Z604: Big Data Analytics for Web and Text

Project Report

On

Extracting Causal Relationship between Major Events using Wikipedia Data

(Wikipedia Group - 2)

Anwar Shaikh

Jay Nagle

Prathik Rokhade

Sanjana Pukalay

Vinay Vernekar

Table of Contents

Contents

1. Introduction 3
1.1. Research Question 3
1.2. Overview 3
1.3. What is an Event? 3
2. Dataset Description 3
3. Data Description 4
3.1. Data Selection 4
3.2. Data Sampling 4
3.3. Data Indexing 4
4. Methodology 5
4.1. Identifying similar events 5
4.2. Identifying similar assets or entities affected 5
4.3. Creating similarity matrix 6
4.4. Identifying the sentiment associated with an event 6
4.5. Visual aid for Similarity matrix 6
5. Experiment 7
5.1. Finding the sequence of Events 7
5.2. Semantic Role Labelling 7
5.3. Identifying the year of occurrence 8
5.4. Lexical diversity 8
6. Analysis 8
7. Neo4j Database 9
7.1. Event File Generation 9
7.2. Neo4j event properties 9
7.3. Neo4j graph description 9
7.4. Interpretation of reading the Output: 11
8. Evaluation 11
8.1. Why F-Measure and not MAP or NDCG? 12
8.2. Experiment Results 12
9. Conclusion 12

1. Introduction

1.1. Research Question

The research question focuses on identifying “Events” from Wikipedia dataset, extracting sub-events and understanding the correlation and causal relationships between them.

1.2. Overview

In this study, we propose a preliminary solution for finding correlation and causal relationship between Events/Entities that have a chain of Sub-Events. It can be accomplished by building a Timeline of Events according to their year of occurrence.

The timeline of events thus created would help us visualize the correlation between various events and study relationship between them. It will also help in predicting the root cause of a certain event and explore sub-events associated with it.

1.3. What is an Event?

An event is the occurrence of a significant or unusual happening in the past that was caused by a series of reasons which are essentially called sub-events and which itself can act as a sub-event for another major event. For instance, booms, depressions and bubbles are examples of financial events.

An event is basically something which has sequence of cause and effects that combined together and leads to the main event. An object to qualify as an event needs to have the below properties.

1. An origin date or period
2. Sequence of minor activities that build up or contribute together and result in phenomena Example Battles. In case of Battles we have a reason as to why the Battle erupted this is a “Cause”, when it reputed is the “Origin date” or “Year” and what happened in the Battle like who won or lost can be a potential “Effects”. A battle consists of many small wars, these can be sub events which have their own important dates, effects and causes
3. An event cannot be a definition
4. A few examples of the events are as follows. Major historic events like Economic Crisis, Economic Boom, Wars Like world Wars, Spread of Capitalism, freedom struggle of a country etc.

2. Dataset Description

2.1. Data Source

For this study, we used Wikipedia dataset available for educational purposes at: https://dumps.wikimedia.org/enwiki/20151201/

The provided data dump is in well-known BZIP2 format which houses compressed structured data in XML format detailing information about articles, templates, metadata and pages etc. For our research question we used current dataset which has information about latest version only and provides no revision history.

2.2. Data Size

The compressed dataset downloaded from above source is approx. 22 GB in size. The uncompressed data is about 110 GB.

We performed our analysis on a sizable subset of this uncompressed data. For this, we extracted approximately 1130 pages related to Finance and Economy categories.

3. Data Description

3.1. Data Selection

As Wikipedia dump has a huge amount of data from numerous categories, it would have been very difficult to perform analysis on all the data at once and would potentially have yielded sparse results. As we were trying to find correlation between events, it was important that we sampled a consistent corpus.

Once we establish the validity of our model on a uniform subset of data, we would be able to extend the same model to analyse any category from Wikipedia dataset.

3.2. Data Sampling

Out of several alternatives, we decided to work with Financial and Economic crisis categories as it had a fairly large content available and it was easily possible to identify major events from the sample. Some of the other alternatives we considered were War, Medical and Economic Boom categories.

For sampling, we used a self-written multi-threaded Java parser to which we provided a list of categories as a filter to retrieve pages. We had to fine tune our parser and perform several iterations to ensure good quality of data was retrieved and in reasonable time. We managed to reduce sampling time taken from over a day to a few hours.

3.3. Data Indexing

Apache Lucene is a high-performance, full-featured text search engine library written in Java. It is suitable for applications that require full-text search, especially cross-platform. Lucene has been used to create an index on our sampled data. Though considering our NLP analysis we could have used other tool but looking at our research question and the nature of data lucene seems to be fairly good as it allows easy manageable access. In addition to that Lucene was chosen because it provides efficient access to textual data and flexibility to perform analysis (e.g. creating correlation matrix) we intended to perform. The retrieved data from Java parser was fed to this library. By creating a pipeline between sampler and indexer we managed to save considerable amount of time and eliminated repeated redundant access to flat data files.

To perform causal analysis, we needed only plain text from sampled data pages. For this reason, Lucene index was created using English Analyzer and removed stop words, tabular data and other unnecessary information. The index schema has 3 fields viz. PageId, Title and Page Text. In future we can store other fields like incoming-links and out-going links for the page.

4. Methodology

4.1. Identifying similar events

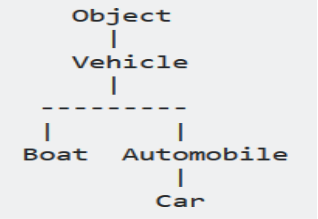
In order to find causality or the affect one event on other events we need to identify similar assets or entities that were affected and in what manner either positive or negative. For the purpose of identifying similar events or assets we used the nouns as nouns represent the names of the assets or the entities, in case of financial data they represent the company's stock, stock indexes, currency, industrial sector affected. For example “Mortgage sector were adversely affected in Subprime crises”. In this sentence, the code will pick up Mortgage as a noun and it will be the entity that was affected in a negative way (as it related economic crisis list of events). Similarly in European crisis, Mortgage assets and bonds were affected. These affected entities or assets are mentioned frequently in the wiki page or contents with reference to the topic and thus have high frequency. This was tested for 91 events and the output was conclusive.

We used NLTK to identify all the Nouns in the text and then used the top frequency nouns. We experimented with different top frequency nouns (50, 40, 30 etc.) and finally based on the content analysis selected 30 to be an optimal cut-off frequency. This was done after going through the results and examining them.

4.2. Identifying similar assets or entities affected

Given the nature of the research question we had to work on the limited data in order to develop a structure that could be further applied in a general manner to any event which consists of a cause, result and a subject of interest. Hence we did not use word to vector as it requires a substantial large amount of data to train. The alternate approach that we followed to look into The Wu & Palmer measure (wup). The Wu & Palmer calculates relatedness by considering the depths of the two synsets in the WordNet taxonomies, along with the depth of the LCS (Least Common Subsumer).

Explanation for Wu & Palmer and Least Common Subsumer :



```
    Object
      |
    Vehicle
      |
  ---------
  |        |
 Boat   Automobile
           |
          Car
```

In this case, "automobile" is the parent (and also ancestor) of "car", while "vehicle" is an ancestor of "car". "Vehicle" is also an ancestor of "boat". In this case, the LCS of "boat" and "car" is "vehicle", since it's the most specific concept which is an ancestor of both "boat" and "car". Note that while "object" is a common subsumer of both "boat" and "car", it is not the least, since there is still a child of "object" (in this case it's "vehicle") which is also a common subsumer of both "car" and "boat". "Automobile" is not the least common subsumer since it's not an ancestor of "boat".

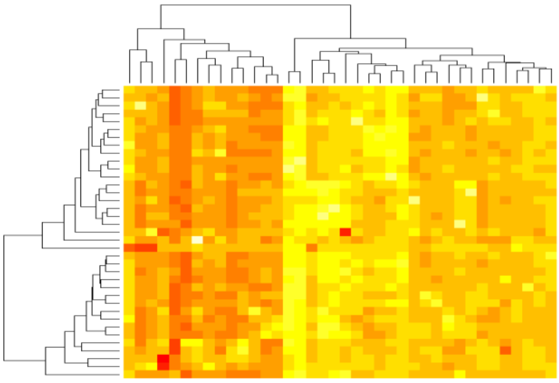
4.3. Creating similarity matrix

After identifying the top frequency Nouns, we calculated the WUP score for each event with other events and created a Similarity matrix. We also normalized the matrix by dividing each event with reference to the main event being compared. Example consider we have events from “A” to “Z”. Event “A” would be compared with other events from “A” to “Z” and the relation or WUP score of “A” with all other events including itself would be divided by the WUP score of “A” to get the normalized WUP value. Similarly “B” would be compared to events from “B” to “Z” and the relation or WUP score of “B” with all other events including itself would be divided by the WUP score of “B” to get the normalized WUP value. This action will happen event wise.

4.4. Identifying the sentiment associated with an event

We assumed that the title or the introductory paragraph of an event to an extent would represent the sentiment of the event. Either Positive or Negative. For this purpose we used the Sentiment analysis algorithm in Graphlab (Code provided). But the sentiment score were incorrect. The lack of wordnet or classifier trained on financial data was one of the cause, as the graph lab classifiers work much better on customer review data. Hence we manually annotated the data for each event after reading the content. This was necessary as two events may have similar asset class or entities but they may be affected in different way. Example in time of economic growth Mortgage or automobile industry will be positively affected whereas in times of recession it would be negatively affected. Hence it was necessary to find a particular sentiment associated with these events in order to properly associate them with each other and parent emotion either positive or negative. This would help us further in storing that data in Neo4j and visualizing it.

4.5. Visual aid for Similarity matrix



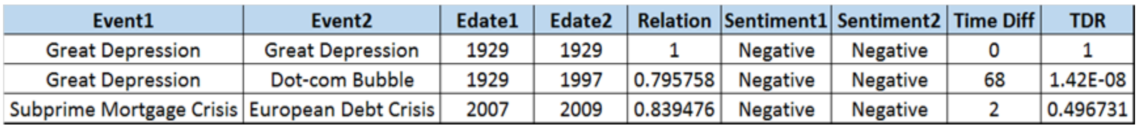
The above graph is a heat map dendrogram based on the WUP similarity matrix. The areas with higher colour of red are high on WUP similarity score and as the WUP value reduces the colour becomes lighter like orange and yellow.

5. Experiment

5.1. Finding the sequence of Events

It was necessary to identify the year when the event took place. This is necessary because a future event cannot affect an event in the past. Another reason we need to identify the sequence of events is that if two events have close proximity in terms of time i.e. years. E.g. If event “A” happened in 1930 and event “B” happened in 2008, then the potential impact of the event of 1930 will be less on event of 2008. Similarly if an event “C” happened in 2007 then the impact of this event on the event of 2008 will be high.

In order to accommodate this we introduced a time based penalty for all events based on the difference in years between any two events. As seen in the table below the difference between the event “Great Depression” and “Dot- com Bubble” is 68 years apart and hence a time related penalty was introduced. The table below has different columns. “Event1” is one of the events and “Event2” is another event. They have respective year of occurrence “Edate1” and “Edate2” and the column “Relation” represents the WUP similarity between these two events. The column “Sentiment1” and “Sentiment2” represents the sentiment related with the respective event. “Time Diff” is the difference in years between the occurrences of these events. The column “TDR” stands for Time Difference adjusted Relation. The “TDR” is calculated using a constant of “1.3” and is raised to the negative power of the “Time Diff” to generate a weight factor. The “Relation” is then multiplied with this weight to get the new “TDR” for each event.



Event1	Event2	Edate1	Edate2	Relation	Sentiment1	Sentiment2	Time Diff	TDR
Great Depression	Great Depression	1929	1929	1	Negative	Negative	0	1
Great Depression	Dot-com Bubble	1929	1997	0.795758	Negative	Negative	68	1.42E-08
Subprime Mortgage Crisis	European Debt Crisis	2007	2009	0.839476	Negative	Negative	2	0.496731

5.2. Semantic Role Labelling

NLTK is a platform for building Python programs to work with human language data. It provides features for text Classification, Tokenization, Stemming. PractNLPTools (Practical Natural Language Processing Tools) is a Python library used for Dependency Parsing, Syntactic Constituent Parsing, Semantic Role Labelling, Part of Speech tagging. In this study, a combination of NLTK and PractNLPTools has been used to perform Semantic Role Labelling (SRL) and POS Tagging on indexed data.

NLTK takes a piece of text as input and analyses the linguistic structure of each sentence and potentially identifies relationship between them. It categorizes the text into various parts of speech such as nouns, verbs, prepositions etc. and also finds probable causality between subjects and objects.

In our study, Semantic Role Labelling has been used to identify sub-events that contribute to the main event.

Sample Input:

“The recession caused demand for energy to shrink in late 2008.”

Annotated Output:

{'A1': 'demand for energy to shrink in late 2008', 'A0': 'The recession', 'V': 'caused'}

5.3. Identifying the year of occurrence

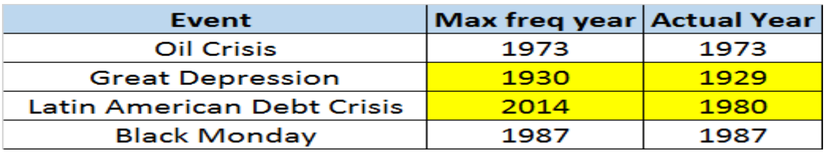
The purpose of this experiment was to explore the possibility of identifying the year of occurrence of an event and see whether the frequency of the year being mentioned in a page can be used as factor to estimate the year of origin of the event.

For this purpose we created a matrix with all the years encountered in the pages as features "Y- axis" and the "X-axis" has the title of the page. Then the frequency of occurrence or mention of each year in the format (YYYY) in a given page title was calculated/counted. Then the feature with highest frequency was picked and it was considered to be the year of occurrence of the event. These figures were compared to the actual year of occurrence (from external source).

The experiment was done on a sample of 91 events with F-1 score of ~65% and the following inferences were made.

1. This method is viable if the event occurred in the first half of the year i.e. preferably before August of each year.
2. The trigger for the event if important or significant can be identified in this way. Because after the major trigger or cause there is an increase in the mention of the year along with the date and the difference between the days (Not years) keeps on reducing.
3. The events with wrong predicted Year were mainly because of an embedded table in the page. These tables mainly represented the amount or value of lost assets adjusted for the present year (specific year when the page was created). These tables help the reader to understand the Net present value the assets or entity that was affected.

The Table below is a representation of some of the events that were identified with wrong year as origin highlighted in yellow



Event	Max freq year	Actual Year
Oil Crisis	1973	1973
Great Depression	1930	1929
Latin American Debt Crisis	2014	1980
Black Monday	1987	1987

5.4. Lexical diversity

As part of the research and midterm presentation there was an experiment conducted to identify the Lexical diversity variations between pages that belong to different categories. The main aim was to find the average Lexical diversity for each category of page and visualize it. However given its limited scope this was not included in the final project.

6. Analysis

Time lags play an important role while analysing the sequence of events which ultimately tells the causal relationship between the events. A thorough analysis of entities was performed and it turned out that the same entities are getting affected in different time lags. For every entity that is taken into consideration, it can be represented in terms of high frequency words associated with that entity. This has been the basic approach of the experiment. High frequency nouns are used for representing an entity.

About 91 main events and 2184 sub events were identified in the subset of Wikipedia data that was employed for the experimentation task. Every event has an average of 10-11 sub causes and effects which when combined together leads to a synergy that leads to the main event.

7. Neo4j Database

7.1. Event File Generation

After calculating WUP similarity, a csv file was generated containing the following factors:

- Event
- Event Description
- Event Cause Nouns (extracted the Nouns as to make the graph readable)
- Event Cause description
- Event Effect Nouns
- Event Effect description
- Event year
- Sentiment
- Similarity Coefficient

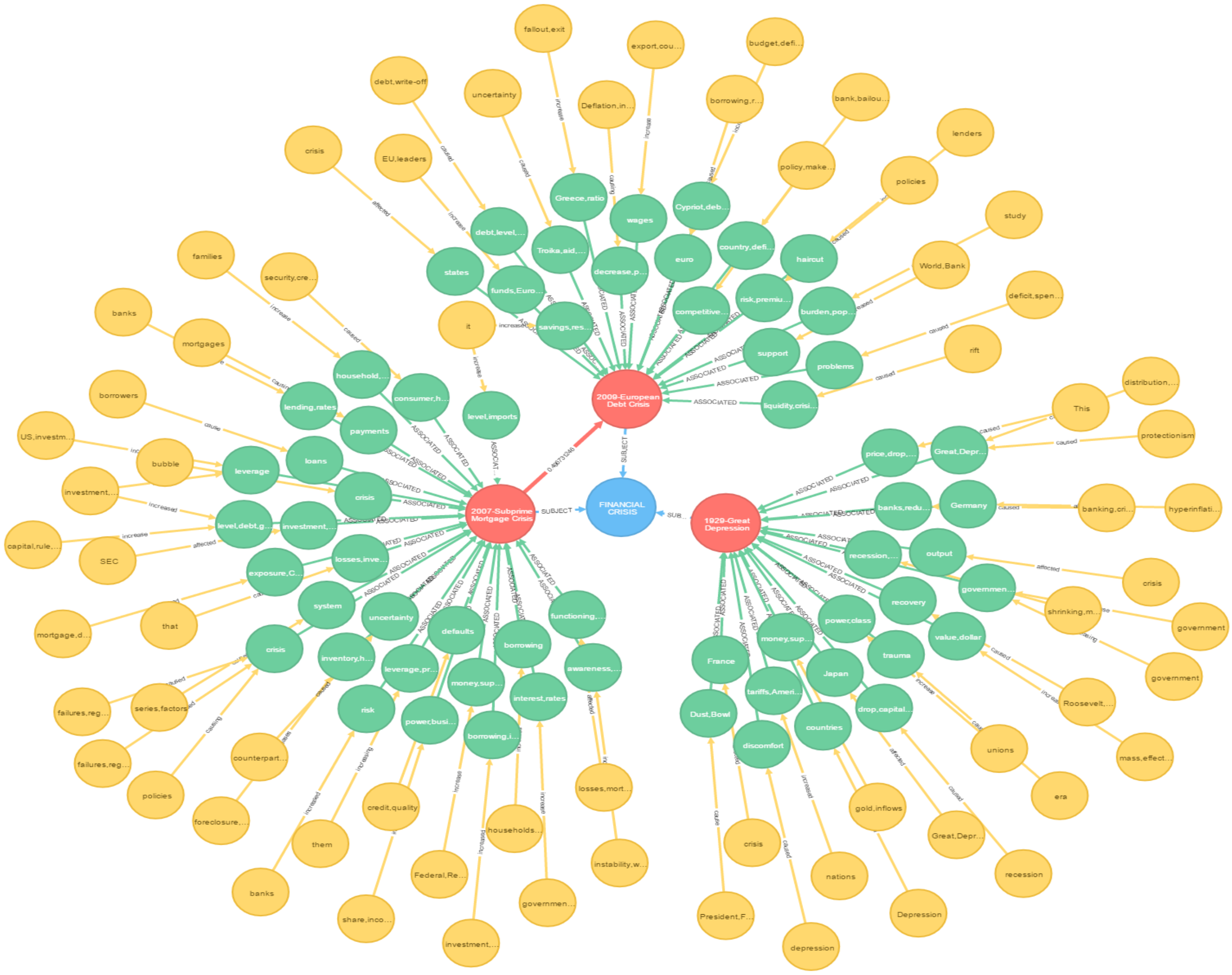
7.2. Neo4j event properties

Following is the properties of respective events in neo4j database.

```
Node_Type: 'SUBJECT'
Properties:
     Node_Name: 'FINANCIAL CRISIS'


Node_Type: MAIN_EVENT
Properties:
    Name: <EVENT_noun>
    Description: <EVENT_description>
    Year: <EVENT_year>
    Sentiment: <EVENT_sentiment>


Node_Type: CAUSE
Properties:
                 Name: <EVENT_Cause_Noun>
                 Description: <EVENT_Cause_description>


Node_Type: EFFECT
Properties:
                 Name: <EVENT_Effect_Noun>
                 Description: <EVENT_Effect_description>


Relationship - (SUBJECT)  <- [:Correlation {value: row.TDR}] - (MAIN_EVENT) <-
[:ASSOCIATED] - (EFFECT) <- [:ACTION {verb: row.Action}] -(CAUSE)
```

7.3. Neo4j graph description

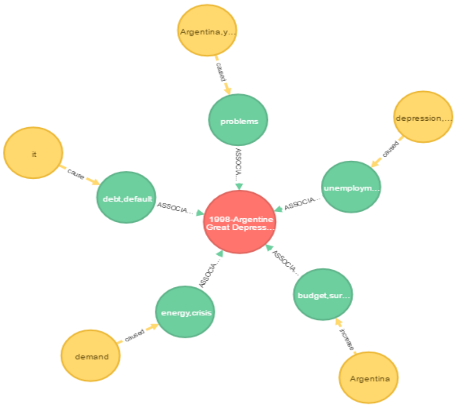
Below is an example of three MAIN EVENT Nodes with its associated CAUSE and EFFECT. Here, a MAIN_EVENT node is connected to other MAIN_EVENT (Red colour) node based on their Similarity Coefficient and the earlier occurred event is pointing to later event. Also, a MAIN_EVENT is connected to other event if it's 'Similarity_Co-efficient' > 0.25.



(Causal Relationship between Events)

BLUE Node - ‘SUBJECT’
RED Nodes - ‘MAIN_EVENT’
GREEN Nodes - ‘EFFECT’
YELLOW Nodes - ‘CAUSE’

The main advantage of Neo4j is that we can maintain a database for EVENT based search. For Example, We can look for the cause and effects of a particular event visually.



(Single Event with its Sub Events)

7.4. Interpretation of reading the Output:

The graph is based on the data processed by Semantic Role Labelling. To read the graph we start with the yellow node. This represents the EFFECT event, hovering over the graph displays the actual content or description, then the link connecting this to the green circle is the VERB or action word. The green circle represents the CAUSE, again hovering over this will display the actual text. Finally this green node is ASSOCIATED with MAIN EVENT highlighted in red.

Reading the graph in such a fashion helps in understanding the sub events and causes and how the flow occurred and how these events combined together to form or result in the main event.

Also, We could apply our algorithm to different topics like 'Financial Boom', 'Wars', 'Healthcare diagnosis' etc. and then scale our project with different related topics like 'Financial crisis' and 'Financial Boom' and layer them under one generic topic like 'Finance' in Neo4j. Similarly there would be other sub topics under generic topics like 'Health Care', 'Technology'.

8. Evaluation

Our preliminary evaluation mechanism is validation against InfoBox data available in Wikipedia pages. The limitation of this approach is that not every page has an InfoBox and no clear causes listed in a structured format.

To overcome this limitation we manually created ground truth for some sample events by reading the page content.

For a better evaluation, a more comprehensive F - Measure approach has been used. F-Score is a measure of test's accuracy. It considers Precision and Recall to compute the F score. Precision is the number of correct positive results divided by the number of positive results whereas Recall is the number of correct positive results divided by the number of positive results that should have been returned. The F score thus computed is a weighted average of Precision and Recall with its best value being 1 and worst being 0. Currently results were calculated are with β = 1, we can adjust the β parameter in order to provide more weightage to precision or recall for the experiment.

In our study, a total of 358 causes (sub-events) were used for evaluation and the resultant best F-measure is as follows:

Label	Causes	Not Causes
Identified Causes	TP = 172	FP = 88
Not Identified	FN = 98	TN = NA

The traditional F-measure or balanced F-score can be calculated as:

$$F_1 = 2 \cdot \frac{\mathrm{precision} \cdot \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}}$$

For the analysed data below are the values obtained:

Precision = 0.6615

Recall = 0.6370

Thus, F score can be computed as:

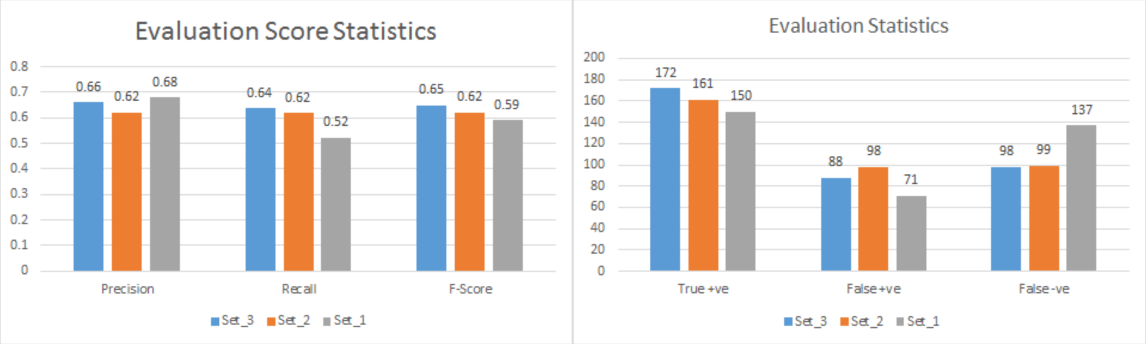
F-Score = 2 * 0.6615 * 0.6370 / (0.6615 + 0.6370) = 0.6490

8.1. Why F-Measure and not MAP or NDCG?

The results we are getting from our experiments are not ranked. Hence ranked evaluation methods like MAP and NDCG are not suitable for our methodology. But in future if we are able to give weights to the causality or association of the event and sub-events then we can use MAP or NDCG evaluation methods.

8.2. Experiment Results

We have experimented with causal verbs used for analysing semantic role labelled data. We have categorized verbs in three sets, Set-1 is has very stringent causal verbs, Set-2 are lenient causal verbs and Set-3 were the balanced causal verbs. While we are getting best score from using Set-3 verbs and Set-1 has low recall due to stringency and Set-2 has low precision due to leniency. These sets are stored in set1.txt, set2.txt and set3.txt files.



9. Conclusion

Through this study a preliminary step has been taken in representation of causal relationship between various events in a given domain. A visual representation was generated in Neo4J which helps in visualizing details of correlation between major events and their constituent sub-events. This can be further extended to represent multiple categories and potentially finding causal relation between events across domains. However, there is considerable noise introduced in Semantic Role Labelling and can be smoothened using more advanced techniques.