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

Reconhecimento de Entidades Mencionadas em Notificações de Atos de Concentração Econômica do Conselho Administrativo de Defesa Econômica

Parte Subjetiva

Renan Fichberg - 7991131

São Paulo, novembro de 2016

Este documento é um *relato informal* do trabalho desenvolvido para a disciplina MAC0499 - Trabalho de Formatura Supervisionado. O leitor não deve esperar nada refinado com a leitura deste documento, mas sim informações gerais e um breve histórico de como o trabalho surgiu bem como seus desafios e aprendizado na opinião e experiência do aluno. Ainda, por ser algo mais informal, me reservei no direito de fazer uso da primeira pessoa do singular ao redigir este documento.

Introdução

Foi próximo ao final do ano de 2014 que eu comecei a pensar a respeito do TCC. Nessa época eu já estava fazendo a contagem de créditos eletivos acumulados do curso e percebi que haviam disciplinas do meu interesse que não teriam como cursar (inevitavelmente, eu me formaria antes de qualquer oportunidade de ter um contato com todas as áreas que me despertam curiosidade). Assim, decidi que eu gostaria de fazer um TCC em uma destas áreas que eram muito do meu interesse mas que eu muito provavelmente não iria me matricular em um disciplina relacionada. No segundo semestre de 2013, eu tive a oportunidade de ter aula com o Prof. Dr Marcelo Finger na disciplina MAC0239 - Métodos Formais em Programação (atual MAC0239 - Introdução à Logica e Verificação de Programas), que acabou sendo o meu primeiro contato com um professor envolvido com as áreas de Inteligência Artificial e Aprendizado de Máquina.

Inicialmente, pensava em fazer uma Iniciação Científica relacionada à Inteligência Artificial, que posteriormente poderia se relacionar com um futuro projeto de TCC, mas por razões diversas a idéia da Iniciação Científica não vingou. Acabei nesse tempo (primeiro semestre de 2015) tornando-me monitor da disciplina MAC2301 - Laboratório de Programação, oferecida para a Escola Politécnica. Ao longo deste período, eu e o professor conversávamos sobre a possibilidade de um TCC.

Finalmente, no final de 2015 e início de 2016, quando eu já possuía os requisitos para me matricular na disciplina MAC0499, o professor me apresentou algumas possibilidades de projeto, entre as quais acabei selecionando o que se tornou o meu tema de TCC. Todos os projetos apresentados eram bem interessantes e distintos uns dos outros, mas para mim, o que pareceu mais desafiador acabou sendo a minha escolha final uma vez que era algo novo e não tinha ninguém mexendo nisso. Foi assim que eu terminei com o tema da minha monografia.

Do Trabalho

O objetivo do trabalho que me foi apresentado era muito simples: queríamos classificar Atos de Concentração entre dois tipos de ritos distintos. Tais ritos poderiam ser ou sumário ou ordinário. Só aqui, uma série de dúvidas já começaram a surgir, uma vez que eu não sabia o que eram estes tais Atos de Concentração e muito menos o que eram estes ritos. Assim, algum tempo depois (um bom tempo depois, na verdade), finalmente eu consegui a assistência de uma pessoa do Conselho Administrativo de Defesa Econômica (CADE), o Kemil, citado nos agradecimentos da minha monografia, para me ajudar com estas questões técnicas e conceituais do direito econômico relacionadas ao CADE e aos Atos de Concentração.

Entretanto, apesar dos objetivos de simples compreensão, o caminho que deveria ser percorrido até que obtivessemos um resultado era consideravelmente longo e não a toa acabou que o escopo do trabalho teve de ser diminuído para a duração da disciplina. Ainda, mesmo com a assistência de uma pessoa do CADE, a dificuldade de trabalhar com uma quantidade esmagadora de dados estava longe de ser uma tarefa simples.

Acabou que no final do mês de setembro de 2016, já relativamente próximo à finalização da disciplina e ainda com uma monografia para ser escrita, o escopo do trabalho foi oficialmente diminuído para a tarefa de extração de informações do conjunto de dados, algo que sozinho já é um bom tema de TCC devido a quantidade imensa de trabalho envolvido.

Desafios

Apesar do curso do BCC ser extremamente desafiador, trabalhar com conjuntos grandes de dados é algo que ao menos eu nunca tive a oportunidade de ter ao longo das disciplinas cursadas. Após ter realizado este TCC, para mim ficou bastante claro da razão de um Cientista de Dados ser bastante valorizado no mercado de trabalho: trabalhar com grandes quantidades de dados é difícil. É necessário conhecer bastante os dados (que podem estar com uma organização caótica) para conseguir extrair o máximo de conhecimento acerca das informações contidas ali e que sejam relevantes ao problema em questão, que foi basicamente o que eu tentei realizar no trabalho desenvolvido.

Estudar conjuntos de dados é uma tarefa bastante complexa e sem receitas de bolo, uma vez que estes dados todos são redigidos por pessoas e estas têm um grau de liberdade bastante alto acerca da forma e do conteúdo que elas redigem, de tal forma que encontrar um padrão exato freqüentemente é impossível. No caso deste trabalho, mesmo que as notificações dos Atos de Concentração tivessem uma estrutura textual semelhante, as informações contidas no texto freqüentemente divergiam para o nosso propósito (de classificação). Para ilustrar o problema, informações tais como o grau de Concentração Horizontal ou de Integração Vertical, que podem ser fundamentais para decidir o tipo de rito, muito vezes não estão presentes nas notificações, ou a menos, não nos documentos públicos. Isso sozinho aumenta consideravelmente o grau do problema.

Outro desafio foi com relação às ferramentas para o uso em documentos escritos em português. Até que o William Collen (também mencionado nos agradecimentos da monografia) me apresentasse ao BRAT, nada do que foi realizado aqui teria sido possível. As poucas ferramentas disponíveis que eu tinha encontrado até então ou eram pagas, ou exigiam outras tarefas tais como detecção de sentenças, tokenização e pos-tagging (e essa última, em particular, é muito trabalhosa) de tal forma que eu não teria tempo de me dedicar ao reconhecimento de entidades mencionadas a tempo de desenvolver algo.

Finalmente, um último desafio foi com relação as anotações: inúmeras vezes eu tive de mudar as entidades e os relacionamentos ora porque soavam redundantes, ora porque a minha cobertura dos dados que julgava pertinente não parecia precisa o suficiente. Ao longo do processo manual de geração de anotações, eu com certeza fiz mais de 25 declarações diferentes para entidades e seus relacionamentos. Ou seja, muitas vezes eu tive de editar (modificar, adicionar ou deletar)

anotações previamenta criadas pois percebi que aquele conjunto de definições não era bom o suficiente para cobrir o conjunto de Atos de Concentração que eu havia selecionado.

Aprendizado

O principal aprendizado foi, com toda a certeza, a experiência de trabalhar com grandes quantidades de dados e buscar extrair o máximo de informações relevantes corretamente dos mesmos. Aprendi bastante sobre Processamento de Linguagem Natural e Reconhecimento de Entidades Mencionadas na prática e tive a oportunidade de estruturar um córpus do começo ao fim, anotá-lo e ver os resultados do meu próprio trabalho experimentalmente, de tal forma que consegui até medir sua eficiência de acordo com métricas tradicionalmente usadas em recuperação de informações. Estou certo que essa atividade não é trivial e muito menos é uma experiência que qualquer Cientista da Computação terá a oportunidade de ter.

Fiquei satisfeito, em particular, pois consegui usar um pouco do conhecimento adquirido na disciplina eletiva MAC0333 - Armazenamento e Recuperação de Informações, oferecida no segundo semestre de 2015 pelo Prof. Dr. Alair Pereira do Lago, na hora de medir os resultados do córpus (o livro usado nesta disciplina até aparece nas referências da monografia, no item nº 18).

Além disso, tem o aprendizado da interdisplinaridade com direito, que também foi uma experiência interessantíssima e me proporcionou um novo conhecimento acerca de um mundo até então totalmente desconhecido para mim. Eu, claro, sabia que existiam mecanismos para garantir a livre concorrência mas não sabia com tantos detalhes da burocracia envolvida em todo este processo.

Finalmente, a oportunidade de desenvolver algo diferente do habitual é uma experiência que gera um aprendizado para a vida. Ao aceitar o projeto eu jamais teria a idéia de que o que eu acabei desenvolvendo fosse tão intimamente ligado à atividade de um cientista de dados. Eu honestamente não esperava que eu tivesse de, literalmente, trabalhar os dados do começo ao fim para extrair as informações supostamente (e aqui eu digo supostamente pois no final a classificação não ocorreu) relevantes. Eu esperava, como o habitual, trabalhar mais com código e fui positivamente surpreendido com um outro mundo: o dos dados.