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 Jarude 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 foram surgindo.

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

"The voice that navigated was definitely that of a machine, and yet you could tell that the machine was a woman, which hurt my mind a little. How can machines have genders? The machine also had an American accent. How can machines have nationalities? This can't be a good idea, making machines talk like real people, can it? Giving machines humanoid identities?"

- Matthew Quick, The Good Luck of Right Now

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[1]. 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 no capítulo 7, seção 7.2.1 - Aprendizado de Máquina.

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.

[1] Acesso à Informação: Conheça o CADE

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[2]. 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 chapter 7, section 7.2.1 - Aprendizado de Máquina.

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.

[2]Acesso à Informação: Conheça o CADE

Glossário de Siglas

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

AC - Ato de Concentração

AM - Aprendizado de Máquina

CADE - Conselho Administrativo de Defesa Econômica

IA - Inteligência Artificial

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 e o problema envolvido. 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 ser 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 orgão 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 mesmo 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: os mercados, de um modo geral, são flutuantes, de tal forma que a sua capacidade de se transformar é altíssima. Um dado mercado pode estar em alta em um dia e em queda no mês seguinte. Devido em grande a parte globalização, mercados constituem entidades de difícil previsão mesmo para *experts* em economia.

À luz do que foi dito, portanto, quanto antes uma resposta for obtida, melhor será, já que ela teoricamente será mais fiel a configuração do mercado no momento da petição.

1.2 Problema

Dado um conjunto de processos judiciais com ritos conhecidos, queremos buscar saber em quais ritos os processos futuros se encaixarão. Temos dois tipos de classes possíveis para os ritos: Ordinário e Sumário. Desta forma, este é claramente um problema de classificação binária e já existem maneiras conhecidas de resolvê-lo eficientemente a partir de técnicas e algoritmos clássicos de AM.

Para conseguirmos desempenhar esta função, porém, é necessário estudarmos as especificidades do problema proposto e, em particular e principalmente, do tipo de dados com que estamos lidando. Informações com relações a isso serão tratadas no capítulo 2 - CADE, na seção 2.2 - Base de Dados Pública. Note que esta etapa é essencial para garantirmos não apenas uma maior eficiência na classificação final, mas também que o algoritmo está sendo treinado sobre documentos totalmente fiéis à realidade.

Acerca do então exposto, surge naturalmente um segundo problema do qual o nosso *approach* por AM irá depender diretamente: construir um Córpus para servir de modelo de treinamento para identificar entidades chaves que podem ser determinantes no julgamento da classe do rito de um futuro processo.

Assim, uma das possibilidades que surge é usar REM para identificar tais entidades e, a partir destas, buscar alguma relação entre as entidades encontradas e um dos tipos de rito, com base nos padrões que foram aprendidos a partir do Córpus de treinamento criado com processos antigos, presentes na base de dados pública do CADE. Focamos a nossa atenção, portanto, em primeiramente resolver o problema da construção do Córpus para treinamento.

1.3 Objetivos

O principal objetivo deste trabalho foi estudar os ACs analisados pelo CADE para poder então, criar um Córpus de treinamento e a partir deste classificar o AC em um dos ritos. Infelizmente, conforme já mencionado no Resumo, o escopo acabou revelando-se grande demais para o tempo disponível, e portanto teve de ser reduzido ao primeiro problema. Mais sobre isso será discutido no capítulo 7, seção 7.1 - Dificuldades Encontradas.

Ainda, houve também um estudo do próprio funcionamento do CADE para entender mais sobre o problema e, em particular, um estudo voltado para a sua coletânea de processos armazenada na sua extensa base de dados pública com o intuito de identificar os tipos de documentos que poderiam ser mais pertinentes à análise dos processos no que diz respeito à classificação final do rito e também à forma que tais processos deveriam ser tratados para que fossem extraídas destes informações relevantes. Todo este conteúdo pode ser encontrado no capítulo 2 - CADE.

Como objetivo também existiu a necessidade de estudar assuntos relacionados a AI, tais como PLN e REM, que serão abordados nos capítulos 3 e 4, respectivamente. Ademais, foram estudadas ferramentas de *software* que trabalham com PLN e REM: OpenNLP e o BRAT. Mais será dito sobre elas no capítulo 5, em suas respectivas seções. Nestas seções será apresentado um estudo de como funcionam as funcionalidades usadas destas ferramentas e também como tais ferramentas foram utilizadas para que os resultados apresentados no capítulo 6 - Resultados fossem alcançados.

Finalmente, foi estudado a partir da técnica estatística de validação cruzada o desempenho do Córpus criado e os tipos de erros que surgiram no processo de geração automatizada de anotações de entidades mencionadas a partir do modelo de treinamento do Córpus, de tal forma que foram percebidas algumas boas práticas com relação ao uso destas ferramentas para

aumentar a eficácia e a acurácia das marcações. Tais percepções serão comentadas, também, no capítulo 6 - Resultados.

Capítulo 2

CADE

Neste capítulo serão abordados assuntos relacionados ao CADE, aos Atos de Concentração que devem ser legalmente submetidos a ele e à base de dados pública que ele possui e que contém tais processos judiciais. Sobre o CADE, será exposto um pouco da sua história e da sua função, para em seguida falarmos sobre o que são os ACs que cabem à sua competência a análise e, finalmente, sobre a base de dados pública que foi usada para obter os Atos de Concentração que compõem o Córpus construído, posteriormente usado para treinarmos um modelo que busca entidades mencionadas de forma automatiza. Mais informações sobre o Córpus serão abordadas no capítulo 3 - PLN e as suas entidades mencionadas no capítulo 4 - REM.

2.1 Quem é o CADE?

Criado pela Lei n° 4.137/62 como um orgão do Ministério da Justiça, o CADE hoje é uma autarquia em regime especial com jurisdição em todo o território nacional. Inicialmente, era da responsabilidade do Conselho a fiscalização da gestão econômica e do regime de contabilidade das empresas, atraves da Lei n° 8.884/1994, o CADE transformou-se em uma autarquia vinculada ao Ministério da Justiça.

Tal Lei definia as atribuições do CADE e de outros órgãos que formavam juntos com o Conselho Administrativo de Defesa Econômica o Sistema Brasileiro de Defesa da Concorrência e tinham como missão garantir a política de defesa da livre concorrência em todo o território nacional. O CADE, em particular, era responsável pelo julgamento dos processos administrativos que tinham relação com condutas anticompetitivas e também por apreciar Atos de Concentração, tais como aquisições, fusões, *joint ventures* e outros que fossem submetidos à sua aprovação.

Com a entrada da Lei nº 12.529/2011 em maio de 1012, esta uma nova Lei de Defesa da Concorrência, houve uma reestruturação do Sistema Brasileiro de Defesa da Concorrência e a política da qual ele era encarregado, de defesa da concorrência, passou por mudanças significativas. Em especial, pela nova legislação, o CADE passou a ser responsável por competências até então dos outros órgãos do Sistema Brasileiro de Defesa da Concorrência: instruir processos administrativos de apuração de infrações à ordem econômica e também de processos de análise de Atos de Concentração. Ainda sobre a Lei nº 12.529/2011, a principal mudança introduzida consistia na exigência de submissão prévia ao CADE de fusões e aquisições de empresas que podem proporcionar efeitos anticompetitivos no mercado, algo que no período anterior a esta Lei poderia ser feito depois destas operações serem consumadas. Para o CADE, passou a existir então um prazo máximo de dozentos e quarenta (240) dias para análise das operações, prorrogáveis por mais noventa (90) dias em casos de operações demasiadamente complexas.

Estruturalmente, com a Lei nº 12.529/2011 em vigor, tambem houveram mudanças: o CADE passou a ser constítuido pelo Tribunal Administrativo de Defesa Econômica, pelo Departamento de Estudos Econômicos e pela Superintendência-Geral. A esta última cabe desempenhar no novo

sistema grande parte das funções realizadas pelos outrora pelos órgãos que compunham junto ao CADE o Sistema Brasileiro de Defesa Econômica antes da entrada da nova Lei de meio de 2012, tais como a investigação e a instrução de processos de repressão ao abuso do poder econômico e a análise dos atos de concentração[2].

2.2 Atos de Concentração Econômica

2.3 Base de Dados Pública

[2]Acesso à Informação: Histórico do CADE

Capítulo 3

PLN

3.1 O que é um Córpus?

3.2 Criação do Córpus

3.3 Tamanho do Córpus

3.3.1 Anotações

3.4 Tokenização

3.5 Detecção de Setenças

Capítulo 4

REM

Capítulo 5

Ferramentas Utilizadas

5.1 BRAT

5.2 OpenNLP

Capítulo 6

Resultados

Capítulo 7

Conclusão

7.1 Dificuldades encontradas

7.2 Próximo Passo

7.2.1 Aprendizado de Máquina

Referências

[1] Acesso à Informação: Conheça o CADE. Disponível em: <http://www.cade.gov.br/acesso-a-informacao/institucional >. Acesso em: 13 de outubro de 2016.
[2] Assessoria de Comunicação Social. Acesso à Informação: Histórico do CADE. Disponível em: <http://www.cade.gov.br/acesso-a-informacao/institucional/historico-do-cade >. Acesso em: 13 de outubro de 2016.