

Comparativo entre o algoritmo de Luhn e o algoritmo GistSumm para sumarização de Documentos

Eduardo Muller¹
Jones Granatyr¹
Otto Robert Lessing¹

Data submissão: 02.06.2014

Data Aceitação: 16.02.2015

Resumo. Este artigo descreve um comparativo entre dois algoritmos da área de mineração de textos, os quais são utilizados na tarefa de sumarização automática de documentos. Foram comparados nos experimentos o algoritmo clássico de Luhn e o algoritmo GistSumm, sendo realizadas dois tipos de avaliação, ambas utilizando o Português do Brasil como idioma alvo. A primeira consistiu em gerar um resumo de um texto fonte com cada algoritmo, e a avaliação foi conduzida utilizando avaliadores humanos que indicaram a coerência nos resumos de cada um. Por outro lado, a segunda foi conduzida por meio de uma avaliação baseada no resumo, no qual os avaliadores responderam perguntas sobre o texto original possuindo como fonte de consulta somente o resumo gerado pelos algoritmos. Após as análises, foi demonstrado que o algoritmo GistSumm possui maior capacidade para gerar resumos que mantenham a ideia principal do texto, sendo classificado com 81,6% de eficiência no primeiro experimento e 90% no segundo experimento.

Abstract. This article describes a comparison between two algorithms in the mining of texts, which are used in the task of automatic summarization of documents. Classic Luhn algorithm and the GistSumm algorithm were compared in the experiments, two types of evaluation being conducted, both using the Portuguese of Brazil as a target language. The first was to generate a summary of a source text with each algorithm, and the evaluation was conducted using human evaluators indicated that the coherence of the summaries of each. On the other hand, the second was conducted by means of an evaluation based summary, in which the evaluators asked about the original text as a reference source having only the summary generated by the algorithms. After analysis, it was shown that GistSumm algorithm has greater capacity to generate summaries to keep the main idea of the text, being rated 81.6% efficiency in the first experiment and 90% in the second experiment.

¹Grupo de Pesquisa de Inteligência Computacional. Universidade do Contestado – UnC. {muller.du@hotmail.com; jones@unc.br; otto@unc.br}

1. Introdução

A Inteligência Artificial é uma subárea da Ciência da Computação e nos últimos anos vem crescendo rapidamente, com o intuito de tornar uma máquina capaz de realizar tarefas que somente humanos até então conseguiriam realizar [10].

A Inteligência Artificial possui muitas linhas de pesquisa, e uma delas é a Mineração de Dados, na qual o principal objetivo é encontrar conhecimento em um montante de informação desconhecida. Dentro deste contexto encontra-se a extração de informações, que consiste em um conjunto de técnicas capazes de encontrar informações ou conhecimento em montantes de dados, muitas vezes desorganizados[3].

Seguindo essa linha de pesquisa existe a Mineração de Texto, que tem o objetivo de encontrar padrões e conhecimento apenas em dados textuais [10]. Para resolver a necessidade de trabalhar com esses textos, existe a sumarização automática de documentos, que é o principal objeto deste trabalho, consistindo em submeter um texto extenso a um algoritmo de sumarização, o qual irá retornar um resumo que contém as ideias principais do texto.

Neste contexto, este trabalho tem o intuito de analisar os resultados de dois algoritmos específicos para a tarefa de sumarização de documentos, ou seja, o algoritmo de Luhn e o algoritmo GistSumm. Para isso, foram feitos dois experimentos, no primeiro foi avaliada a capacidade de extração de frases que mantém as características e/ou ideias originais do texto original. Por outro lado, o segundo experimento consiste na avaliação dos resumos gerados pelos algoritmos, a fim de determinar se os resumos contém informações suficientes para responder perguntas baseadas no texto original.

Esse artigo na prática pode ajudar estudantes ou pesquisadores quando estão a procura de conhecer o assunto que algum documento trata, ou buscam apenas uma síntese de um texto extenso, e precisam realizarem esta tarefa de forma eficiente e rápida. Os testes feitos para esse artigo aplicam-se para o Português do Brasil.

O presente trabalho está dividido na seguinte sequência. A seção 2 apresenta um tópico sobre mineração de dados, a seção 3 aborda extração de informações, a seção 4 mineração de textos, na seção 5 foi escrito sobre sumarização de documentos e posteriormente na seção 6 os algoritmos avaliados, na seção 7 os materiais e métodos utilizados, a seção 8 do artigo mostra os resultados do trabalho, onde são abordados os experimentos e finalizando a com a seção 9, a conclusão.

2. Mineração de Dados

Com o avanço dos sistemas computacionais, o armazenamento de informações em bancos de dados tem possibilitado para as organizações papel de destaque, sendo que um dos principais objetivos das empresas relacionadas com a área da computação tem sido o de armazenar dados. Nas últimas décadas essa tendência ficou bem mais evidente com a queda

nos preços de *dehardwares*, tornando possível armazenar quantidades cada vez maiores de dados com baixo custo.

Com o volume de dados armazenados crescendo diariamente, as técnicas tradicionais de processamento de informações não são mais adequadas para encontrar um determinado dado específico ou conjunto de informações dentro de um montante armazenado. A mineração de dados surgiu no final da década de 80 e pode ser considerada uma das áreas mais promissoras. Dentre as áreas que a mineração de dados pode ser aplicada de forma satisfatória, destacam-se algumas como as áreas bancária, *telemarketing* e medicina [6].

3. Extração de Informações

A extração de informação é uma tarefa que visa encontrar uma informação específica em um grande volume de dados. Neste contexto, os documentos que contém as informações nem sempre apresentam um nível de estruturação, que se referem ao formato que os dados possuem, tais como *tags* em documentos XML (*Extensible Markup Language*) ou JSON (*JavaScript Object Notation*). Essa estruturação facilita o entendimento de um documento de texto, pois com essas *tags* é possível encontrar determinadas informações consultando cada uma delas, não necessitando uma análise complexa em um documento quando não apresenta essas características.

Outro exemplo são documentos divididos por capítulos, parágrafos ou tabelas, que mantêm uma certa formatação ou estrutura textual. Por outro lado existem os textos sem nenhuma formatação, nos quais as palavras estão colocadas de forma desorganizada, sem uma formatação previamente definida. Exemplos são textos extraídos de *blogs*, *sites* ou *e-mails*, os quais não possuem uma linguagem para marcação de conteúdo.

Existem atualmente muito mais dados na forma de textos eletrônicos do que em tempos passados, porém, boa parte dessa informação é ignorada, não sendo realizada nenhuma análise mais detalhada sobre seu conteúdo. Muitos destes textos estão em fóruns *on-line*, redes sociais, *sites* de avaliação de produtos, entre outros.

Dentro deste contexto da quantidade de informações armazenadas, nenhuma pessoa é capaz de ler, entender e sintetizar *megabytes* de texto no seu cotidiano. Baseado nisso, novas pesquisas foram estimuladas no sentido de desenvolver técnicas para exploração e administração da informação, a fim de estabelecer uma ordem na imensidão de textos [8].

Aprofundando-se na área de recuperação de informações, as estratégias mais comuns são a filtragem e a extração. A extração de informações tem muitas aplicações potenciais, como, por exemplo, as informações disponíveis em textos não estruturados podem ser armazenadas em bancos de dados relacionais e os usuários podem examiná-las por meio de consultas em SQL (*Structured Query Language*) [6].

Por outro lado, a filtragem de informação tem o objetivo de classificar as informações depois de extraídas. Por exemplo, um sistema de filtragem de informações que tem como objetivo localizar dados relevantes para um usuário específico. Desta forma, o usuário não

necessita procurar frequentemente por elas, pois esta tarefa será realizada pelo próprio sistema de filtragem[6].

4. Mineração de Textos

A mineração de textos é um conjunto de métodos utilizado para navegar, organizar, encontrar e descobrir informações em bases textuais [2].

A tecnologia de mineração de textos deriva das técnicas de recuperação de informações e da descoberta de informações estruturadas, por meio do uso de bancos de dados e de procedimentos estatísticos [2]. É uma subárea da extração de informações, porém é utilizada somente para análise em textos.

Por mais que possa parecer similar, a mineração de textos é diferente de mecanismos de busca, uma vez que na busca o usuário já sabe o que quer encontrar; enquanto que em mineração de textos o usuário descobrirá conhecimento e padrões até então por ele desconhecidos. Em suma, na busca o usuário pesquisa determinada informação e na mineração é realizada a coleta de novos conhecimentos “escondidos” nos textos.

Mineração de textos é o processo de extrair conhecimento não conhecido previamente a partir de fontes textuais, tais como correio, imprensa, transações, *websites*, *newsgroups*, fóruns, listas de correspondência, redes sociais, dentre outros.

5. Sumarização de Documentos

A sumarização no contexto geral é uma atividade bastante comum. Quando se narra um evento a uma pessoa, costuma-se fazer um resumo do que aconteceu, e não uma narração completa e detalhada. Mesmo não sabendo, as pessoas estão sempre sumarizando. Muito frequentes também são os sumários escritos, como por exemplo, notícias em jornais, artigos de revistas, resumo de textos científicos, entre outros.

Computacionalmente explicando, existem duas formas de se abordar o problema da sumarização, a superficial e a profunda. Neste trabalho foi abordada a primeira, que utiliza métodos estatísticos e/ou empíricos para obter o sumário. Essa técnica é a mais simples de ser implementada e é utilizada por grande parte dos pesquisadores, porém, pode produzir sumários com problemas de coesão e principalmente de coerência, o que pode deixar o resumo sem um sentido lógico da ordem das frases, apresentando deficiências no sentido das frases. Por outro lado, a sumarização profunda realiza uma análise semântica frase a frase no texto, analisando a forma que as frases são construídas e também o relacionamento de uma frase a outra. A Figura 1 apresenta a ideia principal do funcionamento de um sumarizador automático de documentos.

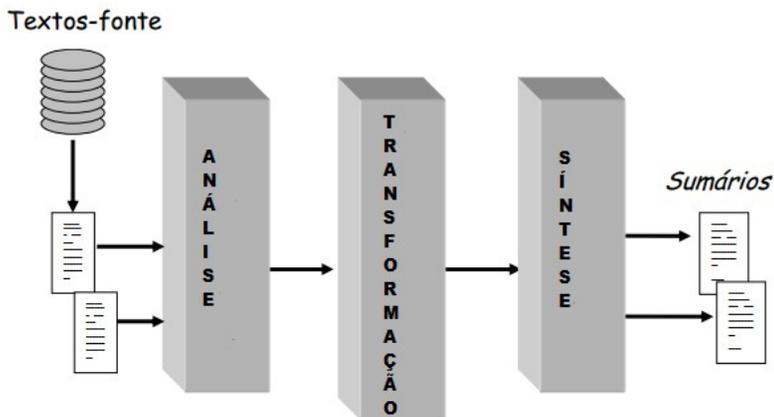


Figura 1. Arquitetura geral de um Sumarizador [12]

Na Figura 1, inicialmente pode-se perceber que o primeiro passo é quando o sistema recebe como entrada um arquivo fonte. Depois esse arquivo é passado para o módulo de análise, onde o algoritmo examina sua forma de estruturação, que é a forma em que o texto se encontra em relação a sua composição de frases. No procedimento seguinte ocorre a transformação e o cálculo da relevância das palavras chaves por meio de técnicas de extração.

Na transformação o algoritmo divide o texto em sentenças, que são as frases; então aplica seu método de extração das palavras chaves, por fim, a partir das palavras extraídas é feito o agrupamento das frases, o que irá resultar nos resumos.

Alguns algoritmos baseiam-se na geração automática de resumos, nos quais os assuntos principais de que trata um documento podem ser determinados por meio da análise dos termos que mais ocorrem no mesmo [14]. As sentenças mais importantes, que são usadas para compor o resumo do texto, são aquelas em que os termos mais frequentes aparecem em maior quantidade dentro de uma mesma frase.

No entanto, como muitos termos aparecem mais de uma vez dentro do texto, é adotado um critério de corte que consiste em considerar somente aqueles que se repetem mais que a média de repetições no documento inteiro. Caso o número de termos selecionados ainda seja muito grande para fins práticos, o valor de corte pode ser aumentado para média de repetições mais meia ou uma vez o desvio padrão (o desvio padrão aplica-se como a raiz quadrada do número de termos). Por exemplo, se o número de termos selecionados de uma frase é 09, aplicando a raiz resulta no número 3. Com isso, para todas as frases as raízes com maiores resultados serão as frases mais importantes.

6. Algoritmos de Sumarização Avaliados

Neste trabalho foram analisados e avaliados os algoritmos de Luhn e o GistSumm, ambos sumarizadores baseados na extração de palavras chaves do texto. O algoritmo de Luhn é um dos trabalhos mais importantes na área de Processamento de Linguagem Natural, em Sumarização de Documentos, é um algoritmo clássico que serviu como base para muitos algoritmos, seu método de extração é baseado na extração de palavras chaves.

O algoritmo GistSumm tem embasamento no trabalho de Luhn, porém é mais completo, seu método de extração de palavras chaves baseado na sentença principal do texto é muito eficiente quando trata-se de gerar resumos com as ideias principais de um texto segundo. Pardo [13], o GistSumm atualmente encontra-se como o estado da arte de sumarização automática de documentos, com uma função para sumarização multi-documentos.

Devido a essas características, acredita-se que o GistSumm apresente resultados melhores que o algoritmo de Luhn, pois segundo o trabalho de Oliveira [10], a eficiência de um algoritmo de sumarização está ligada ao desempenho de seu método de extração de palavras chaves. Para verificar essa hipótese, serão utilizadas tabelas com os resultados, essas tabelas mostram todos os dados tanto em forma numérica e percentual, onde podemos perceber a eficiência e demais características dos algoritmos tais como taxa de erros e acertos. Avaliações de forma semelhante a essas foram feitas nos trabalhos de Luhn [8], e Pardo [13], sendo que somente o trabalho de Pardo aplica-se ao português do Brasil.

6.1 Algoritmo de Luhn

O algoritmo de Luhn analisa as frases mais importantes de um documento, que são aquelas que mais aparecem no texto. Neste contexto, não são consideradas as *stopwords*, que são palavras como artigos, preposições, conjunções, entre outras que aparecem com frequência em um texto, porém são insignificantes em relação ao significado semântico do documento. Em suma, as *stopwords* são utilizadas apenas para dar um sentido gramatical correto na formação das frases. Como o algoritmo faz a sumarização baseada na frequência que as palavras ocorrem, elas são desconsideradas para não confundir o sumarizador.

O algoritmo não procura compreender os dados em um nível semântico, e simplesmente computa resumos com agrupamento de palavras que ocorrem com frequência no texto [8]. A Figura 2 apresenta os passos desde quando o algoritmo de Luhn recebe um texto como parâmetro até a geração final do resumo. A primeira tarefa é identificar as frases, calculando a similaridade entre todas as frases do texto.

Na segunda abordagem outra tarefa é fazer o cálculo da pontuação, levando em conta que o resumo nunca pode ter mais pontos e vírgulas que o texto original. Terminadas a primeira e segunda abordagens mostradas na Figura 2, o algoritmo finalmente extrai as sentenças escolhidas, as agrupa e mostra como resultado final o resumo.

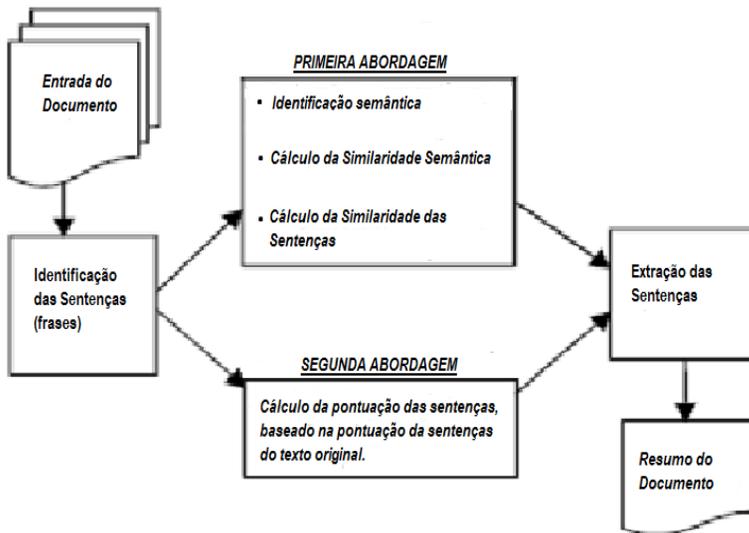


Figura 2. Modelo de Funcionamento do Algoritmo de Luhn [12]

6.2 Gistsumm

O *GistSumm* (*GistSumarizer*) é um sumarizador extrativo que usa técnicas estatísticas para determinar a ideia central dos textos por ele sumarizados. Baseia-se na simulação da sumarização humana, primeiro identificando a ideia principal do texto e, então, acrescenta informações adicionais ou complementares [12]. Essas informações adicionais podem ser a segunda ou terceira frase mais importante do texto, seguindo em ordem crescente de acordo com a quantidade de frases que se deseja extrair do texto.

Dessa forma, o sumarizador primeiro procura a sentença que melhor expressa a ideia principal do texto e baseado nela são escolhidas as demais sentenças, que vão compor o extrato textual. Algumas etapas do funcionamento do *GistSumm* são as seguintes: primeiro o *GistSumm* realiza a identificação da sentença principal com o uso de métodos estatísticos simples, e por segundo conhecendo-se as sentenças principais é possível produzir extratos coerentes [8]. Mesmo quando a sentença escolhida não for a sentença principal e há uma aproximação significativa da mesma, o extrato já pode ser gerado [6].

O *GistSumm* compreende três processos principais, e mais alguns secundários, os quais são mostrados na Figura 3 e descritos a seguir, os principais são: segmentação textual, ranqueamento de sentenças e seleção de sentenças [13].

A segmentação textual delimita as sentenças do texto-fonte e procura pelos sinais de pontuação. O ranqueamento é uma ordenação a partir de pesos obtidos na aplicação de métodos estatísticos, sendo feita a análise léxica, extração das *stopword* e aplicação do método de ranqueamento.

Por fim, a seleção de sentença escolhe as sentenças mais relevantes, por meio de seus métodos extrativos como descritos anteriormente, para deste modo gerar o sumário do documento analisado. Neste momento, o texto é transformado de acordo com a taxa de compressão definida. A taxa de compressão é a porcentagem do texto original que o algoritmo pode extrair para o resumo. Estes procedimentos são mostrados na Figura 3.

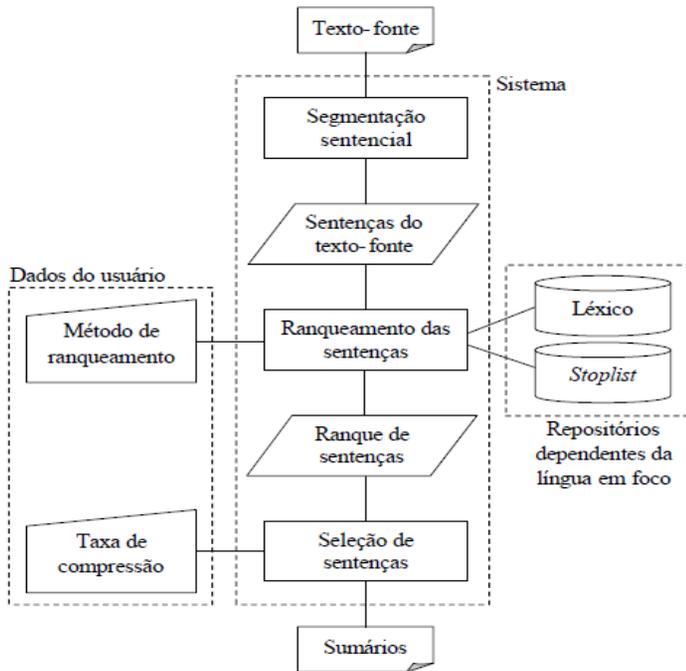


Figura 3. Modelo de funcionamento do Algoritmo GistSumm [13]

7. Materiais e Métodos

Esta seção descreve os materiais utilizados para a produção dos experimentos e também a abordagem conduzida para a realização dos testes e avaliações.

7.1 Materiais

Foi utilizada a linguagem de programação *Python* na implementação feita por Russell (2011) do algoritmo de Luhn, juntamente com a biblioteca NLTK (*Natural Language*

Toolkit). A biblioteca NLTK é uma biblioteca de funções para a construção de programas na linguagem Python, sendo utilizada para trabalhar com dados de linguagem humana[4].

A NLTK estabelece uma infraestrutura que pode ser utilizada para criar programas de processamento de linguagem natural. Fornece classes básicas para representar dados relevantes ao processamento de linguagem natural, como interfaces padrão para executar tarefas de classificação gramatical, análise sintática e classificação textual[4].

O algoritmo GistSumm foi disponibilizado pelo próprio autor Thiago Alexandre Salgueiro Pardo [13]. Sendo disponibilizada uma versão da implementação do algoritmo com execuções através de linha de comando.

7.2 Métodos

Primeiramente foi realizada uma revisão de literatura para tomar conhecimento das técnicas de mineração de dados disponíveis para sumarização de documentos. Posterior a isso, foi feita a execução do algoritmo de Luhn e configuração do algoritmo *GistSumm* para então realizar os testes com vários documentos de texto, com o objetivo de entender a forma de funcionamento de cada um.

Na sequência foi feita uma comparação entre as técnicas aplicadas por cada algoritmo por meio da construção de uma tabela comparativa de resultados, com o intuito de visualizar as principais diferenças entre eles.

Por fim foram feitos dois experimentos como dois algoritmos. O primeiro com o objetivo de identificar se os resumos mantinham as características essenciais do texto, como a ideia principal. O segundo visa determinar se é possível responder perguntas com base em um texto original, baseando se somente no resumo gerado pelos algoritmos.

7.3 Experimento 1

Para obter resultados com os algoritmos, foram utilizadas notícias jornalísticas retiradas do site da Globo do portal de notícias G1 [16]. Foram feitos testes com diversas configurações nos parâmetros modificáveis dos algoritmos, tais como taxa de compressão, quantidade de palavras chaves e quantidade de sentenças a serem retornadas, de maneira que os dois algoritmos ficassem equiparados entre si.

A taxa de compressão é a quantidade que pode ser extraída do texto para o resumo, segundo Pardo[8] deve ser entre 20 a 50% do texto para gerar um resumo eficiente.

Neste experimento, foram utilizados 2 textos apresentados a seguir, sendo que cada número no texto corresponde a uma sentença ou frase. Foi utilizado uma taxa de compressão de 80% e com configuração para extração de duas sentenças.

Segundo o autor Pardo [8] uma taxa de compressão de 80% é suficiente para construir um resumo com as ideias principais de um texto, ou seja em um texto de 10

sentenças seriam extraídas 2 sentenças, essa taxa foi aplicada ao experimento 1 pois nele o objetivo é extrair apenas a ideia principal do texto.

Texto 1:

[1Sonda indiana revela grande quantidade de água na Lua.] [2A descoberta animou a comunidade científica.] [3Pode revelar mistérios sobre a vida fora da terra e ser um passo em direção a um antigo e, até agora, distante sonho da Nasa: construir uma estação espacial na lua.] [4As fotos divulgadas pela Nasa ainda são a única amostra da descoberta.] [5Uma parte da imagem revela a presença de moléculas de H₂O, a fórmula química da água, sobre partículas de poeira, encontradas na superfície da lua.] [6A evidência da presença de água foi detectada pela sonda indiana Chandrayan 1.] [7Os dados foram analisados por cientistas americanos que já tinham descoberto vestígios de gelo em crateras próximas a um dos polos da lua.] [8A descoberta surpreendeu os cientistas que chegaram a duvidar do que foi encontrado, mas novos testes confirmaram uma quantidade de água maior do que eles imaginavam.] [9No mês que vem, uma sonda da NASA vai pousar na lua e recolher fragmentos do solo para análise.] [10Os cientistas dizem que esta é a prova inequívoca de que existe água em mais de 50% da superfície lunar.]

Esse texto foi dividido em sentenças e possui 183 palavras. Quando foi submetido ao algoritmo de Luhn configurado com taxa de compressão de 80%, obteve-se então um resumo com 33 palavras.

Esse mesmo texto também foi submetido ao *GistSumm* com a mesma taxa de compressão. Sendo assim o algoritmo retornou um resumo de 22 palavras, no qual apareceram menos palavras do que com o algoritmo de Luhn, pois as sentenças extraídas são mais curtas.

Texto 2:

[1Cientistas anunciaram nesta quarta-feira (06) ter encontrado um "manto de invisibilidade" que permite ao vírus da imunodeficiência humana adquirida (HIV), causador da Aids, invadir as células humanas sem ser notado e se replicar sem ativar o sistema imunológico.][2Eles também conseguiram expor o vírus em células cultivadas em laboratório usando um medicamento experimental, feito que pode levar a tratamentos novos e mais eficazes contra o HIV.][3A equipe identificou duas moléculas dentro das células humanas que são recrutadas pelo HIV após a infecção para ajudar a protegê-lo e, assim, atrasar a resposta imunológica.][4De acordo com o comunicado, os cientistas administraram uma droga experimental, com base em ciclosporina, amplamente usada pra evitar a rejeição de órgãos em pacientes transplantados porque reprime o sistema imune.][5Os cientistas descobriram que a medicação conseguiu evitar que o vírus utilizasse estas moléculas como disfarce.][6A equipe usou uma

versão modificada da droga, que bloqueia os efeitos das duas moléculas-disfarce sem anular a atividade imunológica, destacou o comunicado.][7"A esperança é que um dia nós consigamos desenvolver um tratamento que ajude o corpo a se livrar do vírus antes que a infecção consiga tomar conta", afirmou o co-autor do estudo, *Greg Towers*, da Universidade *College* de Londres, em um comunicado do fundo *Wellcome Trust*, que co-financiou a pesquisa.]

Esse texto também foi dividido em sentenças e possui 213 palavras. Quando foi submetido ao algoritmo de Luhn configurado com taxa de compressão de 80%, obteve-se então um resumo com 61 palavras. Esse mesmo texto também foi submetido ao *GistSumm* e foi determinado para ele a mesma taxa de compressão para os dois algoritmos. Sendo assim o algoritmo retornou um resumo de 43 palavras. Pode-se observar que o número de palavras extraídas varia de acordo com a quantidade de palavras que a sentença possui, não alterando o objetivo do sumarizador, que é extrair a ideia principal do texto, independente de em qual sentença do texto ele esteja.

7.3.1 Sentenças Extraídas

Quando submetido ao algoritmo de Luhn, do Texto 1 foram extraídas as sentenças 6 e 10, do texto 2 foram as sentenças 1 e 6, de acordo com a numeração no texto do Quadro 1. De acordo com o algoritmo, essas duas sentenças respectivamente em ambos os textos são as mais importantes e as que trazem a ideia principal dele.

Por outro lado, a geração do resumo com *GistSumm* apontou as sentenças 1 e 6 no Texto 1, e as sentenças 3 e 5 no Texto 2 como as mais importantes dos textos. O Quadro 1 apresenta as diferenças sentenciais entre os resumos gerados.

Textos Testados	Sentenças Extraídas com o Algoritmo de Luhn	Sentenças Extraídas com o Algoritmo GistSumm
Texto 1:	[6 A evidência da presença de água foi detectada pela sonda indiana Chandrayan 1.] [10 Os cientistas dizem que esta é a prova inequívoca de que existe água em mais de 50% da superfície lunar.]	[1 Sonda indiana revela grande quantidade de água na Lua.] [6 A evidência da presença de água foi detectada pela sonda indiana Chandrayan 1.]
Texto 2:	[1 Cientistas anunciaram nesta quarta-feira (06) ter encontrado um "manto de invisibilidade" que permite ao vírus da imunodeficiência humana adquirida (HIV), causador da Aids, invadir as células humanas sem ser notado e se replicar sem ativar o sistema imunológico.] [6 A equipe usou uma versão modificada da droga, que	[3 A equipe identificou duas moléculas dentro das células humanas que são recrutadas pelo HIV após a infecção para ajudar a protegê-lo e, assim, atrasar a resposta imunológica.] [5 Os cientistas descobriram que a medicação conseguiu evitar que o vírus utilizasse estas moléculas como disfarce.]

	bloqueia os efeitos das duas moléculas-disfarce sem anular a atividade imunológica, destacou o comunicado.]	
--	--	--

Quadro 1. Sentenças extraídas no primeiro experimento

7.3.2 Validação com avaliadores humanos

Um dos maiores problemas na área de sumarização automática de documentos é como realizar a avaliação dos resumos [13]. Para isso, foram utilizados avaliadores humanos para verificar a qualidade e coerência dos resumos obtidos. As pessoas receberam o texto original e os respectivos resumos gerados por cada um dos algoritmos, sendo solicitado que os avaliadores fizessem a leitura do texto original e de ambos os resumos, sendo feita a seguinte pergunta: Avalie a capacidade que o algoritmo teve em extrair a ideia principal do texto, e como resposta os avaliadores tinham as opções bom, regular e ruim. Essa forma de avaliação foi baseada no trabalho de Rino [13].

A classificação como “bom” indica que o resumo manteve as características principais do texto original, estando de forma ordenada e coerente em relação ao texto original. A classificação como “regular” indica que o resumo apresenta algumas das características principais do texto, porém elas não são agrupadas no resumo na mesma ordem que estavam no texto fonte ou então se apresentam de maneira confusa. Por fim, a classificação como “ruim” indica que as informações no resumo não trazem a ideia principal do texto.

7.4 Experimento 2

Essa avaliação foi realizada para descobrir a qualidade do resumo obtido, que segundo Filho [5], busca identificar se é possível obter todas as informações relevantes do texto original no resumo. Para isso foram utilizados dois textos distintos e elaboradas cinco perguntas com base em cada um. Posteriormente foram gerados os resumos com os algoritmos, buscando-se extrair além da ideia principal, algumas ideias secundárias, foi necessário extrair mais informações do que no experimento 1, então aplicou-se a taxa de 50%, onde extraiu-se metade do texto, logo que o objetivo era verificar se os avaliadores são capazes de responder as perguntas analisando somente o resumo gerado pelos algoritmos, sem ter acesso ao texto original. Foram utilizadas 30 pessoas nessa avaliação, sendo repassado a elas cinco perguntas sobre o texto original. Na sequência são mostrados os textos originais onde as perguntas foram extraídas [7] e [9] respectivamente.

Texto 1:

[1 Morre neste domingo, aos 76 anos, em sua casa em Toronto, no Canadá, o ex-boxeador americano Rubin "Hurricane" Carter, que causou uma comoção mundial ao ser preso injustamente.] [2 Carter passou 19 anos na prisão, acusado por um triplo homicídio em Nova Jersey, nos Estados Unidos, em 1966.] [3 A injustiça teria sido motivada por racismo.] [4 A sua história inspirou a música "Hurricane", do cantor e

compositor Bob Dylan, o filme "The Hurricane" ("Hurricane - O Furacão"), com Denzel Washington como protagonista, e ainda uma série de livros.] [5Rubin "Hurricane" foi condenado sem provas em 1967, e foi novamente julgado em 1976, sendo mais uma vez incriminado pelos três assassinatos.]

[60 ex-boxeador ficou famoso por travar uma dura batalha para provar a sua inocência.][7Após uma campanha que contou com o apoio de estrelas como Bob Dylan e Muhammad Ali, Carter ganhou a liberdade em 1985 com a retirada do processo e a anulação da pena.][80 juiz que cuidou do caso na época afirmou que conclusões sobre a prisão do ex-pugilista foram "com base no racismo e não na razão, assim como na ocultação da verdade".]

[9"Hurricane" teve a sua promissora carreira nos ringues interrompida ao ser surpreendido pela polícia quando andava de carro com um amigo.][10Antes de ser preso, ele era o favorito ao cinturão dos pesos médios de 1966, aos 29 anos.][11Duas testemunhas que teriam visto o triplo homicídio em um bar da cidade americana e confirmaram Carter e um amigo como os autores do crime.][120 amigo ficou preso por 15 anos.] [13Carter lutava contra um câncer na próstata e morreu enquanto dormia, ele escreveu um livro quando estava preso, chamado de "The Sixteenth Round".]

No resumo com o GistSumm foram extraídas as sentenças 1, 5, 8, 9 e 10e com o algoritmo de Luhn foram extraídas sentenças 4, 6,8, 9, e 11. O Quadro 2 apresenta os resumos gerados por cada um dos algoritmos.

Sentenças Extraídas com o Algoritmo de Luhn	Sentenças Extraídas com o Algoritmo GistSumm
<p>4 A sua história inspirou a música "Hurricane", do cantor e compositor Bob Dylan, o filme "The Hurricane" ("Hurricane - O Furacão"), com Denzel Washington como protagonista, e ainda uma série de livros.</p>	<p>1 Morre neste domingo, aos 76 anos, em sua casa em Toronto, no Canadá, o ex-boxeador americano Rubin "Hurricane" Carter, que causou uma comoção mundial ao ser preso injustamente.</p>
<p>6 O ex-boxeador ficou famoso por travar uma dura batalha para provar a sua inocência.</p>	<p>5 Rubin "Hurricane" foi condenado sem provas em 1967, e foi novamente julgado em 1976, sendo mais uma vez incriminado pelos três assassinatos.</p>
<p>8 O juiz que cuidou do caso na época afirmou que conclusões sobre a prisão do ex-pugilista foram "com base no racismo e não na razão, assim como na ocultação da verdade".</p>	<p>8 O juiz que cuidou do caso na época afirmou que conclusões sobre a prisão do ex-pugilista foram "com base no racismo e não na razão, assim como na ocultação da verdade".</p>
<p>9 "Hurricane" teve a sua promissora carreira nos ringues interrompida ao</p>	<p>9 "Hurricane" teve a sua promissora</p>

<p>ser surpreendido pela polícia quando andava de carro com um amigo.</p> <p>11 Duas testemunhas que teriam visto o triplo homicídio em um bar da cidade americana e confirmaram Carter e um amigo como os autores do crime.</p>	<p>carreira nos ringues interrompida ao ser surpreendido pela polícia quando andava de carro com um amigo.</p> <p>10 Antes de ser preso, ele era o favorito ao cinturão dos pesos médios de 1966, aos 29 anos.</p>
---	---

Quadro 2. Sentenças Extraídas do Texto1 no Experimento 2.

As perguntas foram definidas de acordo com as frases que tinham maior influência no título dos textos, e nas informações principais de acordo com uma sumarização humana feita pelos autores. Foram feitas cinco perguntas aos avaliadores, sendo elas e suas possíveis respostas de acordo com o Quadro 3: (as respostas corretas seguem assinaladas)

1. O texto conta a história de quem?			
(X)Rubin Carter	()Denzel Washington	()Bob Dylan	()Não sei
2. Segundo o texto o que aconteceu com ele?			
()Foi Morto	(X)Foi Condenado	()Foi Crucificado	()Não sei
3. O que ele fazia?			
()Era Cantor	(X)Era Pugilista	()Era Compositor	()Não sei
4. Onde se passou a história?			
()Califórnia	(X)Nova Jersey	()Nova York	()Não sei
5. Como ele se encontra atualmente?			
()Vivo	(X)Morto	()No ringue	()Não sei

Quadro 3. Perguntas para o Texto 1 no Experimento2.

Foi passado aos avaliadores a orientação de que escolhessem apenas uma resposta para cada pergunta, e que respondessem apenas com as informações ocorrentes no resumo à eles passado.

Texto 2:

[1 Um asteróide recém descoberto do tamanho aproximado de uma casa passará relativamente perto da Terra neste domingo (7), aproximando-se dos satélites de comunicação que circundam o planeta, disseram cientistas.] [2 A Nasa afirmou que o asteroide, conhecido como 2014 RC, não representa uma ameaça, embora em seu ponto de maior aproximação irá estar a cerca de um décimo da distância da lua, ou a cerca de 40 mil quilômetros da Terra.] [3 Os satélites de comunicação e meteorológicos geralmente ficam em órbita a cerca de 36 mil quilômetros acima do planeta.] [4 "Embora este objeto celestial não

pareça ameaçar a Terra nem os satélites, sua aproximação cria uma oportunidade única para os pesquisadores observarem e aprenderem mais sobre os asteroides", informou a Nasa em um comunicado divulgado na terça-feira.] [5 Com um diâmetro aproximado de 18 metros, o asteroide 2014 RC estará tênue demais para ser visível a olho nu, mas astrônomos amadores podem conseguir um vislumbre seu enquanto passa voando, disse a agência espacial.] [6 O 2014 RC tem um diâmetro um pouco menor que os 20 metros do asteroide que explodiu sobre Chelyabinsk, na Rússia, no ano passado.] [7 A onda de choque da explosão, que se estimou ter tido 30 vezes mais energia que a bomba atômica de Hiroshima, estraçalhou janelas e danificou edifícios.] [8 Mais de mil pessoas ficaram feridas pelos estilhaços de vidro e destroços.] [9 No mesmo dia da explosão de Chelyabinsk, outro asteroide grande chegou a 27.630 quilômetros da Terra, ou seja, ele poderia ter atingido os satélites de comunicação e meteorológicos.] [10 O mais recente visitante celestial ao planeta foi avistado em 31 de agosto pelo programa de monitoramento espacial Catalina Sky Survey, perto da cidade de Tucson, no Estado do Arizona, e confirmado na noite seguinte pelo telescópio Pan-STARRS 1 no Havaí.] [11 O ponto máximo de aproximação do asteroide será sobre a Nova Zelândia as 15h18 (horário de Brasília) do domingo, disse a Nasa.] [12 Atualmente a Nasa acompanha mais de 11 mil asteróides em órbitas que passam relativamente perto da Terra.]

No resumo com o GistSumm foram extraídas as sentenças 1, 2, 3, 4, e 9 e com o algoritmo de Luhn foram extraídas sentenças 1, 3, 5, 6 e 9. O Quadro 3 apresenta os resumos gerados por cada um dos algoritmos.

Sentenças Extraídas com o Algoritmo de Luhn	Sentenças Extraídas com o Algoritmo GistSumm
<p>1 Um asteroide recém descoberto do tamanho aproximado de uma casa passará relativamente perto da Terra neste domingo (7), aproximando-se dos satélites de comunicação que circundam o planeta, disseram cientistas.</p> <p>3 Os satélites de comunicação e meteorológicos geralmente ficam em órbita a cerca de 36 mil quilômetros acima do planeta.</p> <p>5 Com um diâmetro aproximado de 18 metros, o asteroide 2014 RC estará tênue demais para ser visível a olho nu, mas astrônomos amadores podem conseguir um vislumbre seu enquanto passa voando, disse a agência espacial.</p>	<p>1 Um asteroide recém descoberto do tamanho aproximado de uma casa passará relativamente perto da Terra neste domingo (7), aproximando-se dos satélites de comunicação que circundam o planeta, disseram cientistas.</p> <p>2 A Nasa afirmou que o asteroide, conhecido como 2014 RC, não representa uma ameaça, embora em seu ponto de maior aproximação irá estar a cerca de um décimo da distância da lua, ou a cerca de 40 mil quilômetros da Terra.</p> <p>3 Os satélites de comunicação e meteorológicos geralmente ficam em órbita a cerca de 36 mil quilômetros</p>

<p>6 O 2014 RC tem um diâmetro um pouco menor que os 20 metros do asteroide que explodiu sobre Chelyabinsk, na Rússia, no ano passado.</p> <p>9 No mesmo dia da explosão de Chelyabinsk, outro asteroide grande chegou a 27.630 quilômetros da Terra, ou seja, ele poderia ter atingido os satélites de comunicação e meteorológicos.</p>	<p>acima do planeta.</p> <p>4 "Embora este objeto celestial não pareça ameaçar a Terra nem os satélites, sua aproximação cria uma oportunidade única para os pesquisadores observarem e aprenderem mais sobre os asteroides", informou a Nasa em um comunicado divulgado na terça-feira.</p> <p>9 No mesmo dia da explosão de Chelyabinsk, outro asteroide grande chegou a 27.630 quilômetros da Terra, ou seja, ele poderia ter atingido os satélites de comunicação e meteorológicos.</p>
---	---

Quadro 4. Sentenças Extraídas do Texto2 no Experimento 2.

Foram feitas cinco perguntas aos avaliadores, usando os mesmos critérios do Texto1, sendo elas e suas possíveis respostas: (as respostas corretas seguem assinaladas)

<p>1. Qual é o tema principal do texto? <input checked="" type="checkbox"/>Um asteróide()Um satélite ()A NASA ()Não sei</p> <p>2. Qual é o nome do personagem principal do texto? ()Pan-STARRS <input checked="" type="checkbox"/>2014 RC ()Catalina Sky Survey()Não sei</p> <p>3. Segundo o texto ano passado houve a explosão de um Asteróide, onde foi? ()Tucson ()Nova Zelândia <input checked="" type="checkbox"/>Chelyabinski ()Não sei</p> <p>4. O 2014 RC pode ser considerado uma ameaça ao planeta? ()Sim <input checked="" type="checkbox"/>Não ()Não sei</p> <p>5. Existem objetos que ficam em orbita a cerca de 36 mil km da Terra, que objetos são esses? <input checked="" type="checkbox"/>Satélieis ()Lua ()Meteóros ()Não sei</p>
--

Quadro 5. Perguntas para o Texto 2 no Experimento2.

8. Resultados

Esta seção aborda os resultados obtidos com ambos algoritmos, detalhando os testes e resultados obtidos em cada experimento realizado.

A Tabela 1 descreve a avaliação feita por humanos no experimento 1, sendo consultadas 30 pessoas para cada teste. Para essa avaliação 4 testes, sendo 2 testes para cada algoritmo, foram consultados 120 avaliadores nesse teste.

Classificação/ Algoritmo	Bom	Regular	Ruim	Total
	Luhn	82	27	11
	68,3%	22,5%	9,2%	
GistSumm	98	18	4	120
	81,6%	15%	3,4%	

Tabela 1. Avaliação dos algoritmos avaliados por humanos no Experimento 1.

Pode-se observar que segundo as avaliações feitas pelos avaliadores, o algoritmo GistSumm apresentou resultados um pouco melhores do que o algoritmo de Luhn. Uma hipótese para isso é devido ao algoritmo possuir um método de extração de palavras chaves melhor, pois segundo [12], seu método baseado na extração da principal sentença faz com que seus resumos extraiam as frases mais relacionadas ao texto, mantendo assim as suas características principais.

Na Tabela 2, é possível observar os resultados referentes ao algoritmo de Luhn para os dois textos, sendo 30 avaliadores em cada.

Luhn	Pergunta 1	Pergunta 2	Pergunta 3	Pergunta 4	Pergunta 5
Respostas Certas	60	30	60	0	30
Respostas Erradas	0	30	0	60	30
Total	60	60	60	60	60

Tabela 2. Resultados do Algoritmo de Luhn.

Na Tabela 3 pode-se observar os resultados do algoritmo GistSumm para os 2 textos, sendo 30 avaliadores em cada. Cada campo da tabela é composto pelo número de avaliadores divididos em avaliadores que responderam certo e errado cada pergunta.

GistSumm	Pergunta 1	Pergunta 2	Pergunta 3	Pergunta 4	Pergunta 5
Respostas Certas	60	60	60	30	60
Respostas Erradas	0	0	0	30	0
Total	60	60	60	60	60

Tabela 3. Resultados do Algoritmo GistSumm

Esses testes foram realizados de forma que o texto original foi reduzido para metade, ou seja, foi extraído para formar os resumos de 50% do texto. Nessas condições, na Tabela 5 é possível verificar que o resumo com o algoritmo GistSumm permitiu que os avaliadores acertassem 90% das perguntas, enquanto que com o algoritmo de Luhn os avaliadores obtiveram 60% de respostas certas.

É importante ressaltar que resultados diferentes podem ser obtidos em outros textos, visto que os textos podem apresentar diferentes formas de agrupamento e palavras chaves. Outra questão importante é o fato de que o foco principal da notícia é a morte de uma pessoa no texto 1, e o GistSumm foi capaz de extrair esta informação, conforme pode ser observado no Quadro 2, enquanto que com o algoritmo de Luhn as pessoas responderam que a notícia se trata da prisão da pessoa.

A Tabela 4 avalia a porcentagem de erro e de acerto que os dois algoritmos, o de Luhn e o GistSumm tiveram nos testes. Ela foi construída através da média de erros e acertos das Tabelas 2 e 3.

Algoritmo	Taxa de Erros	Taxa de Acertos
Luhn	40%	60%
GistSumm	10%	90%

Tabela 4. Porcentagem de erros e acertos.

9. Conclusão

Ambos os algoritmos de sumarização de documentos testados tiveram resultados satisfatórios, ou seja, isso ocorre quando o resumo mantém a ideia principal do texto. Neste sentido, o algoritmo GistSumm foi mais coerente em relação ao algoritmo de Luhn.

Com base em trabalhos anteriores de autores relacionados, como o de Luhn [8], que realizou avaliações com pessoas de modo similar ao presente artigo, descobriu-se a necessidade de submeter esses resumos à classificadores humanos, uma vez que essa avaliação é bastante subjetiva, pois um avaliador humano pode ter um opinião diferente de outro, e, por esse motivo que foram consultados 30 avaliadores, sendo feita a média dos resultados de suas avaliações.

Para as avaliações foram consultados acadêmicos da UnC – Universidade do Contestado, Campus de Porto União – SC. Na avaliação pelas sentenças extraídas do experimento 1, esperava-se que o GistSumm, segundo Pardo[11], devido a possuir mais técnicas de extração em sua composição, apresentasse resultados mais significativos se comparados ao algoritmo de Luhn. No entanto, os resultados apontaram resultados equiparados, com um nível de qualidade razoavelmente melhor.

Analisando os resultados da avaliação do experimento 1, observou-se que dos 30 avaliadores 81,6% classificaram o GistSumm como bom, resultados um pouco melhores se comparados ao algoritmo de Luhn que teve 68,3% das avaliações boas. Esses resultados são referentes às Tabelas 1 e 2. Na avaliação baseada no resumo do experimento 2, pode-se notar que o algoritmo GistSumm apresentou uma taxa de acerto das questões propostas aos avaliadores de 90%, foi superior ao algoritmo de Luhn que apresentou uma taxa de acerto de 60%, esses dados estão demonstrados nas Tabelas 3, 4 e 5.

Como trabalhos futuros, sugere-se a avaliação da sumarização multi-documentos, assim como abordar trabalhos com algoritmos de sumarização através de técnicas utilizando meta-heurísticas.

Referencias

- [1] Agostini, V.; Camargo, R.T.; Pardo, T.A.S.; Di Felippo. Alinhamento manual de textos e sumários em um corpus jornalístico multidocumento. XI encontro de linguística de corpus. São Carlos/SP. 2012.
- [2] Aranha, Christian; Passos, Emmanuel. A Tecnologia de Mineração de Textos. RESI-Revista Eletrônica de Sistemas de Informação. 2006.
- [3] Benevenuto, Fabrício; Almeida, Jussara M.; Silva, Altigran S.. Coleta e Análise de Grandes Bases de Dados de Redes Sociais Online. In IEEE Multimedia Computing and Networking (MMCN). 2008.
- [4] Bird, Steven; Klein, Ewan; Loper, Edward. Natural Language Processing with Python – Analyzing Text with the Natural Language Toolkit. Primeira Edição. Editora O’Reilly. 2009.
- [5] Camilo, Cassio Oliveira; Silva, João Carlos da. Mineração de Dados: Conceitos, Tarefas, Métodos e Ferramentas. Universidade Federal de Goiás. Instituto de Informática. 2006.
- [6] Filho, Pedro Paulo Balage; Pardo, Thiago Alexandre Salgueiro; Nunes, Maria das Graças Volpe. Sumarização Automática de Textos Científicos: Estudo de Caso com o Sistema GistSumm. NILC - ICMC-USP. São Carlos, São Paulo. 2007.
- [7] GloboEsporte.com. Rubin “Hurricane” Carter, ex-boxeador que inspirou Bob Dylan, morre aos 76 anos. Última atualização em 20/04/2014. Disponível em <http://globoesporte.globo.com/boxe/noticia/2014/04/rubin-hurricane-carter-ex-boxeador-que-inspirou-bob-dylan-morre-aos-76.html>. Acesso em 21 abr. 2014.

- [8] Luhn, H.P. The Automatic Creation of Literature Abstracts. IBM Journal. 1958.
- [9] Matson, John. http://www2.uol.com.br/sciam/noticias/asteroide_recem-descoberto_passa_pela_terra.html Acessado em 15 set. 2014.
- [10] Oliveira, Marcelo Arrantes; Guelpeli, Marcos Vinicius. BLMSumm – Métodos de Busca Local e Metaheurísticas na Sumarização de Textos. Centro Universitário de Barra Mansa (UBM). Barra Mansa, Rio de Janeiro. 2011.
- [11] Rezende, Solange O.; Marcacini, Ricardo M.; Moura Maria F.. O uso da Mineração de Textos para Extração e Organização Não Supervisionada de Conhecimento. Revista de Sistemas de Informação da FSMA n. 7, pp. 7-21. 2011.
- [12] Russel, Matthew A.. Mineração de dados da Web Social. Primeira edição, Novatec. São Paulo, 2011.
- [13] 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.
- [14] Silla Jr., Carlos N., Kaestner, Celso A.A.. kNNSumm: Um Sumarizador Automático de Documentos Utilizando Aprendizado Baseado em Instâncias. Pontifícia Universidade Católica do Paraná (PUC-PR). Curitiba, Paraná. 2004.
- [15] Silva, Luiz Cláudio; Sampaio, Renelson R.. Uso de Grafos de Termos para Análise do Conteúdo de Documentos Técnicos. Faculdade de Tecnologia SENAI Climatec. Salvador, Bahia. 2013.
- [16] Teles, Lilia; Nova York; Última atualização em 25/09/2009; Sonda indiana revela grande quantidade de água na Lua; Acesso em 14 nov. 2013.
- [17] Turney, Peter D.. Learning to Extract Keyphrases from Text. National Research Council, Institute for Information Technology, Technical Report ERB-1057. 1999.
- [18] Wives, Leandro Krug. Tecnologias de descoberta de conhecimento em textos aplicadas à inteligência competitiva. Universidade Federal do Rio Grande do Sul, Instituto de informática, Programa de pós-graduação em Computação, Porto Alegre. 2012.