**UNIVERSIDADE FEDERAL DO RIO DE JANEIRO**

**ESCOLA POLITÉCNICA**

**DEPARTAMENTO DE ENGENHARIA ELETRÔNICA E DE COMPUTAÇÃO**

**PROPOSTA DE TRABALHO DA MATERIA: BIG DATA**

Aluno: Juan Moliner Malaxechevarría

juanmolimalax@gmail.com

Orientador: Prof. Alexandre de Assis Bento Lima, D. Sc.

assis@cos.ufrj.br

**1. TÍTULO**

Estudo da cotização em bolsa das empresas do Dow Jones(USA) mediante análise de sentimentos de tweets.

**2. ÊNFASE:**

Tratamento de dados, big data

**3. TEMA**

O tema do trabalho é estudar o uso de ferramentas de tratamento de dados de grande escala, Big Data e a aplicação de elas ao mercado de valores usando como fonte principal para obtenção de dados publicações em Twitter.

**4. DELIMITAÇÃO**

O escopo do projeto será desenvolver um aplicativo com interface web que procure tweets relacionados com as diferentes firmas, e mediante um tratamento do linguagem natural apresente o sentimento general respeito a firma. O aplicativo será desenvolvido com Apache Spark e outras ferramentas logo indicadas de tal forma que poderia ser rodado em um cluster com um volumem de trabalho muito maior ao volumem com o qual será desenvolvido.

Durante o desenvolvimento poderão ser incluídas outras variáveis de relevância na evolução da cotização da firma (eg: cotização de ‘commodities’, valor da moeda (eg: $R-EUR),...). Caso seja viável dentro do tempo e dos recursos, poderia ser implementado um algoritmo que em função aos dados antes discutidos indique operações de compra-venda rentáveis (algorithmic trading), incluso baseado em Machine Learning.

**5. JUSTIFICATIVA**

O desenvolvimento das ferramentas de tratamento de dados recentemente tem possibilitado um processamento de dados a uma velocidade, volumem e variedade completamente impensáveis há uma década. Dentro de esse marco surgiu o ‘algorithmic trading’ que utiliza ferramentas de Big Data para desenvolver algoritmos capazes de predizer a cotização em bolsa. Em 2006, uma terceira parte das operações nos mercados Europeios e Americanos foi feita por programas o algoritmos automatizados.

Este trabalho pretende assim tratar as 3 ‘Vs’ do denominado Big Data dentro de um ambiente no qual estas ferramentas tem se tornado cruciais. Aquelas três ‘Vs’ são:

* Velocidade: o aplicativo tratará um stream de tweets em tempo real o quasi-real
* Variedade: o conteúdo dos tweets não tem limites em termos de linguagem.
* Volumem: a quantidade de tweets disponíveis é muito alta

**7. FERRAMENTAS**

* Spark 2.1.0
* Hadoop 2.7.3
* Microsoft Azure
* Node.js

Rio de Janeiro, 1 de Maio de 2017

|  |  |
| --- | --- |
| \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Juan Moliner Malaxechevarria  Aluno | \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Alexandre de Assis Bento Lima  Orientador |