Crime Density using News Article Analysis

Shailendra Kr. Gupta - 2016CSB1059
Shreyanshu Shekhar - 2016CSB1060

Supervisor:

Neeraj Goel
Mukesh Saini

Problem Statement

- The main objective of this project is to provide a crime score/crime heat map for every location of different cities in India.
- Use Case:
  - Finding safest route/place
  - Dynamically organising police force
  - Predicting the future occurrence of crime

Related Work

- Extracting crime information from online newspaper articles By Rexy Arulanandam Bastin et al.
- Crime analytics: Analysis of crimes through newspaper articles By Isuru Jayaweera et. al.
- Safe Routes Based on Tweet Sentiments By Jaewoo Kim et. al.
- Crime News Analysis: Location and Story Detection By Mehedee Hassan et. al.
- Analyzing Newspaper Crime Reports for Identification of Safe Transit Paths By Vasu Sharma et. al.

Extracting crime information from online newspaper articles

- **Aim:** Identify theft location by classify each sentence as CLS or NO-CLS
- Four different NER methods of Location entity extraction:
    - NLTK pretrained named entity chunker - ne_chunk() method of NLTK
    - Stanford NER - Java based model. Uses CRF model for identification
    - NLTK chunker class using Gazetteer - uses Gazetteer corpus
        - Locations from all around the globe
    - **LBJ Tagger - Neural Network based**
- Deciding on features and labelling dataset
- Training CRF model using labelled data
- Classification of sentences as CLS or NO-CLS using CRF model
- **LIMITATIONS**:
    - Duplicacy of articles not eliminated
    - Small sample size(Around 70 articles)

Crime analytics: Analysis of crimes through newspaper article

- **Aim:** Web based portal for
  - Hot Spot Detection - location wise crime heat map
  - Crime Comparison - comparing different crimes over a given period
  - Crime pattern visualization - analyzing a particular crime over a given period
- Crawler to crawl news articles
- Classification of news articles as crime or non-crime
  - LibSVM
  - SMOTE used to sample minority class
- Entity Extractor:
  - Combination of ANNIE POS and Stanford POS tagger
  - Google Maps API used for location identification
- Duplicate articles identification using entities
  - SimHash values using entities

Safe Routes Based on Tweet Sentiments

- Data filtering
    - Public Geotagged tweets
    - Mentions, replies & retweets
- Sentiment Analysis
    - Sentiment value of each tweet is determined on a scale from -1 to 1
- Regional Clustering
- Router finding and visualizing

Crime News Analysis: Location and Story Detection

- Document classification (SVM)
- Name Entity Recognition
  - Person
  - Location
  - Organization
- Feature selection and extraction
  - Representing text documents as numeric vectors (TF-IDF)
  - Document Clustering (hierarchical clustering)
    - Cosine Similarity

Analyzing Newspaper Crime Reports for Identification of Safe Transit Paths

- Data collection
- Crime classification
  - Term document matrix
  - Latent Semantic Analysis → KNN
- Identification of location
  - Named Entity Recognition
- Mapping crime intensities
- Identifying safest path

Proposed Plan

- Collecting crime data
- Finding ground truth
- Classifying crime and non-crime articles
- Finding location of occurrence of crime
- Analyzing duplicate articles
- Calculation crime score for each location

Challenges and Blockers

- Data and ground truth (unlabeled data)
- Rating severity of crime (very subjective)
- Finding location of crime (similar location and person name)
- Unavailability of news articles for remote areas
- Repeated articles of same crime
- Different coverage of a crime at different locations
- English media has different view of crime than regional media

Progress with data

- Crawler
  - Newspaper3k
  - 50k and still collecting
  - News sites:
    - TOI
    - Hindu
    - NDTV
    - News18
    - India Today
    - Hindustan Times

- Interface
  - http://172.26.5.254/login.php
  - Php based web interface
  - Total 737 tagged
    - 336 Crime
    - 401 Non-Crime

News Classification as Crime or Non-Crime

- Words selection
    - Major Crime Words
    - Synonyms addition
    - Assigning score to each word
- Finding similar meaning words in the article by averaging
    - WUP(Wu Palmer) similarity - based on taxonomy depth
    - PATH similarity - shortest path
    - LCH (Leacock Chordorow) similarity - shortest path + taxonomy depth
- Final **CrimeClassificationScore** calculation using the assigned score to synonyms
- Threshold to segregate crime and non-crime articles(empirically)

Result: Crime classification

Actual → **Pred ↓**	**CRIME NEWS**	**NON-CRIME NEWS**	**TOTAL NEWS**
CRIME NEWS	333 (TP)	178 (FP)	**511**
NON-CRIME NEWS	6 (FN)	215 (TN)	**221**
TOTAL	**339**	**393**	**732**

Accuracy = TP + TN / (TP + FP + TN + FN) = 0.749

Precision = TP / (TP + FP) = 0.652

Recall = TP / (TP + FN) = 0.982

Location Extraction

- Created some Lists
  - Created List of locations in INDIA(Location list)
  - Created List of some commonly used tags in location names(Tag list)
    - Eg. Patel **nagar**, Pahar**ganj**, Anand **vihar**, Chandni **chowk**, etc.
  - Created List of prepositions used before location entities(Preposition list)
    - Eg. in, near, from, at, etc.
- Extracted entities from text using NLTK pretrained chunker
- Selection of Named Entities as location
  - Entities that are present in the list of locations
  - Entities that are have any location tags
  - Entities that have any preposition from preposition list before it and belongs to either location list or have any location tag from tag list.

Results of location extraction

- Prediction:
  - Total crime articles: 344
  - Total articles with location prediction > **50%** : 147 (42.7%)
  - Total Crime Locations in all articles: 636
  - Total Locations predicted: 363 (57.1%)

Post mid-sem Work

- Improve our naive classifiers
  - Crime classification
  - Location extraction
- Identifying duplicate news
- Finalizing a crime score assigning strategy based on retrieved data for each location

Questions and Answers

Thanks!