Classifying Refugee News Reports

Data Warehousing and Computing Lab

Euan Dowers, Robert Lange, Nandan Rao
Supervisors: G. Bartolozzi and C. Brownlees
Barcelona Graduate School of Economics

December 1, 2016

Outline

1. **Aim and Motivation**
   Working with IOM and Refugee News Flood
2. **Database Management**
   UI, Data Types and MongoDB
3. **Text Processing**
   String Cleaning, Vectorization and TF-IDF Representations
4. **Modelling the Data**
   Clustering and Cross-Validated Classifiers

Motivation

- Ever since the start of the refugee crisis there has been a steady increase in news reports and rumors regarding missing migrants.
    - Not even factoring in Donald Trump's Twitter activity
- In order to efficiently allocate resources and to help people in nedd, it is crucial to determine hot spots based on reliable data.
- Cooperation with the International Organisation for Migration (IOM)

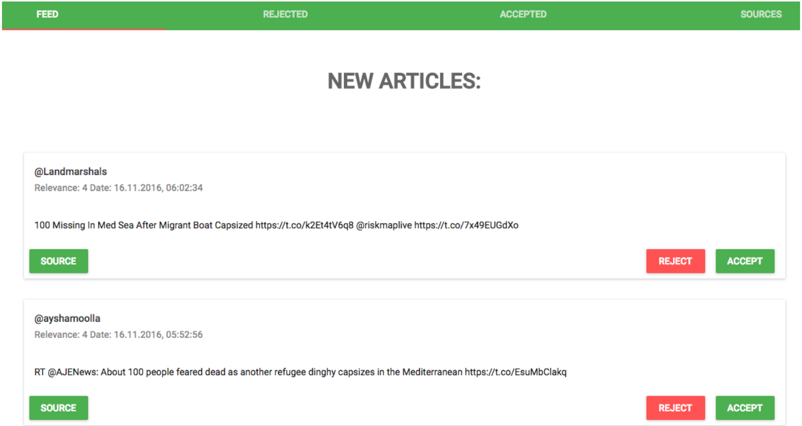
Data Types, Challenges and a Solution

- Data Sources:
  - → Google Alert News Feeds
  - → Twitter Feeds
  - → Missing Migrant Project (MMP) data
- Datamanagement Challenges:
  - → One schema is not enough (different data types)
- Datamanagement Solution: MongoDB

MongoDB

- MongoDB has several advantages:

UI and Automated Labelling Process



NEW ARTICLES:

@Landmarshals
Relevance: 4 Date: 16.11.2016, 06:02:34

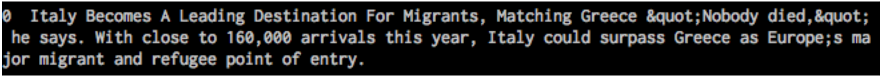
100 Missing In Med Sea After Migrant Boat Capsized https://t.co/k2Et4tV6q8 @riskmaplive https://t.co/7x49EUGdXo

@ayshamoolla
Relevance: 4 Date: 16.11.2016, 05:52:56

RT @AJENews: About 100 people feared dead as another refugee dinghy capsizes in the Mediterranean https://t.co/EsuMbClakq

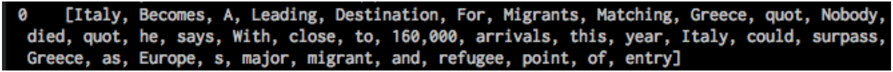
▸ Link

Cleaning the strings



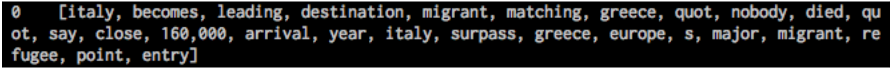
```
0  Italy Becomes A Leading Destination For Migrants, Matching Greece &quot;Nobody died,&quot;
 he says. With close to 160,000 arrivals this year, Italy could surpass Greece as Europe;s ma
jor migrant and refugee point of entry.
```

- Splitting text into tokens



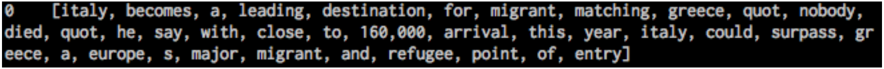
```
 0    [Italy, Becomes, A, Leading, Destination, For, Migrants, Matching, Greece, quot, Nobody,
  died, quot, he, says, With, close, to, 160,000, arrivals, this, year, Italy, could, surpass,
  Greece, as, Europe, s, major, migrant, and, refugee, point, of, entry]
```

- Removing stopwords



```
 0    [italy, becomes, leading, destination, migrant, matching, greece, quot, nobody, died, qu
 ot, say, close, 160,000, arrival, year, italy, surpass, greece, europe, s, major, migrant, re
 fugee, point, entry]
```

- Stemming the words and converting them into lemmas



```
0    [italy, becomes, a, leading, destination, for, migrant, matching, greece, quot, nobody,
died, quot, he, say, with, close, to, 160,000, arrival, this, year, italy, could, surpass, gr
eece, a, europe, s, major, migrant, and, refugee, point, of, entry]
```

Constructing a Vectorized Representation

- tf-idf: term frequency-inverse document frequency
- Deciding on dimensionality: bi-grams, tri-grams, etc.
  → Which representations do really matter?

The Problem

- Easy/accelerated classification of relevant and irrelevant news
- Problems:
  1. Redundancy: Many observations cover the same events
  2. Sensitivity: Hard classification problem
- Solutions:
  1. Hierarchical clustering using DBSCAN
  2. Ensemble Methods: Random Forest

Clustering using DBSCAN - Density-based spatial clustering of applications with noise

- Density-based clustering algorithm: core points, (density-)reachable points and outliers
- Core point forms cluster together with all reachable points (core or non-core).
- Clusters contain at least one core point; non-core points can be part of a cluster, but they form its "edge", since they cannot be used to reach more points.
- Applied to TF-IDF matrix and parametrized with difference tolerance

Building a First Classification Model

- Many potential classifiers available: Logistic Regression, Naive Bayes, SVM, Decision Tree, Random Forest and Neural Networks
- Idea: Start with MVP (minimal viable product) to grasp the problem → Generative Model: Naive Bayes

Problems in Classification

- Hyperparameter choices: 5-fold cross-validation and parameter grid search
- Adding non-parametric complexity: Random Forest

Improving Classification

- Hyperparameter choices: 5-fold cross-validation and parameter grid search
- Adding non-parametric complexity: Random Forest

Conclusion

- Open research/work:
  1. Better understanding of the decision boundary problem
- *Any Questions?*
- *Thank you for your attention!*