MC102 - Algoritmos e Programação de Computadores
Turmas QRSTWY
Instituto de Computação - Unicamp
Professores: *Hélio Pedrini e Zanoni Dias*
Monitores: *Andre Rodrigues Oliveira, Gustavo Rodrigues Galvão, Javier Alvaro Vargas Muñoz e Thierry Pinheiro Moreira*

Lab 12a - Tweet Extractor

Prazo de entrega: 08/06/2015 às 13h59m59s
Peso: 8

Linguagem de marcação é um sistema para anotação de um texto de modo que ele seja sintaticamente distinguível. Diferente das linguagens de programação, as linguagens de marcação definem formatos, maneiras de exibição e padrões dentro de um documento qualquer. As linguagens de marcação são usadas, por exemplo, na indústria editorial para marcar a formatação (exibição gráfica) de páginas. Se o código de marcação for padronizado, ou puder ser processado por um programa de computador, então garante-se o intercâmbio de uma publicação complexa entre autores, editores e impressoras. Um exemplo famoso de linguagem de marcação é o HTML (*HyperText Markup Language*), que é utilizado para produzir páginas na Web.

Um documento HTML é um conjunto de *elementos*, usualmente demarcados por duas *tags*: a *tag inicial*, no formato `<nome-do-elemento atributo(s)>`, e uma *tag final*, no formato `</nome-do-elemento>`. Um *atributo* define uma característica ou propriedade de um elemento, e, sempre que existir, é incluído na tag inicial de um elemento utilizando a sintaxe `nome_do_atributo="valor"`. Além disso, um elemento pode ter vários atributos, separados por espaços em branco. O conteúdo de um elemento geralmente é um texto que fica entre as tags inicial e final, e nele também podem existir outros elementos. É garantido que não existem caracteres "<" e ">" em outras partes de um código HTML que não sejam delimitando tags. O exemplo abaixo é um trecho de um documento HTML:

```
<p> Texto simples, agora <b>em
negrito</b> e agora <i>em itálico</i>.
Também pode ser <b><i>em negrito e
itálico ao mesmo tempo</i></b>. </p>

<p align="right"> Este parágrafo está à
direita! </p>

<p align="center"> Este parágrafo está
no centro e
<font style="color:red">este texto está
em vermelho</font>. </p>
```

Texto simples, agora **em negrito** e agora *em itálico*. Também pode ser ***em negrito e itálico ao mesmo tempo***.

Este parágrafo está à direita!

Este parágrafo está no centro e este texto está em vermelho.

Neste laboratório, dado uma página HTML de um usuário do Twitter, sua tarefa será extrair apenas os textos dos tweets do usuário.

Dicas:

- No código HTML da página do Twitter, os tweets do usuário estão dentro de uma tag `p` cujo atributo `class` possui valor `"ProfileTweet-text js-tweet-text u-dir"`.

- Como aspas simples (') e aspas duplas (") são caracteres de sintaxe HTML, quando vamos utilizá-los em um texto dentro de um arquivo HTML é adequado convertê-los em *entidades*. Uma entidade tem como prefixo o caractere `'&'` ("e" comercial), e como sufixo o caractere `';'` (ponto-e-vírgula) e, entre o prefixo e o sufixo, um identificador do caractere. Um caractere pode possuir mais de uma entidade, porém, as entidades dos caracteres aspas simples e aspas duplas utilizadas pelo Twitter são, respectivamente, `&#39;` e `&quot;`. Seu programa deve extrair os tweets do usuário convertendo estas duas entidades pelo caractere correspondente. Além das duas entidades citadas, outras entidades poderão estar presentes nos tweets, e devem apenas ser ignoradas (ou seja, devem ser suprimidas no texto final).
- Não se preocupe caso o tweet seja um "retweet", contenha links ou imagens. Seu programa deve apenas imprimir os textos correspondentes que estiverem entre as tags.
- Abra os arquivos HTMLs correspondentes às entradas abaixo num editor de texto e procure as regiões onde estão localizados o tweets, conforme a primeira dica acima, de forma que você veja o código completo de um tweet e possa compará-lo com a saída esperada (veja nos exemplos abaixo).

Importante:

- Utilize o código fornecido na seção "Arquivos auxiliares" como ponto de partida. Você deve incluir o que for necessário para resolver este laboratório e não deve alterar nada do que já existe nesse código.
  - No código, além da função `main`, é fornecida a função `converte_entidade`, que descobre se uma entidade corresponde aos caracteres aspas simples ou aspas duplas, imprimindo-os diretamente no arquivo de saída (ignorando qualquer outra entidade). Para mais detalhes sobre esta função, veja os comentários no código fornecido.
- Para testar este laboratório, a linha de comando é a seguinte:

  ```
  ./lab12a arqXX.in arqXX.out
  ```

  sendo `arqXX.in` o arquivo que contém o código HTML de uma página do Twitter (corresponde a `nomearqin` citado abaixo) e `arqXX.out` o arquivo onde os tweets serão escritos (corresponde a `nomearqout` citado abaixo) e que deve ser comparado com o arquivo `arqXX.res` correspondente. Note que os arquivos com extensão "`.in`" e "`.res`" são os arquivos fornecidos na seção "Testes".
- Neste laboratório você não deve o usar o Script para Auxílio nos Testes dos Laboratórios, já que ele não está preparado para lidar com o modo de execução mencionado acima.

Entrada

- A entrada é composta por duas sequências de caracteres `nomearqin` e `nomearqout` fornecidas na linha de comando tais que:
  - `nomearqin` é igual ao nome do arquivo contendo um código HTML de uma página do Twitter, com $1 \leq |\texttt{nomearqin}| \leq 25$.
  - `nomearqout` é igual ao nome do arquivo em que seu programa deverá imprimir os tweets, com $1 \leq |\texttt{nomearqout}| \leq 25$.

Saída

- Seu programa deve imprimir todos os tweets encontrados na página HTML, um por linha.

Exemplos

Atenção: devido o tamanho dos arquivos HTML o campo de entrada das tabelas abaixo contém o link

para o conteúdo real.

#	Arquivo nomearqin	Conteúdo do arquivo nomearqout
1	arq01.in	Check this off your bucket list. http://ofa.bo/s4MQ pic.twitter.com/0Dr0V3czMH "Time and time again, @SenatorReid stood up to special interests and made sure every one of his constituents had a voice." —President Obama Wouldn't it be nice to fly to D.C. to meet President Obama? Here's your chance: http://ofa.bo/g4Gw If you believe in this grassroots movement for change, make an investment in its future—chip in today: http://ofa.bo/q4J2 "There's nothing we can't do if the American people decide it's time." —President Obama "We're also our brother's keeper ... our sister's keeper. We're also a country that believes everyone gets a fair shot." —President Obama "Wall Street reform, what we passed five years ago, is protecting working families and taxpayers." —President Obama "I want to invest in basic research so the jobs and industry of the future take root here." —President Obama "The deficit has come down by two-thirds since I've been president." —President Obama "I want to put more people back to work—rebuilding our roads and our bridges, modern ports, faster trains, faster internet" —President Obama "These ideas are not about ideology—the reason we proposed these ideas is because we know they work." —President Obama "Two years of community college should be as free and universal as high school is today." —President Obama #CollegeOpportunity "Thanks to the hard work of the American people, America's coming back." —President Obama "More than 16 million Americans have gained the security of health insurance." —President Obama #BetterWithObamacare "In America, if you work hard, you can get ahead." —President Obama LIVE: President Obama is speaking about the economy at Lawson State Community College in Birmingham, Alabama. http://ofa.bo/f4Kq Tune in at 4:10 p.m. ET to watch President Obama speak about the economy at Lawson State Community College: http://ofa.bo/r4L1 Add your name. Get entered to win a trip for two to D.C. Meet the President. It's that easy. http://ofa.bo/p4Oa Get ready—enter for your chance to meet President Obama: http://ofa.bo/r4KO pic.twitter.com/ilxzLTtOmG Read how these three women's lives and careers are #BetterWithObamacare: http://ofa.bo/r4Jx
2	arq02.in	Since 2010, children’s and mothers’ lives have been saved at the fastest rate ever: http://b-gat.es/1xw4Klb pic.twitter.com/P7stuwjFpY The new book “Becoming Steve Jobs” (http://b-gat.es/19NhQQn) has me thinking of my old friend. A true visionary. #TBT pic.twitter.com/SpSXXRUTLD Very interesting. A new test could help us understand how malaria evolves: http://b-gat.es/19Nh6uA pic.twitter.com/p2zBdYzwuk The world needs to arm itself for a different type of war—a war against germs: http://b-gat.es/1FTNAyO pic.twitter.com/eFOiuo9ipv 7 things we'll need to strengthen before the next epidemic: http://b-gat.es/18TyE7l pic.twitter.com/3eTpsSMTCy The next epidemic could be 1,000x worse than Ebola. Here’s what I hope we do about it: http://b-gat.es/1ALLQ5A pic.twitter.com/LUhzdi4vBq My favorite business book and a guide to lying with statistics. 6 books I recommend: http://b-gat.es/1Ofy4kN #TED2015 pic.twitter.com/Nuqm7LMvgq This suit is hot enough in Vancouver. What would it be like treating patients in W Africa? http://b-gat.es/197GgCY

		pic.twitter.com/5BY1LWGFIK Microbes—not missiles—are what could kill 10M people: http://b-gat.es/1EwAR0W pic.twitter.com/9dTd2MHraj Getting new tools to the people fighting Ebola is harder than it needs to be: http://b-gat.es/1ALGF5A pic.twitter.com/ikkKdo9BR4 "Failure to prepare could allow the next epidemic to be dramatically more devastating than Ebola." — @BillGates http://t.ted.com/N5EFpHB We're not ready for the next epidemic. Here's what it will take: http://b-gat.es/1O9oRug pic.twitter.com/S8mrxF7Do5 Ebola suits make everything more difficult. Even writing a simple message like this: http://b-gat.es/1FFcty8 pic.twitter.com/jC4dpYEHh0 In the mock Ebola ward at #ted2015. It is amazing to me that people can get anything done in these suits. pic.twitter.com/ZneG0EGUat How to take off Ebola protective equipment. At the #TED2015 Suiting Up for Ebola experience. https://vine.co/v/OVx3b6LlgOw So @BillGates set up an Ebola clinic at #TED2015 and I got to pretend to be a aid worker http://recode.net/2015/03/18/inside-the-mock-ebola-hospital-bill-gates-set-up-at-ted-video/… pic.twitter.com/M6X3LpnA6i The world isn't prepared to handle a massive epidemic. But we can get there: http://b-gat.es/1DBrrpX http://b-gat.es/1bghrGE Thanks for taking the time to read it, @Atul_Gawande. Your work was a real inspiration for it. What did we learn from Ebola? We need to be better WHEN the next epidemic hits: http://b-gat.es/19zudiL I think @WarrenBuffett's a better investor today than ever before. Here's why: http://b-gat.es/1FtsLdS pic.twitter.com/llDKe7hKu5
3	arq03.in	Happy New Year ! pic.twitter.com/quJvHqWw8A HAPPY ANNIVERSARY, LINUX! Free Intro to Linux Online Course Starts Today http://www.linux.com/news/featured-blogs/200-libby-clark/782685-free-intro-to-linux-online-course-starts-today… via @linuxfoundation Happy SysAdmin Day! Linux Nears Total Domination of the Top500 Supercomputers http://www.linux.com/news/enterprise/high-performance/147-high-performance/778179--linux-nears-total-domination-of-the-top500-supercomputers… via @linuxfoundation Thank you for wishing me Happy Birthday :) The Top 10 Best Linux Videos of 2013 http://www.linux.com/news/featured-blogs/200-libby-clark/752470-best-linux-videos-of-2013… via @linuxfoundation Linux for Workgroups Linux 3.11s feature set now confirmed http://www.h-online.com/open/news/item/Linux-for-Workgroups-Linux-3-11-s-feature-set-now-confirmed-1917712.html/from/twitter… LinuxCon Japan -Presentations http://events.linuxfoundation.org/events/linuxcon-japan/program/presentations… 6 Key New Features in Linux 3.9 http://www.linux.com/news/software/linux-kernel/716624-6-key-new-features-in-linux-39… via @linuxfoundation 8,000 developers from 800 companies have contributed to the Linux kernel http://bit.ly/Ha9sK1 Linux Foundation Training Prepares the International Space Station for Linux Migration https://www.linux.com/news/featured-blogs/191-linux-training/711318-linux-foundation-training-prepares-the-international-space-station-for-linux-migration… via @linuxfoundation Support #Linux and its creator Linus Torvalds by joining @linuxfoundation: http://bit.ly/Ra12D9 Friday Funnies http://www.linux.com/news/friday-funnies/good-old-days… via @linuxfoundation 'The Linux kernel' website updated https://www.kernel.org/ Sonar Project Wants to Bring Linux to Everyone http://www.linux.com/news/software/applications/701409-sonar-project-wants-to-bring-linux-to-everyone… via @linuxfoundation

		@Linus__Torvalds well respected love the work been a supporter of open source for many years now great work Divers ahoy! Lots of improvements, including much better support for multiple dive computers, new deco calculations. Intelligence is the ability to avoid doing work, yet getting the work done.
4	arq04.in	A presidenta @dilmabr lamentou a queda do avião da Germanwings e apresentou seus sentimentos e votos de luto http://goo.gl/V2yCaC As centrais sindicais construíram conosco a política de valorização do #SalárioMínimo Tenho consciência da importância dos parlamentares e tbm das centrais sindicais, q tiveram um papel relevante nesse País. #SalárioMínimo É um momento especial e agradeço aos parlamentares, pq foi com a compreensão e participação q estamos dando esse passo. #SalárioMínimo A valorização do #SalárioMínimo beneficia um conjunto imenso de trabalhadores, p/ q tenham acesso aos bens e condições mínimas de vida. O crescimento econômico não se dá em detrimento do trabalhador, e não se dará c/ a redução de políticas sociais. #SalárioMínimo É importante o Brasil continuar com a política de valorização do #SalárioMínimo. Esse País tem uma economia sólida, não temos nenhum desequilíbrio. O País está passando por uma dificuldade conjuntural. #SalárioMínimo Dilma: É importante lembrar que nesse período tivemos um reajuste em torno de 70%. #SalárioMínimo "Temos uma situação de sistemático reajuste e valorização do #salariomínimo", afirma @dilmabr. Acompanhe na @TVNBR: http://conteudo.ebcservicos.com.br/streaming/nbr Essa política, que representou um ganho real do salário dos trabalhadores mais pobres desse País, passa a ser realidade. #SalárioMínimo Agora, enviamos outro projeto, para o período de 2015 a 2019. #SalárioMínimo Em 2011, enviamos projeto que ia de 2011 a 2015, englobando este ano. #SalárioMínimo A política nacional de valorização do #SalárioMínimo foi implantada pelo governo Lula e vigia ano a ano. #AoVivo Presidenta @dilmabr assina Medida Provisória da Política do #SalárioMínimo: http://goo.gl/kMjifR A assessoria tuita agora trechos do discurso da presidenta durante assinatura da MP da Política Nacional do #SalárioMínimo Confira no vídeo do @minesporte o detalhamento das ações que estão em andamento, em diversas áreas: http://goo.gl/lYKr5U #Faltam500dias #Faltam500dias p/ os Jogos @Rio2016 e o Brasil está em estágio avançado de preparação. Sem dúvida será um grande evento e uma bela festa! Dilma participou da abertura da colheita de arroz ecológico e inaugurou unidade de armazenagem e secagem da COOTAP. pic.twitter.com/YaTQajBFil É importante provar q é possível, sim, ter desenvolvimento sustentável, de alta qualidade, baseado em assentamento da reforma agrária.
5	arq05.in	Boa noite a todos. #VestibularFatec O prazo para isenção ou redução de 50% da taxa de inscrição do Vestibular da Fatec é até 07/04 http://bit.ly/1E00fhK Acreditar e investir nos jovens, na inovação e no conhecimento científico é uma grande alavanca para o desenvolvimento social e econômico. A edição do ano passado foi um sucesso. Além de trabalhos das Etecs e Fatecs, a feira contou com projetos de outros estados e países. A 9ª Feira Tecnológica do @paulasouzasp será em outubro. As inscrições estão abertas até 15/05 http://bit.ly/19nWlVd pic.twitter.com/0u8ft0GN6g

		Comentei também os avanços das obras do metrô, nossos projetos em habitação e a melhora nos índices de criminalidade que divulgamos ontem. Falei sobre abastecimento de água, economia, combate à dengue, educação e o bônus que estamos pagando aos professores. Foi um prazer conversar hoje pela manhã com o Eli Correa e seus ouvintes da Rádio @capital1040 http://bit.ly/1CS9x0C pic.twitter.com/1nEnhNtj0e Vamos trabalhar para que o Senado também aprove rapidamente esta lei pic.twitter.com/Zmo3rXG5Ao Esta lei adota a linha da proposta que enviamos, transformando o homicídio de policiais em homicídio qualificado, com pena superior ao comum Hoje a Câmara de Deputados aprovou um importante projeto que torna mais rígidas as penas p/crimes contra policiais e agentes de seg pública. Boa noite a todos. Fui surpreendido pelo mascote do grupamento Águia de Campinas, que me recepcionou com uma mensagem de boas vindas pic.twitter.com/84qeBInxmx Também tivemos uma queda de quase 40% nos casos de latrocínio e de mais de 27% nos de roubo de veículos comparando fevereiro de 2015 e 2104 Tivemos o menor número de furtos desde 2002 para o mês de fevereiro http://bit.ly/1N9wvn5 Mais uma ótima notícia p/ a segurança: todos os índices criminais sofreram queda no 1º bimestre de 2015 em relação ao mesmo período de 2014. Na sexta, a equipe do helicóptero Águia 10 ajudou uma jovem a chegar a tempo p/uma cirurgia de transplante de fígado http://bit.ly/1BIHGdU Isso permite que 85% da população do Estado esteja ao alcance de um dos nossos helicópteros em no máximo 20 minutos. SP tem a maior organização de policiamento aéreo da América Latina e do hemisfério sul e uma das maiores do mundo. Em uma situação de emergência, a rapidez na ação e o preparo dos profissionais fazem toda a diferença pic.twitter.com/Mki4hCBIk1