

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 Pessa, 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, Felipe 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 Econômica (CADE) é um órgão 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 dispõe 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 dividido 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, constitui-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 correteza 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 N^o 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

As siglas descritas a seguir aparecem em partes diversas desta monografia. Confira abaixo os seus significados:

AC - Ato de Concentração

AI - *Artificial Intelligence*

AM - Aprendizado de Máquina

CADE - Conselho Administrativo de Defesa Econômica

AI - Inteligência Artificial

ML - *Machine Learning*

NER - *Named Entity Recognition*

NLP - *Natural Language Processing*

PLN - Processamento de Linguagem Natural

REM - Reconhecimento de Entidades Renomadas

Índice

Capítulo 1

Introdução

Neste primeiro capítulo será abordado um pouco da motivação para o desenvolvimento deste trabalho bem como os seus objetivos. Todo e qualquer conteúdo mais técnico mencionado aqui será melhor explorado nos capítulos sucessivos, desde assuntos relacionados a áreas da computação, tais como Reconhecimento de Entidades Mencionadas, até o funcionamento do Conselho Administrativo de Defesa Econômica, seus processos e a sua base de dados. Assim sendo, o leitor não deve se preocupar com eventuais dúvidas técnicas que possam surgir com a leitura deste capítulo meramente introdutório.

1.1 Motivação

O tema do projeto é atraente uma vez que está diretamente ligado à realidade da nossa sociedade. O CADE tem um papel fundamental para manter a concorrência de mercado entre competidores de todos os portes, atuando como um órgão regulador legal. Sua existência é particularmente importante para dar alguma garantia aos pequenos negócios de não serem engolidos por *players* veteranos, que já atuam em determinado mercado há mais tempo e, portanto, já dominam fatias consideráveis do público de interesse.

É conhecido também o fato de que processos judiciais tendem a ser demorados e mesmo que os Atos de Concentração tratados pelo CADE, objetos de estudo deste trabalho, sejam mais rápidos quando comparados a outros de diferente natureza, tentar torná-los ainda mais rápidos definitivamente é algo bem vindo, uma vez que tais processos judiciais podem demorar até seis meses para terem um tipo de rito escolhido.

A idéia, portanto, é justamente buscar formas de automatizar os processos em andamento de tal forma que exista um ganho tanto para o CADE quanto para a sociedade. Se tais processos pudessem ser acelerados, em sua totalidade ou partes do *pipeline* envolvido na análise, futuramente a mesma solução poderia ser replicada para resolver problemas similares com outros tipos de processos judiciais ou mesmo em outras áreas do conhecimento.

1.1.1 Retorno

Para o CADE, isso representaria a possibilidade de resolver mais processos em um mesmo intervalo de tempo, e para organizações ou cidadãos isso representaria terem uma resposta mais rápida para planejarem as suas próximas ações. Em especial, é importante ressaltar que estamos lidando com o mercado, que é uma entidade abstrata muito volátil, isto é: o mercado tem uma capacidade de se transformar muito rápida, portanto, quanto antes uma resposta for obtida, melhor, pois ela será, teoricamente, mais fiel ao estado do mercado no momento da petição.

1.2 Objetivos

1.3 Conselho Administrativo de Defesa Econômica (CADE)

1.3.1 Base de Dados pública

1.4 Processamento de Linguagem Natural

1.4.1 Criação do Córpus

1.4.2 Tamanho do Córpus

Anotações

1.4.3 Tokenização

1.4.4 Detecção de Setenças

1.5 Reconhecimento de Entidades Mencionadas

1.6 Ferramentas Utilizadas

1.6.1 BRAT

1.6.2 OpenNLP

1.7 Resultados

1.8 Conclusão

1.8.1 Dificuldades encontradas

Aprendizado de Máquina

1.9 Referências