Gabriela Cavalcanti de Albuquerque Ayres

Análise de sentimentos de clientes bancários

em rede social Twitter

Trabalho de Conclusão de Curso apresentado na

Escola de Economia de São Paulo

da Fundação Getúlio Vargas, para obtenção do

certificado de conclusão de curso de especialização

em Business Analytics e Big Data

São Paulo, Novembro de 2018

Gabriela Cavalcanti de Albuquerque Ayres

Análise de sentimentos de clientes bancários

em rede social Twitter

|  |  |  |
| --- | --- | --- |
| **Orientador** | **Avaliação** | |
|  |  | |
| Prof. Gustavo Mirapalheta | Satisfatório | Não Satisfatório |
|  | □ | □ |

**Resumo**

Este estudo teve como objetivo de conhecer e explorar o nível de satisfação dos clientes em relação aos produtos oferecidos pelo banco Next através de rede social Twitter. O público em geral do banco é um público jovem e voltado para millenials. Em geral é um público que gosta de receber “mimos” e não há interesse em pagamento de taxas para serviços. Temos uma hipótese que os clientes do Next querem primeiramente um cartão de crédito “Na faixa” e que tentam conseguir o cartão de crédito no Nubank e caso sejam rejeitados acabam por procurar outras opções disponíveis no mercado e acabam vindo para o Next. Eles não sabem que o Next é um banco digital até baixarem o app e começarem a fazer uso da conta e outros recursos.

Para ajudar nessa análise, foram lidos os tweets dos clientes pelo @falanext e criadas nuvens de palavras para identificar o que os clientes falam a respeito do Next e se quando falam, outros bancos também são mencionados. Foi utilizada a técnica de análise de sentimentos para identificar os sentimentos dos clientes ao falarem sobre o Next para saber se estão satisfeitos ou não com o banco e se poderiam ajudar a fazer recomendações para outros colegas e amigos.

Os principais resultados confirmaram que o uso da nuvem de palavras se mostrou útil e alguns tweets realmente mencionam outros bancos, mas muitos agradecem quando recebem o cartão e alguns mimos, como descontos em Uber e cinema. Pudemos perceber que os clientes em sua grande maioria estão satisfeitos com os serviços, pois as palavras utilizadas nos tweets possuem maior característica positiva.

Esse trabalho pode ser utilizado para direcionar melhor os produtos e serviços dos bancos para os clientes e ajudar a evitar a evasão de clientes positivos, tentando identificar o que falam a respeito do banco e tomar ações preventivas.

**Palavras chave**: Estratégia de oferta de produtos e serviços, algoritmos de classificação, text mining, análise de sentimentos, nuvem de palavras.

Índice

[Introdução 6](#_Toc529814311)

[Objetivo 8](#_Toc529814312)

[Dados Utilizados 9](#_Toc529814313)

[Abordagem Metodológica 10](#_Toc529814314)

[Referências 16](#_Toc529814315)

[Próximos Passos 16](#_Toc529814316)

[Requisitos de pacotes 17](#_Toc529814317)

Índice de Figuras

[Figura 1 - Visão de Arquitetura da Técnica Proposta 11](#_Toc529814270)

[Figura 2 - Extração de dados do Twitter 11](#_Toc529814271)

[Figura 3 - Tweets após a preparação e limpeza de dados 12](#_Toc529814272)

[Figura 4 - Nuvem de Palavras de Tweets 12](#_Toc529814273)

[Figura 5 - Palavras mais utilizadas nos tweets 13](#_Toc529814274)

[Figura 6 - Lista de tweets após a classificação 13](#_Toc529814275)

[Figura 7 - Percentual de classificação dos tweets 14](#_Toc529814276)

[Figura 8 - WordCloud de tweets classificados como positivos 14](#_Toc529814277)

[Figura 9 - Palavras mais utilizadas nos tweets positivos 15](#_Toc529814278)

[Figura 10 - WordCloud de tweets classificados como negativos 15](#_Toc529814279)

[Figura 11 Palavras mais utilizadas nos tweets negativos 16](#_Toc529814280)

# Introdução

O uso da internet passou a fazer parte da vida das pessoas. Nas instituições financeiras, as pessoas não querem mais ir ao banco, elas desejam resolver tudo de forma simples, rápida e fácil pela internet, sem muitas burocracias. Os dados estão sendo coletados e armazenados em um ritmo dramático, sendo a maioria dos dados em formato semi-estruturado, e muitos podem conter informações úteis. Cada vez mais as pessoas tendem a compartilhar suas experiências sobre produtos e serviços prestados e fazerem recomendações. Eles usam as redes sociais como Facebook, Twitter, Whattsup, entre outros para expressar seus sentimentos, compartilhar suas experiências e logo perceberam que suas reclamações são rapidamente resolvidas quando expostas nas redes sociais. Para um melhor atendimento ao cliente e a experiência dos serviços, tornou-se uma prática comum revisar os comentários dos clientes nas redes sociais. Entender os comentários dos clientes sobre os produtos e serviços é muito útil tanto para a instituição financeira quanto para as pessoas que desejam se tornar clientes. Ler todos os comentários um por um não é tão eficiente quando o número de comentários é muito grande, uma vez que o conteúdo dos comentários também pode trazer confusões. Algumas análises podem conter frases longas, gírias e expressões idiomáticas, por esse motivo, é bem difícil ler e entender o significado dos comentários.

Se alguém ler apenas alguns números de resenhas e chegar a uma decisão, a decisão poderá ser tendenciosa. Por causa dessas razões, uma melhor técnica de mineração de dados para avaliar os comentários dos clientes que estão em formato semi-estruturado é muito importante. Não apenas os comentários sobre o banco em que são clientes, mas também sobre outros bancos para extrair sua opinião verdadeira. Esse problema vem sendo estudado por muitos pesquisadores nos últimos anos. A área de pesquisa é chamada de mineração de opinião e análise de sentimentos. Existem duas tarefas principais desta área de pesquisa. Eles estão encontrando características dos produtos ou serviços que foram comentadas pelos clientes e decidindo se os comentários são positivos ou negativos. Ambas as tarefas são muito desafiadoras e diferentes pesquisas foram realizadas nesse aspecto. Embora ambas as tarefas sejam cobertas por várias abordagens de pesquisa, há algumas áreas a serem melhoradas. Alguns deles estão identificando verbos, frases verbais e algumas sentenças condicionais. A exploração do uso de “smileys” na avaliação de opiniões, melhorando as técnicas existentes são alguns dos trabalhos futuros que foram identificados. Emoções são nossos sentimentos e pensamentos subjetivos têm sido estudados em vários campos, pois estão intimamente relacionadas aos sentimentos. A força de um sentimento ou opinião é tipicamente ligada à intensidade de certas emoções. Nas mídias sociais, as pessoas costumam expressar suas emoções usando diferentes "smileys" e se tornou uma tendência hoje. Assim, usar tanto o léxico do sentimento como “smileys” juntos em um algoritmo para avaliar a opinião dos clientes nas redes sociais será mais bem-sucedida. Considerando todos os potenciais de sinergia de mineração de dados, o desenvolvimento de algoritmos de mineração de dados para a extração de opinião pode ser identificado como uma área de pesquisa realmente importante. Essa pesquisa foi conduzida utilizando a análise de opinião baseada em recursos das opiniões e comentários de clientes, utilizando o léxico em português Re-Li, mas acreditamos que poderá ser melhorado no futuro se adicionado o recurso de “smileys” e algumas expressões verbais e gírias, muito usadas nas redes sociais.

# Objetivo

A empresa alvo deste estudo iniciou suas atividades na metade de 2017, com o objetivo de atingir um público jovem, bancado pelo tradicional banco Bradesco. Inicialmente, o Next oferecia apenas planos pagos, ainda que com um período de testes gratuito. O banco lançou uma conta corrente sem tarifas e com cartão de crédito internacional. O Next é um dos bancos digitais mais vantajosos hoje em dia. Por padrão, a conta digital do Next oferece um cartão de débito, que também pode ter a função crédito caso o usuário seja aprovado em uma análise de crédito. Além disso, o Next tem conta poupança e limite de cheque especial compartilhado com empréstimos e cartão de crédito, que é disponibilizado na hora na conta corrente. Ainda que o Next não tenha agência própria, é possível ser atendido em agências do Bradesco e usar os caixas eletrônicos do banco tradicional para sacar dinheiro. Por fim, correntistas do Next têm direito a aplicar dinheiro em um fundo de investimento de renda fixa com taxa de administração anual de 1% que tem títulos atrelados à [taxa Selic](https://tecnoblog.net/233808/nuconta-como-funciona-vale-a-pena/).

A conta digital do Next é gerenciada pelo aplicativo para iPhone e [Android](https://tecnoblog.net/curso/formacao/desenvolvedor-android/) e tem alguns diferenciais interessantes:

* **Flow**: um recurso para controlar como você gasta o dinheiro no Next;
* **Vaquinha**: para quem quer juntar dinheiro com amigos para um objetivo em específico;
* **Objetivo**: é possível definir quanto se quer economizar e o dinheiro é aplicado no fundo de investimento;
* **Mimos**: alguns descontos que vêm com cartões de crédito do Next, como R$ 20 de desconto no [Uber](https://tecnoblog.net/cupom/desconto-uber), 50% de desconto no ingresso do Cinemark, [entre outros](https://next.me/mimos).

O Next possui outros concorrentes como o banco Inter, o Nubank e o Original e cada vez é mais importante entender as características do seu público alvo e suas necessidades, para um atendimento de qualidade e garantia de satisfação, assim como evitar a evasão de clientes positivos.

O público do Next é um público jovem que usa muito as redes sociais e normalmente se utiliza dos recursos digitais para se expressar. Portanto é preciso entender o feedback desses clientes e responder as seguintes perguntas:

* O que trouxe o cliente para o Next? Algum produto específico?
* O que mantêm o cliente no Next?
* O que o cliente mais gosta e menos gosta do Next?

Com essas respostas, o banco será capaz de direcionar melhores ofertas e oferecer produtos e serviços de forma diferenciada, conseguir atrair novos clientes e manter sua carteira de clientes de forma crescente.

# Dados Utilizados

Foi criada uma conta de desenvolvedor no [Twitter](https://apps.twitter.com/) para possibilitar a coleta dos comentários de clientes através da API Search utilizando a string @falanext.

Após a criação da conta, foi necessário criar alguns dados para viabilizar a autenticação do usuário:

* Consumer Key (API Key)
* Consumer Secret (API Secret)
* Access Token
* Access Token Secret

Um conjunto de tweets foi recuperado de um período até 7 dias, tendo uma média de 500 tweets no total a cada extração realizada.

Os tweets normalmente possuem uma grande quantidade de caracteres especiais contidos em suas mensagens, tornando necessário um processo de limpeza dos dados, antes de iniciar sua análise e interpretação.

# Abordagem Metodológica

Esse trabalho foi realizado em linguagem de programação Python, devido ser esta a linguagem adotada pelo banco Bradesco para projetos voltados a Big Data e Analytics.

A técnica proposta desta pesquisa é uma técnica de mineração de dados baseada em classificação e para que a pesquisa pudesse ser adequadamente realizada e evitar o overfit, a amostra de 500 tweets foi dividida em learn e test, sendo que a amostra de test foi definida em 0,3 ou 30% do total de tweets. Foram realizados testes com os algoritmos de classificação: Random Forest, Naive Bayes, Boosting e Dummy e posteriormente avaliamos qual dos algoritmos teve um melhor nível de acurácia para então iniciar o processo de classificação dos tweets. Durante o desenvolvimento, obtivemos os seguintes níveis de acurácia para os algoritmos utilizados:

* Random Forest: 0,7
* Naive Bayes: 0,63
* Boosting: 0,66
* Dummy: 0,55

A classificação tem o objetivo de criar categorias positivas e negativas para o conjunto de tweets que foi coletado. Para realizar a classificação foi utilizado o léxico Re-Li [1] em português que separa as palavras por:

* Adjetivos (positivos e negativos)
* Expressões idiomáticas (positivos e negativos)
* Substantivos (positivos e negativos)
* Verbos (positivos e negativos)

Além disso, foram criadas nuvens de palavras para identificar quais as palavras que foram mais mencionadas nos tweets com o objetivo de identificar os produtos que foram mais mencionados ou as maiores satisfações e também os principais problemas com reclamações dos clientes.

Para que as técnicas pudessem ser aplicadas foi necessária a realização de um processo de preparação e limpeza dos dados que consistiu em:

* Retirada de expressões regulares (regex)
* Decodificação HTML
* Retirada de caracteres especiais (UTF-8)
* Retirada de menções a empresas ou clientes com @nome
* Retirada de links URL
* Retirada de hashtags

Após a preparação e limpeza dos dados foram retiradas palavras consideradas “StopWords” que consistem em artigos, preposições e algumas palavras que não agregam valor na pesquisa e análise que será feita.

A Figura 1 fornece a visão geral da arquitetura de nosso sistema de análise dos tweets.

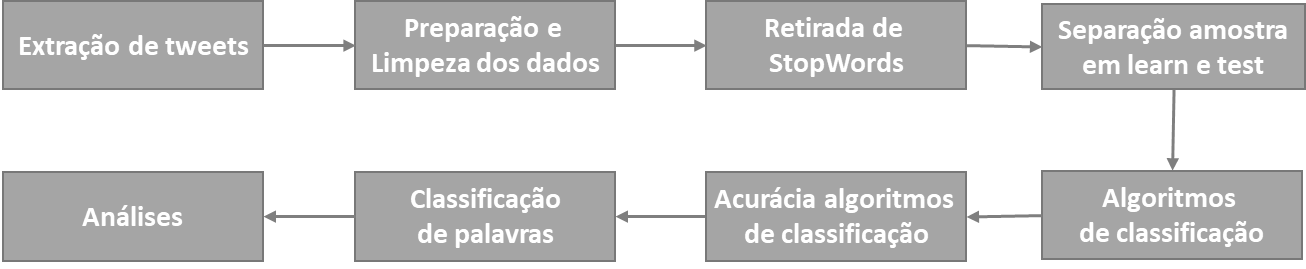


Figura 1 - Visão de Arquitetura da Técnica Proposta

A Figura 2 fornece um exemplo da extração dos tweets de forma bruta ainda sem os tratamentos de limpeza dos dados aplicados.



Figura 2 - Extração de dados do Twitter

A Figura 3 demonstra o resultado dos tweets após a realização da preparação e limpeza dos dados.

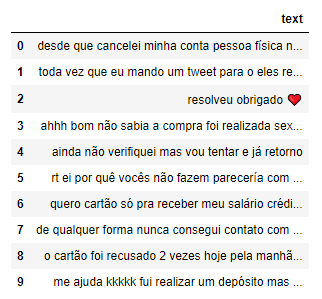


Figura 3 - Tweets após a preparação e limpeza de dados

Para retirada das StopWords foi utilizado o pacote nltk em português.



A Figura 4 é a nuvem de palavras após a extração de uma média de 500 tweets do período de 7 dias.

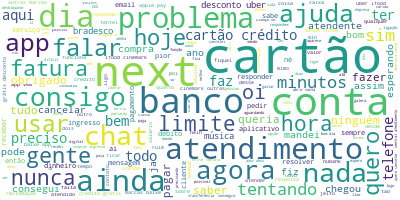


Figura 4 - Nuvem de Palavras de Tweets

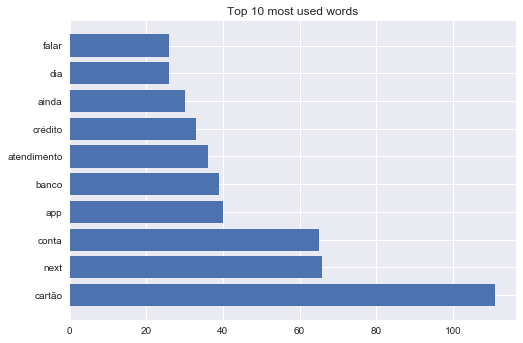


Figura - Palavras mais utilizadas nos tweets

Foi utilizada a biblioteca sklearn para os algoritmos de classificação: Random Forest, Naive Bayes, Boosting e Dummy.

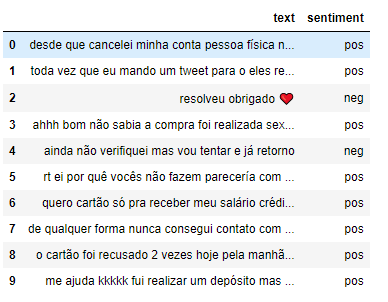


Figura 6 - Lista de tweets após a classificação

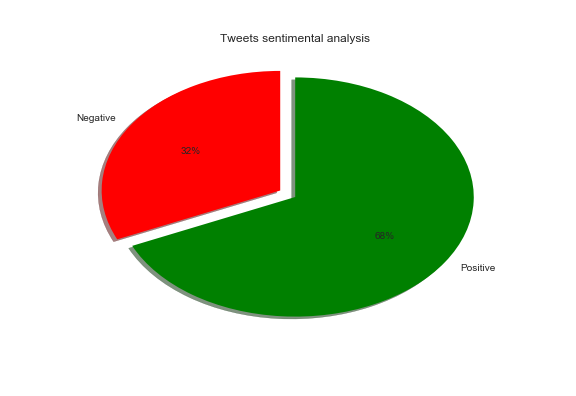


Figura 7 - Percentual de classificação dos tweets



Figura 8 - WordCloud de tweets classificados como positivos

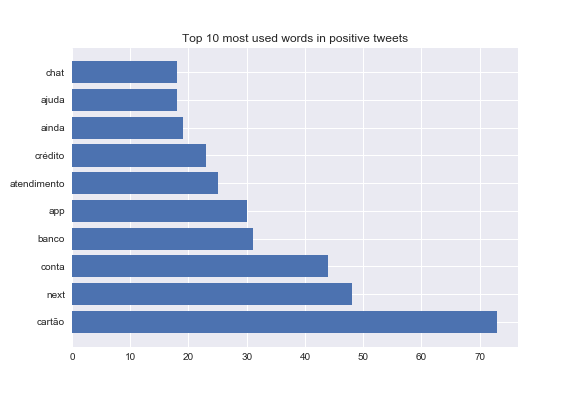


Figura 9 - Palavras mais utilizadas nos tweets positivos

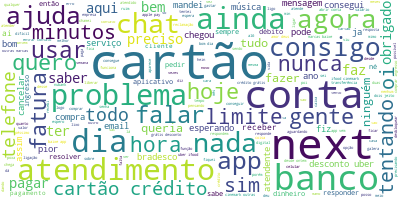


Figura 10 - WordCloud de tweets classificados como negativos

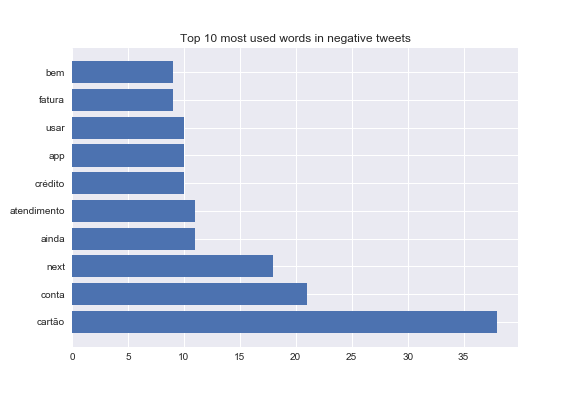


Figura 11 Palavras mais utilizadas nos tweets negativos

# Referências

**Re-Li - Léxico em português**

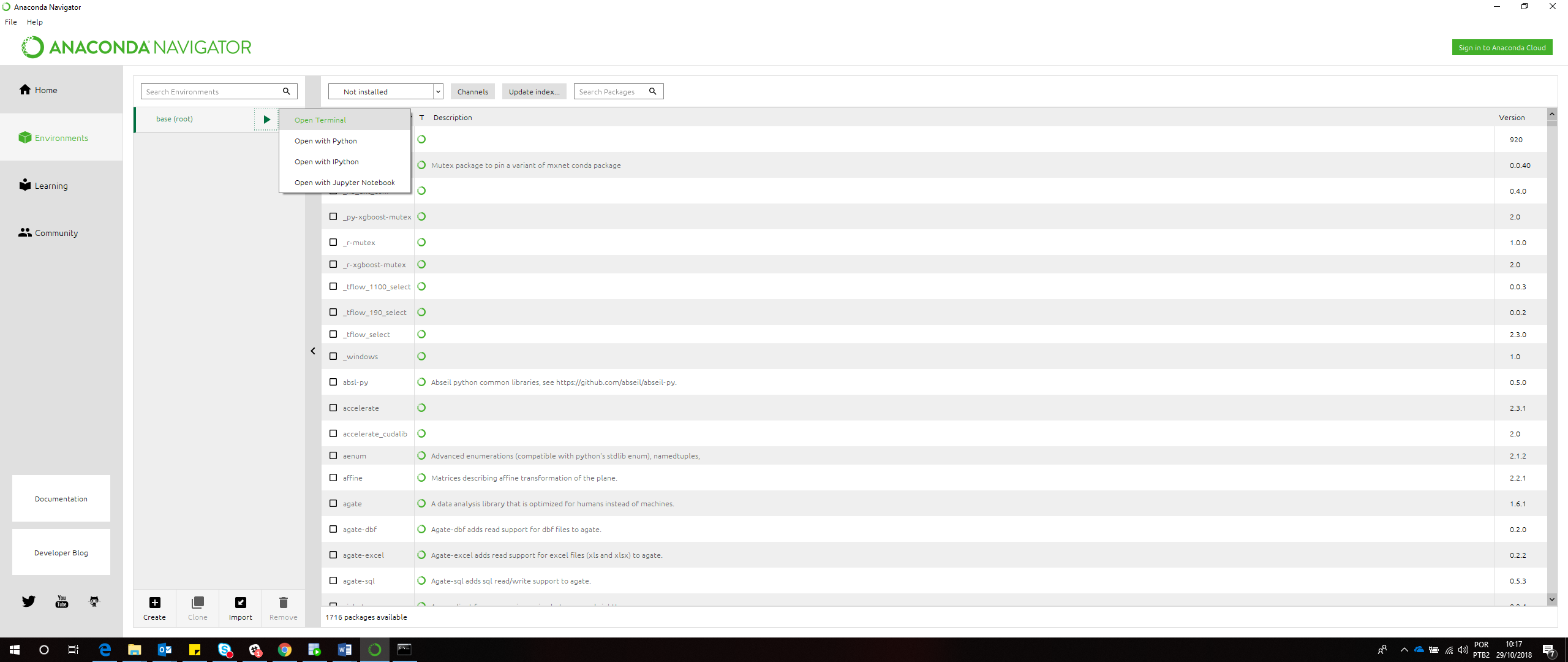
<https://www.linguateca.pt/Repositorio/ReLi/>

# Próximos Passos

# Anexos

## Requisitos de pacotes

**Instalação dos pacotes Python**



**Textblob:**

conda install -c conda-forge textblob

**Tweepy:**

conda install -c conda-forge tweepy

**NLTK:**

conda install -c conda-forge nltk

**WordCloud:**

conda install -c conda-forge wordcloud

**HTMLParser:**

conda install -c conda-forge HTMLParser

**BeautifulSoup**

conda install -c conda-forge beautifulsoup