UNIVERSIDADE FEDERAL DO CEARÁ
CENTRO DE HUMANIDADES II
DEPARTAMENTO DE CIÊNCIAS DA INFORMAÇÃO
CURSO DE BIBLIOTECONOMIA

JOSÉ CARLOS CANDIDO

O PROCESSAMENTO DE DADOS DE PÁGINAS HTML OBTIDAS NA WEB: UMA DISCUSSÃO SOBRE O QUÃO DIFÍCIL É EXTRAIR DADOS DOS ESPAÇOS DIGITAIS

FORTALEZA
2022

JOSÉ CARLOS CANDIDO

O PROCESSAMENTO DE DADOS DE PÁGINAS HTML OBTIDAS NA WEB: UMA DISCUSSÃO SOBRE O QUÃO DIFÍCIL É EXTRAIR DADOS DOS ESPAÇOS DIGITAIS

Monografia apresentada, como requisito parcial para conclusão do curso de Biblioteconomia, da Universidade Federal do Ceará.

Orientador: Prof. Dr. Osvaldo de Souza

FORTALEZA
2022

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca de Ciências e Tecnologia

C223p Candido, José Carlos.

O PROCESSAMENTO DE DADOS DE PÁGINAS HTML OBTIDAS NA WEB : DISCUSSÃO SOBRE O QUÃO DIFÍCIL É EXTRAIR DADOS DOS ESPAÇOS DIGITAIS / José Candido. – 2022.

36 f. : il.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Centro de Humanidades, Curso de Biblioteconomia, Fortaleza, 2022.

Orientação: Prof. Dr. Osvaldo de Souza.

1. HTML. 2. Textos Jornalísticos. 3. Jano. I. Título.

CDD 020

JOSÉ CARLOS CANDIDO

O PROCESSAMENTO DE DADOS DE PÁGINAS HTML OBTIDAS NA WEB: UMA DISCUSSÃO SOBRE O QUÃO DIFÍCIL É EXTRAIR DADOS DOS ESPAÇOS DIGITAIS

Monografia apresentada como requisito parcial para conclusão do curso de Biblioteconomia, da Universidade Federal do Ceará.

Aprovada em: ___/___2022

BANCA EXAMINADORA

_____

Professor Dr. Osvaldo de Souza

Orientador (UFC)

_____

Professor Dr. Hamilton Rodrigues Tabosa

Examinador (UFC)

_____

Professor Dr. Arnoldo Nunes da Silva

Examinador (UFC)

Dedico este trabalho aos meus pais, que sempre me suportaram em meus empreendimentos.

AGRADECIMENTOS

A TotalClipping de Notícias pelo fornecimento do acesso gratuito à plataforma JANO, a partir do qual vários dados utilizados neste trabalho foram obtidos.

RESUMO

Este trabalho é um estudo exploratório sobre a coleta de conteúdos jornalísticos em páginas na *web* e como tal busca se relacionada com a *Hypertext markup language* (HTML), especificamente no contexto das definições de seu padrão, o *HTML Standard*, bem como no contexto de como o HTML é utilizado na prática no desenvolvimento de sites jornalísticos. O objetivo deste trabalho é apresentar estruturais ideais de páginas HTML, a partir dos padrões definidos pelas suas entidades mantenedoras, e propor heurísticas para extração de dados das mesmas. Para realização das atividades, a pesquisa se utiliza de dados e relatórios coletados pela plataforma JANO, serviço disponibilizado na web que providencia coleta e análise de dados em páginas web como também em outras mídias. A partir dos dados coletados, foi feita uma análise sobre o uso do HTML no desenvolvimento de páginas de sites jornalísticos, também foi feita uma proposta de estrutura para tais páginas com base no padrão HTML e foram apresentadas heurísticas de como remediar os problemas encontrados nas páginas analisadas.

Palavras-chave: HTML; Textos Jornalísticos; JANO

ABSTRACT

This work is an exploratory study on the capture of journalistic content on web pages and how such search is related to the Hypertext markup language (HTML), specifically in the context of the definitions of its standard, as well as in the context of how HTML is utilized in practice in the development of journalistic websites. The objective of this work is to present ideal HTML page structures, based on the standards defined by its maintaining entities, and propose heuristics to extract date from those. To conduct the activities, the research uses data and reports collected by the JANO platform, a service available on the web that provides data collection and analysis on web pages as well as other media. From the collected data, an analysis was made on the use of HTML in the development of pages of journalistic sites, a proposal of structure for such pages based on the HTML standard also was made and heuristics on how to remedy the problems found in the analyzed pages were presented.

Keywords: HTML; Journalistic texts; JANO.

LISTA DE ILUSTRAÇÕES

LISTA DE ABREVIATURAS E SIGLAS

CERN - Conseil Européen pour la Recherche Nucléaire

HTML - Hypertext markup language

IETF - Internet Engineering Task Force

TDICs - Tecnologias Digitais da Informação e Comunicação

URIs - Uniform Resources Identifiers

W3C - Wide Web Consortium

WHATWG - Hypertext Application Technology Working Group

WWW - World wide web

XHTML - Extensible HyperText Markup Language

XML - Extensible Markup Language

SUMÁRIO

1 INTRODUÇÃO

A nossa atual era informacional é marcada por dois pontos prevalentes: as vastas quantidades de informações que são criadas a todo momento e a velocidade cujas elas são distribuídas e consumidas pelo mundo. Ambos os pontos acontecem graças à constante e rápida evolução das Tecnologias Digitais da Informação e Comunicação (TDICs) nas últimas décadas. Além disso, a melhora na acessibilidade econômica dos usuários a tais tecnologias também é fator catalisador de tal era. A internet hoje é ubíqua na vida de grande parte do mundo, e sua influência só vem a crescer.

Informações desde as mais generalizadas, de uso popular, até as mais especializadas se encontram hoje em dia nas pontas dos dedos daqueles que as procuram. São eles os usuários da informação no contexto moderno, que foram treinados pelas suas experiências a esperar a ter acesso vinte e quatro horas à informação, em qualquer lugar e de forma rápida. Esses usuários atuam também com papel duplo, pois a barreira cada vez mais baixa para produção de conteúdo online, seja de simples páginas agregadoras de textos até os mais elaborados conteúdos multimídia, vem permitindo-os a experienciar o processo de criação.

Apesar de tantos benefícios, os efeitos reais desse grande aumento na produção de conteúdo online vêm também se acumulando. Nos primórdios da web, podia-se ter certeza de que era possível realizar uma busca em 100% do conteúdo disponível naquela época. Já hoje, o percentual da web que pode ser, realisticamente, consultado em uma busca só vem a cair. Apontar como problema o grande aumento na produção de conteúdo pode fazer sentido instintivo, mas não pintaria uma imagem completa da questão. Sim, a produção online vem aumentando consideravelmente nos últimos anos, e, mais impactante que este fenômeno é o fato de que a maioria das tecnologias na base dessas criações não foram construídas, ou modificadas em seu tempo de existência, para permitir uma boa recuperação de suas informações.

Em especificamente, a Linguagem de Marcação de Hipertexto, do inglês *Hypertext markup language* (HTML), cuja história está firmemente interligada com a história da *world wide web* (WWW), que é elemento base para criação de parte mais do que significativa dos documentos online, falha em impor diretrizes em seu uso que poderiam ser utilizadas para tornar a *web* um ambiente voltado à recuperação da informação.

1.1 A motivação

A Biblioteconomia está inserida no contexto da *guarda* dos dados que existem nas publicações registradas. Esse é o primeiro entendimento que se obtém no curso de Biblioteconomia da UFC, todavia, no desenrolar do curso, e com o contato com vários docentes, alguns dos quais, pesquisadores da área, percebe-se que a Biblioteconomia na verdade está, ou deveria estar envolvida com todo o processo informacional, destacando-se os desdobramentos desse processo no leque das vastas possibilidades ofertadas pela Internet, e nessa, os serviços encontrados na WWW. Todavia, entendemos esse leque como vasto e confuso, justamente pelo seu tamanho e pela falta de uma organização estrutural aplicada aos dados contidos na infinidade de páginas hiperlink. Penso que falta à WWW um catálogo, um ponto de entrada, a partir do qual tudo possa ser encontrado sem dificuldades. Infelizmente esse catálogo não existe, o que existe é um conjunto de outros serviços chamados de *buscadores* ou ainda de *meta-buscadores*, os quais coletam dados automaticamente e intensamente, de tudo que encontram na WWW.

Esses serviços, no entanto, no máximo apontam a direção, mas não dão respostas e não apresentam como resultados das buscas, a informação exata que o potencial usuário procura.

1.2 Objeto de Estudo

Neste estudo, trabalhamos com as páginas HTML encontradas na WWW, com foco no setor de comunicação jornalística, e deste setor, elege-se como objeto de estudo as principais emissoras de conteúdo jornalístico no Ceará, compreendendo os jornais online e blogs jornalísticos profissionais.

1.3 Objetivo do Trabalho

O presente estudo tem por objetivo geral analisar publicações jornalísticas na WEB, a fim de apresentar estruturais ideais de páginas HTML e desenvolver heurísticas para extração de dados de tais páginas.

1.3.1 Objetivos específicos

Para o alcance do objetivo geral deste trabalho, desdobramos os esforços para alcançar os seguintes objetivos específicos:

a) Determinar quais são as principais emissoras de conteúdo jornalístico no Ceará, presentes na WWW, compreendendo os jornais online e blogs jornalísticos profissionais.
b) No extrato considerado nesta pesquisa, determinar o ciclo de vida de uma publicação jornalística, visando conhecer o quanto as publicações são modificadas ao longo deste ciclo.
c) Determinar o volume de publicações que ocorrem em um período de 30 dias.
d) Determinar as principais categorias de estruturas técnicas presentes na composição do HTML e que contenham de fato o conteúdo jornalístico
e) Determinar dentro das categorias definidas no objetivo anterior, as estruturas técnicas usadas na composição do HTML, as principais dificuldades no tratamento dos dados visando a obtenção do conteúdo jornalístico.
f) Formular estrutura ideal de um arquivo HTML com base nos resultados os objetivos (D) e (E).
g) Apresentar heurísticas de coleta de dados a partir das categorias e estruturas apontadas nos objetivos (D), (E) e (F).

1.4 Organização da monografia

Esta monografia foi organizada em 5 capítulos. No primeiro capítulo introduzimos a problemática, os atores envolvidos e os objetivos deste trabalho. No segundo capítulo trazemos a necessária conceituação teórica a sustentar o presente trabalho. O terceiro capítulo descreve a metodologia deste trabalho, enquanto o quarto capítulo apresenta os dados obtidos na pesquisa. O quinto capítulo traz as análises e conclusões iniciais. Por fim, o sexto capítulo traz as conclusões finais.

2 CONCEITUAÇÃO TEÓRICA

Neste capítulo apresentamos os termos e conceitos centrais ao entendimento do estudo presente nesta monografia. Inicialmente abordamos o HTML - sua história, desenvolvimento e atual estado - e na sequência apresentamos alguns conceitos sobre sua contraparte do lado do usuário final: navegadores da web.

2.1 HTML

A história do HTML está intrinsecamente conectada com a história da web. A linguagem HTML foi criada por Tim Berners-Lee, em 1990. Junto com outras tecnologias como os *Uniform Resources Identifiers* (URIs), compôs o que ficou conhecido como a Web 1.0. (CHOUDHURY, 2014).

Como o nome introduz, HTML é uma linguagem utilizada para definir a marcação dos elementos que compõem documentos a serem exibidos em um navegador de web (MUSCIANO e KENNEDY, 2000). Notavelmente, o HTML não deve ser pensado como uma ferramenta para produção visual de documentos. A real área de atuação da linguagem está na estruturação informacional. Ela nos possibilita definir áreas e elementos cujos navegadores entenderão como tendo finalidades específicas.

A estrutura do HTML é definida por três elementos principais: *tags*, atributos e conteúdos.

Figura 1 - Exemplo de documento HTML

```
<html>
     <head>
          title>Título da Página</title>
     </head>

     <body>
          <p>
               Exemplo de estrutura simples de uma página HTML na <i>web</i>
          </p>

          <p hidden="true">Este parágrafo está ocultado</p>
     </body>
</html>
```

Fonte: Autor.

Segundo WHATWG(2022b), *tags* são geralmente representadas por um elemento de abertura e um elemento de fechamento, este elemento posterior podendo ser omitido em algumas situações. No exemplo acima podemos notar as *tags*: *html*, *head*, *title*, body, *p* e *i*. Atributos são características agregadas a uma determinada *tag*. A segunda *tag p* possui o atributo "*hidden"* com o valor "*true*", indiciamento que o elemento estará oculto na representação visual do documento por um navegador. Conteúdos são os dados pertencentes a uma determinada *tag*. O conteúdo da *tag i* no exemplo acima é "web", enquanto o conteúdo da tag *body* são as duas *tags p* incluindo seus conteúdos. Cada um desses elementos é extensivamente definido e curado pelo padrão HTML.

Dentro do padrão HTML, o grupo WHATWG (2022b) disponibiliza um breve histórico da linguagem: Nos primeiros cinco anos de existência do HTML, suas definições eram mantidas por membros do *Conseil Européen pour la Recherche Nucléaire* (CERN) e não muito depois pela *Internet Engineering Task Force* (IETF). Em 1994, com a criação do *World Wide Web Consortium* (W3C), o HTML muda de mãos novamente. De então até 1997 - as versões 3.0, 3.2 e 4.0 do HTML foram desenvolvidas em rápida sucessão.

No ano seguinte, em 1998, o W3C decidiu parar a evolução da linguagem HTML para focar-se no que seria sua nova e melhorada sucessora: *Extensible HyperText Markup Language* (XHTML). Uma versão que estende e modifica as funcionalidades do HTML por meio de outra linguagem de marcação: a *Extensible Markup Language* (XML).

Uma das principais diferenças entre HTML e XHTML se dá na diferença de tratamento que as duas linguagens guardam para documentos malformados, documentos contendo erros no uso das estruturas definidas pelos seus respectivos padrões. Em HTML, navegadores tentam projetar o máximo possível a partir de um documento malformado. Enquanto no XHTML, devido a sua relação com XML, caso determinado documento seja malformado, o processo se finaliza antes mesmo de chegar na etapa de projeção.

Nos anos seguintes até meados de 2004, o desenvolvimento do HTML se encontra em um limbo. Por um lado, a entidade mantenedora do padrão HTML da época decidiu por abandonar seu desenvolvimento, e por outro temos as empresas desenvolvedoras de navegador que lutam, sem sucesso, para estender a linguagem em sua versão 4.0.

Em 2004, durante um workshop da W3C organizado pela Adobe, uma proposta pela reabertura do desenvolvimento do padrão HTML foi apresentada conjuntamente pelas empresas Mozilla e Opera. Membros da W3C vetaram a proposta, com votação final de 8 a favor e 11 contra (W3C, 2004). O motivo dado: a proposta vai contra a decisão prévia da instituição de que a evolução da web se encontra com o XHTML.

Pouco tempo depois - as empresas Apple, Mozilla e Opera anunciaram que iriam dar elas mesmas continuidade ao desenvolvimento do padrão HTML sob a bandeira *Web Hypertext Application Technology Working Group* (WHATWG).

Com o grupo WHATWG direcionado por membros das empresas desenvolvedoras de navegadores - Apple, Mozilla, Opera e futuramente também Microsoft e Google - não é surpresa que os navegadores respectivos de cada empresa escolheram em suas evoluções trabalhar com os padrões de HTML agora definidos pela WHATWG, ao invés de seguir a W3C e seu plano de uso para XHTML.

O grande problema do padrão XHTML observado pelo grupo WHATWG se dá pelo viés da retrocompatibilidade. Para eles, a adoção do XHTML demonstrava perigo real de perda de conteúdos antigos na web. A retrocompatibilidade é valorizada tanto pelo grupo que tal conceito é nomeado como um dos pilares principais da organização.

Em 2006 o grupo W3C decidiu retornar a trabalhar com HTML e de forma conjunta com a WHATWG. Anos depois, em 2011, ambos os grupos perceberam que cada um estava trabalhando com objetivos diferentes. Por um lado, a W3C tinha como objetivo publicar uma versão finalizada do padrão "HTML5", enquanto a WHATWG visava trabalhar com o que chamam de "padrão vivo", onde ao invés de decidir padrões com versões definitivas se tinha apenas um padrão que seria trabalhado ao passar dos anos.

Em 2019 os dois grupos assinaram um termo de colaboração de desenvolvimento de uma única versão do HTML (W3C, 2019). Esta versão seguiria os moldes do "padrão vivo" da WHATWG.

Hoje o "padrão vivo" do HTML pode ser encontrado no site da WHATWG, *https://html.spec.whatwg.org*, disponível em acesso aberto sob a licença *Creative Commons Attribution 4.0 International License (CC BY 4.0)*.

Já observamos acima, mas devemos também frisar o fato que o padrão HTML está à mercê da implementação por meio das desenvolvedoras de navegadores. O padrão pode ser público e de livre acesso, mas cabe a cada empresa decidir quanto, como e até mesmo se determinados pontos do padrão serão implementados em seus navegadores. Tal questão gerou o nascimento de documentos como o site *https://caniuse.com/*, onde pode-se pesquisar sobre como anda o processo de implementação do padrão HTML em determinados navegadores. Além disso, problemas também são gerados para desenvolvedores de websites, que devem não só ter em mente o HTML na hora de sua utilização, mas como também as características específicas de cada navegador, já que certas diferenças podem acarretar diferenças de grande cunho onde um mesmo site pode ter estrutura diferente se aberto em navegadores distintos.

2.2 JANO

O JANO (JANO, 2022) é um serviço disponibilizado na web através do endereço *http://51.161.52.61*, e que provê serviços de coleta e análise de dados. O processo de coleta é automático e o usuário pode indicar termos e assuntos de seu interesse. O usuário pode também indicar quais fontes de notícias serão coletadas automaticamente.

A partir de uma definição de quais websites a serão monitorados - a plataforma JANO captura as páginas desejadas, incluindo opcionalmente páginas cujos links estão listados nelas, em intervalos de 10 minutos. Ao salvar uma página, o sistema analisa seu conteúdo com o objetivo de identificar páginas com conteúdo previamente capturado para evitar duplicatas. Ao processar a página, o JANO produz: título da página, resumo automático do assunto principal, palavras-chave e sentimento do conteúdo. O sentimento gerado pelo sistema pode ser positivo, negativo ou neutro, de acordo com a natureza do texto ingerido.

Com as páginas coletadas, o sistema dá continuidade no trabalho com funcionalidades de filtros de assunto. Eles são configurados a partir da definição de a quais websites ele será aplicado e o conjunto de termos a serem buscados, bem como a forma de busca. Atualmente o JANO trabalha com filtros de detecção de humor, detecção de termo, contador de ocorrências de termo e detecção de múltiplos termos.

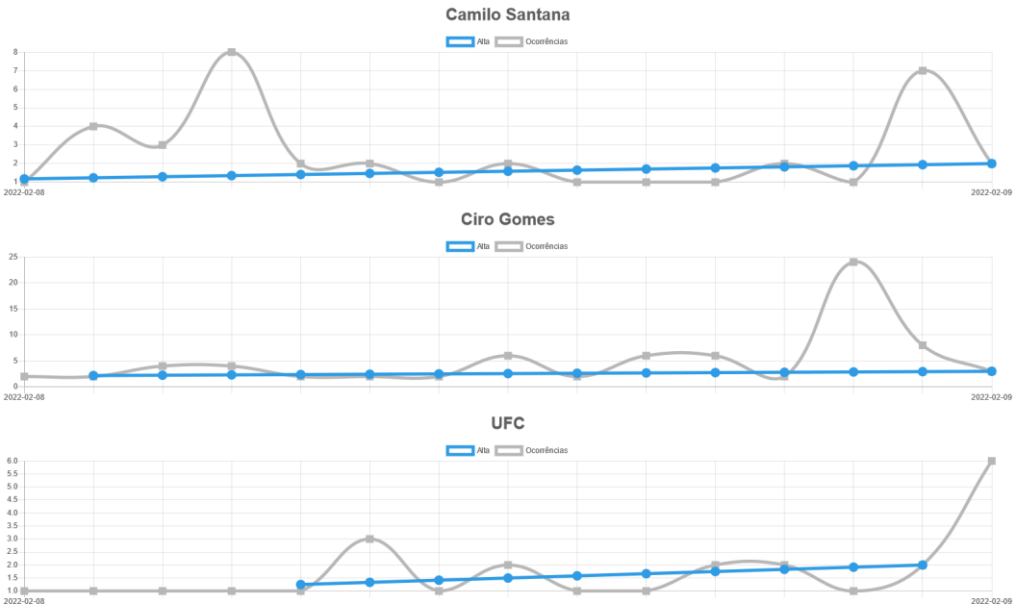
Na ativação de um filtro, dois processos ocorrem: o primeiro é a realização de ações ligadas ao filtro, enviar dados sobre a página onde o filtro foi ativado para um determinado e-mail de cliente, por exemplo. o segundo processo diz respeito à criação de registros específicos sobre a ativação, que serão utilizados na criação dos relatórios disponibilizados pelo sistema.

Um dos relatórios que o JANO disponibiliza, apresentando na figura 2, se chama "Ao vivo", ele apresenta dados de monitoramento de sentimentos gerais e subdivididos por emissora. Outro tipo de relatório, que pode ser observado na figura 3, apresenta tendências de alta nos últimos dias baseado nos termos de filtragem cadastrados.

Figura 2 - Relatório "Ao vivo" do JANO

Fonte: JANO, 2022

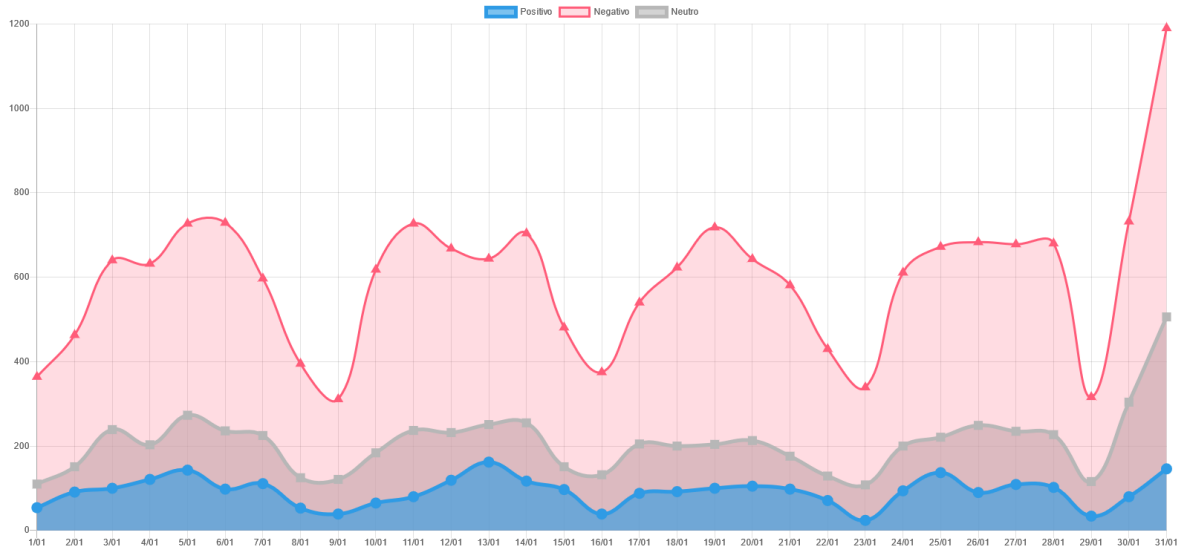
Figura 3 - Relatório de tendências do JANO



Fonte: JANO, 2022

Também é disponibilizado pelo JANO um gráfico com dados mensais de humor, separados por dia, figura 4.

Figura 4 - Relatório de humor mensal do JANO

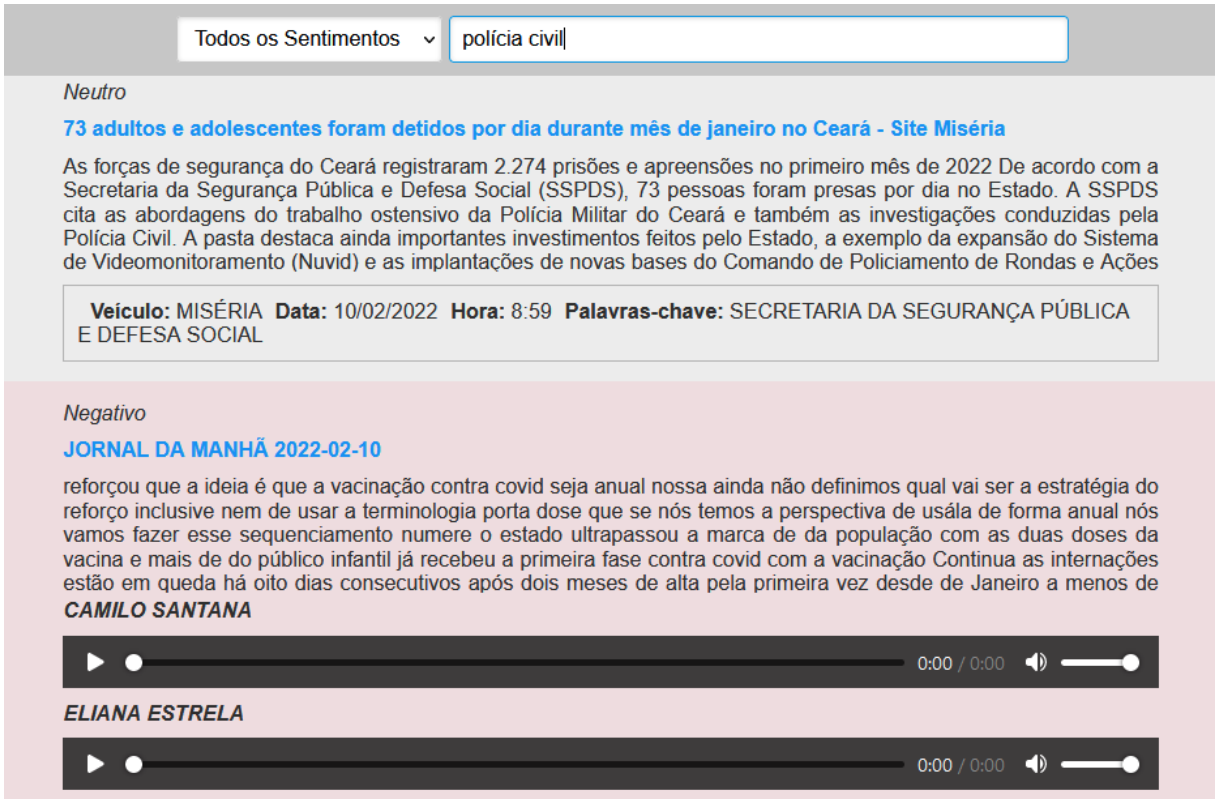


Fonte: JANO, 2022

Além de gráficos com dados agregados, a plataforma também disponibiliza para visualização os resumos automáticos de cada página, gerados no processo inicial de captação, como podemos ver na figura 5. Esta visualização, bem como todos os outros relatórios do sistema oferecem a possibilidade de filtragem por termos específicos.

A plataforma JANO não só trabalha com páginas da web, mas também oferece a possibilidade de captura e filtragem automática de programação da rádio e televisão. Para realização de suas atividades com tais mídias, o sistema se utiliza de funcionalidades de transcrição de fala para texto.

Figura 5 - Relatório de humor mensal do JANO



Fonte: JANO, 2022

3 METODOLOGIA

Considerando-se a necessidade de um contato primário com o campo do estudo e como o objeto do estudo, foi adotado uma estratégia exploratória, com o propósito de possibilitar ao pesquisador maior familiaridade com o tema e a liberdade para adotar as ferramentas mais adequadas, de acordo com a evolução, como afirma Gil, este tipo de pesquisa:

> "tem como finalidade proporcionar maiores informações sobre determinado assunto, facilitar a delimitação de um tema de trabalho. Normalmente constitue a primeira etapa de uma investigação mais ampla. Desenvolve-se com o objetivo de proporcionar uma visão geral, de tipo aproximativo, acerca de determinado fato" (GIL, 2006, p. 43).

Levando em conta que o objeto do estudo se constitui em documento digital, adota-se como procedimento metodológico a pesquisa documental, da qual nos fala GIL, "Primeiramente, há que se considerar que os documentos constituem fonte rica e estável de dados. Como documentos subsistem ao longo do tempo, tornam-se a mais importante fonte de dados em qualquer pesquisa de natureza histórica, (2002, p.45-46). E (FONSECA, 2002, p. 32), pois a pesquisa documental lança mão de fontes mais diversificadas e dispersas, com ou sem tratamento analítico, tais como: tabelas estatísticas, jornais, revistas, relatórios, documentos oficiais, cartas, filmes, fotografias, pinturas, tapeçarias, relatórios de empresas, vídeos de programas de televisão etc.

Quanto a análise, esta é realizada no presente trabalho de forma qualitativa no primeiro momento, para então no segundo momento, a análise puramente quantitativa.

Os desdobramentos do método tiveram início com uma busca no serviço de busca do Google (GOOGLE, 2022) e Bing (BING, 2022). A seleção desses dois buscadores deu-se pelo fato de serem os dois maiores buscadores disponíveis gratuitamente na WWW. Na sequência, utilizaram-se as seguintes expressões de busca: "Jornais do Ceará", "Jornal" + "Ceará", "informação Jornalística", "Portal de notícias" e "últimas notícias". Utilizando-se essas expressões foram encontrados os seguintes totais de possíveis fontes jornalísticas:

Tabela 1 - Resultado da busca inicial por emissoras jornalísticas

Expressão de Busca	**Buscador**	**Quantidade de resultados**
Jornais do Ceará	www.google.com.br	16,300,000
Jornal + Ceará	www.google.com.br	74,900,000
“informação Jornalística” + Ceará	www.google.com.br	28,800
“Portal de notícias” + Ceará	www.google.com.br	447,000
“últimas notícias” + Ceará	www.google.com.br	51,500,000
Jornais do Ceará	www.bing.com.br	5.290.000
Jornal + Ceará	www.bing.com.br	26.600.000
“informação Jornalística” + Ceará	www.bing.com.br	7.630
“Portal de notícias” + Ceará	www.bing.com.br	196.000
“últimas notícias” + Ceará	www.bing.com.br	1.030.000

Fonte: Autor, a partir de dados obtidos durante a pesquisa.

Observando-se os números de resultados encontrados, percebe-se intuitivamente a impossibilidade de analisar todas essas entradas na resposta, portanto, no sentido de delimitar o corpus da pesquisa, optou-se por analisar apenas os 100 primeiros registros obtidos em cada uma das respostas das consultas realizadas nos buscadores. Realizando-se esse procedimento, obteve-se como resultado a seguinte lista de emissoras / fontes de dados jornalísticos:

Tabela 2 - Emissoras de conteúdo jornalístico considerados na pesquisa

Emissora Jornalística	**Endereço na WWW**
AVOL - ANTÔNIO VIANA	http://antonioviana.com.br
BLOG DO LAURIBERTO	https://blogdolauriberto.com
BLOG DO ROBERTO MOREIRA	https://blogrobertomoreira.com
CEARÁ AGORA	https://cearaagora.com.br
CNEWS7	https://cn7.com.br
ELIOMAR DE LIMA	https://opovo.com.br/blogsecolunas/eliomardelima
FLÁVIO PINTO	https://flaviopintonews.com.br
FOCUS.JOR	https://focus.jor.br

Fonte: Autor a partir dos dados da plataforma JANO, 2022.

Tabela 2 - Emissoras de conteúdo jornalístico considerados na pesquisa (Cont.)

Emissora Jornalística	**Endereço na WWW**
IGUATU NOTÍCIAS	https://iguatunoticias.com
MISÉRIA	https://miseria.com.br
O ESTADO (CEARÁ)	https://oestadoce.com.br
O GLOBO	https://oglobo.globo.com
O OTIMISTA	https://ootimista.com.br
O POVO ONLINE	https://opovo.com.br
PORTAL A VOZ DE SANTA QUITÉRIA	https://avozdesantaquiteria.com.br
PORTAL DIÁRIO DO NORDESTE	https://diariodonordeste.verdesmares.com.br
POLÍTICA COM K	https://politicacomk.com.br/
SOBRAL DE PRIMA	http://sobraldeprima.blogspot.com/
UOL	https://uol.com.br/
G1 CEARÁ	https://g1.globo.com/ce/ceara
GC MAIS	https://gcmais.com.br

Fonte: Autor a partir dos dados da plataforma JANO, 2022.

Justifica-se a decisão de optar pelos 100 listados nas respostas dos respectivos buscadores, em cada uma das expressões de busca, pelo fato de que a resposta desses buscadores é influenciada pelo grau de relevância das respostas, que é resultado do volume de buscas realizados pelos usuários dos buscadores em associação com o volume de atualizações das respectivas páginas HTML dos jornais online, blogs jornalísticos profissionais.

A partir dos dados apresentados na tabela 2, teve início a fase de coleta de dados relativos às páginas HTML. Para realizar a coleta foi utilizado os recursos da plataforma JANO (JANO, 2020), e através dela, no recorte de tempo de 01 de dezembro de 2021 até 31 de dezembro de 2021, com capturas a cada 10 minutos, foram obtidos os dados que estão sumarizados na tabela 3.

Tabela 3 - Resultado da busca inicial por emissoras jornalísticas

Emissora	**Páginas HTML coletadas**	**Páginas HTML únicas**	**Páginas HTML republicadas**	**Percentual de republicação**
AVOL - ANTÔNIO VIANA	107473	2175	105298	97.98%
BLOG DO LAURIBERTO	48201	536	47665	98.89%
BLOG DO ROBERTO MOREIRA	34475	1889	32586	94.52%
CEARÁ AGORA	109367	827	108540	99.24%
CNEWS7	43647	378	43269	99.13%
ELIOMAR DE LIMA	52428	688	51740	98.69%
FLÁVIO PINTO	41544	451	41093	98.91%
FOCUS.JOR	54955	1192	53763	97.83%
G1 CEARÁ	204527	625	203902	99.69%
GC MAIS	69572	2209	67363	96.82%
IGUATU NOTÍCIAS	21580	235	21345	98.91%
MISÉRIA	78193	816	77377	98.96%
O ESTADO (CEARÁ)	161262	1824	159438	98.87%
O GLOBO	125219	4275	120944	96.58%
O OTIMISTA	88475	2266	86209	97.44%
O POVO ONLINE	159420	3286	156134	97.94%
PORTAL A VOZ DE SANTA QUITÉRIA	44821	832	43989	98.14%
PORTAL DIÁRIO DO NORDESTE	10817	125	10692	98.84%
POLÍTICA COM K	104475	45	104430	99.96%
SOBRAL DE PRIMA	20114	127	19987	99.37%
UOL	131029	4178	126851	96.81%

Fonte: Autor a partir dos dados da plataforma JANO, 2022.

Tendo completado o trabalho necessário para a coleta das páginas HTML, no recorte temporal e no conjunto de emissoras consideradas nesta pesquisa, teve início a fase de análise das páginas HTML. Dessa análise foram obtidos elementos necessários para responder ao primeiro objetivo específico, aqui relembrado: “Determinar quais são as principais emissoras de conteúdo jornalístico no Ceará, presentes na WWW, compreendendo os jornais online, blogs jornalísticos profissionais.”. Portanto, quanto a esse objetivo específico, as principais emissoras são aquelas listadas na tabela 3.

Quanto aos objetivos específicos (B) e (C) os dados necessários para esclarecimento estão contidos na Tabela 3.

Na sequência os dados, cuja sumarização se vê na tabela 3, foram analisados com a seguinte sistemática:

a) Determinar o conjunto de palavras-comando que fazem parte da estrutura normal do HTML, a fim de que seja possível identificar apenas o conteúdo útil das páginas, descartando-se o que é estrutural nesses documentos.

b) Uma vez descartados as palavras-comando dos documentos HTML, excluir da análise qualquer página HTML que contenha menos de 20 palavras.

c) O conjunto documental restante após o passo (b) foi analisado para identificar quais as estruturas de palavras-comando eram recorrentes entre os documentos que compunham o conjunto documental e dessa forma identificar as estruturas que são usadas na prática, para conter os dados úteis do conteúdo jornalístico.

d) Por fim, foi realizada uma análise para determinar quais estratégias de solução algorítmica poderiam ser utilizadas para extrair os textos de maneira automática. Essa análise é descrita mais adiante neste texto.

Tabela 4 - Principais estruturas técnicas das páginas HTML

Estruturas técnicas	**Uso preferencial**
Estruturas de Metadados	
<title></title>	Título do documento HTML, estruturado de forma que seja de fácil reconhecimento mesmo fora de contexto. Elemento opcional e limitado a um único uso por página.
Estruturas de Seções	
<article></article>	Representa uma seção completa e independente do documento. Elemento reusável e de redistribuição independente. Pode representar uma mensagem em fórum, um artigo de notícia ou jornal, um comentário ou um *widget*.
<section></section>	Representa uma seção genérica de um documento. Definida como um agrupamento temático de conteúdo, geralmente com cabeçalho. Esta estrutura não é para ser entendida como container genérico de conteúdo. Tal papel pertence à estrutura . Geralmente, esta estrutura é para ser usada apenas para conteúdos que serão explicitamente listados no delineamento da estrutura de um documento.
<aside></aside>	Representa uma seção de relação tangencial, e que pode ser considerada como separada, ao conteúdo ao redor de si. Utilizada para citações, barras laterais e propagandas.
             	Essas seis estruturas são utilizadas para representar o cabeçalho das seções onde elas se encontram. São numeradas a partir de seu nível de importância, onde <h1> é o cabeçalho mais importante e <h6> o menos importante.
Estruturas de Agrupamento	
<main></main>	Representa o elemento dominante de um documento. Apenas um elemento <main> pode existir por documento. Deve ser - obrigatoriamente - hierarquicamente correto, ou seja, deve possuir como elementos predecessores apenas os elementos: <html>; <body; e <form>, sem nomes acessíveis ou elementos autônomos customizados.

Fonte: Autor a partir dos dados de WHATWG, 2022b.

Para responder ao objetivo específico (D) serão utilizados os dados da tabela 4. Por fim, quanto ao objetivo específico (E), “Determinar nas categorias de estruturas técnicas usadas na composição do HTML, as principais dificuldades no tratamento dos dados visando a obtenção do conteúdo jornalístico.”, foram aplicados os passos descritos na análise de estratégia de solução algorítmica.

3.1 Análise de estratégias de solução algorítmica

Para determinar o nível de esforço na extração automática de texto jornalístico das páginas HTML foi lançado mão das seguintes técnicas:

a) Elencar um conjunto padrão ideal de métricas e heurísticas para a extração de texto jornalístico. Esse conjunto de métricas e heurísticas deveriam funcionar para todo caso no qual as regras de construção de páginas HTML fossem seguidas corretamente, empregando-se as palavras-comandos específicas para cada fim.
b) Analisar quais estruturas técnicas das páginas HTML usadas na prática, conforme os dados obtidos na pesquisa, falham na extração de texto jornalístico usando-se o conjunto padrão ideal de métricas.
c) Para cada grupo de falha encontrada, definir uma estratégia que poderia ser utilizada para contornar a falha.

Os dados obtidos nesta análise foram utilizados nas discussões do quarto e quinto capítulos.

4 PESQUISA

Este capítulo é dividido em duas partes. Inicialmente iremos propor uma estrutura de HTML ideal a partir das definições do padrão definido pela WHATWG, bem como uma heurística que seria usada para identificação e captura de conteúdos jornalísticos de páginas que utilizariam tal estrutura.

4.1 Casos ideais

Ao analisar o padrão HTML (WHATWG, 2022b), pode-se encontrar mais de cem *tags* descritas tanto em estrutura como em seu uso. Das tais - identificamos 12, apresentadas na tabela 4, cujas definições cabem aos nossos objetivos.

Pensando nas *tags* selecionadas, visionamos como estrutura ideal de um documento HTML, o seguinte exemplo.

Figura 6 - Exemplo de estrutura HTML ideal baseada nas *tags* selecionadas

```
<html>
    <head>
        <title>Título da Página</title>
    </head>
    <body>
        <main>
            <article>
                <h1>Título da notícia</h1>
                <p> Corpo da notícia. </p>
                <p> Cont. do corpo da notícia. </p>
                <aside>Banner de propaganda 1</aside>
                <h2>Subcapítulo da notícia</h2>
                <p> Cont. do corpo da notícia. </p>
                <p> Cont. do corpo da notícia. </p>
                <aside>Banner de propaganda 2</aside>
                <aside>Conjunto de links para notícias relacionadas</aside>
            </article>
        </main>
    </body>
</html>
```

Fonte: Autor.

Com base na estrutura acima, podemos utilizar a seguinte heurística para realizar a extração de texto jornalístico de uma determinada página:

1) Capturar o conteúdo da *tag* <*title*>.
2) Descartar todos os elementos que se encontram fora da *tag* <*main*>.
3) Descartar dentro das estruturas restantes todos os elementos que não forem as *tags* <*p*> e das tags de cabeçalho, incluindo seus respectivos conteúdos.
4) Procurar por estruturas <*article*>.
5) Se apenas uma estrutura <*article*> existir:
    a) Entender a página como contendo apenas uma notícia.
    b) Procurar nos elementos restantes o elemento de cabeçalho de maior nível e utilizar seu conteúdo juntamente com o conteúdo da tag <*title*>, previamente capturado, para definir o tema principal da página.
    c) Dos elementos restantes remover todas as *tags*, mantendo o conteúdo e a separação entre elementos de cabeçalhos e parágrafos.
    d) O texto restante será o conteúdo da notícia da página caso possua mais de 20 palavras.
6) Se mais de uma estrutura <*article*> existir:
    a) Entender a página como contendo várias notícias, cada uma sendo contida dentro de uma tag <*article*>.
    b) Para cada elemento <*article*>, realizar o seguinte procedimento:
        i) Procurar pelo elemento de cabeçalho de maior nível e utilizar seu conteúdo juntamente com o conteúdo da tag <*title*>, previamente capturado, para definir o tema principal da notícia.
        ii) Dos elementos restantes remover todas as *tags*, mantendo o conteúdo e a separação entre elementos de cabeçalhos e parágrafos.
        iii) O texto restante será o conteúdo da notícia da respectiva *tag* <*article*> caso possua mais de 20 palavras.

4.2 O estilo e estrutura do HTML utilizado nas páginas obtidas na pesquisa

Analisando a estrutura de páginas dos 24 websites de notícias escolhidos para este trabalho, observamos os seguintes fatos:

Dentre os 24 sites, três apresentam mau uso da *tag* <*title*>. Dois deles utilizando a tag de forma duplicada e um deles utiliza a *tag* preenchida com texto sem indicações do tema principal da página, optando por apenas colocar o título do site como conteúdo da *tag*.

Dez dos 24 sites, não utilizam a *tag* <*article*> em sua estruturação. Dos que utilizam, apenas dois utilizam a tag para identificar o corpo de uma notícia. Os outros oito utilizam a tag para marcar sessões genéricas ou para marcar links de navegação para outros artigos.

Apenas um site utiliza a *tag* <*section*> como marcação para conteúdo de notícia, algo que não é expressamente contra o que é delineado pelo padrão HTML, mas entendemos aqui como erro devido a existência e funcionalidade específica da tag <*article*>. Dos 23 restantes, nove não utilizam a *tag* e a utilizam de forma correta.

Apenas oito websites utilizam a *tag* <*main*>. Dois entre os oito utilizam a *tag* incorretamente, pois empregam ela mais de uma vez por página.

Apenas dez sites utilizam a *tag* <*aside*>. Dentre eles, seis utilizam a tag como marcação para o menu lateral. Dois utilizam para identificar conteúdo extra no final da página, como barra de informações sobre cookies ou como adicionar a página na barra de favoritos. E os últimos dois sites utilizam a *tag* para demarcar um bloco com links para conteúdos relacionados à notícia da página atual.

A tabela 5 apresenta um resumo desses achados da pesquisa.

Tabela 5 - Resumo dos problemas encontrados no uso do HTML nos dados da pesquisa

Tag Utilizada	**Uso esperado**	**Uso encontrado**	**Qtd. de sites com problema**	**Gravidade do problema**
<title>	Apenas uma instância por documento.	Múltiplas instâncias por documento	2	Moderada
	Conteúdo indica assunto principal da página	Conteúdo indicando apenas nome do site sem especificação de página	1	Alta
<section>	Utilizada para seções temáticas genéricas	Utilizada para representação de conteúdo principal da página	1	Alta
<article>	Utilizada para representar conjunto completo de elementos compositores de um bloco de notícia	Não utilizada	10	Alta
		Utilizada para marcar apenas seções genéricas	8	Moderada
<main>	Utilizada em apenas uma instância para indicar conteúdo principal da página	Não utilizada	16	Baixa
		Múltiplas instâncias por documento	2	Baixa
<aside>	Utilizada para marcar elementos não relacionados a uma notícia, mas que entrelaça seu conteúdo. Ex.: Anúncios no meio de uma notícia	Não utilizada para esta finalidade	22	Moderada

Fonte: Autor a partir dos dados da pesquisa.

Para gravidade do problema, presente na Tabela 5, adotamos as seguintes definições: A gravidade baixa define questões menores que não afetam a identificação de grupos de informação ou a velocidade de processamento de uma determinada página. A gravidade moderada indica problemas que afetam a identificação de grupos de informação dentro da estrutura completa. A gravidade alta indica problemas não só de mal uso de uma estrutura, como a falta completo de seu uso ou outras questões que afetam significativamente a identificação de grupos de informação dentro da estrutura completa.

Uma rápida análise dos dados contidos na Tabela 5 revelam que **todas** as 22 fontes de conteúdo jornalísticos analisados no presente estudo, apresentam **ao menos um problema moderado e** aproximadamente 45% (quarenta e cinco por cento) das fontes apresentam 2 problemas de **gravidade alta.**

Ainda com relação aos problemas apresentados na Tabela 5, foi construído um conjunto de heurísticas para ser usado na solução computacional para solucionar o problema. Este conjunto pode ser visto na Tabela 6.

Tabela 6 - Heurísticas para automatizar a solução dos problemas

Tag Utilizada	**Uso encontrado**	**Heurística desenvolvida**
<title>	Múltiplas instâncias por documento	Tratar o conteúdo de múltiplas *tags* <*title*> como contíguo na ordem que aparecerem no documento.
	Conteúdo indicando apenas nome do site sem especificação de página	Ignorar *tag* <*title*> e tentar extrair tema principal da página a partir da tag de cabeçalho com maior rank(h1>h6).
<section>	Utilizada para representação de conteúdo principal da página	Analisar quantidades de *tags* <*p*> dentro do conteúdo de uma tag <*section*> para determinar se ela contém o corpo de uma notícia.
<article>	Não utilizada	Ao invés de procurar por esta tag, procurar pelo menor elemento HTML que possui a maior quantidade de *tags* <*p*> em seu conteúdo.
	Utilizada para marcar apenas seções genéricas	Descartar tags <*article*> caso em seu conteúdo não existirem tags <p>. Caso *tags* <*p*> existam, avaliar a quantidade de palavras em todas elas para determinar se elas representam conteúdo principal da notícia ou não.
<main>	Não utilizada	Ignorar a procura pela *tag* <*main*> e continuar com a busca por *tags* <*article*>.
	Múltiplas instâncias por documento	Tratar o conteúdo de múltiplas *tags* <*main*> como contíguo na ordem que aparecerem no documento.
<aside>	Não utilizada para finalidade esperada	Dentro do conteúdo de uma notícia, procurar por elementos genéricos diferentes das *tags* <*p*>

Tag Utilizada	Uso encontrado	Heurística desenvolvida
		e das de cabeçalho para remoção.

Fonte: Autor a partir da análise dos dados da pesquisa.

No que diz respeito aos problemas de baixo impacto encontrados, ambos os casos estão relacionados com a *tag* <*main*>. A função desta *tag* no cenário ideal ajudaria apenas na limitação da área de trabalho, logo, sua ausência ou múltiplo uso geraria apenas um problema de performance, já que o algoritmo de processamento terá de percorrer todo o documento ou dois grupos de elementos tratados como um só, ao invés de uma parte específica já denominada como central.

Já para problemas moderados, temos três casos. O primeiro diz respeito ao uso de múltiplas *tags* <*title*>, que deveria ser de instância única. Adotamos como solução o tratamento dos conteúdos de *tags* repetidas como contíguos na ordem que aparecem no documento. Em casos onde não só múltiplas *tags* <*title*> existem, mas também estão preenchidas com conteúdos de assuntos muito divergentes, a identificação de temas pode ser bastante difícil. Mas em todos os casos estudados, em documentos onde as *tags* se repetem, seus conteúdos são idênticos ou diferem-se apenas por sufixos identificadores de site no final de seu conteúdo.

O segundo problema moderado está ligado com o uso de *tags* <*article*> para definir seções genéricas de uma página e não o grupo de elementos de uma notícia. Neste caso, a identificação de *tags* <*p*> dentro de <*article*> possibilita a fácil confirmação da natureza da tag, já que a utilização de *tags* <*p*> se encontrada limitada para o conteúdo de notícias nos sites analisados.

O último problema moderado está ligado à falta de uso da *tag* <*aside*> para identificação de conteúdo tangencial, anúncios por exemplo, dentro do corpo de uma notícia. Além de proporcionar uma queda significativa de performance, já que várias *tags* terão de ser analisadas para remoção, existe também a possibilidade de deleção ou permanência de *tags* extras devido à grande variância de possibilidades nesta área.

Também temos uma quantidade de três problemas de gravidade alta. O primeiro deles se trata da má utilização da *tag* <*title*>, sendo preenchida apenas com o nome do site sem nenhuma especificação do documento específico atual. Neste caso, o assunto principal da página tentativamente seria extraído da *tag* de cabeçalho com maior rank(h1>h6). Mas como

as tags de cabeçalho são genéricas, a precisão da definição de assunto pode cair consideravelmente nesses casos.

Os dois outros problemas de alta gravidade estão interligados. são eles: o não uso das *tags* <*article*> e o consequente uso de *tags* <*section*> para demarcar o conteúdo principal de uma página. Nesses casos, novamente a velocidade do processamento tende a cair pois mais elementos serão analisados e de forma mais profunda. Neste caso em específico, todas as *tags* <*section*> serão analisadas para detectar a existência do conteúdo principal através das *tags* <*p*>.

Devido a variedade tanto na natureza quanto na extensão de cada problema, ainda existirão casos em que uma certa quantidade de capturas irá falhar, mas devido à aplicação das heurísticas definidas acima, tal quantidade cairá consideravelmente.

5 ANÁLISE DOS DADOS

Aqui apresentaremos nossos achados em relação aos dois lados discutidos no capítulo anterior: o padrão HTML em seu estado ideal e como o HTML é utilizado na prática.

5.1 Padrão HTML

Ao estudar o padrão HTML e sua história, o ponto mais notável é a relação entre padrão e as empresas desenvolvedoras de navegadores. O padrão HTML não é algo imposto de cima para baixo, mas sim algo desenvolvido de forma colaborativa e aberta - e aceito entre os pares - especialmente após tomada de responsabilidade a partir do grupo WHATWG. As discussões abertas podem ser facilmente acessadas a partir da página do GitHub, *website* host de projetos de desenvolvimento de software e provedor de serviços de versionamento, do grupo (WHATWG, 2022a).

No que diz respeito ao padrão HTML (WHATWG, 2022b) e como ele define suas estruturas, para fins deste trabalho consideramos as estruturas delineadas com grande satisfação. O padrão disponibiliza e define estruturas de grande serventia para os processos de identificação e coleta de conteúdo jornalísticos. Além disso - as notas, exemplos e sugestões de uso vão também em direção a esses objetivos.

Nota-se também, que as *tags* selecionadas se encontram completamente implementadas por todos os navegadores modernos de maior relevância - Chrome, Firefox, Safari, Opera, Edge e suas respectivas versões para dispositivos móveis - em suas versões mais recentes. Assim o uso dessas estruturas por parte dos desenvolvedores de websites se apresenta sem fricção adicional.

5.2 O HTML da vida real

Os problemas na identificação e captação de textos jornalísticos encontrados durante a produção deste trabalho tem como origem a mesma fonte: o modo de uso do HTML por meio dos desenvolvedores de website. Como foi observado no capítulo anterior um número significativo dos sites selecionados para a pesquisa - ou não utiliza uma ou mais das *tags* levantadas, ou utiliza tais tags de forma incorreta.

Supomos que tal fato ocorre devido a dois grandes motivos:

- Primeiramente, na visão de um desenvolvedor de website, tags como <*aside*>, <*article*> e <*main*> não possuem utilidade alguma caso o mesmo não tenha como parte em sua visão fatores além dos que apenas influenciaram diretamente o consumidor final de seu website - o leitor, o usuário. Pois, como definido pelo padrão html, tais *tags* não possuem funcionalidade prioritariamente visual, logo a utilização delas seria trabalho extra sem garantia de retorno direto algum.
- Como segundo motivo, creditamos a existências de tags de uso genérico existentes no padrão, como a tag <*div*>. Por serem genéricas e maleáveis, os desenvolvedores dão preferências ao uso de tais tags, pois assim diminuem o número de ferramentas à sua disposição e, por consequência, acabam diminuindo a carga cognitiva necessária para realização de suas atividades

Ambas as questões são de difícil solução, pois se tratam de questões impregnadas na prática, e até mesmo ensino, da área de desenvolvimento de websites. Enquanto os indivíduos da área não colocarem uma boa estrutura HTML - não só para acesso por parte do usuário final, mas também para coleta por parte de mecanismos automáticos - os problemas descritos neste trabalho continuarão a ocorrer.

6 CONCLUSÕES

Com a realização deste trabalho observamos inicialmente o quão bem-posicionada a linguagem HTML se encontra para estruturar páginas da web, apesar das limitações que a seguem desde o berço. Para aqueles que tem como objetivo desenvolver websites com estruturas que facilitam seu processamento e captação por ferramentas automáticas, o padrão HTML disponibiliza amplas ferramentas para atingir tal objetivo.

E por outro lado pudemos entender as deficiências no uso do HTML na prática por parte de desenvolvedores de sites jornalísticos. Erros esses que nos apontam um cenário onde o pensamento dos responsáveis pela criação de tais páginas está voltado apenas para o usuário final e não para o contexto informacional da web.

Além disso, durante a realização deste trabalho encontramos uma possível linha de questionamento para a produção de outro: estudar como se dá o ensino e aprendizado de HTML e se o modo atual de realização de tais atividades seriam causa das deficiências no uso da linguagem encontradas.

Por fim, apresentamos a dificuldade em desenvolver heurísticas generalizadas para a identificação e captação de textos jornalísticos, justamente devido às deficiências no uso do HTML nas páginas selecionadas. Com problemas de tantos diferentes níveis e escalas a generalização de uma solução completa se torna impossibilitada, forçando-nos a pensar em soluções específicas para cada caso. Concluímos, portanto, que a tarefa de extrair dados dos espaços digitais é complexa, pois os documentos não são estruturados de maneira a identificar facilmente o conteúdo real dele, mensurado como o texto útil e informação presente no mesmo. Da forma como se encontra hoje, a WWW, cuja maioria dos conteúdos disponíveis é estruturado em HTML, torna-se um campo de grande dificuldade de compreensão e extração de dados através de sistemas computacionais.

REFERÊNCIAS

BING. **Bing**, 2020. Disponível em: https://bing.com. Acesso em: 03/01/2022.

CHOUDHURY, N. **World Wide Web and its Journey from Web 1.0 to Web 4.0**. International Journal of Computer Science and Information Technologies, v. 5, n. 6, p. 8096-8100, 2011. Disponível em: http://ijcsit.com/docs/Volume%205/vol5issue06/ijcsit20140506265.pdf. Acesso em: 20/12/2021.

FONSECA, J. J. S. **Metodologia da pesquisa científica**. Fortaleza: UEC, 2002. Apostila.

GIL, A. C. **Métodos e técnicas de pesquisa social**. 5.ed. São Paulo: Atlas, 1999. Disponível em: https://docero.com.br/doc/nxs1n8x. Acesso em 31/12/2021.

GOOGLE. **Google**, 2022. Disponível em: https://google.com. Acesso em: 03/01/2022.

JACKSI, K. ABASS, S. M. **Development History Of The World Wide Web**. International Journal Of Scientific & Technology Research, v. 9, n. 9, p 75-79, Setembro, 2019. Disponível em: https://www.researchgate.net/profile/Karwan-Jacksi/publication/336073851_Development_History_Of_The_World_Wide_Web/links/5d8d1f8f92851c33e94064cb/Development-History-Of-The-World-Wide-Web.pdf. Acesso em: 13/12/2021.

JANO. **JANO**. Disponível em: http://51.161.52.61. Acesso em: 10/01/2022.

MUSCIANO, C. KENNEDY, B. **HTML & XHTML**: The Definitive Guide. 4. ed. O'Reilly, 2000.

W3C. **Memorandum of Understanding Between W3C and WHATWG**, 2019. Disponível em: https://www.w3.org/2019/04/WHATWG-W3C-MOU.html. Acesso em: 05/01/2022.

W3C. **The W3C Workshop on Web Applications and Compound Documents**, 2004. Disponível em: https://www.w3.org/2004/04/webapps-cdf-ws/summary. Acesso em: 05/11/2021

WHATWG. **whatwg / html: HTML Standard**, 2022a. Disponível em: https://github.com/whatwg/html/. Acesso em: 05/01/2022.

WHATWG. **Html Standard**, 2022b. Disponível em: https://html.spec.whatwg.org. Acesso em: 05/01/2022.

WHATWG. **Html Standard FAQ**, 2022c. Disponível em: https://github.com/whatwg/html/blob/main/FAQ.md. Acesso em: 05/01/2022.

WHATWG. **WHATWG - FAQ**, 2022d. Disponível em: https://whatwg.org/faq. Acesso em: 05/01/2022.