MAC0215 - Proposta

Aluno: Victor Hugo Miranda Pinto
Supervisor: Alfredo Goldman vel Lejbman
Co-supervisor: Renato Cordeiro Ferreira

2º Semestre de 2018

1 Resumo

Hackathons ("hacker marathons") são eventos em que os participantes desenvolvem protótipos de software ou hardware num tempo restrito. O HackathonUSP é um hackathon universitário que tem como objetivo ser acessível para novos alunos e estimular a inovação tecnológica na Universidade de São Paulo (USP). Este projeto propõe a criação de um algoritmo para seleção automática de participantes do evento, utilizando dados dos inscritos e o número de participantes desejado. Serão investigados diferentes modelos e técnicas de aprendizado de máquina para desenvolver o algoritmo, cujas estruturas (meta-parâmetros) serão baseados nos critérios de decisão utilizados na seleção manual e as distribuições (parâmetros) serão treinadas a partir de dados das edições anteriores do evento. Dessa maneira, espera-se tornar a seleção mais rápida e transparente, beneficiando as próximas edições do HackathonUSP e outros eventos do tipo.

2 Introdução

Nesta proposta, serão introduzidos alguns conceitos fundamentais para a pesquisa. Em seguida, quais os objetivos do trabalho para este semestre. Detalhamos a metodologia que seguiremos ao longo da pesquisa, seguida por um plano de trabalho que apresenta a evolução esperada durante o semestre, inclusive estimativas de tempo para concluir cada etapa. Por fim, descrevemos o método de acompanhamento do trabalho sendo desenvolvido.

2.1 Hackathons

O termo *hackathon* vem da combinação das palavras *hacking* e *marathon*. Hackathons são eventos contínuos em que os participantes se organizam em pequenos times com o objetivo de criar um protótipo de software ou hardware sob um limite de tempo (24 horas em seu formato mais comum) [KPR+15]. Realizados com objetivos e de maneiras diferentes, hackathons são feitos tanto por empresas (normalmente de tecnologia) quanto por instituições de ensino, sociedade civil ou governo.

Muitas empresas de tecnologia realizam hackathons internos, ou seja, apenas para seus funcionários, com o objetivo de gerar inovação dentro da empresa ou para fazer com que os desenvolvedores trabalhem em projetos diferentes dos do seu dia a dia. Dependendo dos protótipos desenvolvidos durante o evento, a empresa pode decidir implantar a ideia. É comum que essas empresas também realizem hackathons abertos, em geral com o objetivo de promover um novo produto na comunidade desenvolvedora. Em geral, esses eventos também incluem algum tipo de premiação para motivar a participação, constituindo portanto uma competição.

No contexto de instituições de ensino superior, são realizados hackathons *universitários*. Esse tipo de evento possui objetivos diferentes daqueles das empresas, já que costumam não ter fins lucrativos e buscam incentivar a criatividade e a inovação no ambiente universitário. Hackathons universitários são muito efetivos como ferramenta educacional para o crescimento de futuros desenvolvedores. Esses eventos fornecem um ambiente para aprendizado e experimentação com novas tecnologias. Ao passo que incentivam os alunos a criarem novas soluções para a comunidade interna [Kay17]. Além disso, propiciam a integração entre alunos de diferentes cursos, promovendo a interdisciplinaridade, a troca de conhecimentos e experiências.

2.2 HackathonUSP

O HackathonUSP é o um dos maiores hackathons universitários do Brasil, com número crescente de inscritos e participantes. Ele tem como público-alvo os alunos de graduação e pós da Universidade de São Paulo. É organizado pelo grupo de extensão USPCodeLab (UCL) em parceria com o Núcleo de Empreendedorismo da USP (NEU) e o Hardware Livre USP (HL). O objetivo do evento é ser um hackathon acessível, o primeiro evento do tipo para alunos da Universidade, ao passo que ajude a própria comunidade USP a gerar soluções para as suas demandas.

Atualmente, o evento é realizado em edições semestrais isoladas, sempre na cidade de São Paulo, no Campus Capital da USP. A partir de 2019, o *USPCodeLab* planeja expandir o evento de maneira que ele ocorra no formato de "liga", sendo que a primeira fase ocorrerá paralelamente nos diferentes campi da USP, durante o 1º semestre, e a 2ª fase reunirá os times mais bem colocados numa final no campus Capital, durante o 2º semestre. Essa expansão tem o potencial de aumentar o alcance do evento na universidade, pois a participação de alunos dos campi do interior é reduzida pelo deslocamento necessário para que eles participem do evento

2.3 Seleção de participantes

Um dos elementos mais decisivos para o sucesso do HackathonUSP é a seleção dos seus participantes. Como o evento historicamente possui mais inscritos que vagas, a etapa de seleção é importante para maximizar a participatividade (i.e., evitar a desistência) e a diversidade (i.e., evitar o viés de público existente na computação) do evento, possibilitando o impacto proposto pelos organizadores. Entre os dados que compõem o formulário de inscrição, sempre está presente a pergunta sobre uma possível equipe, onde o inscrito informa se já possui uma equipe e quais os seus integrantes. Esse dado é utilizada no processo de seleção para evitar que algum membro do time fique fora do evento (o que, pela experiência dos organizadores, aumenta a chance do time como um todo desistir de participar).

O método utilizado na seleção atual pode ser descrito da seguinte maneira:

Sejam N o número de inscritos no evento, M o número de participantes e D a taxa de desistência entre os selecionados para o evento. Então:

1. Selecionar os primeiros M inscritos, por ordem de inscrição.
2. Selecionar mais inscritos para obter M participantes efetivos contando com um percentual de desistência D (em geral, 150% de M para D igual a 33%), levando em consideração:
    - o completamento de times: selecionando um inscrito que participa de um time cujos demais integrantes foram selecionados nos primeiros M;
    - a paridade de gêneros: selecionando mulheres que não tenham times para equilibrar a proporção de gênero entre os selecionados (a área da computação conta com um percentual pequeno de mulheres comparativamente ao de homens [FA90]);

- o desempenho em edições anteriores: como último critério de completamento, selecionar times cujos membros tiveram destaque em edições anteriores mesmo que nenhum deles estivesse entre os M primeiros inscritos.

Na edição mais recente do evento, o HackathonUSP 2018.1, a organização recebeu mais de 360 inscritos. Seguindo a tendência de crescimento desse número, observada com o passar das edições, o processo de seleção dos participantes tem se tornado cada vez mais desafiador.

2.4 Hacknizer

O Hacknizer é uma plataforma para organizar e hospedar hackathons, visando prover uma solução integrada e completa para a criação desses eventos. Ele está sendo construído pelo *USPCodeLab* e seu desenvolvimento começou durante a *USPCodeLab Winter School 2018*, um evento do tipo *dev.camp* (parte do ciclo *dev.journey*), organizado pelo grupo com o objetivo de experimentar novas tecnologias disponíveis para o desenvolvimento Web.

A arquitetura do sistema segue o padrão de microsserviços, ou seja, a estrutura de *backend* é constituída por módulos cujo desenvolvimento, implantação e replicação são semi-independentes, permitindo melhor escalabilidade do sistema [New15]. Os serviços existentes são:

1. **Hackathons**: responsável por gerenciar informações referentes a cada série de hackathons e suas respectivas edições hospedadas no sistema.
2. ***Auth***: responsável por autenticação de usuários e autorização de acesso, além do armazenamento de dados básicos dos usuários (nome, e-mail, etc.).
3. **Gerador de PWAs**: responsável pela criação de *Progressive Web Apps* (PWAs) para cada hackathon, ou seja, páginas web instaláveis como aplicativos para cada evento.
4. **Participantes**: responsável pelo gerenciamento dos usuários associados aos hackathons, incluindo seus papéis numa dada edição (organizador, participante, mentor, jurado ou convidado).
5. **Grupos**: responsável pelo gerenciamento dos grupos de participantes dentro de um hackathon e pela submissão de projetos.
6. **Seleção**: responsável pela seleção manual de participantes a serem convidados para um certo hackathon a partir da lista de inscritos.
7. ***API Gateway***: responsável por receber requisições dos clientes da plataforma (Web e mobile) e direcionar essas requisições para os microsserviços corretos afim de construir uma única resposta para enviar de volta ao cliente.

O projeto foi desenvolvido utilizando ferramentas e tecnologias tais como:

- ***Docker***: usado para conteinerização dos módulos.
- ***Kubernetes***: usado para gerenciamento de *clusters*.
- ***GraphQL***: especificação para design de APIs.
- ***Node***: linguagem e ambiente de programação para *backend*.
- ***Koa***: *microframework* de *Node* para construir servidores HTTP.

- ***Apollo Server***: biblioteca para construir servidores GraphQL.
- ***PostgreSQL***: sistema gerenciador de bancos de dados relacionais para consulta.
- ***Kafka***: transmissor de mensagens para consistência entre serviços.
- ***Prisma***: mapeador do esquema de APIs GraphQL para o PostgreSQL.

3 Objetivos

Os objetivos dessa pesquisa são:

- **Automatizar a seleção de participantes**
  Buscar uma maneira de automatizar o processo de seleção já existente para o HackathonUSP. Com isso, será possível aumentar a transparência na organização do evento — pedido recorrente dos inscritos — e diminuir o trabalho envolvendo tal seleção — dado os planos de expansão do evento para mais unidades da USP e o consequente aumento no volume de inscritos.
- **Integrar o algoritmo de seleção com o Hacknizer**
  Complementar o serviço de **Seleção** do Hacknizer, adicionando um componente que recebe a lista de inscritos para um dado hackathon e é capaz de gerar uma recomendação dos participantes para o evento, aplicando o algoritmo desenvolvido no passo anterior.
- **Investigar possíveis melhorias no algoritmo de seleção para o HackathonUSP**
  Estabelecer um critério de seleção que minimize o grau de desistência ao passo que maximize a diversidade (gênero, ano de ingresso, curso, unidade) entre os participantes do evento.

4 Metodologia

- **Tratamento dos dados**
  A análise que será feita para esta pesquisa se baseia nos dados de inscrição obtidos dos inscritos nas edições anteriores do HackathonUSP. Esses dados foram preenchidos individualmente pelos inscritos e possuem campos não estruturados (de modo geral, entradas para texto sem formatação exigida). Esses dados serão uniformizados e agrupados segundo o modelo de dados do Hacknizer.
- **Desenvolvimento do algoritmo**
  Para o desenvolvimento do algoritmo de seleção e estudo do conjunto de dados disponíveis, a linguagem *Python* será utilizada devido à disponibilidade da biblioteca *scikit-learn* que contém a implementação de diversos algoritmos para aprendizagem de máquina, e de *Jupyter Notebooks*, uma aplicação web de código aberto que permite a criação e compartilhamento de documentos contendo código *Python* e visualizações de dados. Em particular, vamos testar alguns classificadores: Regressão Linear, Regressão Logística, Árvores de Decisão e Florestas Aleatórias.

  Para medir a qualidade das seleções, será considerada a distância entre a classificação gerada pelo algoritmo em comparação com a seleção feita por membros do *USPCodeLab* manualmente (conforme descrito na Subseção 2.3), já conhecidas de edições anteriores do HackathonUSP. Serão testados diferentes **meta-parâmetros** dos modelos (e.g., quais as perguntas a serem consideradas na hora da montagem das florestas). Em seguida, será analisado qual seria o melhor conjunto de **parâmetros** para o treinamento dos modelos. Isso demandará

a criação de características (*features*) derivadas dos dados brutos, tais como as conexões entre times, coletada de forma não estruturada nos formulários.

- **Integração com o Hacknizer**
  Para integrar o algoritmo com o Hacknizer, será desenvolvido um componente que fará parte do serviço de **Seleção** da plataforma. Essa integração será feita via *Kafka*, ferramenta da Apache para *streaming* de dados. O componente também será conteinerizada via *Docker* e sua implantação poderá ser feita junto com os demais serviços do Hacknizer. A ferramentas *Kafka* foi escolhida porque todos os outros serviços do Hacknizer a utilizam (comunicando-se segundo um padrão de sincronização baseado em eventos [New15]), ao passo que *Python* com *Flask* se integrará com a parte envolvendo aprendizagem de máquina.

5 Plano de Trabalho

Os seguintes passos serão necessários para cumprir com os objetivos propostos:

- **Limpeza e organização dos dados a serem analisados (25 horas)**
  Nessa primeira fase da pesquisa, com previsão de durar por volta de um mês e meio, será feita a coleta e agregação do conjunto de dados utilizado para a pesquisa, além da normalização desses dados e a geração de visualizações para primeiros *insights*. Esses dados serão coletados dos formulários de inscrições das cinco edições anteriores do evento, contendo dados de um total de 1200 interessados.

- **Estudo de diferentes técnicas de aprendizado de máquina (15 horas)**
  Para entender melhor os modelos que podemos utilizar no desenvolvimento do algoritmo, será feito o estudo de modelos e técnicas de validação para aprendizado de máquina. Em particular, focaremos em testar modelos simples tais como: Regressão Linear, Regressão Logística, Árvores de Decisão e Florestas Aleatórias

- **Desenvolvimento do algoritmo (30 horas)**
  Na segunda fase da pesquisa, serão definidos os critérios preliminares de decisão para o algoritmo de seleção, buscando decidir os **meta-parâmetros** e criar os **parâmetros derivados** para construir e treinar os modelos.

- **Desenvolvimento do serviço (25 horas)**
  Após definir o algoritmo a ser utilizado, será desenvolvido um novo componente do serviço de seleção do Hacknizer, utilizando Flask para criar um componente conectado via Kafka que receberá uma lista de inscritos para uma edição de um determinado hackathon e, aplicando o algoritmo desenvolvido, poderá gerar uma lista sugerindo os participantes para tal hackathon.

- **Feedback (5 horas)**
  O HackathonUSP 2018.2, segunda edição da série neste ano, irá ocorrer nos dias 9 e 10 de novembro. O serviço desenvolvido será utilizado para obter uma sugestão dos participantes a serem chamados para o evento. Com isso, será possível obter feedback do algoritmo criado, tanto com os participantes quanto com os organizadores após o evento. Além disso, uma validação imediata será a comparação direta entre a sugestão gerada pelo algoritmo contra a seleção feita manualmente por outros membros do *USPCodeLab*, verificando o quanto o algoritmo de fato pode ser utilizado para automatizar a seleção de participantes no HackathonUSP sem prejudicar os objetivos previstos pela organização do evento. Com esses resultados, será elaborado o relatório final da pesquisa para este semestre.

6 Método de acompanhamento

Na página https://victorhmp.github.io/mac0215-blog/ divulgarei semanalmente as discussões realizadas com meus supervisores e os avanços na pesquisa. Além da página, o código sendo desenvolvido estará no repositório do Hacknizer no GitLab: https://gitlab.com/uspcodelab/projects/hacknizer contento todo o código sendo desenvolvido.

Meus supervisores serão: Professor Alfredo Goldman, que pode ser contatado pelo email: gold@ime.usp.br, e o aluno de mestrado Renato Cordeiro Ferreira, que pode ser contatado pelo email: renatocf@gmail.com.

Referências

[FA90] Karen A. Frenkel and Karen A. Women and computing. *Communications of the ACM*, 33(11):34–46, 11 1990. 2

[Kay17] Christina Kayastha. Enabling innovation through community and competition. In *2017 IEEE Women in Engineering (WIE) Forum USA East*, pages 1–4. IEEE, 11 2017. 2

[KPR+15] Marko Komssi, Danielle Pichlis, Mikko Raatikainen, Klas Kindstrom, and Janne Jarvinen. What are Hackathons for? *IEEE Software*, 32(5):60–67, 9 2015. 1

[New15] S. Newman. *Building Microservices: Designing Fine-Grained Systems*. O'Reilly Media, 2015. 3, 5