

UNIVERSIDADE DE SÃO PAULO



IME-USP

INSTITUTO DE MATEMÁTICA E ESTATÍSTICA

MAC0499 TRABALHO DE CONCLUSÃO DE CURSO

Atividades concluídas e futuras

Autor:

Jackson J. de Souza

NºUSP: 6796969

Professor orientador:

Daniel M. Batista

1 de Dezembro de 2013

1 Introdução

Este informe tem por objetivo descrever de forma sucinta as atividades realizadas no TCC até agora e mostrar o que falta ser feito para concluir o planejamento de atividades. As atividades descritas estão separadas por atividades que geraram frutos ao trabalho e aquelas cujo resultado não pôde ser incorporado. No caso das atividades que não agregaram valor ao trabalho elas são acompanhadas de uma explicação do porque de não terem sido incorporadas. A intenção da descrição das atividades realizadas é dar uma noção de todo o esforço e tempo gastos no TCC.

2 Sobre o trabalho

O objetivo deste trabalho é fazer um estudo empírico das mensagens de segurança no Twitter escritas na língua inglesa para detectar alertas de segurança virtual. Para tal será feita uma comparação de desempenho entre os classificadores *Support vector machines* (SVM) e *Naive Bayes* para a detecção dos alertas de segurança computacional. Este estudo serve de apoio às teses de doutorado do Luiz A. F. Santos e do Rodrigo Campiolo que estão relacionadas com a detecção antecipada de anomalias em redes de computadores.

A realização das atividades pode ser dividida em duas fases:

1. Pesquisa e planejamento
2. Produção

Após eu ter entendido o problema e ter proposto uma abordagem (aprendizagem de máquina) pra ele eu precisei pensar em como processar os documentos de forma a construir um sistema para classificar os documentos. Como eu nunca tinha feito classificação de texto e não sabia como esse tipo de problema é implementado seja pelo mercado ou pela academia eu fiquei pesquisando conceitos de classificação de texto. Além disso, consultei o JEF e a Nina sobre o que eles achavam adequado para um TCC quanto ao escopo e à dificuldade da proposta de abordagem do problema. Também tive uma conversa com dois orientandos de doutorado do Daniel que trabalham na detecção de alertas de segurança virtual no Twitter. Na reunião pude tirar algumas dúvidas e conversamos sobre vários

tópicos envolvendo o problema como classificador a ser usado, pré-processamento dos tuítes, ferramenta a ser usada, etc. Ao final do primeiro semestre letivo eu consegui me definir a ferramenta a ser utilizada e descobri por meio do João Eduardo Ferreira (JEF), que deu uma introdução a Recuperação de Informação na disciplina de BD, uma boa bibliografia que possui uma abordagem consagrada pra pré-processamento e classificação de texto tornando possível planejar a realização do trabalho.

As atividades realizadas no segundo semestre tiveram caráter predominantemente prático. Ou seja, produziram informações e resultados para o trabalho. Por isso mesmo, a maioria delas são atividades de produção. As atividades podem ser separadas em atividades frutíferas e infrutíferas.

2.1 Atividades frutíferas

- Script que auxilia a categorização dos tuítes
- Script que armazena e ordena todos os diferentes caracteres encontrados nos tuítes (para fins de pré-processamento e tokenização)
- Script que lê os tuítes separados por classes e gera um arquivo que é lido pela Weka pra fazer a classificação
- Base de dados (tuítes separados por classes) com um total de 1500 tuítes
- Classificação dos tuítes coletados usando a Weka
- Realização de pesquisa entre pessoas da área da computação para avaliar os conhecimentos da comunidade da computação sobre alerta de segurança virtual
- Estudo dos modelos estatísticos *Naive Baiyes* e *SVM*
- Pesquisa sobre segurança virtual e adequação da definição de alerta de segurança virtual

Apesar da pesquisa sobre segurança virtual tratar de um dos conceitos do trabalho como os modelos estatísticos comparados na classificação dos tuítes e estar escrita na monografia eu resolvi listá-la entre as atividades pra enfatizar o tempo

e esforço que foram necessários para definir os conceitos envolvendo segurança virtual além de ameaça (e alerta) de segurança neste contexto. Ainda sobre as atividades frutíferas vale comentar que o atual estágio de escrita da monografia não condiz com o que já foi estudado. Ela está bastante incompleta ainda, pois não houve tempo para descrever os modelos estatísticos de classificação e explicar como foram extraídos os dados categorizados e classificados.

2.2 Atividades infrutíferas

- Pesquisa sobre abordagens para classificação de línguas em texto
- Coleta de corpora das 19 línguas mais faladas para construir um classificador de idiomas em texto
- Construção de classificador de línguas em texto
- Script que lê urls válidas e captura a manchete (*headline*) das páginas referenciadas pelas urls

No que diz respeito às atividades infrutíferas eu dediquei bastante tempo à pesquisa (o que inclui a leitura de artigos sobre métodos e comparações entre classificadores) e implementação de um classificador de línguas em texto, pois a base de dados de tuítes não está 100% na língua inglesa pela forma como o Twitter define a língua do tuíte. Embora o classificador não seja necessário para fazer a categorização dos tuítes, eu resolvi tentar implementar o classificador para ganhar conhecimento e experiência sobre o problema de detecção de línguas. Por isso, resolvi implementar um classificador de língua pra filtrar a base de dados de tuítes e para possivelmente aproveitar esse classificador em um sistema beta que classifica tuítes. Contudo, o fato de ter conseguido fazer o classificador funcionar apenas para línguas com alfabeto latino me fez deixar de lado o classificador de lado até concluir as atividades mais importantes do TCC.

Note que cada tuíte (dado) categorizado é lido por um ser humano capaz de aferir a classe dele. No caso isso inclui diferenciar um texto (tuíte) escrito em língua inglesa dos escritos em outras línguas. Além disso, a proposta do TCC não é construir um sistema classificador de tuítes e sim realizar um estudo que compara dois modelos utilizados para classificar os tuítes.

Eu também tinha pensado em ler as manchetes de urls presentes em cada tuíte que aparecessem no texto, mas isso iria exigir um trabalho extra e uma decisão conceitual que me deixou desconfortável. Por exemplo, se eu iria substituir o tuíte por uma manchete, concatenar o tuíte às manchetes, a influência que essas decisões poderiam surtir sobre o classificador como um possível viés, etc. Além disso, algumas páginas possuem links quebrados e alguns dos links apontam para uma página que redireciona para a da notícia. Ao tentar aplicar esse problema percebi que eu precisaria gastar bastante tempo nesse problema realizando muitos testes e aprendendo uma forma de lidar com a página que serve de redirecionamento para a notícia real como acontece com links para artigos da revista Forbes. No final das contas percebi que a detecção das manchetes dos tuítes seriam bastante úteis para ajudar um sistema classificador de tuítes a fazer uma boa predição.

2.3 Atividades que precisam ser realizadas

Para finalizar o trabalho faltam algumas atividades. Entre elas estão:

- Continuar categorização de tuítes até ter um montante de 12000 tuítes
- Estudar o modelo estatístico SVM multiclases (mencionar conversa com o Fujita)
- Revisar conceitos estatísticos que apoiam a efetividade da classificação
- Análise estatística de pesquisa realizada com pessoas da área da computação sobre alerta de segurança virtual
- Refinamento do pré-processamento de texto
- Refinamento da extração de características do texto para realização da classificação
- Continuar a escrita da monografia

Algumas atividades ainda requerem um pouco de estudo, mas não muito como: os refinamentos que consistem basicamente em remover caracteres encontrados nos tuítes que não agregam valor às características do classificador e a extrair mais informações para melhorar a classificação dos tuítes do que é feito atualmente.

Também preciso estudar o modelo SVM multiclass devido à indicação do supervisor da disciplina de que seria melhor fazer a comparação entre os modelos adotados utilizando o mesmo número de classes em ambos. Vou inclusive, segundo sugestão, me reunir com o professor André Fujita pra conversar sobre qual abordagem do SVM multiclass seria interessante para o problema de classificação deste trabalho.

Para a análise estatística dos dados eu preciso conversar com um professor do departamento de estatística para saber como fazer a análise corretamente, pois os conceitos usados fogem do escopo das disciplinas obrigatórias de estatística oferecidas para o BCC.

2.4 Considerações finais

Antes de encerrar vale mencionar que não posso apresentar um plano de trabalho para concluir o TCC, pois ele pode variar de acordo com o período de recuperação da disciplina.

A documentação do material que eu tenho até agora e enviei ao supervisor é insuficiente e ela sonega a forma como foi usada a Weka pra gerar os resultados preliminares. Além disso, a base de dados é grande demais para ser enviada como trabalho final, mas ela está disponível no repositório usado para a realização do trabalho.

Para mais detalhes sobre o material que foi estudado ao longo do ano até agora vale a pena dar uma olhada na movimentação do repositório que tenho utilizado para armazenar todo o material estudado e produzido para a realização do trabalho.

Segue o link do repositório: <https://bitbucket.org/jacksonjsouza/tcc>