Jefferson Carlos de Mendonça

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

São Paulo

Agosto 2016, v-1.0.1

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

Jefferson Carlos de Mendonça
Aluno

Edson Satoshi Gomi
Tutor/Orientador

São Paulo
Agosto 2016, v-1.0.1

Lista de tabelas

Lista de abreviaturas e siglas

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

Sumário

Introdução

Não raro a evolução na área computacional se desenvolve em ritmo acelerado, novos algoritmos, técnicas para processamento de imagem, computação distribuída, inteligência artificial etc. Com as *Linguagens de Programação* não é diferente, elas aprimoram suas API's constantemente, construindo e desconstruindo métodos e muitas vezes quebram 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 profissionais é prover conhecimento não só para especialistas com perfil para atuar em linhas de pesquisa, onde o conhecimento é mais profundo, mas também para aqueles que, após a formação irão compor o mercado de trabalho, em que o conhecimento pode ser mais raso, porém mais dinâmico.

E para dar vazão a este dinamismo profissionais da área de computação diariamente recorrem a cursos na *internet*, livros, revistas e claro *websites*. Dentre estas mídias, merece destaque o fórum *Stack Overflow*[1], maior comunidade *online* para programadores aprender, compartilhar conhecimento e progredir na carreira. Também é possível fazer um *tour*[2] pelo site, 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, nosso objetivo é entender os principais questionamentos dos usuários e detectar padrões nas dúvidas mais frequentes. Estes insumos serão disponibilizados para que as 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/company/about>
[2] <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*, *Big Data* 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 reduzir a distância entre as expectativas do mercado de trabalho com o que é lecionado na sala 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[1], imputando toda a informação extraída em uma fila de processamento.

Na segunda etapa será proposto um algoritmo de linguagem de máquina (HULTH, 2003) para extrair os assuntos principais (TURNEY, 2000) das dúvidas postadas no *site* e armazenar as informações obtidas. Em adição também serão armazados a quantidade de vizualizações, pontuação e o número de respostas da referida pergunta.

Com os dados processados e armazenados em um banco para grandes massa de dados[2], faremos uso de técnicas de mineração de dados para detectar padrões nas dúvidas mais frequentes da comunidade e por fim será criado um *ranking* (MIHALCEA; TARAU, 2004) das mais relevantes para o ensino da Linguagem de Programação *Java*.

[1] <https://api.stackexchange.com/docs>

[2] 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 computadores desenvolvidos para este projeto sejam executados em ambiente de alta performance, serão utilizados servidores na nuvem nas seguintes etapas: consumo dos *endpoints REST* providos pelo *Stack Overflow* e entrada de dados na fila de processamento, *queue*.

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

Referências

BISHOP, C. *Pattern Recognition and Machine Learning*. 1. ed. Cambridge, MA, UK: Springer, 2007. (Information Science and Statistics). Nenhuma citação no texto.

CASTRO, L. N. de; FERRARI, D. G. *Introdução a Mineração de Dados - Conceitos básicos, algoritmos e aplicações*. 1. ed. [S.l.]: Saraiva, 2016. Nenhuma citação no texto.

ELATIA, S.; IPPERCIEL, D.; ZAïANE, O. R. *Data Mining and Learning Analytics: Applications in Educational Research*. 1. ed. [S.l.]: Wiley, 2016. (Wiley Series on Methods and Applications in Data Mining). Nenhuma citação no texto.

ERL, T.; KHATTAK, W.; BUHLER, P. *Big Data Fundamentals: Concepts, Drivers & Techniques*. 1. ed. [S.l.]: Prentice Hall, 2015. (The Prentice Hall Service Technology Series). Nenhuma citação no texto.

FLACH, P. *Machine Learning: The Art and Science of Algorithms that Make Sense of Dat*. 1. ed. New York, NY, USA: Cambridge University Press, 2012. Nenhuma citação no texto.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2. ed. [S.l.]: Springer, 2009. (Information Science and Statistics). Nenhuma citação no texto.

HULTH, A. Improved automatic keyword extraction given more linguistic knowledge. In: *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2003. (EMNLP '03), p. 216–223. Disponível em: <http://dx.doi.org/10.3115/1119355.1119383>. Citado na página 13.

JAMES, G.; WITTEN, D. et al. *An Introduction to Statistical Learning: with Applications in R*. 1. ed. [S.l.]: Springer, 2013. (Springer Texts in Statistics (Book 103)). Nenhuma citação no texto.

KELLEHER, J. D.; NAMEE, B. M.; ARCY, A. D. *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*. 1. ed. Massachusetts, MA, USA: The MIT Press, 2015. (MIT Press). Nenhuma citação no texto.

MIHALCEA, R.; TARAU, P. Textrank: Bringing order into texts. In: LIN, D.; WU, D. (Ed.). *Proceedings of EMNLP 2004*. Barcelona, Spain: Association for Computational Linguistics, 2004. p. 404–411. Disponível em: <http://www.aclweb.org/anthology/W04-3252>. Citado na página 13.

ROMERO, C. et al. *Data Mining: Practical Machine Learning Tools and Techniques*. 1. ed. Flórida, FL, USA: CRC Press, 2010. (Chapman & HallCRC Data Mining and Knowledge Discovery Series). Nenhuma citação no texto.

ROMERO, C. et al. *Handbook of Educational Data Mining*. 1. ed. Flórida, FL, USA: CRC Press, 2010. (Chapman & HallCRC Data Mining and Knowledge Discovery Series). Nenhuma citação no texto.

TURNEY, P. D. Learning algorithms for keyphrase extraction. *Inf. Retr.*, Kluwer Academic Publishers, Hingham, MA, USA, v. 2, n. 4, p. 303–336, maio 2000. ISSN 1386-4564. Disponível em: <http://dx.doi.org/10.1023/A:1009976227802>. Citado na página 13.

WITTEN, I. H.; FRANK, E.; HALL, M. A. *Data Mining: Practical Machine Learning Tools and Techniques*. 3. ed. Burlington, MA, USA: Morgan Kaufmann, 2011. (Morgan Kaufmann Series in Data Management Systems). Nenhuma citação no texto.