Faculdade de Ciências e Tecnologia da Universidade de Coimbra

Mestrado em Engenharia Informática

2014/2015

Integração de Sistemas

Projeto #1

XML and XML Manipulation, Java Message Service and Message Oriented Middleware

Gonçalo Dias - nº2009111531

João Aguiar - nº2008111986

Introdução

Este projeto consiste na criação de três aplicações.
A primeira aplicação, o ***Web Crawler***, tem por objetivo ler, extrair e tratar informação sobre notícias publicadas no website www.cnn.com, e posteriormente enviar essa informação sob a forma de mensagens XML para um Java Message Service Topic. Este irá então servir outras duas aplicações que o irão subscrever.
O ***HTML Summary Creator*** é uma aplicação que subscreve o Topico JMS e cria ficheiros html com o conteúdo das noticias, gerados a partir das mensagens XML. Por último o ***Stats Producer*** é uma aplicação que guarda o número de notícias que têm menos de 12 horas.

Repartição do trabalho

Ambos os membros contribuíram para todas as componentes do projeto, tendo havido no entanto uma repartição de tarefas pelos dois.
Gonçalo Dias: Web Crawler, XSD/Java Classes, HTML SUMMARY CREATOR
João Aguiar: Web Crawler, JMS, STATS PRODUCER

De seguida é detalhada a implementação destas três aplicações.

Web Crawler

Começou-se por definir qual a informação que iria ser extraída do website. De modo a simplificar a recolha de informação, foram recolhidas as noticias que apresentavam o padrão de formatação mais completo e mais comum. Como tal, foram ignoradas notícias que apenas continham videos ou galerias de imagens. Foram extraídas as "Top Stories" de nove seções do site: World, US, Asia, Africa, Latin America, Middle East, Europe, Business e World Sport.

De cada uma das notícias foram recolhidos os seguintes dados: titulo, url, autor, data, highlights, texto da noticia e os urls das possíveis fotos que as acompanhavam.

Os elementos foram extraídos dando uso à ferramenta Jsoup, e após extraídos foram usados para popular os objetos Java **TopicType** e **NewsType**, gerados a partir de um ficheiro **xsd** criado por nós para este efeito e também para posteriormente validar o XML gerado. A geração das classes Java foi feita com a ferramenta **xjc.**

O Crawler vai assim gerar nove mensagens XML, cada uma constituída por um tópico (correspondente à secção em questão) e um conjunto de noticias que pertencem esse tópico. Após isto, as mensagens são enviadas para o JMS Topic.

Caso o JMS Topic não esteja disponível, o Crawler guarda as mensagens que tem para enviar em ficheiros e durante um período de tempo tenta conectar-se e enviar a informação contida nesses ficheiros para o Topic.

JMS

Foi criado um servidor JMS responsável por receber as noticias retiradas pelo Crawler e por enviar a dois clientes de forma a que estes nunca perdessem nenhuma noticia caso não estivessem online.
Para conseguirmos este fim configuramos o JMS para receber um tópico a que chamámos de news.
O Crawler é responsável por enviar as noticias para “jms/topic/news”.
Os clientes criam durable subscriptions. Isto permite a que sempre que conectem-se recebam as noticias que “perderam” quando estiveram ausentes.

HTML Summary Creator

Esta aplicação está sempre a correr, como “cliente” do JMS Topic, à espera de mensagens vindas do mesmo.
As mensagens recebidas são guardadas em ficheiros XML e validadas contra o XSD Schema já criado. Caso sejam válidos, os ficheiros xml são convertidos para Html dando uso a uma template XSLT criada para o efeito . O XSLT vai permitir assim a criação dos documentos html, um por cada tipo de notícia.

Stats Producer

Este aplicação simples cria uma durable subscription no jms. Assim sempre que recebe noticias é responsável por filtra-las e dizer quantas tem menos 12 horas. E escrever num ficheiro.
Para este fim usamos o url como identificador único de cada noticia (serve para garantir que não temos noticias duplicadas) e implementamos classes serializable que nos permite guardar esta informação. Temos depois uma função que corre sempre que recebemos uma noticia para verificar se alguma noticia já é invalida a essa altura.