**Análise semântica de comentários**

A análise de padrões numéricos, como todas as análises desenvolvidas nesse projeto, se baseia no **princípio de ondas de Elliot,** que afirma que os fenômenos sociais e a especulação financeira são cíclicos.

Nessa análise utilizamos um motor de aprendizado de máquina para treinar um classificador capaz de avaliar se as mensagens publicadas em uma rede social implicam uma tendência de aumento ou de queda no preço da Bitcoin.

Nesse projeto analisamos publicações na rede social **Reddit**, na sessão específica de debates sobre a Bitcoin.

Segue abaixo o roteiro detalhado da análise

**Passo 1:** **Definimos um padrão de tratamento de texto.**

Nessa análise usaremos um classificador para classificar o conteúdo publicado na rede social.

Na implementação desse tipo de classificador, é comum que haja um tratamento no texto de entrada.

Nosso padrão de tratamento inclui:

* Separação de mensagens longas em sentenças menores, utilizando Spacy
* Remoção de alguns caracteres especiais e símbolos.
* Lematização
* E Remoção de Stop Words

Por exemplo, essa é uma mensagem anterior ao tratamento:

<Inserir mensagem anterior ao tratamento>

E essa é a mesma mensagem após o tratamento

<Inserir mensagem após o tratamento>

**Passo 2: Método de classificação**

A seguir, verificamos todas as datas em que temos mensagens registradas em nosso banco de dados.

E classificamos se a data é uma data Positiva ou Negativa para a nossa classificação.

A classificação será positiva para a data X se no dia seguinte a data X o valor de cotação da Bitcoin era pelo menos 5% maior.

A classificação será negativa para a data X se no dia seguinte a data X o valor de cotação da Bitcoin era pelo menos 5% maior.

Datas que não se encaixarem nos critérios acima serão ignoradas na classificação.

Segue a classificação de janeiro de 2020.

<Inserir tabelas com o seguinte: Data de janeiro; Quantidade de mensagens; Cotação no dia; Cotação no dia seguinte; Classificação>

A classificação de uma mensagem será igual a classificação da data da mensagem, pelo método apresentado acima.

**Passo 3: Treinamento do Classificador e avaliação de sua precisão**

Utilizamos a biblioteca skicit-learn para criar o nosso classificador.

O conjunto de dados usado no treinamento do classificador será chamado de dataset.

Dividimos de forma aleatória o nosso dataset em dois datasets menores.

O primeiro dataset terá 90% dos dados. E será usado para treinar o classificador.

O segundo dataset terá apenas 10% dos dados, e sua classificação será omitida.

Usamos o nosso classificador para classificar o segundo dataset.

Então, comparamos o resultado fornecido pelo classificador com a classificação real do segundo dataset, que é conhecida.

Essa comparação nos fornece cinco métricas que estão presentes na página principal

* Verdadeiros Positivos (VP)
* Falsos positivos (FP)
* Verdadeiros Negativos (VN)
* Falsos Negativos (FN)
* Precisão

Onde a Precisão é obtida pela seguinte equação:

Atualmente, as métricas do nosso classificador são:

* VP: <Inserir VP>
* FP: <Inserir VN>
* VN: <Inserir VN>
* FN: <Inserir FN>
* Precisão: <Inserir Precisão>

**Passo 4: Aplicação do classificador e obtenção do resultado**

Por fim, listamos e tratamos as mensagens do dia atual com o mesmo processo definido no passo 1.

Segue abaixo uma amostra com 30 sentenças obtidas a partir das mensagens do dia de hoje:

<Listagem de mensagens do dia de hoje, depois de tratar>

Enfim, aplicamos essas mensagens ao nosso classificador, e obtemos o seguinte resultado:

<Inserir resultado>

A análise consolidada, disponível no dahsboard, terá aumento de 5% a análise por repetição de comentário e a análise semântica tiverem resultado positivo.

Por analogia, A análise consolidada terá queda de 5% se ambas as análise por repetição de comentário e a análise semântica tiverem resultado negativo.