FACULDADE ANGLO-AMERICANO - FAA
CIÊNCIA DA COMPUTAÇÃO
TRABALHO DE CONCLUSÃO DE CURSO

PROPOSTA DE TRABALHO DE CONCLUSÃO DE CURSO

PROPOSTA DE ANÁLISE DE DADOS GOVERNAMENTAIS UTILIZANDO A LINGUAGEM DE PROGRAMAÇÃO PYTHON E SUAS BIBLIOTECAS

THIAGO MEDEIROS DE SOUZA

ORIENTADOR:
CO-ORIENTADOR: JOÃO PAULO DE LIMA BARBOSA

FOZ DO IGUAÇU, 2015

1. Área e linha de pesquisa

Área de Pesquisa: Metodologia e Técnicas da Computação
Código: 1.03.03.00-6
Linha de Pesquisa: Sistemas de Informação
Código: 1.03.03.04-9

2. Introdução e Justificativa

Negócios estão relacionados a dados, e a cada segundo que passa, mais dados são gerados. Certamente, é preciso que exista algum modo em que esses dados possam ter um uso. Devido a complexidade e a velocidade como esses dados são produzidos, a tarefa de analisar essas informações se torna difícil por ser muito vaga. Vaga, pois não existe um problema em específico que precise de uma solução. Vaga também, porque não existe uma questão em particular que precise de uma resposta. O que é comum a todos é o propósito geral: melhorar o negócio. E tudo o que se têm são os dados.

Quando se possui um arquivo com 50 gigabytes automaticamente indica muita informação, porém não é claro o quê de informação esse arquivo carrega. Portanto o primeiro passo é dar uma olhada.

Dar uma olhada, significa literalmente olhar, ou seja, plotar os dados de diferentes maneiras e formas, olhar para os gráficos, planilhas. Neste momento novas informações aparecerão, o modo como os dados estão distribuídos, ou a maneira em que cada quantidade de dados variam com outros montes de dados, ou uma outra porção de informações que se apresenta diferente do restante, ou até mesmo a falta desses dados.

Essas observações levarão a uma reflexão que resultará em possibilidades ou probabilidades concretas para se excercer uma atividade. No primeiro momento essas informações eram amorfas e agora se transformam em ideias. Para que essas ideias se tornem um trabalho futuro é preciso capturá-las e interpretá-las através de um modelo matemático. Esse modelo é uma descrição matemática de um sistema sobre estudo. Um modelo é muito mais que apenas a descrição dos

dados, modelo incorpora o entendimento de todo o processo do sistema que produziu os dados. Logo este modelo consegue fazer previsões sobre os conhecimentos análisados.[1]

Para conseguir fazer melhores previsões é preciso desenvolver métodos mais sofisticados antes de formular um modelo relevante. Com isso a dificuldade aumenta e então é necessário implementar um modelo computacional que conseguirá simular os possíveis resultados.

Dados é um termo deliberadamente vago que agrega várias formas comuns de dados, como por exemplo matrizes (vetores multidimensionais), tabelas ou planilhas aonde cada coluna pode ter um tipo diferente de informação (caracteres, numéricos, data, entre outros). Essas tabelas podem se relacionar através de colunas chaves apresentada no modelo relacional de Edgar Frank Codd.

Certamente que todos esses exemplos citados não demonstram a totalidade e nem toda a abordagem para a palavra dados. Não é sempre que grande percentual de um conjunto de dados pode ser transformados em uma forma estruturada aonde é possível ser analisados e modelados.

O ramo da ciência que estuda e analisa esses dados, e que tem por finalidade gerar ideias através das interpretações realizadas e com elas ser capaz de tomar decisões, é conhecido como Data Science. Joel Grus afirma *et al*[2] que Data Science é a intersecção entre habilidades computacionais, conhecimento matemático e estatísco e domínio sobre o assunto (Hacking skills, math and statistics knowledge, substantive expertise).

Cientistas de dados precisam visualizar os dados com o objetivo de produzir resultados claros e serem capazes de informar ao seus mantenedores sobre a situação atual e a qualquer momento. Este é o verdadeiro valor que um cientista nessa area precisa prover.

Portanto, Python é a ferramenta escolhida pela maioria desses profissionais. Essa escolha se deve, não somente a alta produtividade que a linguagem fornece, mas também por ela ser uma ferramenta comum a diferentes times e organizações. Python é uma linguagem de programação livre e multiplataforma, possui uma

[1] JANERT K. P. Data Analysis with Open Source Tools: O'Reilly, 2011.
[2] GRUS J. Data Science from Scratch - First Principles with Python: O'Reilly, 2015.

excelente documentação e está sobre cuidado de uma enorme comunidade aonde é possível obter ajuda e melhores soluções para problemas durante a codificação. Tem como grande vantagem a facilidade de aprendizado, porque foi desenvolvida para ser simples e descomplicada.

Python é uma linguagem interpretada, dinâmicamente tipada e com grande precisão e sintaxe eficiente. Tem grande popularidade para analisar dados devido ao enorme poder que suas bibliotecas principais possuem (NumPy, SciPy, pandas, matplotlib, IPython). A linguagem apresenta alta produtividade para prototipação, desenvolvimento de sistemas menores e reaproveitáveis.[3]

- **NumPy**: NumPy é o acrônimo para Numerical Python, é o pacote fundamental para computação científica em Python. Essa biblioteca provê a criação de matrizes de uma maneira eficiente e rápida, assim como a possibilidade de leitura e escrita nessas matrizes. Operações de álgebra linear, transformada de Fourier e a geração de números aleatórios. Possui também ferramentas para integração com as linguagens C, C++ e Fortran;
- **SciPy**: SciPy é uma coleção de pacotes científicos;
- **pandas**: pandas fornece uma estrutura de dados e funções designadas para trabalhar com dados estruturados em um modo prático, rápido e expressivo. É um dos ingredientes mais críticos que dá a habilidade ao Python para produzir um ambiente de análise de dados. Esta biblioteca consegue combinar uma alta performance de matrizes computacionais geradas pela NumPy com a flexibilidade de manipulação de dados através de planilhas e tabelas utilizando banco de dados relacionais (como por exemplo instruções em SQL);[4]
- **matplotlib**: É a biblioteca mais popular do Python para plotar gráficos e visualização de dados em 2 dimensões. Os desenhos que o matplotlib é capaz de criar, são também interativos; é possível aproximar (zoom in) o

[3] Why is Python a language of choice for data scientists? Quora. Disponível em: <http://www.quora.com/Why-is-Python-a-language-of-choice-for-data-scientists>. Acesso em 15 de agosto de 2015. Tradução pesoal.
[4] pandas. Pandas Documentation. Disponível em: <http://pandas.pydata.org/pandas-docs/stable/>. Acesso em 15 de agosto de 2015. Tradução pessoal.

gráfico, assim como girar usando a barra de ferramentas da janela;

- **IPython**: É o conjunto de ferramentas para padrões científicos. Provê um robusto e produtivo ambiente para a interação e exploração computacional. Aumenta o design do Python shell para acelerar a escrita, efetuar testes e debugs nos códigos Python.

É necessário citar algumas outras tecnologias que não serão abordadas neste projeto. A linguagem de programação R, Matlab e Octave que são aplicaçãos voltadas para cálculo numérico, e também os softwares que auxiliam na álgebra computacional como Mathematica e Sage. Todas estas são ferramentas alternativas ao Python.

Para que os dados se tornem uma informação útil é preciso saber como coletar, processar, modelar e visualizar. Portanto este estudo é tratado como uma area interdisciplinar que depende uma arquitetura de software apropriada, técnicas de processamento massivo de dados, algoritmos de redução de dimensionalidade, modelagem estatística e computacional, visualização de dados, entre outros.

3. Objetivo

Os objetivos que almeja-se com este projeto são:

- Análise e interpretação de dados governamentais;
- Desenvolvimento de algoritmos para a mineração de dados.

4. Descrição do plano de trabalho e cronograma de execução

As atividades a serem executadas no decorrer do projeto visando o exito do mesmo, estão listados a seguir:

- Estudo e Pesquisa: aquisição dos conhecimentos pertinentes e necessários para o desenvolvimento do projeto.
- Análise de Requisitos: levantamento dos requisitos do projeto.
- Geração de Documentação: desenvolvimento das documentações para

especificação do projeto.

- Implementação: desenvolvimento do códigos para a análise de dados.
- Testes: execução dos testes que irão garantir a qualidade das informações a serem geradas.
- Elaboração de Artigos: parte do tempo destinado ao projeto será para desenvolver artigos visando a publicação em eventos da área.
- Apresentação de resultados: etapas destinadas à apresentação dos resultados parciais e finais.

Semanas	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Estudo e Pesquisa	X	X	X	X	X	X	X	X	
Análise de Requisitos	X	X	X	X	X	X	X	X	
Geração de Documento	X	X	X	X	X	X	X	X	X
Implementação		X	X	X	X	X	X	X	X
Testes		X	X	X	X	X	X	X	X
Elaboração de Artigos					X			X	X
Apresentação de resultados					X				X

Tabela 1 - Cronograma de execução

5. Materiais e métodos

Para o desenvolvimento deste projeto, as seguintes ferramentas, softwares e hardwares, serão utilizados:

- Computadores do tipo PC ou Laptop;
- Qualquer sistema operacional;
- Editor de texto;

- Python 2.7;
- Base Científica do Python: NumPy, SciPy, matplotlib e IPython;
- Dependências do IPython: tornado e pyzmq;
- pandas (versão 0.8.2 ou superior);
- Terminal bash;
- Navegadores Google Chrome ou Mozilla Firefox, em suas últimas versões;
- Editor e compilador Latex.

O início do projeto se dará com o estudo intensivo e leitura dos livros que são citados como referência bibliográfica a fim de definir um escopo global. Na etapa de desenvolvimento do projeto será realizado testes em conjunto para validar os dados a serem analisados.

A análise de requisitos e documentações serão desenvolvidos em editores de texto para a organização e compilação final do trabalho em Latex.

6. Disciplina a serem utilizadas

- Estatística e Probabilidade;
- Álgebra Linear;
- Algoritmos e Estrutura de Dados II;
- Inteligência Artificial.

7. Forma de análise dos resultados

Para que uma análise dos dados seja considerada aceitável ou correta, o aplicativo deve conseguir gerar gráficos, planilhas ou até marcações em mapas geográficos.

8. Produtos a serem gerados

Com o desenvolvimento e evolução do projeto, espera-se gerar os seguintes produtos:

- Proposta de ferramentas e métodos para a análise de dados;
- Apresentação e interpretação dos dados analisados;
- Artigos técnicos sobre o tema.

9. Referências bibliográficas

MCKINNEY W. Python for Data Analysis: O'Reilly, 2013.

JANERT K. P. Data Analysis with Open Source Tools: O'Reilly, 2011.

GRUS J. Data Science from Scratch - First Principles with Python: O'Reilly, 2015.

Intro to Pandas Data Structures. Greg Reda. Disponível em: <http://www.gregreda.com/2013/10/26/intro-to-pandas-data-structures/>. Acesso em 15 de agosto de 2015.

pandas. **Pandas Documentation**. Disponível em: <http://pandas.pydata.org/pandas-docs/stable/>. Acesso em 15 de agosto de 2015.

A Modern Guide to Getting Started with Data Science and Python. While My MCMC Gently Samples. Disponível em: <http://twiecki.github.io/blog/2014/11/18/python-for-data-science/>. Acesso em 15 de agosto de 2015.

Python. **Documentação oficial do Python.** Disponível em: <https://www.python.org/doc/>. Acesso em 15 de agosto de 2015.

Why is Python a language of choice for data scientists? Quora. Disponível em: <http://www.quora.com/Why-is-Python-a-language-of-choice-for-data-scientists>. Acesso em 15 de agosto de 2015.

10. Sintese bibliográfica

MCKINNEY W. Python for Data Analysis: O'Reilly, 2013.

JANERT K. P. Data Analysis with Open Source Tools: O'Reilly, 2011.

GRUS J. Data Science from Scratch - First Principles with Python: O'Reilly, 2015.

JANSSENS J. Data Science at the Command Line - Facing the Future with Time-Tested Tools: O'Reilly, 2015.

RUSSEL A. M. Mining the Social Web - Data Mining Facebook, Twitter, Linkedin, Google+, Github and Mora: O'Reilly, 2014.

HETLAND L. M. Python Algorithms - Mastering Basic Algorithms in the Python Language 2nd ed.: Apress, 2014.

Thiago Medeiros de Souza
Aluno

Professor orientador

João Paulo de Lima Barbosa
Professor co-orientador

Foz do Iguaçu, 15 de agosto de 2015