# Relatório

## Exercício 1

Utilizando o modelo de vetorização do TF-IDF e unigrama teve resultado ligeiramente menor que o modelo binário, como podemos ver pela acurácia, entretanto obteve uma revogação (*recall*) melhor que o obtido pelo algoritmo anterior.

## Exercício 2

Utilizando o modelo de vetorização do TF-IDF e bigrama teve resultado similar que o modelo binário, como podemos ver pela acurácia e as demais métricas, entretanto obteve uma revogação (*recall*) menor que o obtido pelo algoritmo anterior usando TF, mas obteve resultado melhor que utilizando TF-IDF e unigrama.

## Exercício 3

Foram testados 5 algoritmos de classificação com diferentes abordagens como árvore de decisão, floresta randômica, modelo bayesiano, máquina de suporte de máquina e k-vizinhos. Utilizando TF e bigrama, o melhor resultado foi o MultinomialNB seguido pelo SVC e plo RandomForest, no entanto, com alguns aprimoramentos nos algoritmos poderia ocorrer do SVC ser superior ao MultinomialNB. Os outros algoritmos por serem tecnicamentes mais simples obtiveram resultados muito dispares dos três melhores.

## Exercício 4

Os resultados encontrados foram ligeiramente menores que com stopwords, mas aumentou a detecção de valores considerados positivos.

## Exercício 5

O resultados encontrados foram bem piores, caindo cerca de 0.25 em acurácia em relação aos tweets da região de Minas Gerais. Por ser uma coleção de documentos diferentes com um contexto diferente, trouxe um comportamento diferente para analisar o vocabulário do conjunto da reforma da previdência, devido as vezes pela quantidade de palavras, o vocabulário formado, o tipo de encodificação, entre outros.