Quotation Extractor for Portuguese

Projeto Final de Programação

Rafael Reis
rrsilva@inf.puc-rio.br
Orientador: Ruy Milidiú

9 de Dezembro de 2015

Conteúdo

1 Introdução

A tarefa de extração de citações consiste em identificar num texto o conteúdo de uma citação e relacioná-la ao seu autor. Ela possui grande importância na área de recuperação de informação, já que, em jornais e portais de notícias, as citações podem chegar a compor 90% do texto [1].

A figura 1 apresenta um exemplo de associação entre citação e autor. Os autores estão em negrito e a citação em itálico. Os índices subscritos representam os autores.

> **O deputado Delegado Waldir (PSDB-GO)**$_1$ perguntou a **Cunha**$_2$ se ele tinha contas na Suíça. '*Não tenho qualquer tipo de conta em qualquer lugar que não seja a conta que está declarada no meu Imposto de Renda*'$_2$, declarou **o presidente da Câmara**$_2$.

Figura 1: Exemplo de associação entre citação e autor

Os primeiros sistemas que resolviam a tarefa eram baseados em regras complexas. Nos últimos anos, porém, técnicas de aprendizado de máquina têm sido aplicadas com grande sucesso, com a vantagem de possuir capacidade de generalização para outros idiomas.

Este trabalho reproduz o sistema baseado em aprendizado de máquina descrito em [2].

2 Objetivo do Produto

O *software* tem como objetivo criar e avaliar um modelo de aprendizado de máquina, a partir de uma base já processada com técnicas de NLP, que possa ser utilizado na tarefa de extração de citações. Ele deverá ser executado através de linha de comando.

3 Escopo da Solução

Em geral, modelos de aprendizado de máquina são construídos em etapas, abrangendo a preparação dos dados, a modelagem em si e a construção do preditor. Para cumprir a tarefa, o *software* deverá cobrir todas essas etapas, implementando as seguintes necessidades.

3.1 Pré-Processamento dos Dados

- O *software* deverá carregar uma base de dados no formato GloboQuotes (descrito em [2]), excluindo os títulos das notícias. O modelo só deverá levar em consideração o conteúdo das notícias.
- O *software* deverá extrair as citações candidatas e seus possíveis autores para cada notícia da base de dados.

- O produto dessa etapa deverá ser um arquivo CSV, no qual cada linha é composta de uma citação candidata e seu possível autor (em forma binarizada). Esse arquivo servirá de entrada para o algoritmo de aprendizado e para a tarefa de predição.

3.2 Algoritmo de Aprendizado

- O *software* deverá implementar o algoritmo do Perceptron Estruturado, para a construção do modelo de Aprendizado de Máquina.
- O algoritmo do Perceptron Estruturado deverá utilizar como preditor o algoritmo de Agendamento de Intervalos com Pesos (*Weighted Interval Scheduling*): cada citação candidata será um intervalo e seu peso será dado pelas *features* de NLP geradas a partir do possível autor.
- Para a calibração do modelo, o *software* deverá implementar a técnica de validação cruzada *K-fold*, com $K = 5$.

3.3 O preditor em si

O preditor deverá ser baseado no algoritmo de Agendamento de Intervalos com Pesos, como descrito na seção anterior.

3.4 Avaliação da qualidade do preditor

O *software* deverá computar a métricas de *Precision* e *Recall*, tanto na etapa de treino quanto na etapa de teste.

4 Limites e Restrições

A solução implementada será compatível apenas com a base de dados no formato GloboQuotes. A base de dados será fornecida previamente, já dividida em 2 arquivos: um para treino do modelo e outro para teste.

5 Especificações Técnicas

5.1 Linguagem e Ambiente

O *software* será desenvolvido no Mac OS X versão 10.11.1, utilizando a linguagem Python versão 3.5.0 (distribuição Anaconda 2.4.0).

5.2 Organização dos arquivos

O código será organizado seguindo as boas práticas para projetos Python. Ele será dividido em pacotes, nos quais serão armazenados os módulos e seus respectivos testes num diretório próprio:

- **qextractor**
  - **pacote**
    - modulo.y
    - **tests**
      - test_modulo.py

Os pacotes deverão refletir as fases descritas na seção 3. Cada módulo deve ser uma unidade coesa e com responsabilidades bem definidas.

Além disso, as bases de dados utilizadas serão armazenadas num diretório *data*. Os arquivos gerados durante o processamento serão gravados no subdiretório *data/gen*.

Referências

[1] Silvia Pareti et al. "Automatically Detecting and Attributing Indirect Quotations." Em: *EMNLP*. 2013, pp. 989–999.

[2] William Paulo Ducca Fernandes. "Quotation Extraction for Portuguese". Tese de mestrado. 2012.