Jefferson Carlos de Mendonça

Modelo de Mineração de Dados
Análise das Perguntas do Site Stack Overflow

São Paulo

Maio 2016, v-1.0.0

Jefferson Carlos de Mendonça

Modelo de Mineração de Dados
Análise das Perguntas do Site Stack Overflow

Modelo canônico de Projeto de pesquisa em conformidade com as normas ABNT apresentado ao Programa de Pós-Graduação em Engenharia da Computação da Escola Politécnica da Universidade de São Paulo, requisito para a seleção de ingresso ao Curso de Mestrado

Escola Politécnica da Universidade de São Paulo – USP

Faculdade de Engenharia e Computação

Programa de Pós-Graduação

São Paulo

Maio 2016, v-1.0.0

Lista de tabelas

Tabela 1 – Cronograma 17

Lista de abreviaturas e siglas

USP	Universidade de São Paulo
API	Application Programming Interface
REST	Representational State Transfer

Sumário

Introdução 9

1 OBJETIVOS 11

Tópicos de interesse, assunto previsto, relevância para a área específica e aplicabilidade do estudo

2 METODOLOGIA 13

3 RECURSOS 15

4 CRONOGRAMA 17

Considerações finais 19

Introdução

Não raro a evolução na área computacional se desenvolve em ritmo acelarado, novos algoritmos, técnicas para programação distribuída, inteligência artificial etc. Com as *Linguagens de Programação* não é diferente, elas aprimoram suas API's constantemente, construindo e descontruindo métodos, muitas vezes quebrando paradigmas. O que se sabe hoje pode estar obsoleto amanhã, as exigências e tendênicas mudam com frequência, requerindo capacitação atualizada aos profissionais que atuam neste universo.

O grande desafio das universidades que trabalham na formação de quadro profissionais é prover conhecimento não só para os especialistas com perfil para atuar em linhas de pesquisa, onde o conhecimento é mais profundo, mas também para aqueles que irão, após sua formação compor o mercado de trabalho, que muitas vezes é mais razo, porém mais dinâmico.

E para dar vazão a este dinamismo profissionais da área de computação diariamente recorrem a cursos online, livros, revistas e claro *websites*. Dentre estas mídias merece destaque o fórum *Stack Overflow*, maior comunidade online para programadores aprender, compartilhar conhecimento e progredir na carreira. Também é possível fazer um *tour* pelo site[1], onde são apresentados as regras e a mecânica desta poderosa ferramenta para troca de conhecimento.

O objeto deste projeto de pesquisa é propor um algoritmo de linguagem de máquina que consiga catalogar as perguntas desta comunidade, objetivando entender os principais questionamentos dos usuários e detectando padrões nas dúvidas mais frequentes. Estes insumos estarão disponíveis para que universidades e instituições de ensino possam aperfeiçoar seus cursos e treinamentos, estreitando a distância entre o que é lecionado e os requisitos impostos pelo dia-a-dia nas empresas.

[1] <http://stackoverflow.com/tour>

1 Objetivos

Tópicos de interesse, assunto previsto, relevância para a área específica e aplicabilidade do estudo

Os principais questionamentos dos usuários participantes da rede *Stack Overflow*, objeto de pesquisa deste trabalho, será embasado pelos assuntos das áreas: *Educational Data Mining*, *Learning Analytics* e *Machine Learning* e versará sobre os tópicos de interesse: *Mineração de dados* para reconhecer padrões nas dúvidas dos usuários e *algorítmos baseados em aprendizagem de máquina* para arranjar as perguntas sintáticamente equivalentes em grupos.

O universo de dados coletados e categorizados, será de grande relevância para instuições de ensino sobre o tema Programação de Computadores - foco das análises, que poderão aprimorar sua grade curricular e com isso aproximar a distância entre as expectativas do mercado de trabalho com o que é lecionado em salas de aula.

2 Metodologia

O ponto de partida será a reunião de material literário para a composição da estrutura e argumentação sobre o objeto de estudo proposto.

Na prática será desenvolvido um programa de computador para consumir os dados do Stack Overflow utilizando sua API, imputando toda a informação extraída em uma fila de processamento.

Na segunda etapa será proposto um algortimo de linguagem de máquina para extrair os assuntos principais das dúvidas postadas no *site* e armazenar os dados obtidos. Em adição também serão armazados a quantidade de vizualizações, pontuação ou número de respostas da referida pergunta.

Com os dados processados e armazenados em um banco para grandes massa de dados[1] e então utilizaremos técnicas de mineração de dados para detectar padrões nas dúvidas mais frequentes da comunidade, por fim será criado um *ranking* das mais relevantes para o ensino da Linguagem de Programação *Java*.

[1] De acordo com a empresa Stack Overflow o site recebe cerca de 101 milhões de visitantes únicos mensalmente e conta com 3.7 milhões de perguntas respondidas

3 Recursos

Para que os algoritmos propostos, bem como os programas de computares desenvolvidos para este projeto sejam executados em ambiente de alta performance, serão utilizados servidores na nuvem nas seguites etapas: consumo dos serviços REST providos pelo Stack Overflow, entrada de dados na fila de processamento e para o armazenamento das informações em banco de dados.

4 Cronograma

A Tabela 1 apresenta o cronograma estimado para a conclusão do projeto proposto.

Tabela 1: Lista de atividades.

Atividade	**Tópico**	**Tempo**
Linguagem de Máquina	Algoritmos para extração de textos	32
Mineração de Dados	Tabulação de dados	32
Big Data	Técnicas de *Map Reduce*	24
Infra	Criar infraestrutura na nuvem	2
Stack Overflow	Estudo da API da comunidade	4
Programa	Extração dos dados via API	4
Monografia	Revisar e complementar a monografia desenvolvida	3
Apresentação	Material para apresentar a dissertação	3

Observação: Tempo Estimado em semanas

Considerações finais

Sed consequat tellus et tortor. Ut tempor laoreet quam. Nullam id wisi a libero tristique semper. Nullam nisl massa, rutrum ut, egestas semper, mollis id, leo. Nulla ac massa eu risus blandit mattis. Mauris ut nunc. In hac habitasse platea dictumst. Aliquam eget tortor. Quisque dapibus pede in erat. Nunc enim. In dui nulla, commodo at, consectetuer nec, malesuada nec, elit. Aliquam ornare tellus eu urna. Sed nec metus. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas.

Phasellus id magna. Duis malesuada interdum arcu. Integer metus. Morbi pulvinar pellentesque mi. Suspendisse sed est eu magna molestie egestas. Quisque mi lorem, pulvinar eget, egestas quis, luctus at, ante. Proin auctor vehicula purus. Fusce ac nisl aliquam ante hendrerit pellentesque. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Morbi wisi. Etiam arcu mauris, facilisis sed, eleifend non, nonummy ut, pede. Cras ut lacus tempor metus mollis placerat. Vivamus eu tortor vel metus interdum malesuada.

Sed eleifend, eros sit amet faucibus elementum, urna sapien consectetuer mauris, quis egestas leo justo non risus. Morbi non felis ac libero vulputate fringilla. Mauris libero eros, lacinia non, sodales quis, dapibus porttitor, pede. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Morbi dapibus mauris condimentum nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Etiam sit amet erat. Nulla varius. Etiam tincidunt dui vitae turpis. Donec leo. Morbi vulputate convallis est. Integer aliquet. Pellentesque aliquet sodales urna.