Universidade de São Paulo
Instituto de Matemática e Estatística (IME-USP)

Reconhecimento de Entidades Mencionadas para Notificações de Processos Judiciais do Conselho Administrativo de Defesa Econômica

Aluno: Renan Fichberg
Orientador: Prof. Dr. Marcelo Finger

Monografia de Conclusão de Curso realizado para a disciplina
MAC0499 - Trabalho de Formatura Supervisionado

São Paulo, novembro de 2016

Agradecimentos

Este trabalho, apesar de ter apenas um autor, possui muito das experiências e conhecimento que acumulei ao longo do curso de Bacharelado em Ciência da Computação e, reconheço, não seria possível realizá-lo não fosse o aprendizado e o incentivo que tive, de várias pessoas com quem convivi não apenas na universidade, mas fora dela também. Destaco, a seguir, algumas pessoas com quem tive a chance de aprender bastante para chegar até o presente momento:

Primeiramente, aos meus pais Eloy Fichberg e Regina Célia de Oliveira Pinto e aos meus irmãos Felipe Fichberg e Leone Fichberg, que sempre foram as pessoas mais presentes na minha vida, me incentivando a seguir adiante em todos os momentos.

Em seguida, aos meus grandes amigos que acompanharam a minha trajetória de perto, Eduardo Gromatzky Feder e Gabriel Engel Pesso, que sempre foram companheiros em todos os momentos.

Aos meus colegas e amigos de curso Maurício Cardoso, Luiz da Silva Armesto, Renato Cordeiro Ferreira, Pedro de Carvalho Rogrigues, João Marco Maciel da Silva, Gervásio Santos, Renato Massao, Yara Grassi Gouffon, Rafael Raposa, Lucas Hiroshi Hayashida, Victor Sanches Portella, Luciana Abud, Vinícius Vendramini, Ruan Costa, Vinícius Bitencourt Matos e tantos outros que percorreram juntos comigo essa trilha e sempre se mostraram dispostos a ajudar.

Aos meus colegas e amigos do Mezuro, Rafael Reggiani Manzo, Diego Araújo Martinez Camarinha, Fellipe Souto Sampaio, Heitor Reis Ribeiro, Guilherme Rojas, Alessandro Palmeira e Daniel Paulino Alves, que sempre tinham algo de novo a ensinar.

Aos meus grandes amigos do colégio, Jonathan Schiriak, Allon Rozansky, Aaron Zarenczanski e Walter Caspari, pelos bons momentos, eventuais apoios e conselhos.

Aos colaboradores deste trabalho, William Collen, Kemil Raje e o meu orientador Prof. Dr. Marcelo Finger, por toda a paciência que tiveram com as minhas tantas dúvidas e sugestões de estratégias e ferramentas para solucionar os problemas que iam surgindo.

E finalmente, a todos os professores que tive a oportunidade de conhecer e aprender algo. Todos foram essenciais para a minha trajetória.

Resumo

O Conselho Administrativo de Defesa Econônima (CADE) é um orgão independente que reporta ao Ministério da Justiça e possui como missão garantir ao máximo a livre concorrência de mercado em todo o território Brasileiro e realiza as suas funções legais de acordo com a Lei Nº 12.529/2011. O CADE dispõem de uma base de dados bastante extensa, com processos judiciais de vários tipos distintos datados do ano de 1980 até os dias atuais e mais de 100 tipos diferentes de documentos, desde formulários, notificações de processos e cópias escaneadas de documentos diversos até arquivos de áudio e vídeo.

O trabalho foi divido em duas partes distintas: a primeira se constituiu de explorar parte dos vários processos judiciais presentes na base de dados pública do CADE relacionados a Atos de Concentração com finalísticos Sumário ou Ordinário e montar um Córpus com os tipos de documentos que foram julgados pertinentes em primeira análise. Em seguida, a partir de anotações manuais de entidades e seus relacionamentos sobre o Córpus construído, identificar automaticamente as entidades nos processos judiciais futuros que possam ser relevantes a classificação final entre os dois tipos de rito: Sumário ou Ordinário. A segunda parte, por sua vez, constitue-se da classificação do processo judicial em um dos ritos mencionados por meio de algoritmos de Aprendizado de Máquina, considerando as entidades que foram encontradas de forma automatizada nos documentos do processo em questão.

Este trabalho trata especificamente da primeira parte e serão aqui abordados assuntos relacionados a ela, tais como Processamento de Linguagem Natural, Reconhecimento de Entidades Mencionadas e algumas ferramentas de *software* que foram usadas para solucionar o problema em questão. A segunda parte, que infelizmente não foi desenvolvida por falta de tempo, será discutida menos detalhadamente na seção **TODO: COLOCAR SEÇÃO AQUI**.

Por fim, além destes conteúdos, também serão compartilhados experimentos e resultados obtidos por meio de validação cruzada com o Córpus desenvolvido, junto de possíveis estratégias que foram aprendidas ao longo do trabalho que poderiam, talvez, melhorar a precisão e a corretude das anotações.

Abstract

The Administrative Council for Economic Defense (CADE) is an independent agency reporting to the Ministry of Justice and has as mission to ensure to the maximum the free market competition over the entirety of the Brazilian territory and performs its legal functions according to the Law Nº 12.529/2011. The CADE owns an extense enough database, with judicial processes of many distinct types dated from the year of 1980 to the present days and over 100 types of different documents, from formularies, process notifications and scanned copies of diverse documents to audio and video files.

The work was divided in two distinct parts: the first one constituted of exploring part of the many judicial processes within the public database owned by CADE related to Concentrations Acts with final procedure being either "Sumario" or "Ordinario" and build a Corpus with the document types that were considered pertinent in first analysis. After that, considering manual annotations of entities and its relationships done over the built Corpus, identify automatically the entities in the future judicial processes that can be relevant to the final classification between the two types of rite: "Sumario" or "Ordinario". The second part constitutes of the classification of the judicial process in one of the mentioned rites through the use of Machine Learning algorithms, considering the entities that were found automatically in the documents of the given process.

This work covers specifically the first part and will be discussed here subjects related to it, like Natural Language Processing, Named Entity Recognition and some of software tools that were used to solve the introduced problem. The second part, which unfortunately was not developed duo to the lack of time, will be briefly discussed in the section **TODO: COLOCAR SEÇÃO AQUI**.

At last, in addition to these contents, will also be shared experiments and results obtained through the use of cross validation with the created Corpus, along with possible strategies that were learned throughout this project that could, maybe, increase the precision and correctness of the annotations.

Glossário de Siglas

AM - Aprendizado de Máquina
CADE - Conselho Administrativo de Defesa Econômica
REM - Reconhecimento de Entidades Renomadas
PLN - Processamento de Linguagem Natural

Índice

1 Introdução

2 Conselho Administrativo de Defesa Econômica (CADE)

2.1 Base de Dados pública

3 Processamento de Linguagem Natural

3.1 Criação do Córpus

3.2 Tamanho do Córpus

3.2.1 Anotações

3.3 Tokenização

3.4 Detecção de Setenças

4 Reconhecimento de Entidades Mencionadas

5 Ferramentas Utilizadas

5.1 BRAT

5.2 OpenNLP

6 Resultados

7 Conclusão

7.1 Dificuldades encontradas

7.1.1 Aprendizado de Máquina

8 Referências