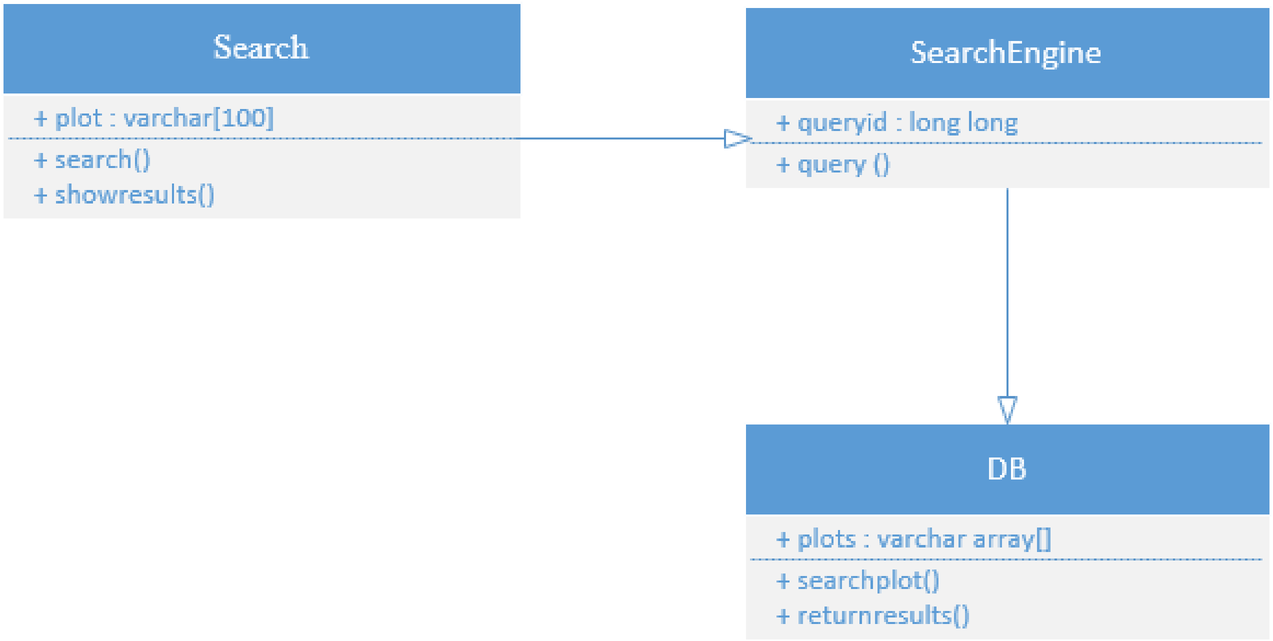
`



General Work-Flow:

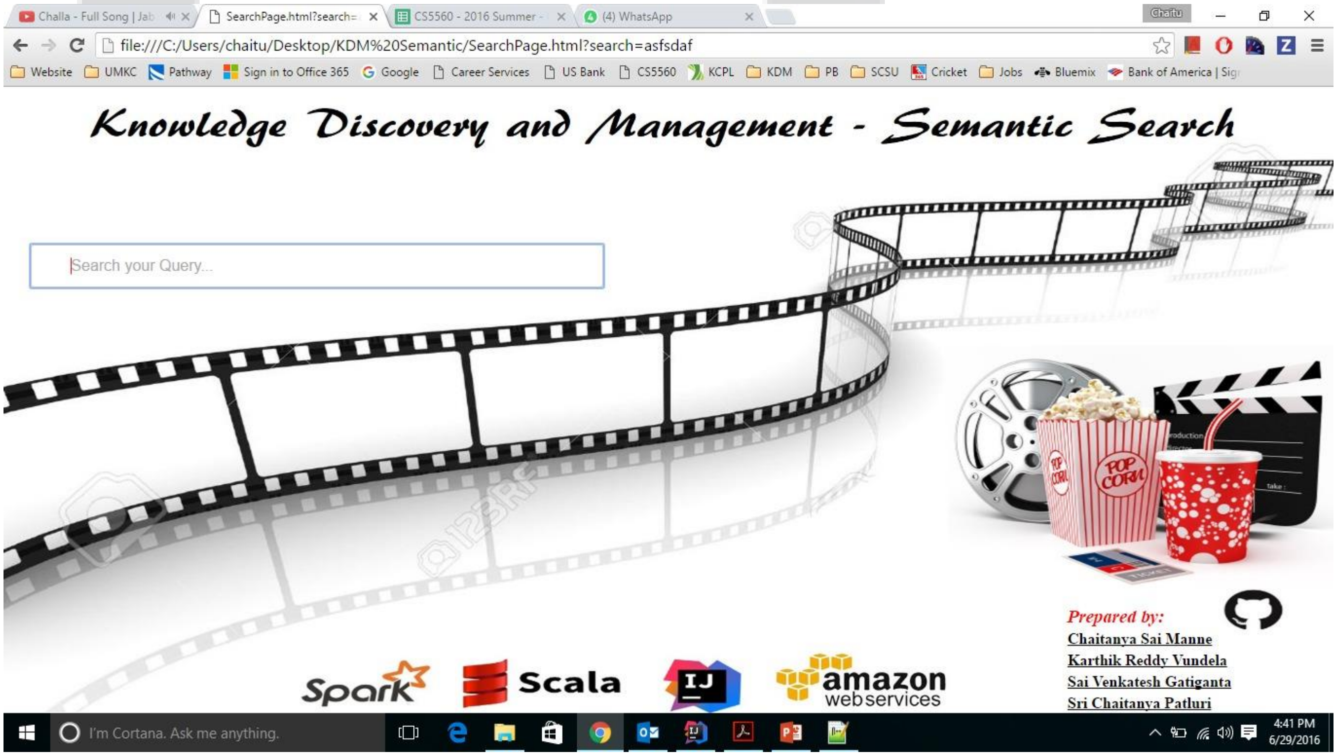


UML Class Diagram:



Working:

Front-End of Search Engine:



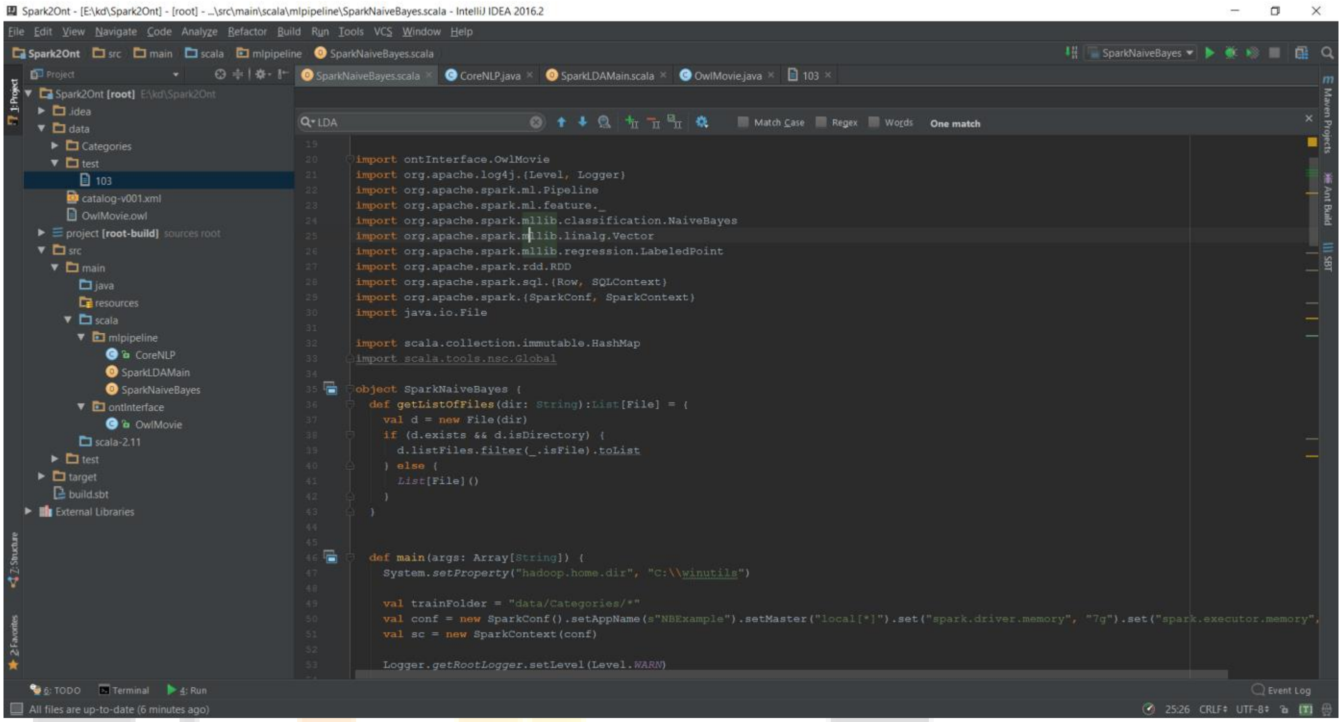
file:///C:/Users/chaitu/Desktop/KDM%20Semantic/SearchPage.html?search=asfsdaf

Knowledge Discovery and Management - Semantic Search

Search your Query...

Prepared by:
Chaitanya Sai Manne
Karthik Reddy Vundela
Sai Venkatesh Gatiganta
Sri Chaitanya Patluri

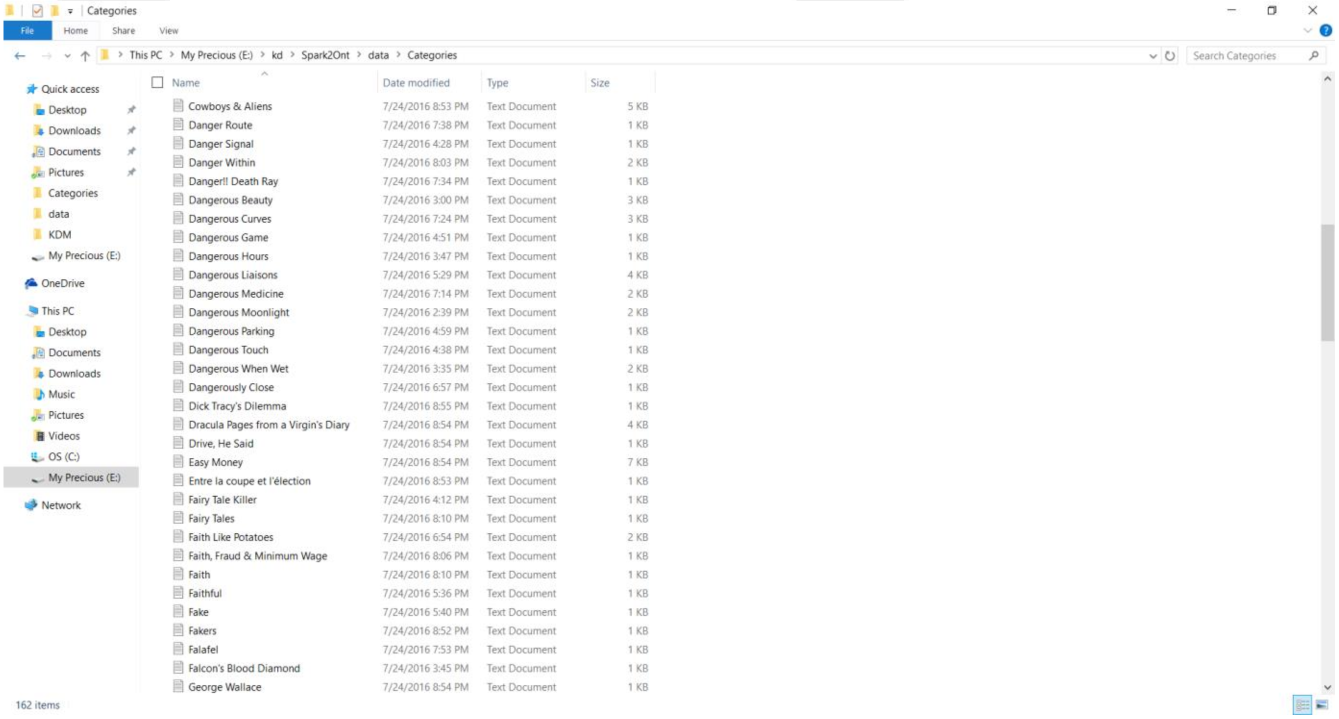
Code Sample:



```scala
import ontInterface.OwlMovie
import org.apache.log4j.{Level, Logger}
import org.apache.spark.ml.Pipeline
import org.apache.spark.ml.feature._
import org.apache.spark.mllib.classification.NaiveBayes
import org.apache.spark.mllib.linalg.Vector
import org.apache.spark.mllib.regression.LabeledPoint
import org.apache.spark.rdd.RDD
import org.apache.spark.sql.{Row, SQLContext}
import org.apache.spark.{SparkConf, SparkContext}
import java.io.File

import scala.collection.immutable.HashMap
import scala.tools.nsc.Global

object SparkNaiveBayes {
  def getListOfFiles(dir: String):List[File] = {
    val d = new File(dir)
    if (d.exists && d.isDirectory) {
      d.listFiles.filter(_.isFile).toList
    } else {
      List[File]()
    }
  }


  def main(args: Array[String]) {
    System.setProperty("hadoop.home.dir", "C:\\winutils")

    val trainFolder = "data/Categories/*"
    val conf = new SparkConf().setAppName(s"NBExample").setMaster("local[*]").set("spark.driver.memory", "7g").set("spark.executor.memory",
    val sc = new SparkContext(conf)

    Logger.getRootLogger.setLevel(Level.WARN)
```

All files are up-to-date (6 minutes ago)

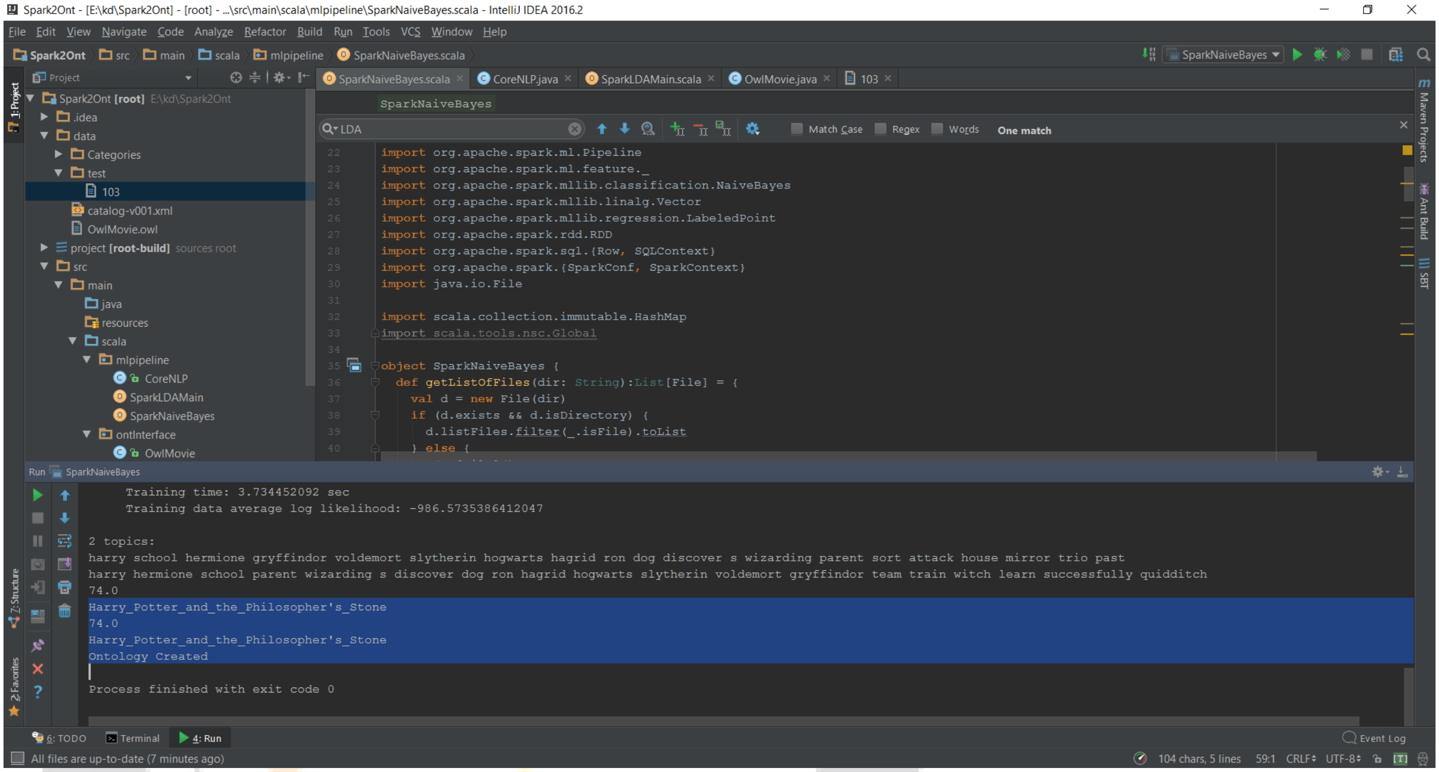
Input Data (Categories):



Name	Date modified	Type	Size
Cowboys & Aliens	7/24/2016 8:53 PM	Text Document	5 KB
Danger Route	7/24/2016 7:38 PM	Text Document	1 KB
Danger Signal	7/24/2016 4:28 PM	Text Document	1 KB
Danger Within	7/24/2016 8:03 PM	Text Document	2 KB
Danger!! Death Ray	7/24/2016 7:34 PM	Text Document	1 KB
Dangerous Beauty	7/24/2016 3:00 PM	Text Document	3 KB
Dangerous Curves	7/24/2016 7:24 PM	Text Document	3 KB
Dangerous Game	7/24/2016 4:51 PM	Text Document	1 KB
Dangerous Hours	7/24/2016 3:47 PM	Text Document	1 KB
Dangerous Liaisons	7/24/2016 5:29 PM	Text Document	4 KB
Dangerous Medicine	7/24/2016 7:14 PM	Text Document	2 KB
Dangerous Moonlight	7/24/2016 2:39 PM	Text Document	2 KB
Dangerous Parking	7/24/2016 4:59 PM	Text Document	1 KB
Dangerous Touch	7/24/2016 4:38 PM	Text Document	1 KB
Dangerous When Wet	7/24/2016 3:35 PM	Text Document	2 KB
Dangerously Close	7/24/2016 6:57 PM	Text Document	1 KB
Dick Tracy's Dilemma	7/24/2016 8:55 PM	Text Document	1 KB
Dracula Pages from a Virgin's Diary	7/24/2016 8:54 PM	Text Document	4 KB
Drive, He Said	7/24/2016 8:54 PM	Text Document	1 KB
Easy Money	7/24/2016 8:54 PM	Text Document	7 KB
Entre la coupe et l'élection	7/24/2016 8:53 PM	Text Document	1 KB
Fairy Tale Killer	7/24/2016 4:12 PM	Text Document	1 KB
Fairy Tales	7/24/2016 8:10 PM	Text Document	1 KB
Faith Like Potatoes	7/24/2016 6:54 PM	Text Document	2 KB
Faith, Fraud & Minimum Wage	7/24/2016 8:06 PM	Text Document	1 KB
Faith	7/24/2016 8:10 PM	Text Document	1 KB
Faithful	7/24/2016 5:36 PM	Text Document	1 KB
Fake	7/24/2016 5:40 PM	Text Document	1 KB
Fakers	7/24/2016 8:52 PM	Text Document	1 KB
Falafel	7/24/2016 7:53 PM	Text Document	1 KB
Falcon's Blood Diamond	7/24/2016 3:45 PM	Text Document	1 KB
George Wallace	7/24/2016 8:54 PM	Text Document	1 KB

162 items

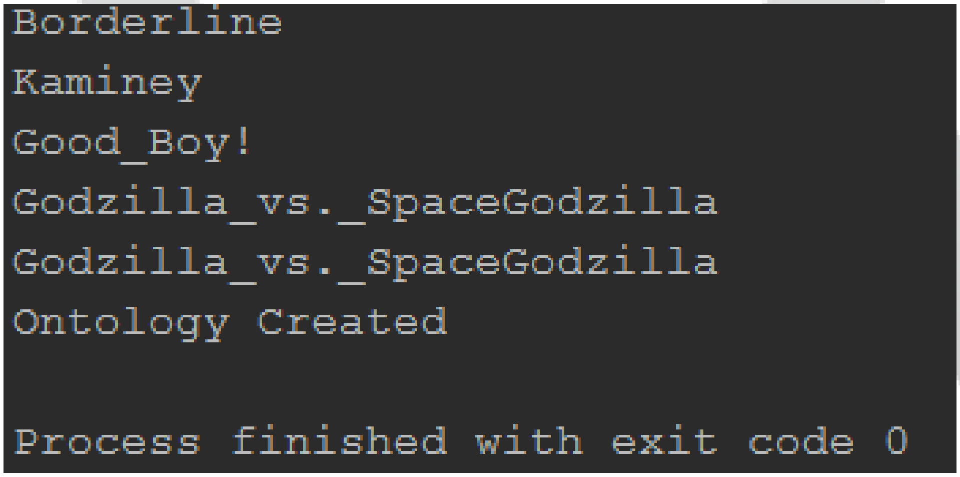
Results:


Spark2Ont - [E:\kd\Spark2Ont] - [root] - ...\src\main\scala\mlpipeline\SparkNaiveBayes.scala - IntelliJ IDEA 2016.2
File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
Spark2Ont src main scala mlpipeline SparkNaiveBayes.scala
SparkNaiveBayes
Project
Spark2Ont [root] E:\kd\Spark2Ont
.idea
data
Categories
test
103
catalog-v001.xml
OwlMovie.owl
project [root-build] sources root
src
main
java
resources
scala
mlpipeline
CoreNLP
SparkLDAMain
SparkNaiveBayes
ontInterface
OwlMovie
SparkNaiveBayes.scala
CoreNLP.java
SparkLDAMain.scala
OwlMovie.java
103
LDA
Match Case
Regex
Words
One match
Run SparkNaiveBayes
6: TODO
Terminal
4: Run
Event Log
All files are up-to-date (7 minutes ago)
104 chars, 5 lines
59:1
CRLF
UTF-8

```
SparkNaiveBayes
import org.apache.spark.ml.Pipeline
import org.apache.spark.ml.feature._
import org.apache.spark.mllib.classification.NaiveBayes
import org.apache.spark.mllib.linalg.Vector
import org.apache.spark.mllib.regression.LabeledPoint
import org.apache.spark.rdd.RDD
import org.apache.spark.sql.{Row, SQLContext}
import org.apache.spark.{SparkConf, SparkContext}
import java.io.File

import scala.collection.immutable.HashMap
import scala.tools.nsc.Global

object SparkNaiveBayes {
  def getListOfFiles(dir: String):List[File] = {
    val d = new File(dir)
    if (d.exists && d.isDirectory) {
      d.listFiles.filter(_.isFile).toList
    } else {
```

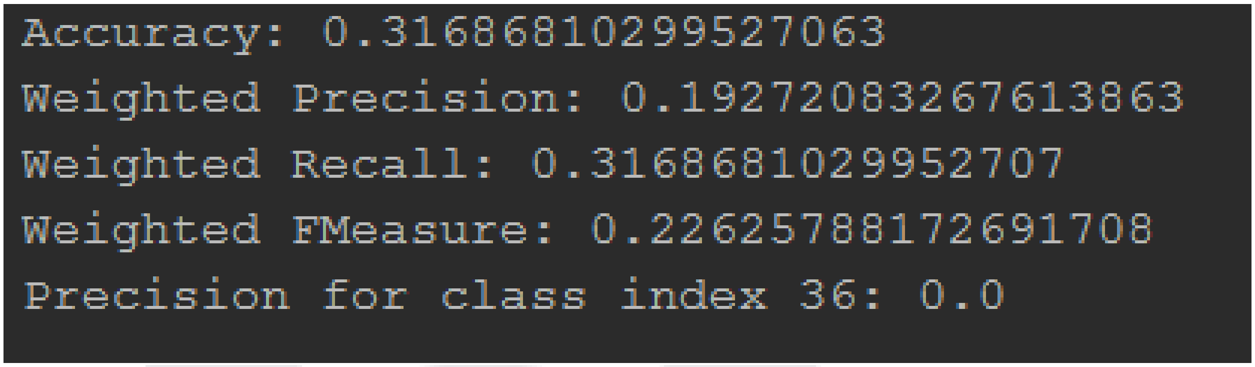
```
     Training time: 3.734452092 sec
     Training data average log likelihood: -986.5735386412047

2 topics:
harry school hermione gryffindor voldemort slytherin hogwarts hagrid ron dog discover s wizarding parent sort attack house mirror trio past
harry hermione school parent wizarding s discover dog ron hagrid hogwarts slytherin voldemort gryffindor team train witch learn successfully quidditch
74.0
Harry_Potter_and_the_Philosopher's_Stone
74.0
Harry_Potter_and_the_Philosopher's_Stone
Ontology Created

Process finished with exit code 0
```

Movie names:



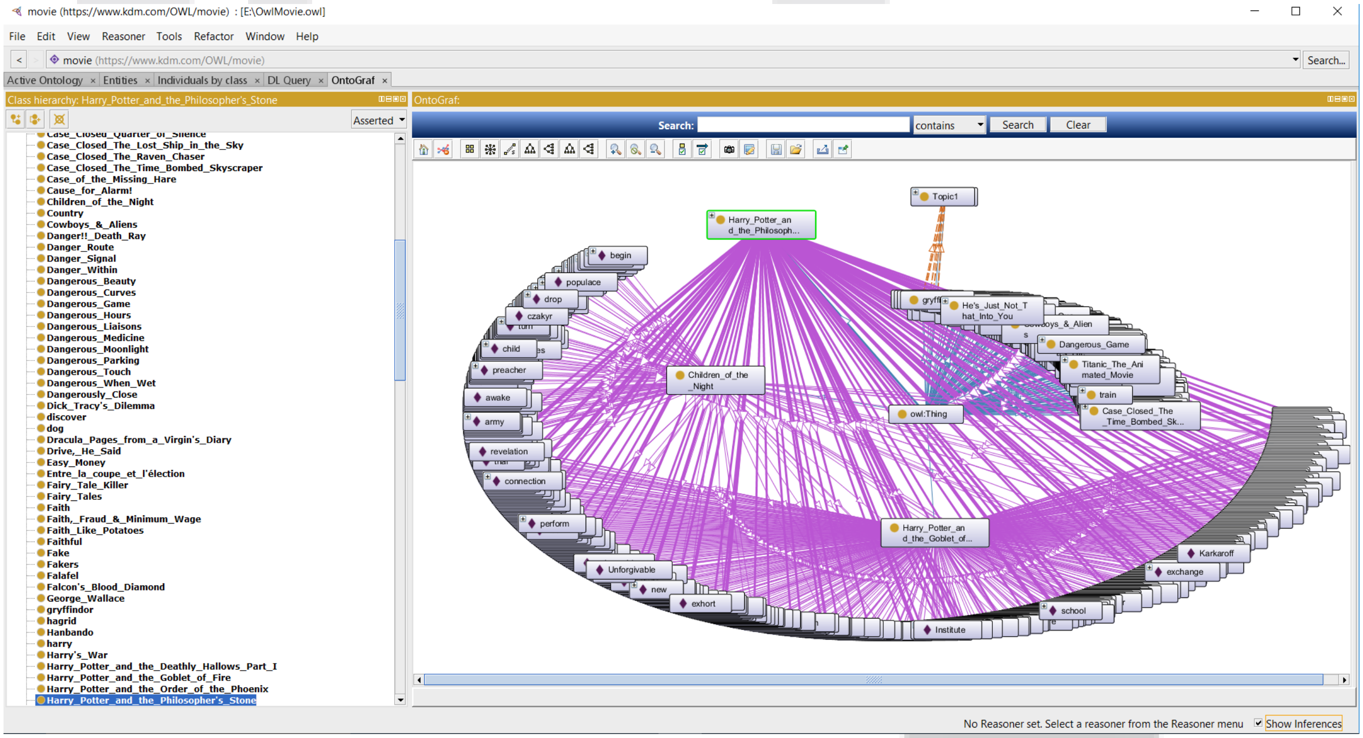
```
Borderline
Kaminey
Good_Boy!
Godzilla_vs._SpaceGodzilla
Godzilla_vs._SpaceGodzilla
Ontology Created

Process finished with exit code 0
```

Evaluation:



```
Accuracy: 0.31686810299527073
Weighted Precision: 0.19272083267613863
Weighted Recall: 0.3168681029952707
Weighted FMeasure: 0.22625788172691708
Precision for class index 36: 0.0
```

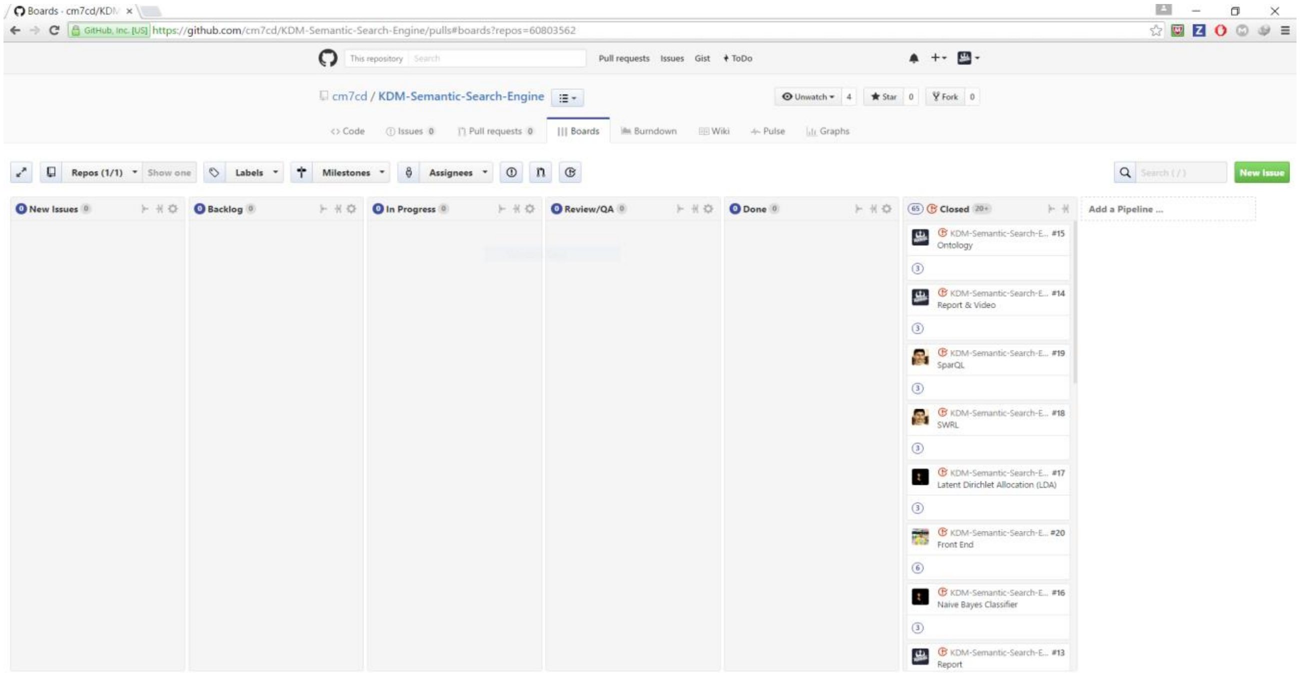
Created Ontology:



Case_Closed_Quarter_of_Silence
Case_Closed_The_Lost_Ship_in_the_Sky
Case_Closed_The_Raven_Chaser
Case_Closed_The_Time_Bombed_Skyscraper
Case_of_the_Missing_Hare
Cause_for_Alarm!
Children_of_the_Night
Country
Cowboys_&_Aliens
Danger!!_Death_Ray
Danger_Route
Danger_Signal
Danger_Within
Dangerous_Beauty
Dangerous_Curves
Dangerous_Game
Dangerous_Hours
Dangerous_Liaisons
Dangerous_Medicine
Dangerous_Moonlight
Dangerous_Parking
Dangerous_Touch
Dangerous_When_Wet
Dangerously_Close
Dick_Tracy's_Dilemma
discover
dog
Dracula_Pages_from_a_Virgin's_Diary
Drive,_He_Said
Easy_Money
Entre_la_coupe_et_l'élection
Fairy_Tale_Killer
Fairy_Tales
Faith
Faith,_Fraud_&_Minimum_Wage
Faith_Like_Potatoes
Faithful
Fake
Fakers
Falafel
Falcon's_Blood_Diamond
George_Wallace
gryffindor
hagrid
Hanbando
harry
Harry's_War
Harry_Potter_and_the_Deathly_Hallows_Part_I
Harry_Potter_and_the_Goblet_of_Fire
Harry_Potter_and_the_Order_of_the_Phoenix
Harry_Potter_and_the_Philosopher's_Stone

Zen-hub Project Management:

Issues Board:



KDM-Semantic-Search-E... #15
Ontology
3

KDM-Semantic-Search-E... #14
Report & Video
3

KDM-Semantic-Search-E... #19
SparQL
3

KDM-Semantic-Search-E... #18
SWRL
3

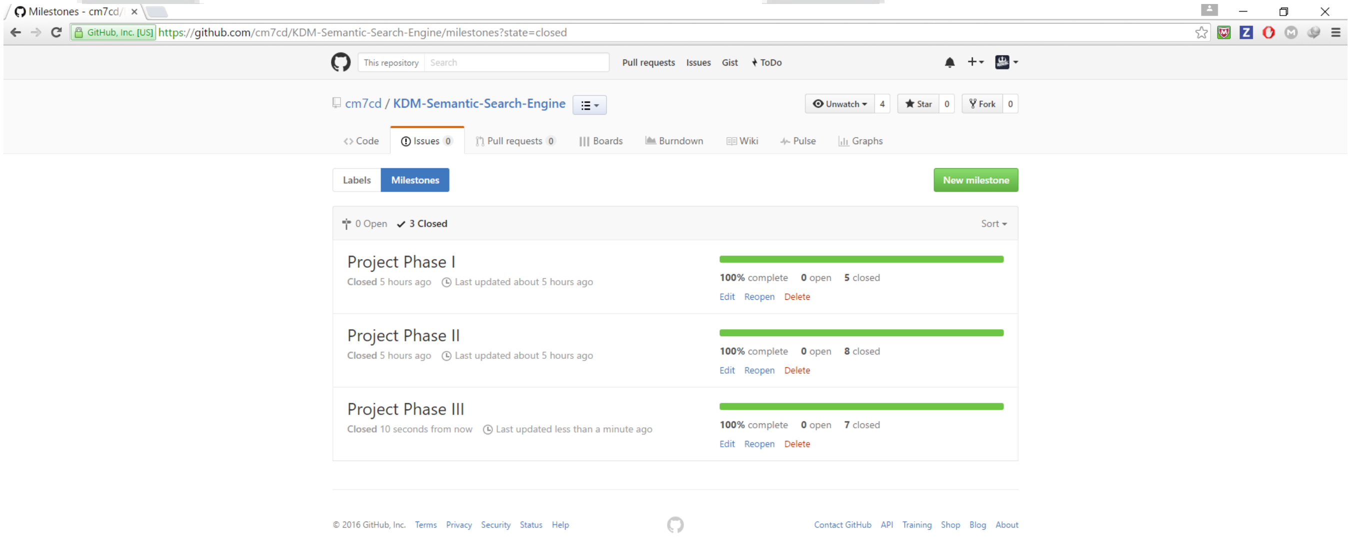
KDM-Semantic-Search-E... #17
Latent Dirichlet Allocation (LDA)
3

KDM-Semantic-Search-E... #20
Front End
6

KDM-Semantic-Search-E... #16
Naive Bayes Classifier
3

KDM-Semantic-Search-E... #13
Report

Milestones:



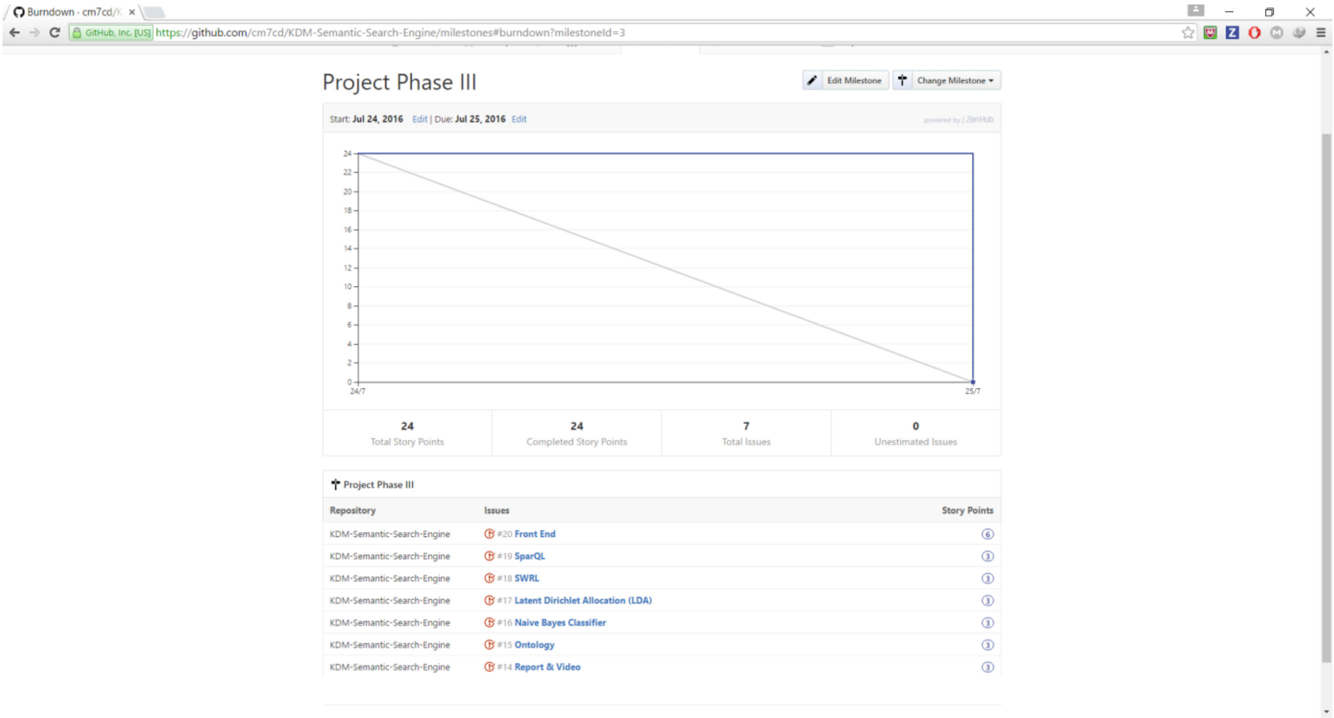
Project Phase I
Closed 5 hours ago · Last updated about 5 hours ago
100% complete 0 open 5 closed
Edit Reopen Delete

Project Phase II
Closed 5 hours ago · Last updated about 5 hours ago
100% complete 0 open 8 closed
Edit Reopen Delete

Project Phase III
Closed 10 seconds from now · Last updated less than a minute ago
100% complete 0 open 7 closed
Edit Reopen Delete

© 2016 GitHub, Inc. Terms Privacy Security Status Help Contact GitHub API Training Shop Blog About

Issues & Burndown:


Burndown · cm7cd/
GitHub, Inc. [US] https://github.com/cm7cd/KDM-Semantic-Search-Engine/milestones#burndown?milestoneId=3
Edit Milestone
Change Milestone
powered by | ZenHub
24
22
20
18
16
14
12
10
8
6
4
2
0
24/7
25/7

Project Phase III

Start: Jul 24, 2016 Edit | Due: Jul 25, 2016 Edit

24	24	7	0
Total Story Points	Completed Story Points	Total Issues	Unestimated Issues

Project Phase III

Repository	Issues	Story Points
KDM-Semantic-Search-Engine	#20 Front End	6
KDM-Semantic-Search-Engine	#19 SparQL	3
KDM-Semantic-Search-Engine	#18 SWRL	3
KDM-Semantic-Search-Engine	#17 Latent Dirichlet Allocation (LDA)	3
KDM-Semantic-Search-Engine	#16 Naive Bayes Classifier	3
KDM-Semantic-Search-Engine	#15 Ontology	3
KDM-Semantic-Search-Engine	#14 Report & Video	3

Contributions:

Sai Venakatesh Gatiganti – Naïve Bayes Classifier, Latent Dirichlet Allocation (LDA)
Karthik Reddy Vundela – SparQL & SWRL
Chaitanya Sai Manne – Front-End, NLP, TF-IDF & ppt.
Sri Chaitanya Patluri – Report, Video & Feature vector.

Future Work:

Our Future Work for the project includes including various other domains like books and research papers for the search engine and adding a dynamic web-crawler which crawls the web for related plots and displaying the results dynamically in real time.

YouTube URL: https://www.youtube.com/watch?v=H3vsaDTTA0M

GitHub URL: https://github.com/cm7cd/KDM-Semantic-Search-Engine/

Presentation URL: https://github.com/cm7cd/KDM-Semantic-Search-Engine/blob/master/Documentation/SemanticSearch.pptx

Bibliography:

- http://jmcauley.ucsd.edu/data/amazon/links.html
- http://nlp.stanford.edu/nlp/
- https://en.wikipedia.org/
- http://wiki.dbpedia.org/