Twitter Police Report

Instituto de Computação - IC/UFF
Daniel Junior

1. Objetivo

Este trabalho tem como objetivo utilizar tweets coletados a partir de palavras chaves e extrair informações úteis para a automatização da construção de um boletim de ocorrências.

A estrutura a ser gerada a partir do texto deve ser a seguinte:

EVENTO	VÍTIMA	CRIMINOSO	LOCAL	QUANDO

Por exemplo, o tweet *'Hóspede é baleado durante assalto em hotel no Caminho das Árvores'*, poderia gerar o seguinte registro:

EVENTO	VÍTIMA	CRIMINOSO	LOCAL	QUANDO
ASSALTO	HÓSPEDE		CAMINHO DAS ÁRVORES	

2. Metodologia

Para alcançar o objetivo o, são necessários as seguintes etapas:

- Coleta de dados
- Pré-Processamento dos dados
- Treinamento do Modelo
- Avaliação Experimental

2.1. Coleta de dados

A coleta dos dados é possível a partir da API (https://developer.twitter.com/en/docs.html) do Twitter. Foi criada uma aplicação Rails para facilitar a captura dos dados com a realização inicial de armazenamento dos dados coletados no banco de dados relacional SQlite. A aplicação se encontra aqui (https://github.com/danieljunior/tweet_collect). A coleta foi realizada em Tweets publicados entre os dias 16/04/2019 e 16/05/2019.

2.1.1 Palavras-Chaves

Umas das formas de se coletar tweets é informando um palavra-chave.

As palavras chaves utilizadas na coleta dos dados foram: **assalto, assaltado, assaltada, roubo, roubado, roubada, assassinado, assassinada, assassinato, tiroteiro, tiros, baleado, baleada**.

O total de tweets coletados a partir dessas palavras chaves foi de **1781**.

2.2 Pré-Processamento dos dados

O pré-processamento dos dados envolve: filtragem e tagueamento dos mesmos.

2.2.1 Filtragem

A coleta inicial dos dados trouxe registros que não estão relacionados diretamente com os eventos que esperamos tratar neste trabalho. Por exemplo: a palavra assalto trouxe tweets relacionados à luta de boxe, roubo referentes a corrupção, etc. Os dados coletados sem filtrar esses casos estão disponíveis no arquivo non_filtered.csv (./data/non_tagged/non_filtered.csv).

Visto isso, se faz necessária a filtragem desses dados.

2.2.2 Tagger

Para ser possível atingir o objetivo deste trabalho, será necessária a aplicação da tarefa *Semantic Role Labeling* (SRL), que precisa que os dados estejam tagueados para treinamento do modelo.

Utilizando o tweet de exemplo da Seção 1, o registro tagueado seria algo parecido com:

< VITIMA > Hóspede **</ VITIMA >** é baleado durante **< CRIME >** assalto **</ CRIME >** em hotel no **< LOCAL >** Caminho das Árvores **</ LOCAL >**

2.3 Treinamento do Modelo

O Modelo de Representação em Embbedings utilizado será o ELMo (https://allennlp.org/elmo).

Esta escolha é baseada nos resultados interessantes para a tarefa de SRL presentes no artigo que apresenta a proposta.

Serão usados configurações (https://s3-us-west-2.amazonaws.com/allennlp/models/elmo/contributed/pt/elmo_pt_options.json) e pesos (https://s3-us-west-2.amazonaws.com/allennlp/models/elmo/contributed/pt/elmo_pt_weights.hdf5) pré-treinados para a Língua Portuguesa que podem ser obtidos na própria página do ELMo.

Também será usado o próprio modelo apresentado no artigo, o que é possível devido a disponibilização do arquivo de configuração (https://github.com/allenai/allennlp/blob/master/training_config/semantic_role_labeler_elmo.jsonnet) do modelo SRL.

Devido a necessidade de realizar manualmento o tagueamento dos dados coletados, a estratégia utilizada será um Aprendizado Semi-Supervisionado. Apenas parte dos dados de treinamento serão tagueados e as predições realizadas para o restante dos dados de treinamento serão utilizados para realimentar o treinamento.

3. Backlog

FINALIZADO	EM ANDAMENTO	A FAZER
COLETA INICIAL DE DADOS	FILTRAGEM DE DADOS	TAGUEAMENTO DOS DADOS FILTRADOS
EMBBEDINGS EM PORTUGUÊS		APRENDIZADO SEMI-SUPERVISIONADO
		AVALIZAÇÃO EXPERIMENTAL
		RELATÓRIO FINAL