**Dangerous events forecasting**

1. **Description of the problem**

Forecasting events like terrorist attacks, conflicts, and any other mass violence can help to assess a risk and take a measure if it is necessary. Nowadays we have a huge events datasets and it is definitely interesting to know in what way we can use this information for predictions.

In my project, I am going to take the central part of EU (due to the fact of recent events) as region for forecasting things, which have so-called « the Goldstein Scale» from «-9.5» to «-10» (the events given above have these values). For more details about the Goldstein Scale you can read the following:

* <http://gdeltproject.org/data/lookups/CAMEO.goldsteinscale.txt>
* <http://gdeltproject.org/data/lookups/CAMEO.eventcodes.txt>

Note: the problem is not new, so you can check:

* <http://data.gdeltproject.org/documentation/ISA.2013.GDELT.pdf>
* <http://foreignpolicy.com/2014/01/03/half-a-billion-clicks-cant-be-wrong/>

1. **Who is a client?**

State organizations, security agencies, anybody who cares about safety precautions.

1. **What data are going to be used?**

Datasets from <http://data.gdeltproject.org/events/index.html>, having the following features:

* have information from different international news sources;
* cover period from 1979 to now;
* describe events as «Actor1 performed an action upon Actor2», where Actor can be a certain person, organization, group, country and so on. Note: everything described by codes;
* assess the event itself giving:
* classifications Verbal Cooperation, Material Cooperation, Verbal Conflict, and Material Conflict;
* GoldsteinScale (mentioned above);
* The average “tone”: the score ranges from -100 (extremely negative) to +100 (extremely positive).
* statistics about the total number of mentions, information sources, documents;
* provide the Internet link (from 2013).

For more details about structure of datasets see <http://data.gdeltproject.org/documentation/GDELT-Data_Format_Codebook.pdf>.

1. **Approaching to solve the problem**

Note: There is no need make text mining, because the datasets have processed information.

I am going to take period from 04.2013 to 12.2005 as a training set and 01.2016 – now for testing set.

Methods: supervised learning.

1. **What are deliverables?**

* Description of:
* problem with details;
* a model for forecasting;
* steps for approaching the problem.
* Python’s code;
* Results analysis including visualizations part.