Knowledge Discovery and Management

Project Phase - II Report

By

Team 1:
Sai Venkatesh Gatiganti (Class ID: 08)
Karthik Reddy Vundela (Class ID: 43)
Chaitanya Sai Manne (Class ID: 20)
Sri Chaitanya Patluri (Class ID: 32)

Objective:

To design a Semantic Search Engine that provides search results on books based on reviews obtained from Amazon & 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 books and movies with similar plot as entered by the user are shown as the search results.

Project Domain: Movies and Books – (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/

Tasks Done

NGram and Word2Vec:

n-gram: a sequential list of n words, often used in information retrieval and language modeling to encode the likelihood that the phrase will appear in the future.

Sample Input:

The following data has been given as the input for all the tasks except for LDA whose input has been mentioned explicitly at that section.

Nine noble families fight for control of the mythical land of Westeros. Political and sexual intrigue is pervasive. Robert Baratheon, King of Westeros, asks his old friend Eddard, Lord Stark, to serve as Hand of the King, or highest official. Secretly warned that the previous Hand was assassinated, Eddard accepts in order of business to investigate further. Meanwhile the Queen's family, the Lannister's, may be hatching a plot to take power. Across the sea, the last members of the previous and deposed ruling family, the Targaryen's, are also scheming to regain the throne. The friction between the houses Stark, Lannister, Baratheon and Targaryen and with the remaining great houses Greyjoy, Tully, Arryn, Tyrell and Martell leads to full-scale war. All while a very ancient evil awakens in the farthest north. Amidst the war and political confusion, a neglected military order of misfits, the Night's Watch, is all that stands between the realms of men and icy horrors beyond.

Sample Output

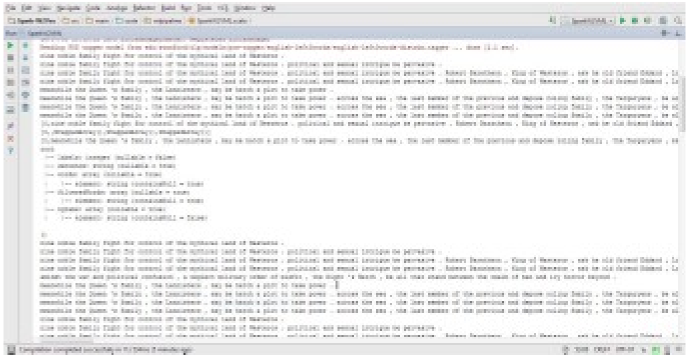
```
nine noble family fight for control of the mythical land of Westeros .
nine noble family fight for control of the mythical land of Westeros . political and sexual intrigue
be pervasive .
nine noble family fight for control of the mythical land of Westeros . political and sexual intrigue
be pervasive . Robert Baratheon , King of Westeros , ask he old friend Eddard , Lord Stark , to
serve as hand of the King , or highest official .
nine noble family fight for control of the mythical land of Westeros . political and sexual intrigue
be pervasive . Robert Baratheon , King of Westeros , ask he old friend Eddard , Lord Stark , to
serve as hand of the King , or highest official . secretly warn that the previous hand be
assassinate , Eddard accept in order of business to investigate further .
meanwhile the Queen 's family , the Lannisters , may be hatch a plot to take power .
meanwhile the Queen 's family , the Lannisters , may be hatch a plot to take power . across the
sea , the last member of the previous and depose ruling family , the Targaryens , be also
scheming to regain the throne .
meanwhile the Queen 's family , the Lannisters , may be hatch a plot to take power . across the
sea , the last member of the previous and depose ruling family , the Targaryens , be also
scheming to regain the throne . the friction between the house Stark , Lannister , Baratheon and
Targaryen and with the remain great house Greyjoy , Tully , Arryn , Tyrell and Martell lead to
full-scale war .
meanwhile the Queen 's family , the Lannisters , may be hatch a plot to take power . across the
sea , the last member of the previous and depose ruling family , the Targaryens , be also
scheming to regain the throne . the friction between the house Stark , Lannister , Baratheon and
Targaryen and with the remain great house Greyjoy , Tully , Arryn , Tyrell and Martell lead to
full-scale war . all while a very ancient evil awaken in the farthest north .
[0,nine noble family fight for control of the mythical land of Westeros . political and sexual
intrigue be pervasive . Robert Baratheon , King of Westeros , ask he old friend Eddard , Lord
Stark , to serve as hand of the King , or highest official . secretly warn that the previous hand be
assassinate , Eddard accept in order of business to investigate further . ,WrappedArray(nine,
noble, family, fight, for, control, of, the, mythical, land, of, westeros, ., political, and, sexual,
intrigue, be, pervasive, ., robert, baratheon, ,, king, of, westeros, ,, ask, he, old, friend, eddard, ,,
lord, stark, ,, to, serve, as, hand, of, the, king, ,, or, highest, official, ., secretly, warn, that, the,
previous, hand, be, assassinate, ,, eddard, accept, in, order, of, business, to, investigate,
further, .),WrappedArray(nine, noble, family, fight, control, mythical, land, westeros, ., political,
sexual, intrigue, pervasive, ., robert, baratheon, ,, king, westeros, ,, ask, old, friend, eddard, ,,
lord, stark, ,, serve, hand, king, ,, highest, official, ., secretly, warn, previous, hand, assassinate, ,,
eddard, accept, order, business, investigate, .),WrappedArray(nine noble, noble family, family
fight, fight control, control mythical, mythical land, land westeros, westeros ., . political, political
sexual, sexual intrigue, intrigue pervasive, pervasive ., . robert, robert baratheon, baratheon ,, ,
king, king westeros, westeros ,, , ask, ask old, old friend, friend eddard, eddard ,, , lord, lord
stark, stark ,, , serve, serve hand, hand king, king ,, , highest, highest official, official ., . secretly,
secretly warn, warn previous, previous hand, hand assassinate, assassinate ,, , eddard, eddard
accept, accept order, order business, business investigate, investigate .)]
[0,,WrappedArray(),WrappedArray(),WrappedArray()]
[0,meanwhile the Queen 's family , the Lannisters , may be hatch a plot to take power . across the
sea , the last member of the previous and depose ruling family , the Targaryens , be also
scheming to regain the throne . the friction between the house Stark , Lannister , Baratheon and
Targaryen and with the remain great house Greyjoy , Tully , Arryn , Tyrell and Martell lead to
full-scale war . all while a very ancient evil awaken in the farthest north .
,WrappedArray(meanwhile, the, queen, 's, family, ,, the, lannisters, ,, may, be, hatch, a, plot, to,
take, power, ., across, the, sea, ,, the, last, member, of, the, previous, and, depose, ruling,
family, ,, the, targaryens, ,, be, also, scheming, to, regain, the, throne, ., the, friction, between,
the, house, stark, ,, lannister, ,, baratheon, and, targaryen, and, with, the, remain, great, house,
greyjoy, ,, tully, ,, arryn, ,, tyrell, and, martell, lead, to, full-scale, war, ., all, while, a, very,
ancient, evil, awaken, in, the, farthest, north, .),WrappedArray(meanwhile, queen, 's, family, ,,
lannisters, ,, may, hatch, plot, take, power, ., across, sea, ,, last, member, previous, depose, ruling,
family, ,, targaryens, ,, also, scheming, regain, throne, ., friction, house, stark, ,, lannister, ,,
baratheon, targaryen, remain, great, house, greyjoy, ,, tully, ,, arryn, ,, tyrell, martell, lead, full-
scale, war, ., ancient, evil, awaken, farthest, north, .),WrappedArray(meanwhile queen, queen 's,
's family, family ,, , lannisters, lannisters ,, , may, may hatch, hatch plot, plot take, take power,
power ., . across, across sea, sea ,, , last, last member, member previous, previous depose, depose
ruling, ruling family, family ,, , targaryens, targaryens ,, , also, also scheming, scheming regain,
regain throne, throne ., . friction, friction house, house stark, stark ,, , lannister, lannister ,, ,
baratheon, baratheon targaryen, targaryen remain, remain great, great house, house greyjoy,
greyjoy ,, , tully, tully ,, , arryn, arryn ,, , tyrell, tyrell martell, martell lead, lead full-scale, full-
scale war, war ., . ancient, ancient evil, evil awaken, awaken farthest, farthest north, north .)]
```

Word2Vec

Word2vec “vectorizes” about words for natural language computer-readable performing operations on words to detect their similarities. Word2vec trains words against other words that neighbour them in the input corpus.

Sample Output

```
(1048576,[1035612],[0.0])
(1048576,
[66092,185129,201038,297264,325189,360210,399040,480792,533116,549206,590226,723078,
736708,787968,816744,826521,978809,1035612,1040706],
[1.0986122886681098,1.0986122886681098,1.0986122886681098,1.0986122886681098,1.098
6122886681098,0.6931471805599453,1.0986122886681098,1.0986122886681098,1.098612288
6681098,1.0986122886681098,1.0986122886681098,0.6931471805599453,1.098612288668109
8,0.4054651081081644,0.6931471805599453,1.0986122886681098,0.6931471805599453,0.0,1.
0986122886681098])
(1048576,
[4200,30084,44133,93484,178985,277042,281561,300909,306775,322412,342796,370708,4013
68,421014,429905,456678,494511,513968,529877,530339,634098,641070,649714,656398,7230
78,731198,751599,787968,813542,816744,859041,924359,1032531,1035612,1041588],
[1.0986122886681098,1.0986122886681098,1.0986122886681098,1.0986122886681098,1.098
6122886681098,1.0986122886681098,1.0986122886681098,0.6931471805599453,1.098612288
6681098,1.0986122886681098,1.0986122886681098,1.0986122886681098,1.098612288668109
8,1.0986122886681098,0.6931471805599453,1.0986122886681098,2.1972245773362196,1.098
6122886681098,2.1972245773362196,1.0986122886681098,1.0986122886681098,1.098612288
6681098,1.0986122886681098,1.0986122886681098,0.6931471805599453,2.197224577336219
6,1.0986122886681098,1.6218604324326575,2.1972245773362196,0.6931471805599453,0.693
1471805599453,0.6931471805599453,1.0986122886681098,0.0,1.0986122886681098])
(1048576,[1035612],[0.0])
```

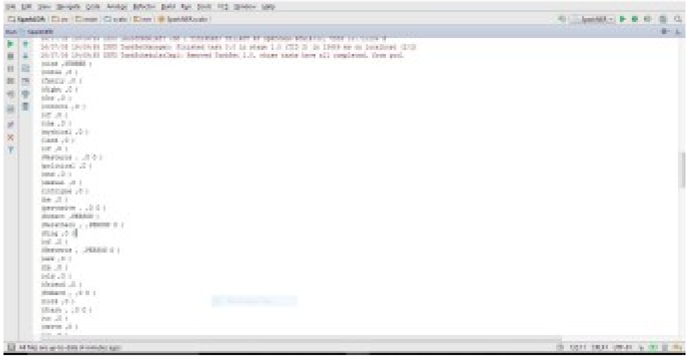


Name Entity Extraction/Relation Extraction

It is a subtask of information that seeks to locate and classify named entities in text into pre-defined categories such as the names of persons, organizations, locations, expressions of times, quantities, monetary values, percentages, etc.

Sample Output

```
(nine ,NUMBER )
(noble ,O )
(family ,O )
(fight ,O )
(for ,O )
(control ,O )
(of ,O )
(the ,O )
(mythical ,O )
(land ,O )
(of ,O )
(Westeros . ,O O )
(political ,O )
(and ,O )
(sexual ,O )
(intrigue ,O )
(be ,O )
(pervasive . ,O O )
(Robert ,PERSON )
(Baratheon , ,PERSON O )
(King ,O )
(of ,O )
(Westeros , ,PERSON O )
```



WordNet

WordNet was developed by Princeton University. It is a lexical database in english. It is similar to that of a thesaurus but has distinct feature of identifying a word based on the contect as well and is very useful for natural language processing.

Sample Ouput

```
definitions for king:
a male sovereign; ruler of a kingdom
a competitor who holds a preeminent position
a very wealthy or powerful businessman
preeminence in a particular category or group or field
United States woman tennis player (born in 1943)
United States guitar player and singer of the blues (born in 1925)
United States charismatic civil rights leader and Baptist minister who campaigned against the segregation of Blacks (1929-1968)
a checker that has been moved to the opponent's first row where it is promoted to a piece that is free to move either forward or backward
one of the four playing cards in a deck bearing the picture of a king
(chess) the weakest but the most important piece

Synonyms for war (pos: v)
assail
attack
bandy
battle
bear down
blitzkrieg
box
chicken-fight
Chickenfight
combat

Hyponyms for sea:
```

South Sea

head sea

Hypernyms for sea:

body of water

water

large indefinite quantity

large indefinite amount

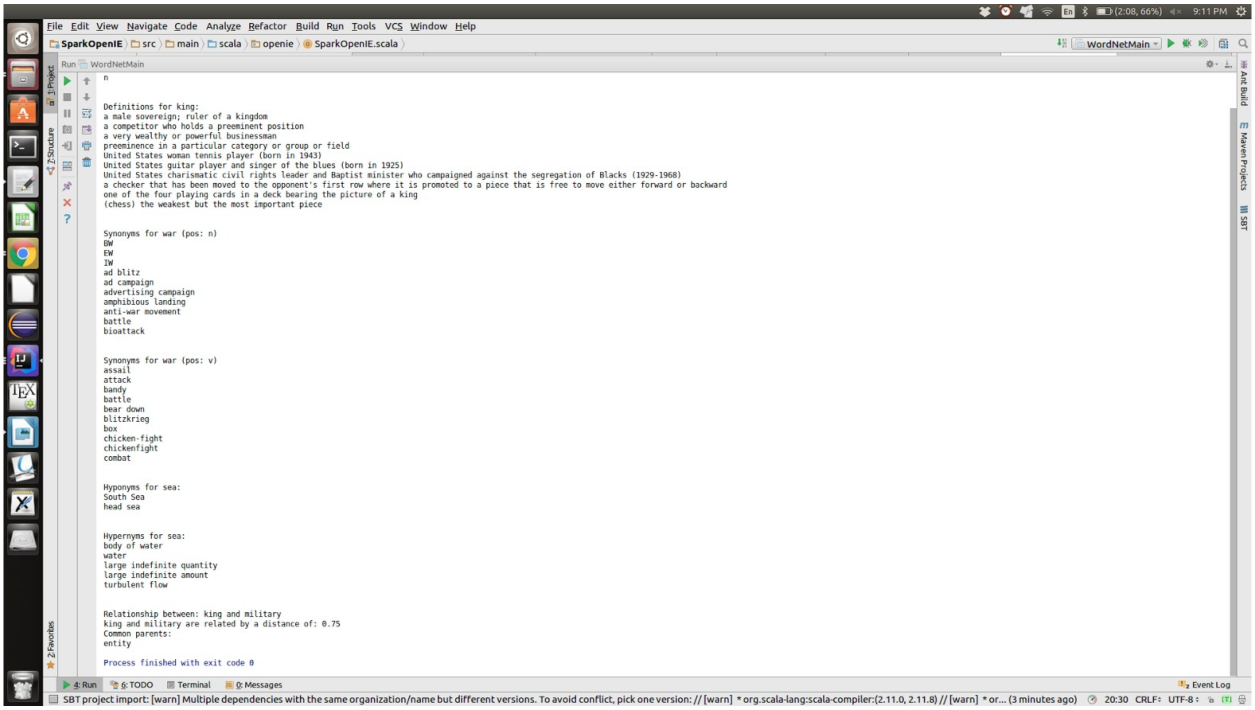
turbulent flow

Relationship between: king and military

king and military are related by a distance of: 0.75

Common parents:

entity



```
n


Definitions for king:
a male sovereign; ruler of a kingdom
a competitor who holds a preeminent position
a very wealthy or powerful businessman
preeminence in a particular category or group or field
United States woman tennis player (born in 1943)
United States guitar player and singer of the blues (born in 1925)
United States charismatic civil rights leader and Baptist minister who campaigned against the segregation of Blacks (1929-1968)
a checker that has been moved to the opponent's first row where it is promoted to a piece that is free to move either forward or backward
one of the four playing cards in a deck bearing the picture of a king
(chess) the weakest but the most important piece


Synonyms for war (pos: n)
BW
EW
IW
ad blitz
ad campaign
advertising campaign
amphibious landing
anti-war movement
battle
bioattack


Synonyms for war (pos: v)
assail
attack
bandy
battle
bear down
blitzkrieg
box
chicken-fight
chickenfight
combat


Hyponyms for sea:
South Sea
head sea


Hypernyms for sea:
body of water
water
large indefinite quantity
large indefinite amount
turbulent flow


Relationship between: king and military
king and military are related by a distance of: 0.75
Common parents:
entity

Process finished with exit code 0
```

SBT project import: [warn] Multiple dependencies with the same organization/name but different versions. To avoid conflict, pick one version: // [warn] * org.scala-lang:scala-compiler:(2.11.0, 2.11.8) // [warn] * or... (3 minutes ago)

File Edit View Navigate Code Analyze Refactor Build Run Tools VCS Window Help
SparkOpenIE src main scala openie SparkOpenIE.scala
WordNetMain
Run WordNetMain
4: Run 6: TODO Terminal 0: Messages
Event Log
20:30 CRLF: UTF-8

LDA

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. It considers all the documents given and extracts the topics out of them.

Input

Ned Stark, Lord of Winterfell, becomes the Hand of the King after the former Hand, Jon Arryn, has passed away. But before Ned goes to the capital, King's Landing, a letter arrives from his wife's sister Lysa, who was the wife of Jon Arryn. There it says that her husband was murdered, and it is up to Ned to find out what's going on. But that isn't everything. The White Walkers have been seen, and they seem to go down south
In the mythical continent of Westeros, several powerful families fight for control of the Seven Kingdoms. As conflict erupts in the kingdoms of men, an ancient enemy rises once again to threaten them all. Meanwhile, the last heirs of a recently usurped dynasty plot to take back their homeland from across the Narrow Sea.

Ned Stark, Lord of Winterfell, becomes the Hand of the King after the former Hand, Jon Arryn, has passed away. But before Ned goes to the capital, King's Landing, a letter arrives from his wife's sister Lysa, who was the wife of Jon Arryn. There it says that her husband was murdered, and it is up to Ned to find out what's going on. But that isn't everything. The White Walkers have been seen, and they seem to go down south
In the mythical continent of Westeros, several powerful families fight for control of the Seven Kingdoms. As conflict erupts in the kingdoms of men, an ancient enemy rises once again to threaten them all. Meanwhile, the last heirs of a recently usurped dynasty plot to take back their homeland from across the Narrow Sea.

While a civil war brews between several noble families in Westeros, the children of the former rulers of the land attempt to rise up to power. Meanwhile a forgotten race, bent on destruction, plans to return after thousands of years in the North.

Output

```
TOPIC 0

family 0.023017295006465092
hand 0.02183801929921515
king 0.017499860756227465
war 0.01697726745883248
```

```
westeros 0.01636905198657341
stark 0.015976305812649007
north 0.012137062587934936
land 0.011923130211216259
noble 0.01189760842786202
power 0.011820341481909457

TOPIC 1

family 0.024097923706312783
eddard 0.019287550009948532
order 0.01927589188706488
political 0.019271936125048704
house 0.019249370977486694
war 0.01913844176829542
previous 0.018952240176458666
baratheon 0.018671246470817096
stark 0.016262502568671146
westeros 0.015066157921142178
king 0.011312642153205814

TOPIC 2

family 0.03846132544150953
westeros 0.03803304617281484
kingdom 0.03568243538234665
rise 0.031123686367345582
ancient 0.02103059368589765
control 0.020960398530144236
plot 0.020913338179132948
fight 0.02085820049546466
sea 0.020755867353489187
mythical 0.020675302925185674
```

TOPIC 3

ned 0.0536248351281768

arryn 0.04097912724738723

king 0.04044884662172056

hand 0.03717547900133889

jon 0.03445662108433512

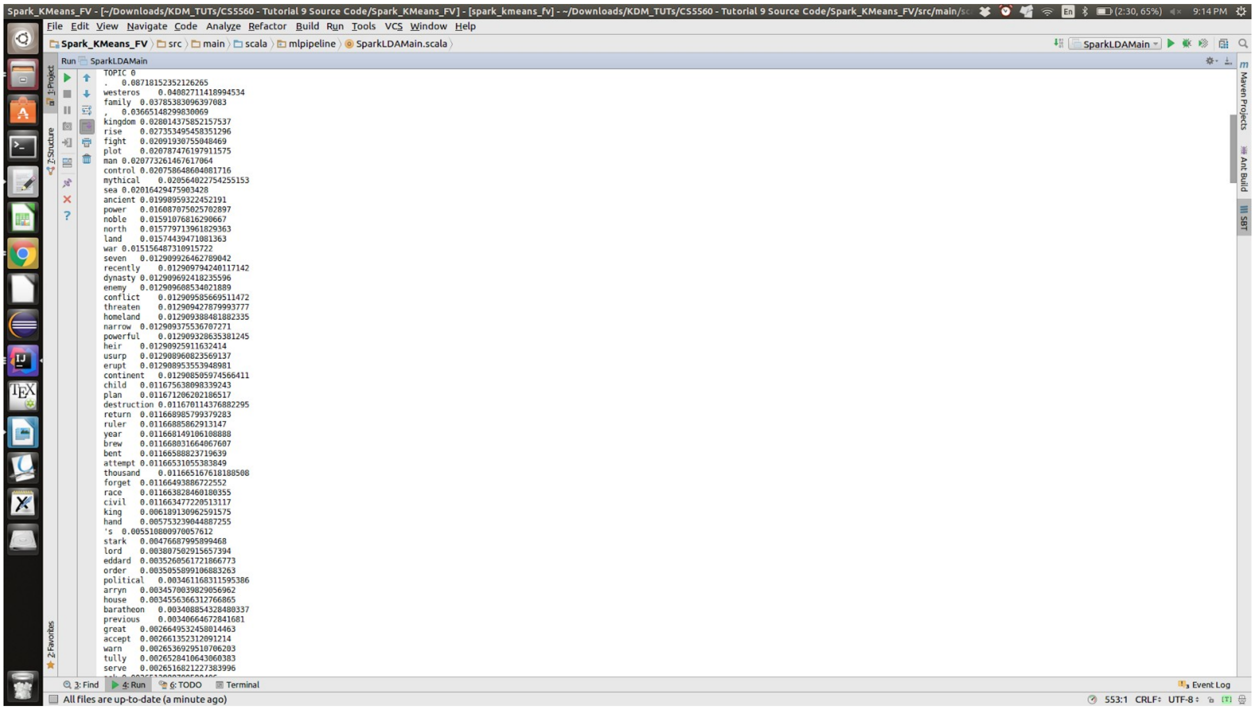
wife 0.03434034953002408

lord 0.017938472834780422

stark 0.015801076493262067

arrive 0.015630130439577238

white 0.015629750809597404



```
TOPIC 0
.	0.08718152352126265
westeros	0.04082711418994534
family	0.03785383096397083
,	0.03665148299830069
kingdom	0.028014375852157537
rise	0.027353495458351296
fight	0.02091930755048469
plot	0.020787476197911575
man	0.020773261467617064
control	0.020758648604081716
mythical	0.020564022754255153
sea	0.02016429475903428
ancient	0.01998959322452191
power	0.016087075025702897
noble	0.01591076816290667
north	0.015779713961829363
land	0.01574439471081363
war	0.015156487310915722
seven	0.012909926462789042
recently	0.012909794240117142
dynasty	0.012909692418235596
enemy	0.012909608534021889
conflict	0.012909585669511472
threaten	0.012909427879993777
homeland	0.012909388481882335
narrow	0.012909375536707271
powerful	0.012909328635381245
heir	0.01290925911632414
usurp	0.012908960823569137
erupt	0.012908953553948981
continent	0.012908505974566411
child	0.011675638098339243
plan	0.011671206202186517
destruction	0.011670114376882295
return	0.011668985799379283
ruler	0.01166885862913147
year	0.011668149106108888
brew	0.011668031664067607
bent	0.01166588823719639
attempt	0.01166531055383849
thousand	0.011665167618188508
forget	0.01166493886722552
race	0.011663828460180355
civil	0.011663477220513117
king	0.006189130962591575
hand	0.005753239044887255
's	0.005510800970057612
stark	0.00476687995899468
lord	0.003807502915657394
eddard	0.0035260561721866773
order	0.0035055899106883263
political	0.003461168311595386
arryn	0.0034570039829056962
house	0.0034556366312766865
baratheon	0.003408854328480337
previous	0.00340664672841681
great	0.0026649532458014463
accept	0.002661352312091214
warn	0.0026536929510706203
tully	0.0026528410643060383
serve	0.0026516821227383996
```

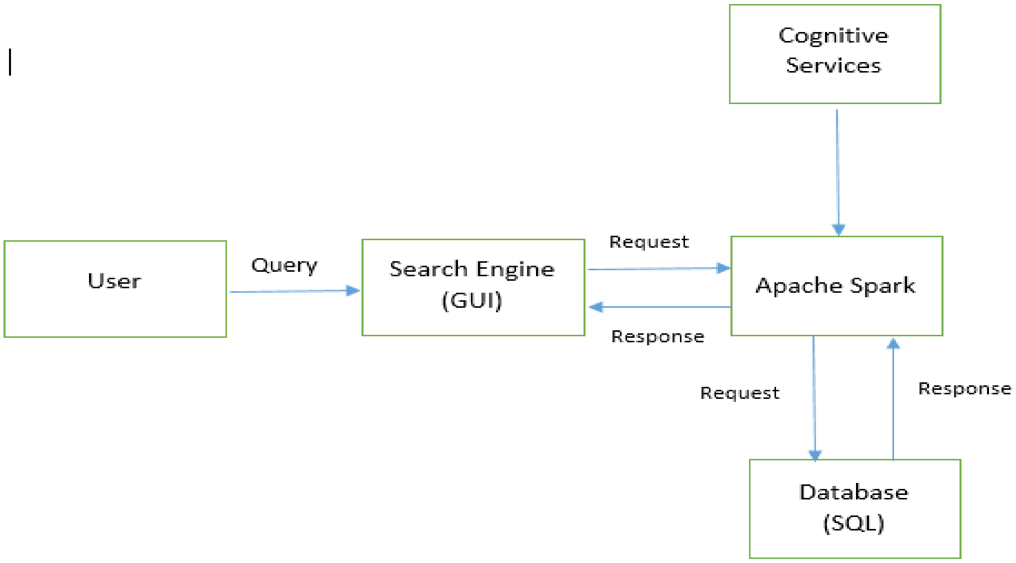
Feature Vector

A feature vector is an n-dimensional vector of numerical features that represent required object. Most of the machine learning algorithms require that the input be given in a numerical format which can further be used for statistical and predictive analysis. We used the word2Vec data input and the feature vector thus obtained is already mentioned in the Word2Vec.

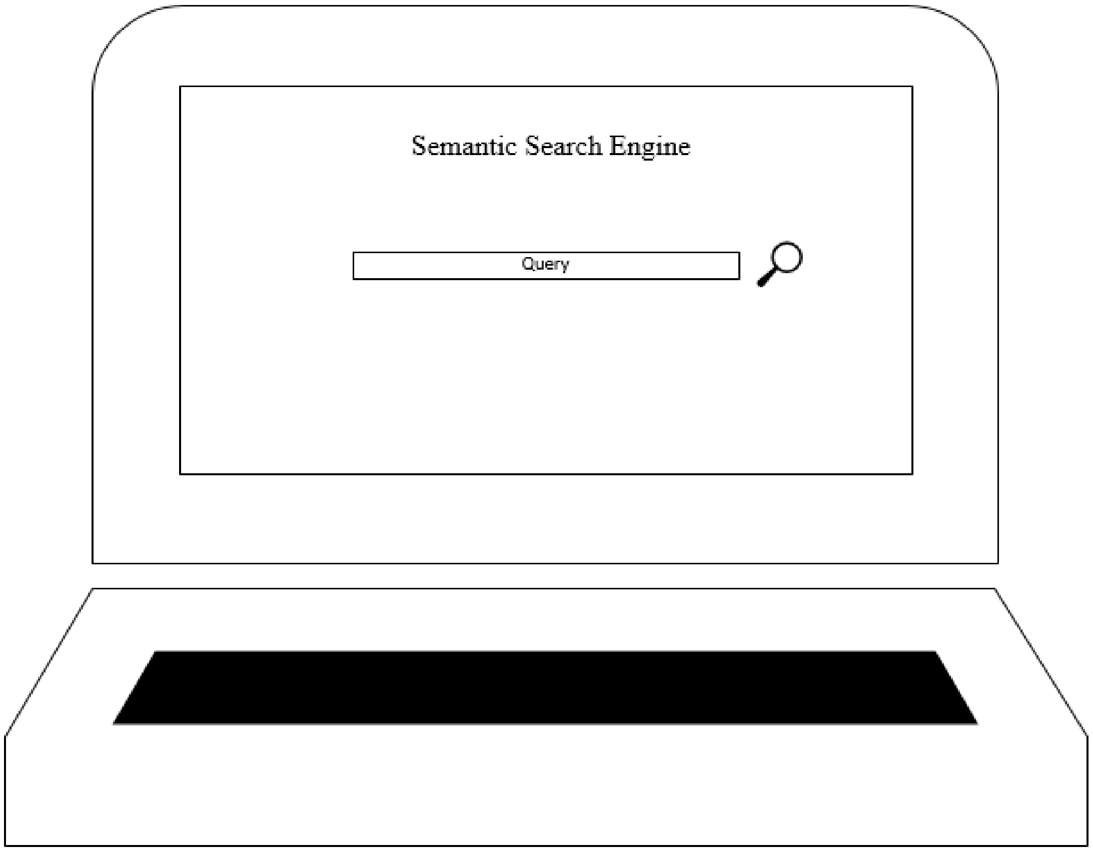
Sample Output

```
(1048576,
[4200,30084,44133,93484,178985,277042,281561,300909,306775,322412,342796,370708,4013
68,421014,429905,456678,494511,513968,529877,530339,634098,641070,649714,656398,7230
78,731198,751599,787968,813542,816744,859041,924359,1032531,1035612,1041588],
[1.0986122886681098,1.0986122886681098,1.0986122886681098,1.0986122886681098,1.098
6122886681098,1.0986122886681098,1.0986122886681098,0.6931471805599453,1.098612288
6681098,1.0986122886681098,1.0986122886681098,1.0986122886681098,1.098612288668109
8,1.0986122886681098,0.6931471805599453,1.0986122886681098,2.1972245773362196,1.098
6122886681098,2.1972245773362196,1.0986122886681098,1.0986122886681098,1.098612288
6681098,1.0986122886681098,1.0986122886681098,0.6931471805599453,2.197224577336219
6,1.0986122886681098,1.6218604324326575,2.1972245773362196,0.6931471805599453,0.693
1471805599453,0.6931471805599453,1.0986122886681098,0.0,1.0986122886681098])
(1048576,[1035612],[0.0])
```

System Architecture:

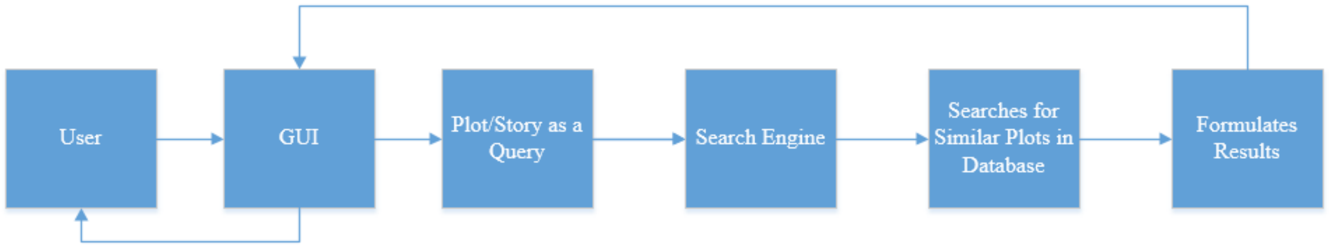


Wireframe of Search Engine:

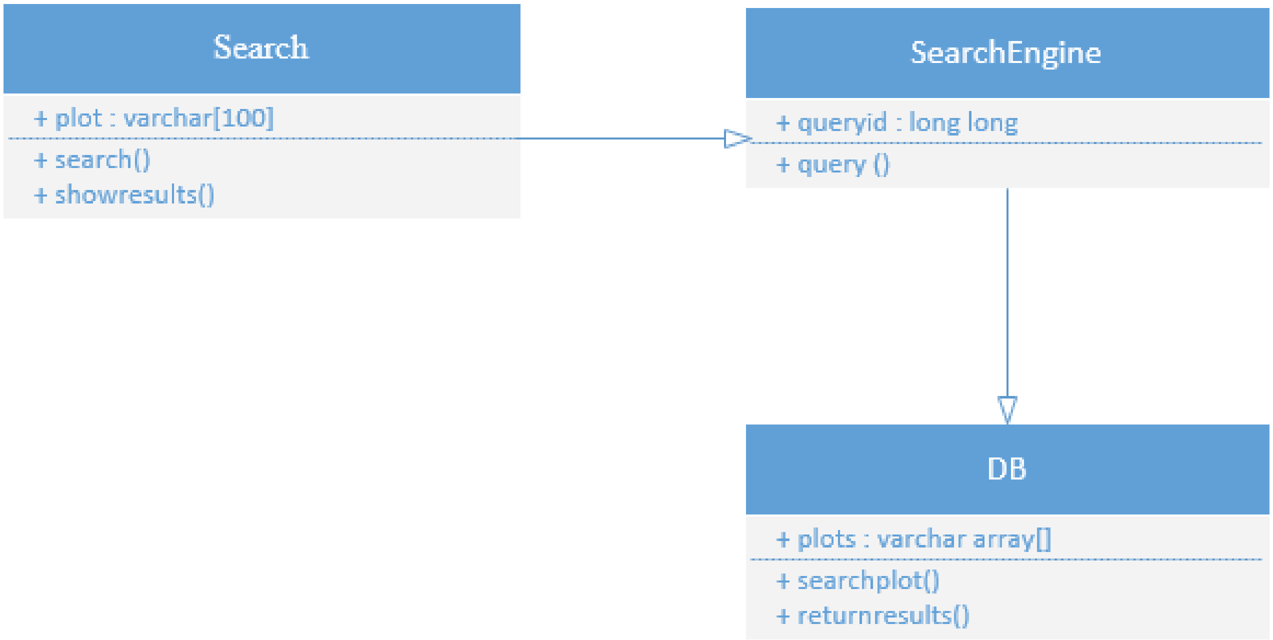


Workflow Diagram:

`

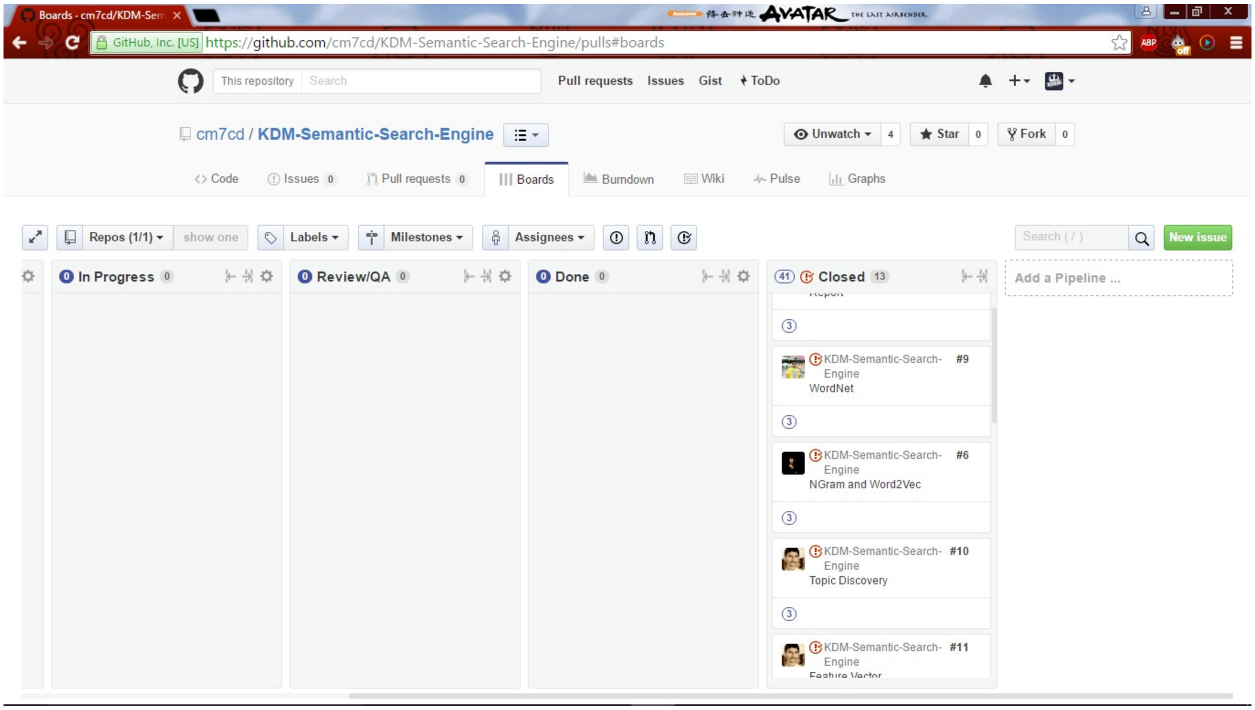


UML Class Diagram:

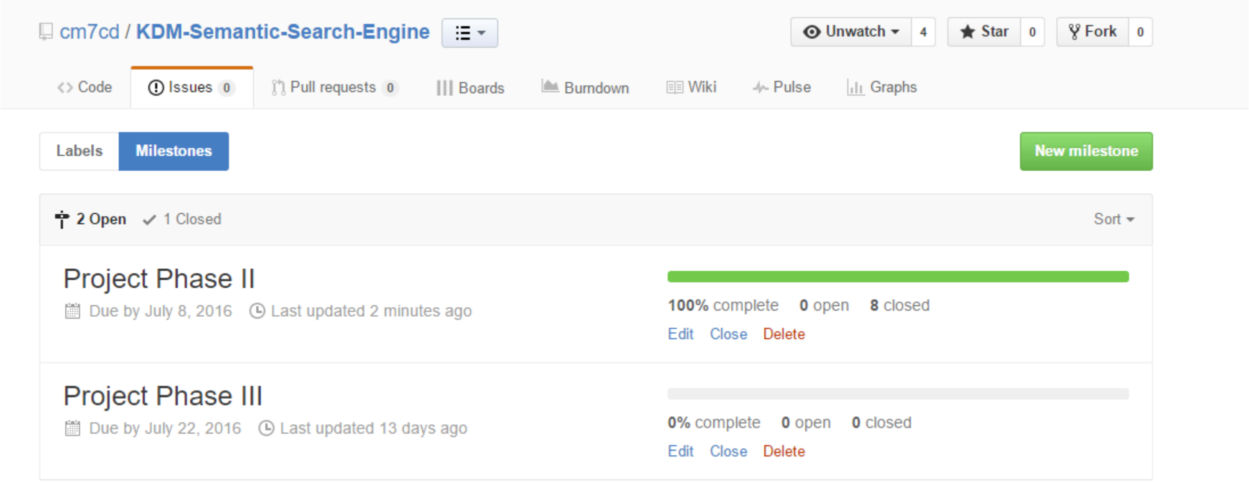


ZenHub Project Management:

ZenHub Board:



Milestones:

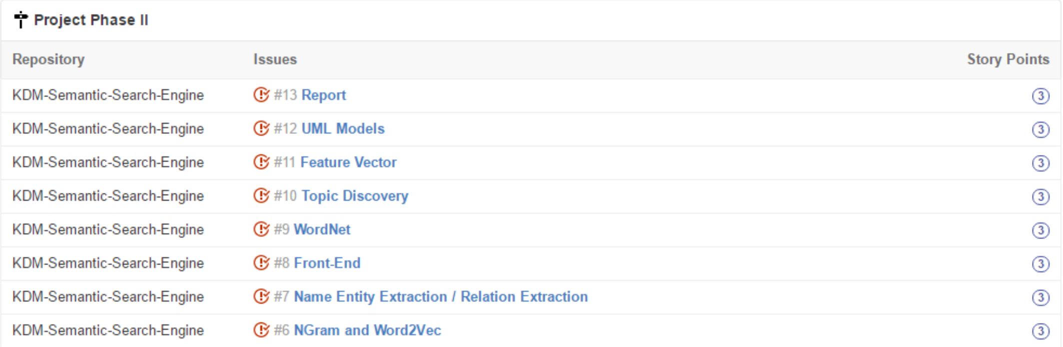


2 Open ✓ 1 Closed

Project Phase II
Due by July 8, 2016 Last updated 2 minutes ago
100% complete 0 open 8 closed

Project Phase III
Due by July 22, 2016 Last updated 13 days ago
0% complete 0 open 0 closed

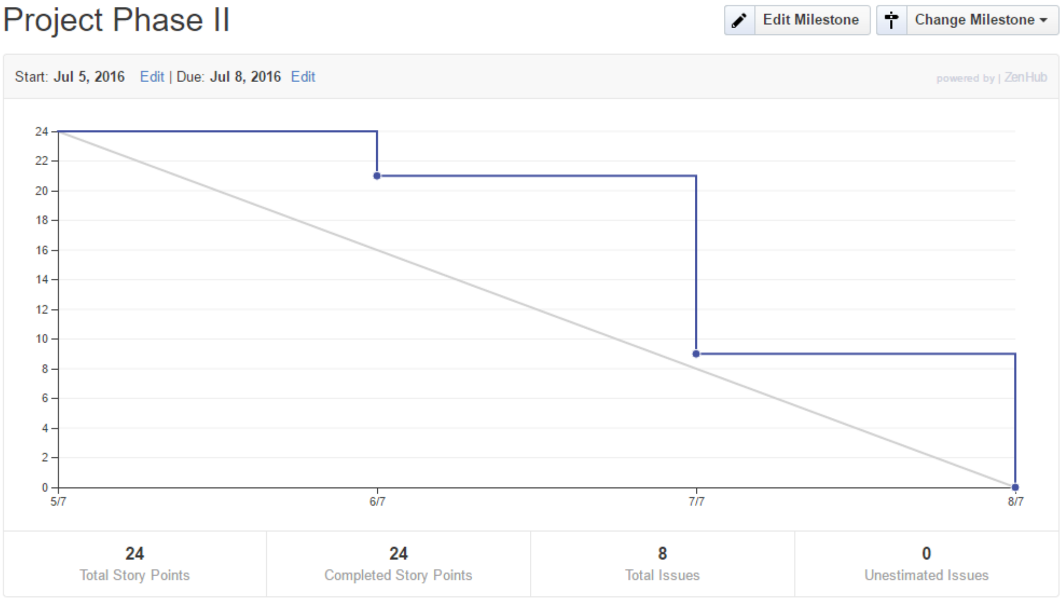
Issues:



Project Phase II

Repository	Issues	Story Points
KDM-Semantic-Search-Engine	#13 Report	3
KDM-Semantic-Search-Engine	#12 UML Models	3
KDM-Semantic-Search-Engine	#11 Feature Vector	3
KDM-Semantic-Search-Engine	#10 Topic Discovery	3
KDM-Semantic-Search-Engine	#9 WordNet	3
KDM-Semantic-Search-Engine	#8 Front-End	3
KDM-Semantic-Search-Engine	#7 Name Entity Extraction / Relation Extraction	3
KDM-Semantic-Search-Engine	#6 NGram and Word2Vec	3

Burndown Chart:



Project Phase II

Start: Jul 5, 2016 | Due: Jul 8, 2016

24	24	8	0
Total Story Points	Completed Story Points	Total Issues	Unestimated Issues

Contributions:

Sai Venakatesh Gatiganti - Name Entity Extraction / Relation Extraction, NGram & Word2Vec
Karthik Reddy Vundela – Feature Vector & Topic Discovery.
Chaitanya Sai Manne – Front-End & Wordnet.
Sri Chaitanya Patluri – UML Models, Workflow, Wireframes & Report.

Project Goal:

Our goal for the next phase is to develop a full-fledged semantic search engine by constructing knowledge graphs and ontologies.

Future Work:

Our Future Work for the project includes including various other domains 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.

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

Bibliography:

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