Sinteticamente, com o algoritmo a ser desenvolvido pretende-se armazenar as notícias relevantes sobre as estratégias de campanha dos candidatos do segundo turno das eleições ao governo do Estado de São Paulo de 2018 - João Dória (PSDM) e Márcio França (PSB) - durante o período das “corridas eleitorais” e as publicações feitas nos seus respectivos perfis do Twitter. Após a localização (*crawling* ou *spidering*), extração e limpeza de dados (*scraping*), armazená-los em uma base de dados. Tendo em vista tais objetivos da aplicação, a seguir estão listadas as principais ferramentas que serão utilizadas para concluí-los, assim como a função e justificativa de cada item.

Python 4.6

Python é uma linguagem de programação criada em 1990 multi-paradigma cujo uso abrange de pequenos projetos pessoais até softwares grandes como Youtube, Instagram, Spotify, etc. Devida sua rapidez e simplicidade, além do seu vasto acervo de bibliotecas e *API*'s de código aberto e gratuitas para uso, atualmente o Python é uma das principais linguagens de programação no ramo da computação científica (*Machine Learning*, *Big Data*, Inteligência Artificial, Ciência de dados).

MySQL 8

O MySQL é um sistema de gerenciamento de banco de dados gratuito que visa oferecer ao usuário uma forma intuitiva e dinâmica na hora de criar, editar, consultar e excluir bancos de dados e os dados gravados nos mesmos.

Urllib 3

O urllib é um biblioteca (*client*) da linguagem Python que propicia ao programador uma série de funções e atalhos que facilitam a localização e extração

de dados das páginas HTML e XML os tornando maleáveis e passíveis à manipulações, configurando o *Data Scraping*.

BeautifulSoup 4-4.8.0

BeautifulSoup é uma biblioteca disponibilizada gratuitamente para a linguagem Python cujo uso se baseia no “parseamento” de dados obtidos na internet. Ou seja, após a localização e extração de dados feita por algum *crawler* (no nosso caso, o urllib 3), ainda é preciso converter esses dados da *web* para dados maleáveis pelo algoritmo em si, dessa forma, podemos iniciar o tratamento, limpeza e alocamento das informações na base de dados.

NTKL e outras bibliotecas

Após a conversão dos dados *web* feita pelo BeautifulSoup, ainda é preciso submeter esse conjunto de dados a uma série de verificações e procedimentos para a normalização de informações e retirada de conteúdos irrelevantes e redundantes, como *tags* HTML/CSS e trechos de códigos javascript (comuns em praticamente qualquer site), e, em alguns casos, a extração de *stopwords* (palavras que não possuem sentido considerável sozinhas, como “a”, “de”, “mas”, “se”) e de radicais das palavras, quando se quer considerar apenas o sentido “original” dos termos (por exemplo, as palavras “governo”, “governabilidade” e “governança” seriam consideradas como uma única palavra. Isso é útil para se reduzir o tamanho na base de dados, porém, em contrapartida, perde-se variedade e especificidade nos resultados).

A biblioteca “NTKL” é responsável por auxiliar o programador a realizar os procedimentos supracitados. Além disso, também foram utilizadas as bibliotecas “RE” e “PYMYSQL” para outras normalizações e conexão da aplicação com o banco de dados, respectivamente.

Anaconda 3

A plataforma de código aberto Anaconda reúne uma série de softwares, bibliotecas e ferramentas úteis para a programação científica a fim de facilitar a integração entre os aplicativos e oferecer recursos exclusivos aos seus usuários. No caso do algoritmo a ser construído, utilizamos a *IDE* Spyder, responsável por reunir uma série de bibliotecas úteis para a linguagem Python (como as listadas anteriormente), simplificando e diminuindo a complexidade dos códigos.

Por fim, após o armazenamento dos resultados no banco de dados, pretende-se tirar conclusões a respeito da forma como os candidatos utilizaram seus perfis do Twitter para se comunicar com seus eleitores e identificar as temáticas mais abordadas de suas campanhas, utilizando para isso as postagens dos perfis dos candidatos na rede social do Twitter. Ademais, também se visa verificar as estratégias que cada candidato usou em sua campanha e se houve o uso de inteligência artificial para o disparo de publicações em redes sociais durante esse período.

Dado tal fato, levamos em consideração todos os *tweets* dos perfis de João Dória Jr. e Márcio França e todas as notícias vinculadas às estratégias desses candidatos nos sites de notícias (nome dos sites aqui) da data __/__/____ até a data __/__/____. Vale ressaltar que os próprios candidatos têm domínio sobre seus perfis do Twitter, logo, há a possibilidade de que nem todos os *tweets* que foram feitos durante o período levado em consideração ainda estejam visíveis para o público. De maneira similar, o mesmo se aplica às notícias, pois as mesmas podem ser alteradas e excluídas posteriormente por seus detentores.

Sintetizando, o objetivo final desta pesquisa é identificar o uso da inteligência artificial como parte da estratégia para disparo de postagens nas redes sociais dos candidatos. Para isso, o *crawler* já mencionado neste documento fará uma varredura nos sites de notícias com o propósito de identificar notícias, artigos e matérias que abordem a temática das táticas de campanha usadas por cada um dos políticos

também já mencionados. A princípio, os filtros/palavras-chave que o algoritmo considerará para determinar a relevância dos textos são: *Fake News*, Inteligência Artificial, eleições SP 2018, João Dória, Márcio França. Não necessariamente todas as palavras-chave terão que estar presentes no texto, mas é evidente que quanto maior a repetição dos termos, maior a relevância daquela notícia para o nosso banco de dados.

Já no caso das postagens dos perfis no Twitter de cada candidato, o algoritmo examinará quais os temas mais frequentes nos *posts* de cada perfil usando a mesma estratégia de contagem de repetição de termos, no entanto, não teremos (ou teremos?) uma lista pré-fixada de palavras para a comparação, como "educação", "segurança", "saúde". Dado tal fato, nosso algoritmo será capaz de descobrir por si só quais as palavras mais frequentes nos posts, excluindo, evidentemente, as chamadas *stopwords*. Com isso, espera-se que tenhamos uma base palpável de dados que nos ajudará a determinar se os *tweets* tiveram algum tipo de direcionamento temático, apelativo, de público-alvo e outros, possivelmente disparados por algoritmos de Inteligência Artificial.

A ser definido (pela orientadora ou em conjunto):

- Quais os sites de notícias a serem analisados pelo algoritmo;
- Intervalo de data das notícias que serão levadas em consideração para a análise;
- Intervalo de data dos tweets que serão levados em consideração para a análise;
- Determinar se devemos fazer uma lista pré-fixada com palavras-chave das temáticas a serem quantificadas (frequência) pelo crawler nos tweets (como "educação", "meio ambiente", "segurança") ou deixaremos o próprio algoritmo capturar as palavra-chave mais mencionadas nos tweets de cada candidato. OBS: No caso de escolhermos a primeira opção, devemos levar em consideração palavras associadas às temáticas também? Por exemplo, termos como "escola", "universidade", "ensino médio", e etc serão consideradas como se estivessem dentro do escopo "educação"?
- Definir quais as variáveis (atributos) do nosso dataset (banco de dados).