Simge Ekiz, s4706757

Robin Wubs, s4135229

A descriptive title

Sentence Classification for Protest Events

A short description of the problem you want to solve and corresponding data.

Event extraction from news articles is a complicated job. Even though event-related entities are correctly detected, ambiguities occur while forming an event from them. We expect to reduce the chance of ambiguities by correctly classifying event occurrence on the sentence level. In this project our focus is finding sentences that has useful information about protest events. Moreover, we aim to classify the sentences as past protest, planned protest or not protest event.

Currently our dataset is consist of 2829 English sentences provided by Koç University in Turkey. Likewise, data is annotated by two annotators by three tags which are (1) not-protest, (2) protest, (3) planned protest. Later on this information will be used to determine if the sentence is worthy for further analysis.

A description of or link to the software that you want to use to solve your problem, including which algorithms you plan to apply.

Planned steps;

- Data cleaning with stopwords, interpunction and caps removal
- Creating a feature set such as bag of words/lemmas, TF/IDF, PoS tags, Named Entities
- Feature selection by removing the attributes that affects less than some threshold.
- Classification

Algorithms that we want to apply and compare;

- Support Vector Machines
- Decision Trees / Random Forest
- K-nearest Neighbours
- Neural Networks

The experiments will be done by using Python programming language and using libraries such as Scikit-learn and Keras.

Pointers to the literature that you use for inspiration and comparison.

- https://www.analyticsvidhya.com/blog/2018/04/a-comprehensive-guide-to-understand-and-implement-text-classification-in-python/
- https://cloud.google.com/blog/products/gcp/problem-solving-with-ml-automatic-document-classification
- https://towardsdatascience.com/machine-learning-nlp-text-classification-using-scikit-learn-python-and-nltk-c52b92a7c73a