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

DOU - Diário Oficial da União

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 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 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 os 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 mês 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[3].

2.2 Atos de Concentração Econômica

Os Atos de Concentração Econômicas são caracterizados por operações que envolvem duas ou mais empresas independentes, conforme descrito no artigo 90 da Lei 12.529/2011. Tais operações podem ser aquisições de controle ou incorporações de uma ou mais empresas por outras ou ainda a celebração de contratos associativos, consórcios ou *joint ventures* entre duas empresas ou mais.

2.2.1 Operações

São as operações, aliadas ao faturamento bruto anual ou volume de negócios no Brasil dos agentes econômicos envolvidos, que caracterizam a necessidade de existência dos Atos de Concentração analisados pelo CADE. Quando o faturamento de uma das empresas envolvidas atinge o patamar mínimo de R$ 750 milhões e o de uma outra, também envolvida na operação, de pelo menos R$ 75 milhões.

Considerando esta informação, é particularmente interessante que a aplicação desenvolvida saiba identificar as operações de um dado processo, especialmente pela razão de que certas operações tendem a seguir mais um ou outro tipo de rito. É relevante ressaltar a observação, no entanto, que o tipo de operação não é uma condição suficiente para identificar o tipo de rito, mas é um bom indicativo para buscarmos o mais provável[4]. Seguem abaixo os possíveis tipos de operações que um AC pode ter:

- **Fusão**: são caracterizadas pela união de duas ou mais empresas distintas para formar um novo agente econômico único.
- **Incorporação**: são caracterizadas pelo ato de uma ou mais empresas incorporar total ou parcialmente outras empresas dentro de uma mesma pessoa jurídica, de tal forma que o incorporado desaparece como pessoa jurídica, mas o incorporador mantém a sua identidade jurídica após a operação.
- **Aquisição**: são caracterizadas pelo ato de uma empresa adquirir o controle total ou parcial da participação acionária de outra empresa.
- ***Joint venture***: são caracterizadas pela criação de uma nova empresa a partir da associação entre duas ou mais empresas, de tal forma que as empresas que se associaram mantém normalmente suas identidades jurídicas pós operação.

2.3 Base de Dados Pública

O CADE possui uma base de dados[5] com processos datados desde 1980 até os dias atuais, de tal forma que para uma pessoa que não sabe ao certo o que está procurando facilmente pode se perder em meio a tantos tipos de processos. Para nós, porém, eram relevantes apenas os tipos de processo "Finalístico: Ato de Concentração Ordinário" e "Finalístico: Ato de Concentração Sumário", uma vez que foram os objetos de estudo deste trabalho.

[3] Acesso à Informação: Histórico do CADE
[4] Perguntas frequentes sobre Atos de Concentração Econômica
[5] Base de Dados Pública do CADE: Pesquisa Processual

Além dos tipos de processo, há outras informações que podem alimentar o sistema de recuperação de informação para que encontremos o que buscamos, tais como buscar um processo pelo seu número ou dentro de um determinado período cronológico. Uma vez selecionado um dos tipos de processo que nos é relevante (um dos Atos de Concentração mencionados no parágrafo anterior), é importante identificarmos os tipos de documentos que nos são relevantes para sabermos onde buscarmos as informações que precisamos.

A lista de tipos de documentos é deveras extensa e para alguém que desconhece o tipo de informação que está contido em cada um destes tipos de documentos, descobrir pode ser uma tarefa bastante demorada. Precisamos, portanto, de alguma heurística para encontrarmos tipos de documentos que sejam potenciais candidatos a serem considerados de alta relevância para nós.

2.3.1 Heurística de Seleção de Tipos de Documentos

Queremos descobrir, dentre os mais de 100 diferentes tipos de documentos presentes na base de dados, aqueles que devem ter as informações mais pertinentes para nós analisarmos os dados e tentarmos descobrir em qual rito determinado futuro processo será classificado. Ignorando os tipos de documentos por um instante e acessando os diferentes Atos de Concentração que aparecem em uma pesquisa "crua" (isto é, com apenas um dos tipos de processo selecionado), comparando-os um a um, é fácil de identificar que a maioria dos ACs, salve pouquíssimas exceções que provavelmente constituem em processos confidenciais, possuem os tipos de documentos "Notificação", "Formulário de Notificação" e "Publicação no DOU".

Notavelmente, tais tipos de documentos, de acordo com as informações presentes na base de dados, sempre estão entre os primeiros documentos submetidos no instante em que um Ato é dado como público. Para nós, o instante em que um AC é dado como público tem o mesmo efeito de encará-lo como inicializado, uma vez que não temos qualquer acesso a documentos confidenciais, e portanto nada podemos inferir sobre eles. Desta forma, chegamos à seguinte heurística para termos um ponto de partida e selecionarmos os tipos de documentos potencialmente mais interessantes, descrita abaixo:

1. Buscamos documentos que frequentemente "abrem" um Ato de Concentração.
2. Olhamos uma quantidade razoável de processos, digamos 50, e vemos em quantos deles tais documentos estão presentes
3. Supomos que tais documentos não são específicos a um AC e portanto devem ter as informações necessárias para que um novo Ato seja consolidado.
4. Como todo Ato obrigatoriamente segue um dos dois tipos de rito, as informações necessárias devem estar presente nos tipos de documentos selecionados.

O item **1** da heurística é particularmente importante pois ele encapsula todo o objetivo da nossa aplicação: *não queremos apenas identificar o mais provável tipo de rito de um determinado Ato de Concentração, mas queremos fazer isso com o mínimo possível de informações*, ou seja: quanto menor o número de análises forem necessárias por parte do CADE para chegar a uma conclusão relacionada ao rito, melhor.

Está idéia está diretamente relacionada ao *pipeline* mencionado no capítulo 1 - Introdução, seção 1.1 - Motivação. Suponhamos aqui para ilustrar a idéia do *pipeline* de análise que um determinado AC pode ter no máximo n fases F_i e que F_1 e F_n sejam suas fases inicial e final, respectivamente. Suponhamos ainda que existam vários tipos de documentos D_j que podem fazer parte do AC, mas que certos documentos só podem aparecer em uma fase F_i específica. Diremos que o valor V_j de um dado documento D_j é tão maior quanto menor for o valor de i, para $i = 1, 2, 3...n$ e que $V_i \in [0, 1]$ de tal forma que $V_1 = 1$ é o valor máximo de um documento

e $V_n = 0$ o valor mínimo. Consideremos, finalmente, o AC de $n = 5$ fases composto dos 10 documentos $D_j, 1 \leq j \leq 10$ tais que:

- $D_1, D_2, D_3 \in F_1$ documentos que abriram o Ato de Concentração.
- $D_4 \in F_2$ documento que foi produzido após análise da fase F_1.
- $D_5, D_6, D_7 \in F_3$ documentos que foram produzidos após a análise da fase F_2.
- $D_8 \in F_4$ documento que foi produzido após a análise da fase F_3.
- $D_9, D_{10} \in F_5$ documentos que foram produzido após a análise da fase F_4. Encerramento do Ato de Concentração. Como o rito já foi decidido pelo Conselho, não possuem valor para nós.

Conseqüentemente, temos a seguinte relação para os valores de cada um dos j documentos considerando o *pipeline* $F_i, 1 \leq i \leq 5$:

$$1 = V_1 = V_2 = V_3 > V_4 > V_5 = V_6 = V_7 > V_8 > V_9 = V_{10} = 0$$

Assim sendo, concluímos que os documentos de maior valor para nós são os mais próximos da fase inicial. Há, porém, uma pergunta que deve ser feita: por qual razão, necessariamente, deveríamos considerar esta interpretação correta? Está questão já foi respondida: tempo. Existe ainda um outro fator que não foi mencionado aqui e será abordado no capítulo 3 - PLN, seção 3.2 - Criação do Córpus, que responde a pergunta de outra maneira e portanto complementa a nossa resposta. Para adiantar a idéia, considere que os 3 tipos de documentos presentes na fase inicial, "Notificação", "Formulário de Notificação" e "Publicação no DOU" possuem diferentes padrões uma vez que a própria estrutura dos documentos são diferentes. Para completar o raciocínio, lembre-se: algoritmos de aprendizado de máquina aprendem com base em padrões aprendidos no modelo de trainamento[6] (ao menos os das ferramentas usadas neste trabalho)!

Abaixo, é exposto um pouco do que cada um destes tipos de documentos contém e a diferença estrutural de cada um deles:

1. **Notificação**: Tem uma estrutura informativa acerca da operação, dos agentes econômicos envolvidos e dos documentos anexos a relevantes ao Ato, com pedidos de acesso restrito para os anexos que contém informações críticas que, na opinião das empresas requerentes, caso os seus concorrentes viessem a conhecer prejudicaria o seu negócio.
2. **Formulário de Notificação**: Tem uma estrutura de perguntas e respostas, onde os agentes econômicos envolvidos respondem ao formulário do CADE. Nem todas as perguntas são respondidas, pois mais uma vez, certas respostas as requerentes querem que se mantenham confidenciais.
3. **Publicação no DOU**: Contém poucas linhas, com informações gerais do Ato tais como as organizações envolvidas, seus advogados, número do processo, operação objeto e setor econômico envolvido, declarando o AC público.

2.4 Ritos de um Ato de Concentração

Até agora, já foi dito muito sobre o objetivo de classificar os Atos em ritos Sumário ou Ordinário, mas nada foi falado a respeito deles. Esta seção, portanto, é dedicada a entendermos um pouco sobre eles: o que são e qual é a sua relevância.

[6] Documentação OpenNLP: REM

Dizemos que um dado processo segue rito ou procedimento Sumário quando ele é simplificado de forma a ser concluído mais rápido. Tal procedimento pode ser aplicado pelo CADE aos casos em que for considerado de pouco potencial ofensivo à concorrência as operações suficientemente simples. Nota que a decisão de enquadramento do pedido de aprovação pelo procedimento Sumário é adotada pelo CADE em casos de conveniência e oportunidade, considerando experiências passadas adquiridas com relação a identificação dos Atos que sejam potencialmente menos agressivos à concorrência.

Existem algumas características que tendem a ser enquadráveis em procedimento Sumário, descrita na Resolução CADE Nº 2, de 29 de maio de 2012, tais como:

I. *Joint ventures* clássicas ou cooperativas, que visa apenas a participação em um mercado cujos produtos e serviços não estejam horizontal ou verticalmente relacionados.

II. Substituição de agente econômico nos casos em que a empresa adquirente não participava, antes do Ato, do mercado envolvido direta ou indiretamente.

III. For provada baixa participação de mercado com sobreposição horizontal.

IV. For provada baixa participação de mercado com integração vertical.

V. Ausência de nexo de causalidade, isto é, concentrações horizontais que resultem em variação de HHI inferior a 200 com uma operação que não gere controle de mais da metade do mercado relevante.

VI. Outros casos que forem considerados simples, a critério da Superintendência-Geral.

Em teoria, todo o procedimento é considerado Ordinário até que se prove o contrário. Na prática, porém, o que se encontra é que a maior parte dos procedimentos são Sumários.

Ao submeter um Ato de Concentração a apreciação do CADE, as requerentes devem também submeter as respostas do Formulário de Notificação, que é diferente dependendo do rito. O procedimento Ordinário possui seções 12 seções a serem respondidas, enquanto o Sumário possui apenas 7. Ainda, estas 7 seções do procedimento Sumário estão contidas no procedimento Ordinário, mostrando que, de fato, o que difere um rito do outro é apenas a complexidade.

Capítulo 3

PLN

Neste capítulo falaremos um pouco sobre Processamento de Linguagem Natural (PLN), um campo da Ciência da Computação já bastante maduro que começou a ser muito explorado a partir do ano de 1950, apesar de ainda antes desta data ser possível encontrarmos trabalhos realizados em PLN. Problemas relacionados a Processamento de Linguagem Natural envolvem o entendimento de linguagens naturais por parte das máquinas ou mesmo geração de linguagem natural (isto é, a conversão de uma representação entendida por computadores em uma representação em linguagem natural).

Estamos particularmente interessados na primeira categoria de problemas de PLN mencionada, uma vez que os Atos de Concentração estão escritos em linguagem natural, mais especificamente o português, e buscamos extrair informações deles. Isso significa, naturalmente, que é necessário que exista algum entendimento por parte da máquina sobre o conteúdo presente nos ACs, usados para construir nosso Córpus. Sendo assim, nosso ponto de partida neste capítulo será justamente o Córpus.

3.1 Córpus

Já falamos muitas vezes a palavra Córpus neste trabalho, mas afinal, o que é um Córpus? Um Córpus é, como o próprio nome em latim sugere, um corpo composto de textos. Uma coleção destes corpos é o que chamamos de *Corpora*. Para trabalhar com um Córpus ou um *Corpora*, é necessário que este que este seja suficientemente extenso.

Mas de quão extenso estamos falando, afinal? Um exemplo de Córpus extenso é o Brown University Standard Corpus of Present-Day American English[7] (ou apenas Brown Corpus), compilado na década de 1960 na universidade de Brown, Providence, Rhode Island, como um Corpus de propósito geral no campo de linguística de corpus. Ele contém 500 exemplares de textos em inglês americano, com cerca de um milhão de palavras. Um Córpus mais modesto é o Susanne Corpus, com aproximadamente cento e trinta mil palavras, que é na realidade um subconjunto do Brown Corpus. Mais será discutido sobre tamanho de Córpus, na seção 3.1.2 - Tamanho do Córpus, deste mesmo capítulo.

De acordo com os autores Christopher D. Manning e Hinrich Schuetze [12], apenas para fazer a ordenação do Brown Corpus e criar uma lista de palavras nos primeiros anos de trabalho na sua construção era necessário 17 horas dedicadas de tempo de processamento, uma vez que os computadores tinham poucos kilobytes de memória. E os problemas não terminavam por aí, uma vez que para trabalhar com documentos deste tamanho também necessitava de discos rígidos grandes o suficiente para armazená-los. Isso significa que, apesar de PLN ser uma área que já vem sendo explorada há algum tempo, a tecnologia poderia facilmente ser o gargalo a depender da estratégia que fosse escolhida para buscar a solução do problema. Felizmente, com

[7] Versões do Brown Corpus podem ser encontradas na internet.

simples computadores atuais podemos realizar estas mesmas tarefas em questão de minutos.

Agora que já sabemos de que se trata um Córpus, falaremos sobre o que precisamos ter em mente para criarmos um Córpus na seção a seguir.

3.1.1 Criação do Córpus

3.1.2 Tamanho do Córpus

3.1.3 Anotações sobre o Córpus

3.2 Tokenização

3.3 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.

[3] Assessoria de Comunicação Social. Perguntas frequentes sobre Atos de Concentração Econômica. Disponível em: < http://www.cade.gov.br/servicos/perguntas-frequentes/perguntas-sobre-atos-de-concentracao-economica >. Acesso em: 14 de outubro de 2016.

[4] Base de Dados pública do CADE: Pesquisa Processual. Disponível em: < http://www.cade.gov.br/assuntos/processos-1 >. Acesso em: 14 de outubro de 2016.

[5] Documentação OpenNLP: REM. Disponível em: < https://opennlp.apache.org/documentation/manual/opennlp.html#tools.namefind >. Acesso em: 14 de outubro de 2016.

[6] CADE. Resolução nº 2, de 29 de maio de 2012. Seção 2, página 3. Disponível em: < http://www.cade.gov.br/assuntos/normas-e-legislacao/resolucao/resolucao-2_2012-analise-atos-concentracao.pdf >. Acesso em: 14 de outubro de 2016.

[7] CADE. Resolução nº 2, de 29 de maio de 2012. Anexo I - Formulário Procedimento Não-Sumário. Disponível em: < http://www.cade.gov.br/assuntos/normas-e-legislacao/resolucao/resolucao-2_2012-analise-atos-concentracao.pdf >. Acesso em: 14 de outubro de 2016.

[8] CADE. Resolução nº 2, de 29 de maio de 2012. Anexo II - Formulário Procedimento Sumário. Disponível em: < http://www.cade.gov.br/assuntos/normas-e-legislacao/resolucao/resolucao-2_2012-analise-atos-concentracao.pdf >. Acesso em: 14 de outubro de 2016.

[9] WIKIPEDIA. Natural Language Processing. Disponível em: < https://en.wikipedia.org/wiki/Natural_language_processing >. Acesso em: 15 de outubro de 2016.

[10] WIKIPEDIA. Brown Corpus. Disponível em: < https://en.wikipedia.org/wiki/Brown_Corpus >. Acesso em: 15 de outubro de 2016.

[11] SAMPSON, Geoffrey. The Susanne Corpus. Disponível em: < http://www.essex.ac.uk/linguistics/external/clmt/w3c/corpus_ling/content/corpora/list/public/susanne.html >. Acesso em: 15 de outubro de 2016.

[12] MANNING, Christopher D., SCHUETZE, Hinrich. Foundations of Statistical Natural Language Processing, MIT Press. Cambridge, MA: May 1999.

As we will see, there are a number of difficult issues in determining what is a word and what is a sentence. In practice these decisions are generally made by imperfect heuristic methods, and it is thus important to remember that the inaccuracies of these methods affect all subsequent results. Livro NLP