Análise e interpretação dos resultados

Campina Grande – PB
2019

SUMÁRIO

1 MOTIVAÇÃO

Diante da necessidade de usar o computador para ler e editar textos, foi preciso "ensinar" o computador a identificar textos em uma linguagem natural, e codificá-los para uma linguagem computacional. Mas como criar programas capazes de interpretar mensagens codificadas em linguagem natural e decifrá-las em uma linguagem de máquina? Com o passar dos anos, houve muitos avanços através de pesquisas na área de Inteligência Artificial (IA), o que possibilitou, a partir de uma de suas subáreas, uma forma para ensinar o computador a identificar palavras, que, no futuro, foi chamado de Processamento de Linguagem Natural, PLN.

O PLN busca soluções para questões computacionais, através de uma aprendizagem automática no processamento de linguagem e se dedica a propor e desenvolver modelos computacionais para a realização de tarefas que dependem da língua humana, escrita como objeto primário. Para isso, linguistas, cientistas da computação, buscam fundamentos em várias disciplinas: Filosofia da Linguagem, Psicologia, Lógica, Inteligência Artificial, Matemática, Ciência da Computação, Linguística Computacional e Linguística. Segundo Gariba [2005] et al. o PLN busca facilitar a interação do software com o usuário, o que possibilita além do melhor entendimento, conseguir exatamente o que se está procurando.

2 OBJETIVOS DO EXPERIMENTO

Analisar dois algoritmos de sumarização automática de documentos tendo em vista observar como cada um se comporta diante do tamanho de conjunto de palavras apresentado, os tipos de textos e seus vocabulários específicos, além das diversas estruturas textuais. Diante disso, perceber o desempenho de cada um deles, através do tempo de execução, a porcentagem de redução das palavras e se apresenta um resultado de um texto com coerência, coesão e com uma quantidade reduzida de erros semânticos e ortográficos.

O primeiro algoritmo a ser analisado é o algoritmo de Luhn, um dos trabalhos mais importantes na área de Processamento de Linguagem Natural, que consiste em métodos estatísticos para cálculo de frequência de palavras e sua distribuição no texto, como critério de significância, gerando palavras-chaves e abstracts.

O segundo algoritmo é denominado algoritmo Marques, proposto pelos criadores deste artigo, que implementa uma nova forma de sumarização automática textual usando uma biblioteca do Python chamada NLTK (*Natural Language Toolkit*), que, ao encontrar as sentenças e palavras mais importantes do texto, gera o resumo a partir delas.

A perspectiva considerada para esse experimento é a do usuário de segundo nível: aqueles que utilizam o serviço. Uma vez que é na visão deste, que se busca entender se o algoritmo cumpriu seu propósito. Baseado na comparação das duas técnicas o usuário pode escolher aquele que lhe proporciona melhor compreensão. O experimento inicialmente pode

ser realizado no laboratório através de um ambiente de desenvolvimento para simulação dos algoritmos e posteriormente no ambiente acadêmico e escolar com alunos.

Resumindo o objetivo:

Analisar dois algoritmos de sumarização automática: o Marques e o de Luhn,
Com a intenção de compará-los
com respeito a coerência dos textos
do ponto de vista do usuário final
no contexto para uso acadêmico e escolar.

3 PLANEJAMENTO EXPERIMENTAL

3.1 Seleção de Contexto

Um ambiente virtualizado em laboratório com variáveis controladas será simulado através de múltiplas entradas de textos para sumarização automática deles, a fim de investigar e avaliar o funcionamento dos algoritmos de Processamento da Linguagem Natural. Para tal é selecionando um grupo de pessoas, no ambiente acadêmico e escolar, para realização da leitura dos resumos provenientes dos mesmos e indicar aquele que apresenta melhor coerência pelo método comparativo, para assim serem computadas as preferências e, a partir disso, qualificá-los.

3.2 Definição das Hipóteses

Uma vez definidos o problema, os objetivos e o contexto, as hipóteses são estabelecidos e os valores de significância para os erros definidos.

Hipótese nula (H0): Não há diferença significativa entre os algoritmos Marques e Luhn em relação a coerência do texto.

Hipótese alternativa (H1): Hipótese alternativa (H1): Há diferença significativa entre os algoritmos de Luhn e o de Marques:

1. O algoritmo de Marques é melhor que o de Luhn.
2. O algoritmo de Luhn é melhor que o de Marques.

Os níveis de significância para erros são:

- · P(erro do tipo I): $\alpha = 0{,}01$

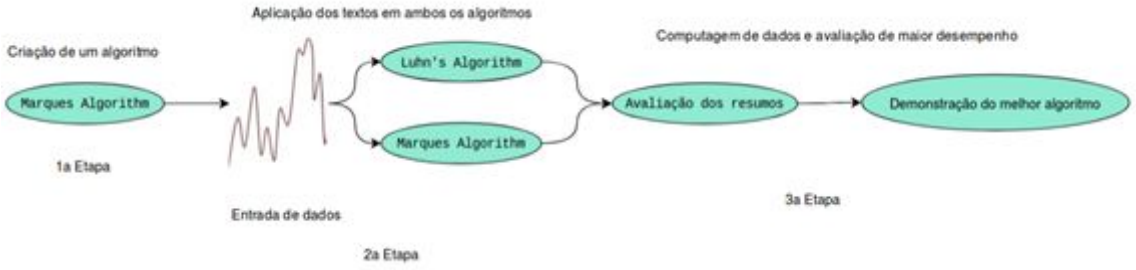
- · P(erro do tipo II): β = 0,05

3.3 Seleção dos Sujeitos

Os sujeitos do experimento são os algoritmos a serem comparados, Marques e Luhn, descritos na Seção 2.

3.4 Design de Experimentos

O experimento consiste nessas três etapas apresentadas na figura 1



A primeira etapa (Figura 1) consiste na criação de um algoritmo de sumarização automática.

A segunda etapa (Figura 1) têm como objetivo realizar a sumarização automática de 10 artigos de diferentes áreas de estudos através de dois algoritmos que utilizam métodos distintos.

Na terceira etapa (Figura 1), o conjunto de resumos resultantes da segunda etapa serão utilizados para fins avaliativos por um grupo de pessoas. E por conseguinte computar a avaliação de cada pessoa e estatisticamente demonstrar o algoritmo que realizou melhor desempenho.

O cenário a ser considerado para os experimentos é a simulação de resumos por uma máquina virtual, com capacidade de processamento e de memória RAM de 8 GB. E posteriormente a análise deles, dentro de um contexto educacional.

3.4.1 O Experimento da 1a Etapa

A primeira etapa consiste na criação de um algoritmo de sumarização textual, que inicia criando uma matriz do texto, separando-o em sentenças, e dentro das sentenças, em palavras. Depois disso, ele identifica palavras que possuem apenas significado sintático dentro da sentença, mas não é relevante para o sentido do texto, como “ou”, “e”, “para”, e as retira da matriz, pois, como essas palavras são frequentes, o algoritmo acabaria dando importância para as mesmas, e isso dificultaria a análise textual. Também retiramos as pontuações do texto, pois o algoritmo as trata como sendo uma palavra, e isso também atrapalha a sumarização.

Após essa limpeza de texto, o algoritmo cria uma distribuição de frequência para a matriz de palavras, para que seja descoberta quais são as mais importantes.

A partir dessa distribuição de frequência, o algoritmo seleciona, a partir de um valor inserido, o quão comum a palavra se apresenta no texto geral para posteriormente classificá-las como significantes ou não, no texto final.

3.4.2 O Experimento da 2a Etapa

A segunda etapa consiste em selecionar 10 artigos no google acadêmico, que incorporam diversas áreas do conhecimento e com complexidades diferentes, para realização de sumarização automática através do algoritmo de Luhn e o de algoritmo criado descrito na primeira etapa.

Foram realizados testes nas configurações dos parâmetros comuns modificáveis nos algoritmos, isto é, na quantidade de sentenças importantes, a fim de equipará-los para a etapa seguinte. Dessa forma, neste experimento foi utilizado o parâmetro de 8 sentenças importantes.

Segundo Pardo[1] para gerar um resumo eficiente deve ser extraído entre 20 a 50% do texto e 80% na taxa de compressão.

As variáveis independentes desta etapa são os textos inseridos e os algoritmos utilizados, enquanto as variáveis dependentes são os resumos gerados e a escolha da quantidade de sentenças importantes.

3.4.3 O Experimento da 3a Etapa

Com o intuito de avaliar a qualidade e coesão dos resumos gerados na segunda etapa, serão escolhidas 30 pessoas com diferentes graus de instrução acadêmica e será distribuído os dois resumos resultantes da inserção de um artigo em ambos os algoritmos para que um grupo de 3 pessoas que não possuem relação, escolhidas através de critérios como: conhecimento e preferência com o tema do texto, possam indicar o resumo que julgou possuir melhor coesão e coerência.

Nesse caso, os resumos tornam-se variáveis independentes enquanto a avaliação de acordo com as métricas de coerência e coesão podem ser consideradas as variáveis dependentes.

Em seguida, será criada uma tabela comparativa de acordo com o resultado da avaliação humana realizada para qualificar os dois algoritmos e assim mostrar aquele que proporciona uma melhor experiência aos leitores.

3.5 Instrumentação

A instrumentação dos experimentos é agrupada por hardware e software. Utilizaremos um computador (Intel Core i5-7200U Dual Core 2.5 GHz com Turbo Max até 3.1 GHz) para executar uma coleção de códigos-fonte escritos em Python 3 que implementam os algoritmos propostos de serem avaliados (Marques e LSA).

3.6 Avaliação da Validade

3.6.1 Validade de Conclusão

Os testes de hipóteses serão realizados considerando um nível de significância $\alpha = 0,1$, para assim dá suporte a teoria de modo a ser confiável. Diferentes tipos de textos serão aplicados aos dois algoritmos. Ambos foram implementados na mesma linguagem de programação. Entretanto, espera-se melhor resultado do algoritmo Marques devido sua natureza abstrata, uma vez que considera elementos adicionais em relação ao algoritmo de Luhn que possui métodos extrativos. Além disso, pode-se obter resultados distorcidos devido o fato de trabalhar com um conjunto amplo de palavras, da complexidade inerente a tentativa de aproximar-se de resumos humanos, além de possuir uma análise comparativa dependente de uma avaliação do grupo experimental passível de parcialidades e julgamento limitado ao conteúdo do texto e sua relação com ele.

3.6.2 Validade Interna

A realização dos experimentos não sofrerá interferência do ambiente de realização devido ao uso exclusivo de uma máquina virtual em que os testes consistem apenas em utilizar textos e testar o resultado da sumarização em algoritmos que representam métodos distintos. Pretende-se realizar os testes utilizando o mesmo hardware, reduzindo ameaças relativas à instrumentação, no mesmo ambiente de desenvolvimento, e em instantes próximos, para que a velocidade da internet interfira de forma mínima na análise da métrica relacionada ao tempo de execução. O tamanho do conjunto de dados, a estrutura na qual é organizada o texto, a morfologia e semântica das palavras são os fatores mais importantes que vão causar impacto no resultado.

3.6.3 Validade de Construção

A proposta apresentada na terceira etapa consiste em avaliar a qualidade dos resumos resultantes dos dois algoritmos apresentados, para tal é selecionado avaliadores humanos do âmbito acadêmico e escolar para através da leitura deles indicar qual apresenta o resultado mais coerente.

3.6.4 Validade Externa

O experimento é dependente da avaliação humana, logo o modo de abordagem para participar do experimento, o interesse daqueles que irão avaliar os resumos e seu entendimento do assunto abordado e dos termos utilizados, podem interferir no resultado.

4 ANÁLISE E INTERPRETAÇÃO DOS RESULTADOS

4.1 Estatística Descritiva

Como descrito na Seção 3.4, o tipo de projeto experimental consiste em uma amostra de 10 artigos científicos de diferentes áreas do conhecimento, que são inseridos nos dois algoritmos de sumarização propostos para análises. Dessa forma, resultando em 2 resumos referentes a cada artigo, que serão distribuídos para um grupo de 30 pessoas para avaliação. Para tal, deveria ser votado o texto que apresentava melhor compreensão de acordo com os quesitos coerência e coesão para o leitor e, caso houvesse uma preferência entre os dois, questionava-se o porquê da escolha. Vale ressaltar que não foi revelado os algoritmos

utilizados, logo o texto 1 em todos os casos representa o resumo proveniente do algoritmo de Marques enquanto o texto 2 o algoritmo de Luhn.

Como o objeto de pesquisa são os algoritmos, a tabela abaixo apresenta informações sobre os mesmos.

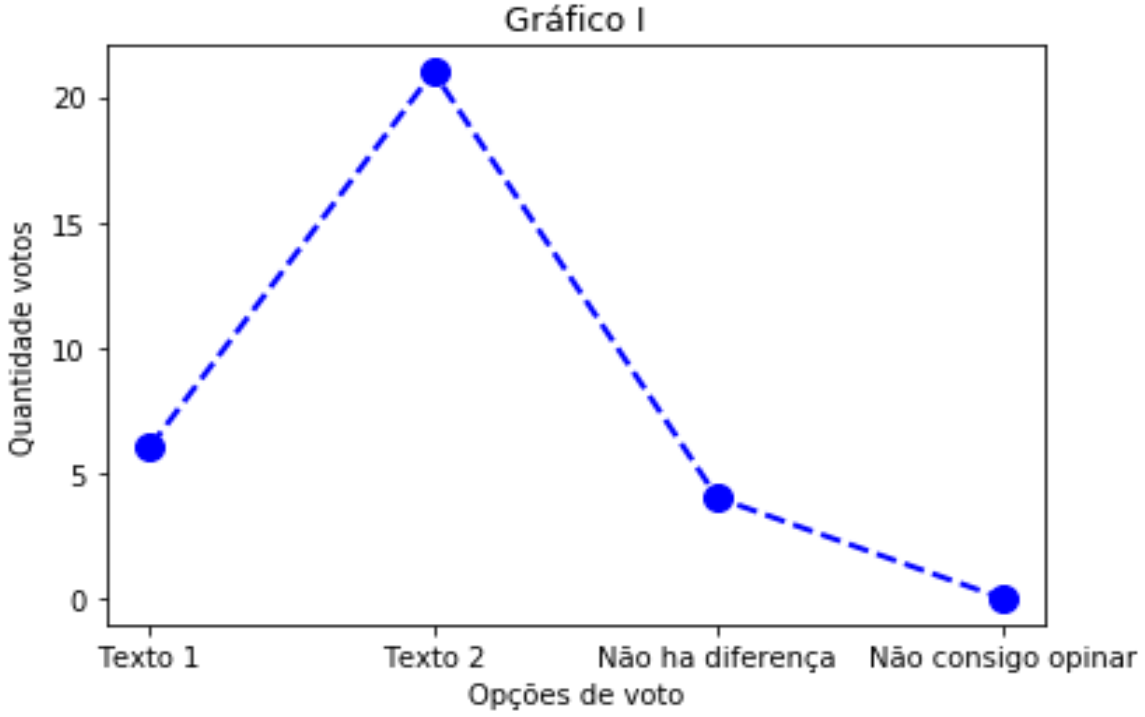
	Tempo de execução texto 1	Tempo de execução texto 2	Quantidade de parágrafos gerados
Teologia	6 segundos	7 segundos	7
Psicologia	4 segundos	4 segundos	5
Moda	5 segundos	4 segundos	6
Meio Ambiente	5 segundos	6 segundos	7
Medicina	5 segundos	4 segundos	7
Matemática	6 segundos	5 segundos	7
História	5 segundos	6 segundos	7
Esporte	6 segundos	5 segundos	8
Economia	7 segundos	6 segundos	6
Direito	7 segundos	8 segundos	8

4.2 Análise Quantitativa

A tabela 2 apresenta o resultado após o período de 15 dias da avaliação de um grupo de leitores em relação aos dois resumos. Pode-se observar as temáticas escolhidas para os artigos e quatro opções de votos como parâmetro avaliativo. Além disso, é importante destacar que a escolha de 3 pessoas para cada tema não foi realizada de modo aleatório, procurou-se encontrar aqueles que possuíam afinidade ou certa aptidão com o assunto tratado no texto. De acordo com os dados apresentados nessa tabela, percebe-se que o resumo gerado pelo algoritmo de Luhn é a preferência das pessoas selecionados.

	Texto 1	Texto 2	Não há diferença entre os textos	Não consigo opinar
Teologia	0	2	1	0
Psicologia	0	3	0	0
Moda	1	2	0	0
Meio Ambiente	1	2	0	0
Medicina	2	1	0	0
Matemática	1	1	1	0
História	0	3	0	0
Esporte	0	3	0	0
Economia	0	3	0	0
Direito	1	0	2	0

Tabela 2 - Quantidade de votos em relação a cada temática.



Graficamente, o somatório da quantidade de votos em cada algoritmo, em relação a amostra absoluta de 30 pessoas. Torna-se ainda mais evidente a escolha pelo algoritmo de Luhn.

4.3 Análise Qualitativa

Como descrito na Seção 4.1, ao fim do questionário é solicitado ao leitor justificar sua escolha. Entretanto, doze dos participantes selecionados preferiram não responder a esta questão. Adiante, segue as respostas coletadas.

Preferência pelo texto 1
“Gosto de coisas tradicionais, claras e simples, por isso escolhi o texto 1.”
“Melhor compreensão do conteúdo exposto.”
“Porque segue uma sequência mais clara de texto, sendo assim ele de forma mais direta e interligada.”
“O texto 1 aparenta possuir uma fluidez um pouco mais atraente para mim. A forma em que o conteúdo é trazido torna-o mais flexível ao entendimento.”

Preferência pelo texto 2
“Devido sua melhor organização e coesão textual.”
“O texto 2 apresenta uma coerência melhor entre os parágrafos, eles se complementam por assim dizer.O texto 1 entretanto, devido a inconsistências, se torna mais confuso. Vale ressaltar, talvez o problema seja do artigo, os dois textos possuem partes confusas dentro dos parágrafos ocasionadas por falta de pontuação.”
“Em se tratando de um exemplar do gênero resumo, considero o texto 2 mais claro e objetivo, com elementos do texto apresentados de maneira mais evidente, além do encadeamento das informações.”
“Melhor detalhamento.”
“Tem melhor coerência e entendimento.”
“Talvez por ter lido primeiro o texto 1, e ambos falarem sobre o mesmo tema, durante a leitura do texto 2 o assunto foi mais facilmente assimilado. O fato do texto 2 possuir uma linguagem não tão rebuscada e sem tantos referentes textuais quanto o primeiro também facilitou a compreensão durante a leitura do mesmo.”

“O segundo texto apresenta uma linguagem de melhor domínio do conhecimento do leitor. Ele define algumas palavras técnica de uma forma mais abrangente.”
“O texto, principalmente, deixou a impressão de ser mais coeso, mantendo-se em um assunto comum entre os primeiros parágrafos, apresentando uma evolução coerente no assunto, embora haja uma talvez perda dessa evolução no parágrafo final. Em consideração, o primeiro causou uma certa confusão sobre o tema tratado, variando de modo abrupto e não coeso entre os fatores do assunto e não deixando uma evolução agradável na leitura do texto.”
“Está mais claro. Compreendi melhor a metodologia.”
“Pois foi mais fácil de compreender e também o texto 2 foi mais direto, já o texto 1 ficou sem uma ideia central, abordou várias coisas sendo que não explicou por completo, logo o texto 2, expôs suas ideias de forma mais compreensível e abordou por completo suas ideias sem deixar nenhuma dúvida.”
“O texto é mais simples, ocasiona mais fácil entendimento.”
“O texto está mais didático para o leitor médio (leitor leigo).”
“Acredito que o ponto do texto 1 ainda vá chegar, estamos em uma transição mais puxada para o 2° texto.”
“Melhor estruturado. Menos informações excessivas.”

5 CONCLUSÕES

De acordo com os testes realizados, a hipótese nula foi rejeitada, isto é, as alternativas são significativamente diferentes. Sendo assim, o algoritmo de Luhn é melhor que o algoritmo de Marques em relação à melhor compreensão, coesão e coerência para experiência de 66,66% dos leitores selecionados. O resultado destoa do esperado inicialmente pelo viés dos pesquisadores. Como limitações, pode-se citar a dependência da avaliação humana, uma vez que não é controlável o grau de interesse e seriedade ao ler os textos propostos, o momento e o tempo que foi realizada a leitura, a ordem dos textos no questionário, além das perguntas. Uma proposta de trabalho futuro é mudar o método de abordagem avaliativa, procurando ao máximo evitar a influência de fatores externos nos resultados.

6 REFERÊNCIAS

- [1] Rino, Lucia Helena Machado; Pardo, Thiago Alexandre Salgueiro. A Sumarização Automática de Textos: Principais Características e Metodologias.NILC/Departamento de Computação. Universidade Federal de São Carlos, São Paulo. 2003.

- Luhn, H.P.The Automatic Creation of Literature Abstracts.IBM Journal. 1958.

- Sumy, Biblioteca contendo algoritmos de sumarização automática na linguágem Python, https://pypi.org/project/sumy

- Lima, Vinicius R. Utilizando processamento de linguagem natural para criar um sumarização automática de textos. 2017.