Ingrid Knochenhauer de Souza
ingridkno@gmail.com

## Introdução

A comunicação textual tem se mostrado cada vez mais importante no dia a dia das pessoas. Com mensagens geradas a todo minuto, se faz necessário a avaliação e classificação das mesmas em diversos campos: mensagens, e-mails, opiniões, tweets… Para isso, o campo da NLP (*Natural Language Processing*) transforma esse dado não-estruturado, que é o texto, em informações analíticas possíveis de serem quantificadas e classificadas. Por isso, tantos modelos estatísticos são trabalhados isoladamente ou em conjunto para diversos tipos de classificação.

Como uma forma de seleção de candidatos a pesquisador, a Senior propõe um desafio considerando uma base de mensagens classificada em "Comum" ou "Spam". Este desafio é composto, por uma primeira etapa, de levantamento de dados sobre essa base e, em uma segunda etapa, a geração de um modelo de classificação automática dessas mensagens. Para este desafio, o modelo de classificação para mensagens “spam” ou “comuns” foi baseado no método estatístico Multinomial de Naive Bayes. Neste método, as mensagens se tornam vetores multinomiais ($p_1$, …, $p_n$) onde $p_i$ é a probabilidade de ocorrência de cada palavra na mensagem. A associação desses vetores a uma classificação prévia permite treinar o modelo para classificar o texto baseado na combinação de probabilidades das palavras de uma mensagem. Portanto, a semelhança da redação de spams é quantificada e reconhecida por este método.

## Metodologia

A primeira etapa foi constituída na avaliação das mensagens classificadas. De início, foram quantificadas as palavras mais frequentes em toda a base de dados. Depois, foram levantados a quantidade de mensagens comuns e spams, bem como indicadores estatísticos do total de palavras para cada mês. Finalizando essa primeira etapa de avaliação, foi quantificado o dia de cada mês que possuía a maior sequência de mensagens comuns.

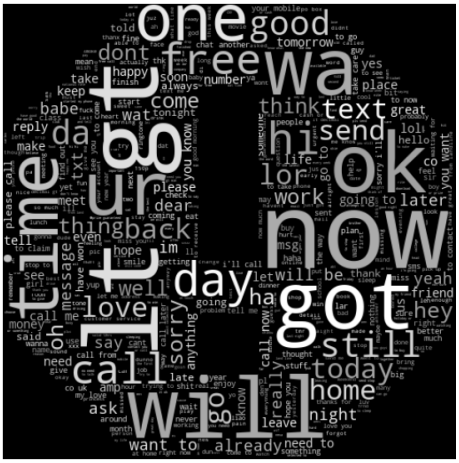
Na segunda etapa, foi aplicado o método Multinomial de Naive Bayes para classificação automática de mensagens em “Spams” ou “Comuns”. Para aplicação e avaliação do método, a base de mensagens foi dividida para treinamento do modelo de classificação (75% dos dados) e para teste de acuracidade do modelo (25% dos dados).

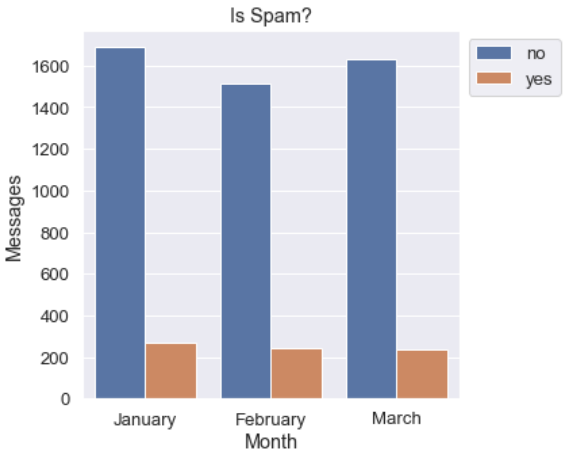
## Resultados

Os resultados são apresentados em duas seções: primeira e segunda etapa.

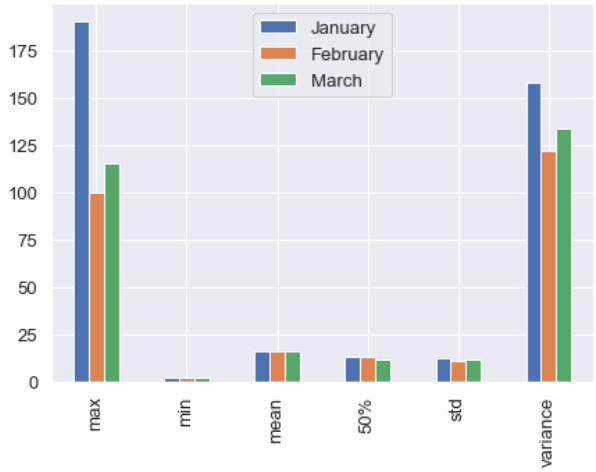
### Primeira Etapa

Os resultados de quantidade de palavras mais frequentes, mensagens “spams” e “comuns”, indicadores estatísticos mensais e dias com mais mensagens em cada mês são apresentados pelas figuras seguintes.




Statistical Indicators




**Dia do Mês com Mais Mensagens**

| Dia | Mês | Quant. Msgs |
|---|---|---|
| 1 | JAN | 69 |
| 13 | FEV | 72 |
| 8 | MAR | 69 |

As palavras mais frequentes, em destaque na primeira figura de nuvem de palavras, foram “now”, “will”, “ok” , “it”, “got”, “gt”. Todas estas são palavras comuns de uma comunicação representando tempo, concordância e termos comuns.

No gráfico “Is Spam?”, foi mostrado que há uma variação maior mensalmente na quantidade de mensagens “comuns”, podendo variar de 1700 a 1500 mensagens. Em relação às mensagens “Spams”, elas se mantiveram constantes na quantidade de 200 mensagens ao longo dos 3 meses de Janeiro, Fevereiro e Março.

Já quanto ao gráfico de Indicadores Estatísticos, os parâmetros de mínimo, média, mediana (50%) e desvio padrão de número total de palavras são muito similares quando comparados mensalmente. No entanto, o parâmetro de máximo e de variância são superiores no mês de Janeiro com uma mensagem beirando a 180 palavras comparadas a 100 e 110 palavras nos meses seguintes, respectivamente.
Os dias com mais mensagens para os meses de Janeiro a Março foram os dias 1º de janeiro, 13 de fevereiro e 8 de março com, respectivamente, 69, 72 e 69 mensagens. Apesar destes representarem dias festivos, como ano novo e dia da mulher, e também o dia anterior ao “valentine’s day” (dia dos namorados), as mensagens não tiveram assuntos relacionados a eles que justificassem sua maior quantidade nestes dias.

**Segunda etapa**

A segunda etapa consistiu em aplicar um método capaz de classificar automaticamente as mensagens como “comum” e “spam”. O método Multinomial Naive Bayes foi utilizado e obtido a acuracidade do modelo para as mensagens separando 75% do dado para treinamento e os outros 25% para teste. A acuracidade do modelo treinado no dado teste chegou a 96,3%, um resultado bem alto. Dentro do erro de 3,7% em que as mensagens foram erroneamente classificadas, o erro se dividiu em 68,6% de mensagens “spam” consideradas “comuns”, e 31,4% das mensagens “comuns” classificadas como “spam”.

**Conclusão**

Nas etapas de análise estatística descritiva e geração de modelo de classificação, os resultados permitiram tirar algumas conclusões. Primeiramente, existiram palavras mais frequentes na troca de mensagens. Constatou-se também que o número de mensagens “comuns” variou mensalmente, em contrapartida, o número de “spams” se manteve constante. A mensalidade no número de palavras, mostrou constantes os parâmetros estatísticos de mínimo, média, mediana e desvio padrão, mas variou no número máximo e variância de palavras. Além disso, apesar dos dias com mais mensagens serem ou estarem próximos de dias festivos, o conteúdo das mensagens não mostrou essa relação. Quanto a utilização do modelo Multinomial Naive Bayes, este se mostrou adequado para classificação de mensagens “comuns” e “spams” com 96,3% de acuracidade e dentro do erro de 3,7%, 68,6% de mensagens “spam” foram consideradas “comuns”.

**Referências**

https://www.datacamp.com/community/tutorials/text-analytics-beginners-nltk
https://en.wikipedia.org/wiki/Naive_Bayes_classifier#Multinomial_na%C3%AFve_Bayes