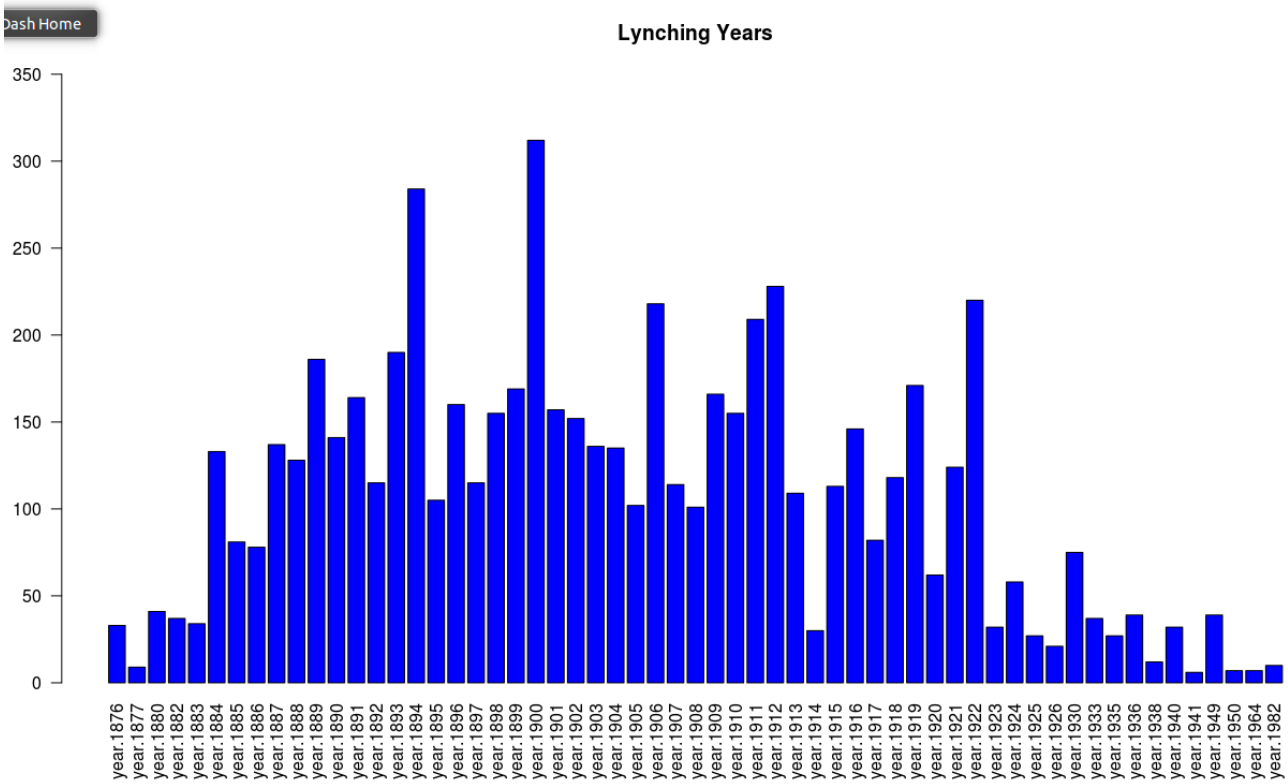
Data Mining: Problem 2 Report

Stephanie Freitag, Simone Salvo, Ahmed Sobih

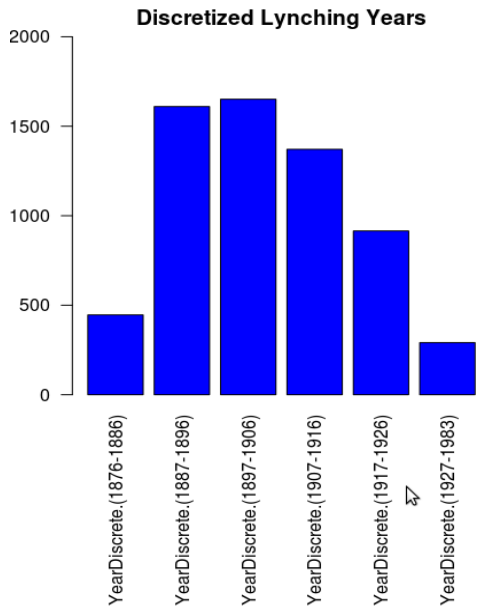
Task 1 (Date Extraction, Item Sets Extractions)

- **Date Extraction and Transformation:**

Date was extracted for all semantic-triplets and discretized as following:
The year ranges between 1876 and 1982. Years were discretized into **equal width intervals** (10 years for each interval except the last grouping 50 years, since there was no much data for the last years). Figure[1,2] show the frequency of semantic triplets per year and the frequency over discretized intervals (10 years).

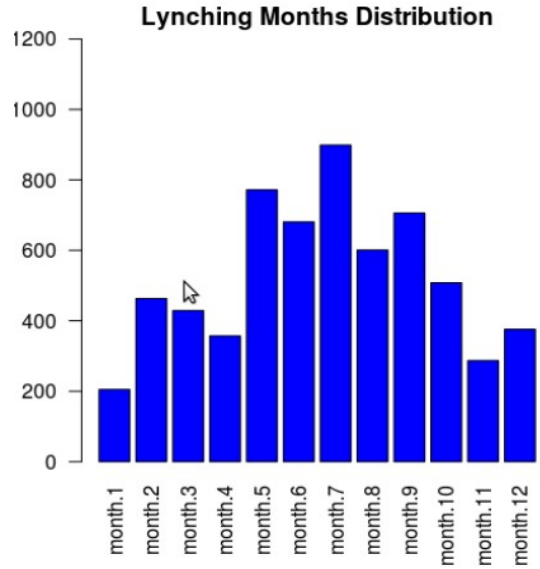


Figure[1]: The frequency of semantic triplets per year



Figure[2]: The frequency of semantic triplets over decades.

We also included the month hoping that we might find some interesting correlation between the month and the rate of lynching. As shown below in figure[3], the distribution of semantic-triplet over months is close to normal distribution, which suggests there might not be unexpected patterns .



Figure[3]: The frequency of semantic triplets per month.

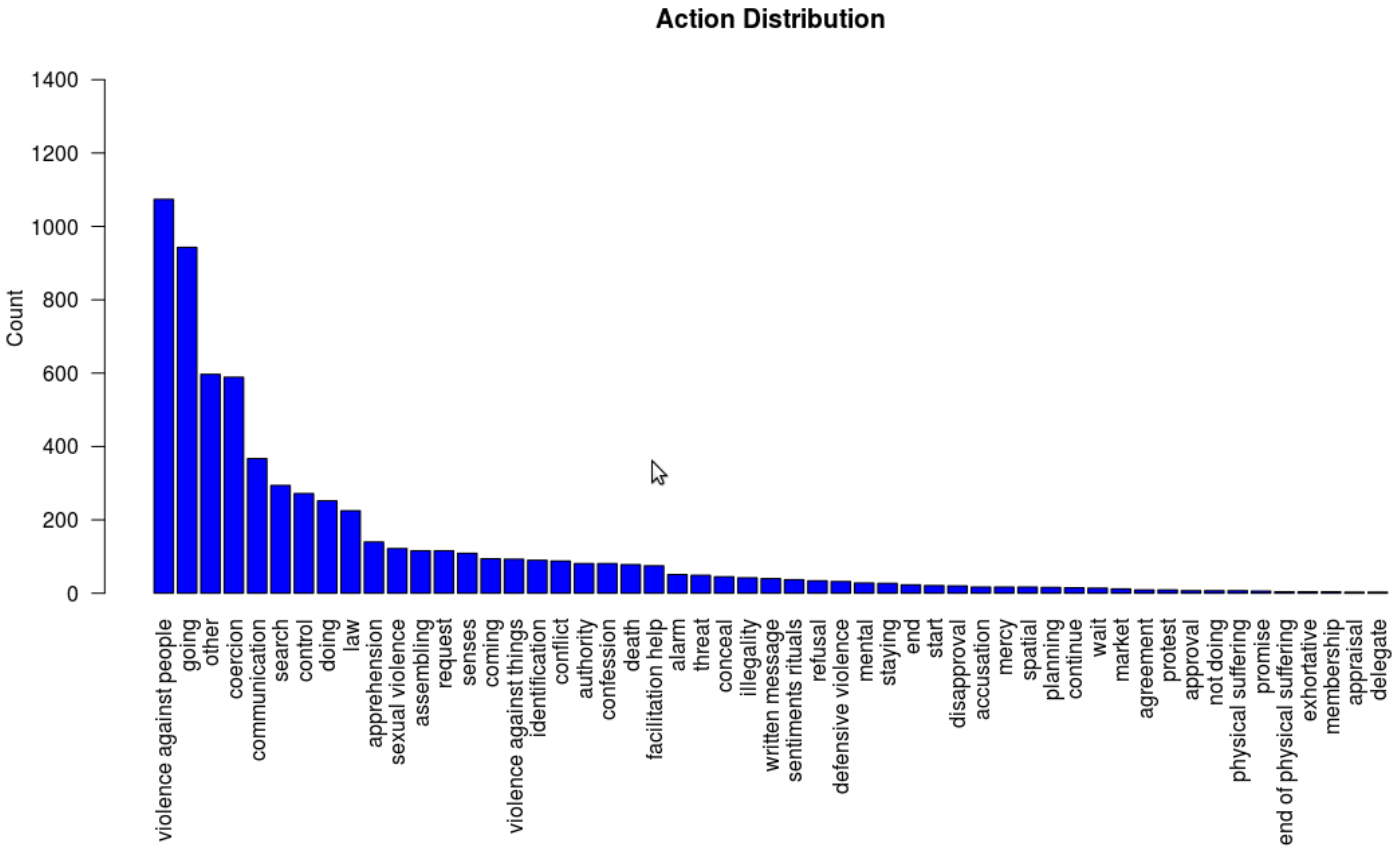
- **Figures 1,2 Item Sets Extractions**

The number of relationships that can be constructed from the two figures are 21 relationships. One of these relationships was not found({Lynched Negro;Violance agaist ppl; Individuals}), since we couldn't find any reference for the word (individuals) in the dataset. For other relationships, our results don't completely match with the paper. It's not clear whether the results in the paper is aggregation on Semantic-Triplet level or not.

How we constructed the semantic-triplets:
We noticed that the possible values for Subject is splitted between more than one field {Subject_actor_aggregate_code,Subject_name_of_individual_actor,Subject_actor_aggregate_individual_colin,Subject_actor_aggregate_collective_colin}. For example, if the subject={Negro}, the value exists in the field " Subject_name_of_individual_actor". Other fields are empty for most records. On the other hand, if the subject is a group of people; e.g. subject= Mob, then the value exists in the field Subject_actor_aggregate_collective_colin, and the rest are empty. The same applies for {Action, Object}. In order to extract the records representing the relationship, we used different fields based on the value for the {Subject, Action, Object} semantic-triplet.
Another thing is the data set contains mistakes. For example, the record with Semantic_Triplet_Id=74503 has value Object_name_of_individual_actor={girl} and Object_ actor_aggregate_code={White girl}. On the other hand, the record with semantic_Triplet_Id=27075 has value Object_name_of_individual_actor={girl} and Object_ actor_aggregate_code={White woman}. By reviewing the story in Semantic_Triplet_Indentifier, it's clear that it's a white girl not a white woman. So we modified some records based on our understanding.
For some relationships, the frequency of subject, action or object is much less than reported in the paper. For example, the relationship {Whites;Violence against ppl; Negro} is reported to be 79 times. However the word (whites) exists only 36 times in all fields. So it's not clear how the relationship {whites, violence against people; negro} can occur 79.

- **Action Analysis:**
  Figure[4] below shows the frequncey of different actions aggregated by semantic-triplet (We have 6516 different sematice-triplet in total) . The table below shows the top actions. The frequency of violence against people and coericion action are among the highest.

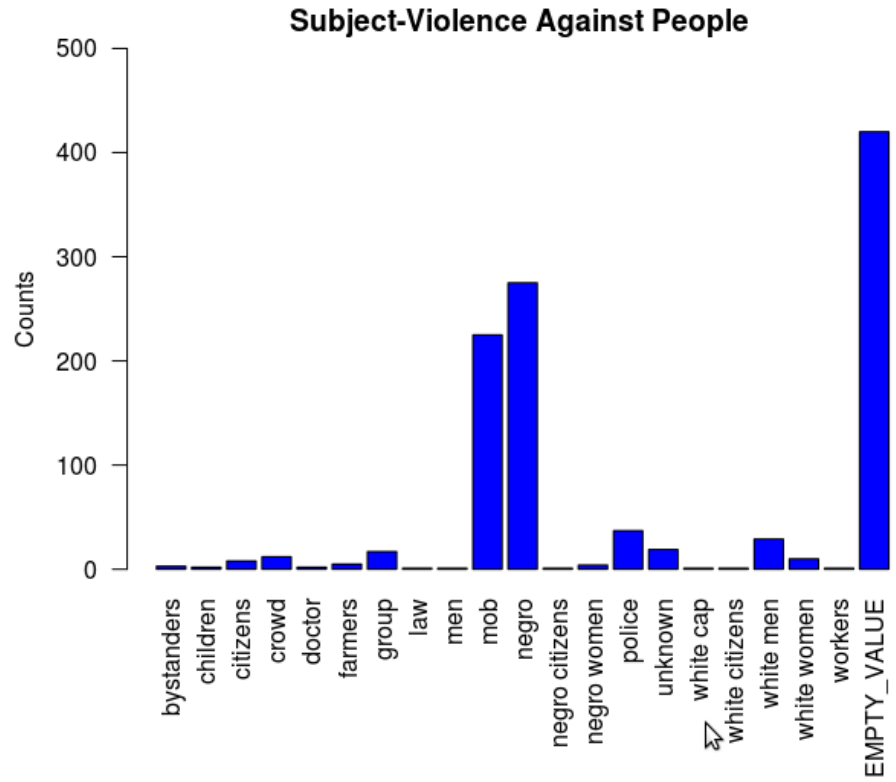
Action	Percentage
Violence against people	16.4%
Going	14.5%
Other	9.2%
Coercion	9%
Communication	5.5%



Figure[4]: The frequncey of different actions aggregated by semantic-triplet.

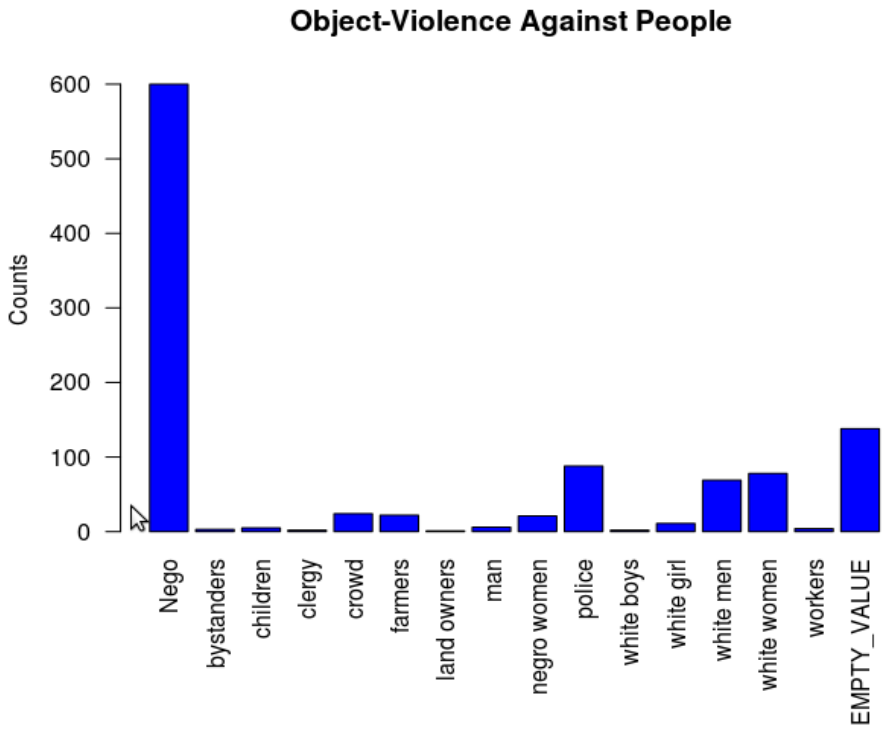
- **Violence against people Analysis:**

Figures[5,6] show the distribution of subjects took the action (violence agaist people) and the distribution for objects whom agaist the action was taken.



Figure[5]: Distribution of subjects took the action (violence agaist people).

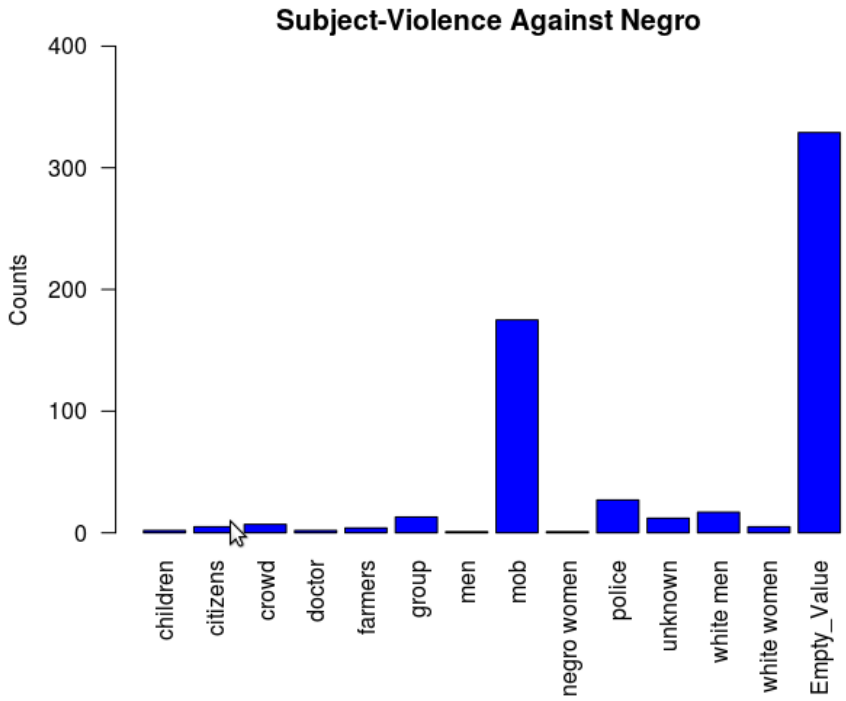
The problem of missing data appears clearly here. 39% of the semantic-triplets with the action violence against people are missing the subject.



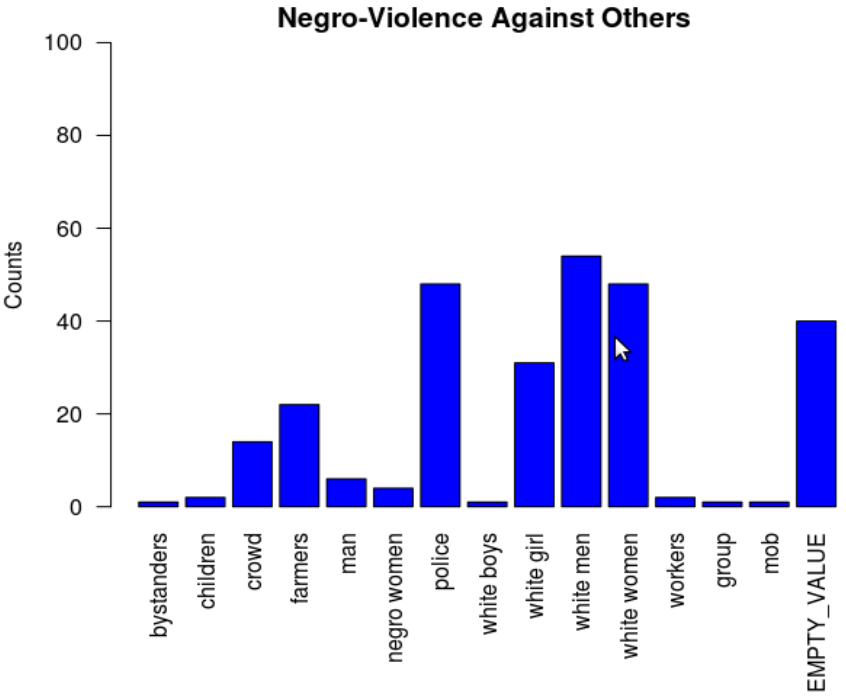
Figure[6]:Distribution of objects whom against the action (violence agaist people) was commited.

The figures show that 25% of the violence against people was committed by negros, while 20% were committed by the Mob. On the other hand, 55% of the violence is known to be against negros.

Figures[7,8] shows the distribution of subjects committed violence against the object negros and the distribution of objects whome against the negros committed violence action.



Figure[7]:Distribution of subject committed (violence agaist) negros.



Figure[8]:Distribution of objects whome agaist negros committed violence.

As the figures show, most of violence against negros was committed by (mob [65%], and police [10%]). On the other hand, most of violence committed by negros was against (white men [23%], police [20%], white women[20%], and white girls [13%]).

Table[1] summarizes the number of records we found for each relationship in the two figures provided by professor Franzosi.

Semantic-Triplet	Our Results (Aggregated On Semantic-Triplet Level)	The paper's Results
{Group;Coercion; Negro}	5	20
{Group;Violence against ppl; Negro}	13	81
{Mob;Coercion; Farmers}	0	9
{Mob;Coercion; Negro}	83	129
{Mob;Coercion; Negro Woman}	4	9
{Mob;Coercion; Police}	21	49
{Mob;Violence against ppl; Negro}	175	338
{Mob;Violence against ppl; Police}	11	43
{Negro;Coercion; White Woman}	13	10
{Negro;Violence against ppl; Police}	48	56
{Negro;Violence against ppl; White Girl}	31	16
{Negro;Violence against ppl; White Men}	54	66
{Negro;Violence against ppl; White Woman}	47	37
{Police;Coercion; Negro}	70	42
{Police;Violence against ppl; Negro}	23	27
{Unkown;Violence against ppl; Negro}	4	22
{White Men;Violence against ppl; Negro}	12	18
{Whites;Coercion; Negro}	0	36
{Whites;Violence against ppl; Negro}	0	79
{Whites;Coercion; Police}	0	16

Table[1]: The number of records we found for the relationships {Subject;Violence-Against-People;Object} and {Subject;Coercion;Object}.

- **Extracting item sets that occur at least twice:**

For extracting the item-sets, we used our implementation of the aprioi algorithm. Below is a description of the implementation.

How it works:

Apriori is an algorithm for frequent item set mining. It proceeds by identifying the frequent individual terms in the database and extending them to larger and larger item sets, as long as those item sets appear sufficiently often in the database.

Implementation:

The algorithm was implemented by using C# language and Visual Studio 2013 as integrated development environment. Such algorithm was developed by two classes: node and main. The first realizes the one node of the graph (we will talk about it soon) and the second realizes essentially the reading of the input, the building of the graph and the output generation.
Each node will contain the number of the frequency, the field/fields, the rows where this kind of field/fields have a matrix cells with 1.

A level of node is a set of nodes that are created by the search of common items between two nodes : the first will be a node in some level composed with one or more fields, the second will be always node of level 1 (the level above the source node) with number of field that realized the node equal to one.

First Step:
Reading a transformed matrix formed just by 0 and 1 with first row containing the name of the column that we will call “field”, the algorithm will create the node that have frequency of “1” greater or equal to the 5% of the number of the total input file rows.
After the first step, we will have the base node linked with a “source” pointer that is just an array of pointer, with that will be possible to go through each node of the graph.

Second step
For each node (A) above the source node will be looking to create other nodes that will contain the items (fields) of A and the fields of B, each new node will contain the relative rows where this “fields union” will have a 1 in the matrix. The search of this new node is based always, also in the above level of node, with the level above the source node (B). This means that hypothetically in the fifty-s level, where the node should be formed of more fields, we will try to create an above level of node by trying to find some row in the matrix that have cells relative to the fields of the node A with 1 and cell relative to the node B (where B is a node in the level above the source node) with 1 also.

Third step

The second step will be iterated for each level of the node, until there will be more than one node in the level above of the current level.

- **Analysis of extracted frequent item sets.**

As shown in table[1], the number of records for each relationship is not high enough to find large number of frequent item sets. Before we run apriori algorithm, some fields were removed. For example, it's logical to remove the subject race field, when we know that the subject is white man or negro. The same applied for the object field.
Table[2] shows the number of item sets found for each relationship.

Semantic-Triplet	Number of Frequent Item Sets
group_coercion_negro	4
group_violence_against_people_negro	28
mob_coercion_negro	285
mob_coercion_negro_women	34
mob_coercion_police	300
mob_violence_against_people_negro	1294
mob_violence_against_people_police	30
negro_coercion_white_woman	27
negro_violence_against_people_police	957
negro_violence_against_people_white_girls	5406
negro_violence_against_people_white_men	433
negro_violence_against_people_white_woman	867
police_coercion_negro	191
police_violence_against_people_negro	59
unknown_violence_against_people_negro	5
white_men_violence_against_people_negro	20

Table[2]: Number of item sets found for each relationship.

Task 2 (Association Rules Extraction):

For extracting association rules, we used our implementation for association rules generation. Below is a description of the implementation.

The algorithm

We implemented the algorithm described under "Rule Generation in Apriori Algorthm" in chapter 6.3.2 in the book. We use it on the data we get as output form the Apriori Frequent item set generation algorithm we used in No.1.

How it works

The file, which contains all freqent item sets and their support, is read and stored into an Arraylist representing F_k, and a HashMap as a look up table for the supports of the different frequent item sets. F_k is traversed entry by entry and the different rules are seperately computed for every $f_k \in F_k$. For generating the rules, the 1-item consequents H_1 are computed and feed into a recursive function. This computes all rules of the form $f_k - h_1 \rightarrow h_1 \{\forall h_1 \in H_1\}$. In a next step the 2-item consequents are computed and the function is called again recursively until all rules for this frequent item set are generated.

How antimonotonicity helps

We used theorem 6.2 of the book for pruning the amount of k-item consequents. From the theorem we know that $f_k - h_1 \rightarrow h_{1.}$ cannot have a lower confidence than $f_k - h_2 \rightarrow h_2 \, if \, |h_2| > |h_1| \text{ and } h_1 \subset h_2$. This means, if the rule $f_k - h_1 \rightarrow h_1$ has a confidence lower than a constant "minconfidence", the confidence of $f_k - h_2 \rightarrow h_2$ is also lower than "minconfidence". We use this knowledge and delete h_1 from H_1 before creating H_2 based on H_1.

Analysis:

Because the frequency of is very low for most triplets, it's difficult to find many meaningful relationships.

Task 3

Although there are numerous measures available for evaluating association patterns, a significant number of them provide conflicting information about the interestingness of a pattern. Thus, selecting the right measure for a given application poses a dilemma because many measures may disagree with each other [1].

Our Selected Measures:

1) The Odds-ratio :

This measure represents the odds for obtaining the different outcomes of a variable.

$$\alpha = \frac{P(A,B)P(\bar{A},\bar{B})}{P(A,\bar{B})P(\bar{A},B)}$$

2) Interest factor:

$$I = \frac{P(A,B)}{P(A)P(B)}$$

3) Jaccard coefficient:

$$J = \frac{P(A,B)}{P(A,B)+P(A,\bar{B})+P(\bar{A},B)}$$

Measuring the correlation of the measures:
The correlation between a pair of measures can be measured in terms of the similarity between their ranking vectors. There are several measures available for computing the similarity between a pair of ranking vectors. This include Spearman's rank coefficient, Pearson's correlation, cosine measure.

References:
[1] Pang-Ning Tan et al, Selecting the right objective measure for association analysis. Selecting the right objective measure for association analysis.