Universidade de Évora

Universidade de Évora

Departamento de Informática

Mestrado em Engenharia Informática

Reconhecimento de entidades em documentos do "AHS - Arquivo Histórico Social"

Marco Emídio

Orientador: *Prof. Paulo Quaresma*

Évora, Agosto de 2011

Tese submetida à Universidade de Évora para obtenção do grau de **Mestre em Engenharia Informática**.

Esta tese não inclui as críticas e sugestões feitas pelo júri.

Universidade de Évora

Universidade de Évora

Departamento de Informática

Mestrado em Engenharia Informática

Reconhecimento de entidades em documentos do "AHS - Arquivo Histórico Social"

Marco Emídio

Orientador: *Prof. Paulo Quaresma*

Évora, Agosto de 2011

Tese submetida à Universidade de Évora para obtenção do grau de **Mestre em Engenharia Informática**.

Aos meus familiares e a todos os meus amigos.

Agradecimentos

Aos meus pais (Luís e Joaquina) e às minhas avós por terem, desde sempre, acreditado em mim. Em toda a minha vida motivaram-me a ser cada vez melhor dando-me as condições óptimas para a minha vida pessoal e académica. Sem eles não era nem metade do que sou hoje.

Ao meu orientador, o Professor Paulo Quaresma, pelo excelente trabalho de orientação! Muito obrigado pela paciência e empenho em colocar-me na direcção correcta.

A todos os meus amigos e colegas, que de alguma maneira, contribuíram para a motivação na realização desta dissertação.

Finalmente e não em último lugar, gostaria de agradecer à minha namorada, Ana Marmelada, pelo apoio e por todas as vezes que disse "Vai trabalhar..."e "Já escreveste alguma coisa da tese hoje?". Obrigado.

Sumário

A presente dissertação visa efectuar a extracção de informação de documentos históricos, provenientes do Arquivo Histórico-Social (AHS), e construir um suporte digital para os mesmos ao abrigo do projecto de investigação científica da Fundação para a Ciência e a Tecnologia (FCT).

Para a extracção de informação dos documentos, aplicou-se a ferramenta Minorthird, que possibilita a extracção de entidades mencionadas dentro dos textos, para posterior avaliação de resultados, possibilitando a pesquisa de elementos chave nos textos introduzidos no arquivo digital.

Os resultados obtidos revelaram-se promissores, tendo-se obtido uma precisão média de 0,8753 e uma cobertura média de 0,5075 na identificação de pessoas, entidades, locais e datas. Os melhores resultados foram obtidos na identificação de entidades, seguida das datas, lugares e pessoas.

O algoritmo Conditional Random Fields (CRF) demonstrou um melhor comportamento para a identificação de entidades, datas e locais, tendo o algoritmo Support Vector Machines (SVM) apresentado melhores resultados para a identificação de pessoas.

Na concepção do arquivo digital, utilizaram-se ferramentas como Archon, Joomla!, estando o portal disponível em `http://arquivo-digital.xdi.uevora.pt/projecto/`.

Entities recognition in documents from "AHS - Arquivo Histórico Social" Abstract

This dissertation aims to perform information extraction of historical documents from AHS, and build a digital archive for it promoted by the FCT scientific research project.

For the document information extraction, a tool called Minorthird was used, which enables extraction of named entities inside texts, for later results evaluation, enabling the search of key elements in the inserted texts in the digital archive.

The results proved to be promising, getting a mean precision of 0,8753 and a mean recall of 0,5075 by identifying persons, entities, places and dates. The best results were obtained by identifying entities, followed by dates, places and persons.

The CRF algorithm presented better performance in identifying entities, dates and places, having the SVM algorithm showed best results recognising persons.

The Archon and Joomla! tools were responsible of creating the digital archive, being the website available at http://arquivo-digital.xdi.uevora.pt/projecto/.

Conteúdo

Lista de Figuras

Lista de Tabelas

Lista de Acrónimos

URL *Uniform Resource Locator*

EI Extracção de Informação

MUC Message Understanding Conferences

AHS Arquivo Histórico-Social

NER Named Entity Recognition

FCT Fundação para a Ciência e a Tecnologia

CMS Content Management System

REM Reconhecimento de Entidades Mencionadas

SVM Support Vector Machines

CRF Conditional Random Fields

CoNLL Conference on Computational Natural Language Learning

HAREM Avaliação de Reconhecimento de Entidades Mencionadas

EM Entidades Mencionadas

CaGE Capturing Geographic Entities

1. Introdução

Nos tempos que decorrem, pode-se afirmar que a capacidade de extrair automaticamente informação de textos/documentos, tal como palavras-chave, é um filtro poderoso, pois possibilita a selecção de documentos potencialmente relevantes e conteúdos neles contidos. Portanto, é importante para o processamento de informação o desenvolvimento da automatização destes processos de selecção.

Reconhecimento de Entidades Nomeadas, do inglês Named Entity Recognition (NER), é uma tarefa importante no Processamento de Linguagem Natural. O NER constitui características importantes que auxiliam na tarefa de Extracção de Informação (EI), podendo assim reconhecer automaticamente elementos em textos tais como pessoas, organizações, locais, tempo e datas relevantes, através de técnicas de aprendizagem automática.

1.1 Objectivos

O objectivo deste trabalho baseia-se na extracção automática de entidades nomeadas de documentos históricos provenientes de estantes do AHSe para construção de um arquivo digital e de um portal para suporte e armazenamento desses mesmos documentos. O que motiva o reconhecimento de entidades nomeadas neste tipo de documentos é o facto de construir uma base que possibilite o reconhecimento de elementos importantes nos documentos que venham a ser introduzidos no arquivo digital, criando assim um motor de pesquisa que facilite a selecção de documentos para posterior leitura dos utilizadores.

1.1.1 Objectivos do projecto FCT

Esta dissertação integra-se num projecto vasto da FCT. Um dos objectivos é a criação do arquivo digital, que irá dar suporte à gestão documental e de conteúdos.

Para fazer uma melhor contextualização, o seguinte texto descreve brevemente os objectivos deste projecto, financiado pela FCT.

"Nas sociedades modernas ocidentais, a diversidade, a crítica e o debate público tornaram-se elementos constitutivos do devir social e da cultura contemporânea. As aspirações emancipatórias, libertárias e de equidade dinamizadas por minorias sociais interagiram com as expectativas de progresso e os reflexos de prudência das grandes massas populacionais, ajudando (com a economia e a acção política institucionalizada) a configurar processos de desenvolvimento social abertos, contraditórios e, por isso mesmo, também relativamente frágeis e irrepetíveis.

O elevado grau de informalidade destas fermentações sociais aconselha a que se faça um esforço no sentido de preservar a memória e de registar as referências de tais iniciativas, núcleos, redes e discursos, restituindo-as devidamente tratadas, de molde a que possam servir à sociedade universal em gestação e, em particular, aos investigadores e agentes culturais interessados. Assim, além dos objectivos sócio-económicos formalmente definíveis como de "desenvolvimento social e promoção geral dos conhecimentos", este projecto visa, mais em particular, a promoção da cidadania pela difusão e salvaguarda da memória e valores do movimento social crítico e alternativo em Portugal.

Ao designar desta maneira o objecto empírico do estudo, estamos a apontar para um universo que historicamente se desencadeou a partir de finais do século XIX e dos países então em vias de industrialização (fundamentalmente Europa e Américas), que não se constituiu em seguida como poder de Estado ou gestor do crescimento económico mas antes se manteve numa atitude distanciada e crítica face às grandes transformações e traumatismos do século XX (totalitarismos, guerras, holocaustos, nacionalismos, terrorismos, enriquecimento material, dominação da ciência e da técnica, etc.) e que hoje assume novas formas, inter-individuais e societais: o "cibernautismo", o "altermundismo", a "resistência ética" ou as "práticas de solidariedade directa" são experimentações contingentes e não-exclusivas de tal "movimento".

No presente projecto, o estudo científico deste movimento social far-se-á a partir dos cinco seguintes planos de análise:

a) O movimento operário histórico, que contestou o industrialismo capitalista e a gestão centralizada e autoritária do Estado moderno;

b) O movimento de questionamento e pesquisa cultural que se desenvolveu em si-

multâneo nos domínios científico, filosófico, literário e artístico;

c) Os movimentos de experimentação e inovação comportamental e social, que tocaram questões e temas tão diversos como a saúde, a mulher, a sexualidade e a procriação, os modos de vida e habitação ou a relação entre os humanos e o mundo natural;

d) A busca e ensaio de formas de economia alternativa, não capitalista nem estatal, que hoje tomam geralmente a designação de "terceiro sector";

e) A criatividade, iniciativas e vivências hoje experimentadas através do "ciber-espaço", constituindo novas redes de sociabilidades horizontais, relativamente autónomas face às exigências produtivistas da economia oficial e refractárias às tentativas de controlo dos aparelhos políticos de enquadramento;

f) E as acções políticas de base, reclamando ou criticando o poder político institucional.

Encontramo-nos hoje numa época em que se concretizam já passos muito importantes e significativos no sentido de uma sociedade da informação e do conhecimento, de contornos globalizantes, mas onde as referências topográficas, temporais e da língua-veículo de comunicação (mas igualmente de pensamento e de acção e, portanto, também de significados) surgem como indispensáveis. O uso dos actuais recursos técnicos da tele-informática deve, pois, servir este propósito.

Foi já esta a preocupação que que levou à criação do Arquivo Histórico-Social (AHS) na Biblioteca Nacional, nos anos 80.

Como metodologia e plano de trabalhos, o presente projecto vai incidir, sobre a documentação constituinte do Arquivo Histórico Social e outra. Além do tratamento e descrição catalográfica destes materiais, que darão origem a um catálogo electrónico elaborado segundo procedimentos técnicos internacionais, serão produzidos também novos conteúdos, igualmente acessíveis através da Internet, entre os quais se perspectivam os seguintes:

- dicionário biográfico de militantes;
- base de dados de entidades associativas;
- bibliografia de monografias e publicações em série;
- roteiros da memória urbana;
- biblioteca digital de trechos de obras;
- exposições temáticas virtuais;
- dados sociais (sobre trabalho, vida urbana, etc.);
- trajectórias pessoais de mobilidade entre Portugal e o Brasil.

Nestes termos, o presente projecto assenta fortemente numa postura inter-disciplinar (abrangendo a sociologia, a ciência política, a história, a geografia, a língua e cultura

portuguesas, e as ciências da informação e documentais), com dimensão internacional (Portugal, Brasil, Europa e línguas portuguesa e inglesa) e explorando a fundo as potencialidades contemporâneas das tecnologias de informação e comunicação (TIC)."

Retirado do link http://arquivo-digital.xdi.uevora.pt/projecto/index.php?option=com_content&view=article&id=1&Itemid=2 do site do projecto.

1.2 Metodologias

Para atingir o objectivo da criação do arquivo digital, considerou-se a integração de ferramentas de gestão de conteúdos com uma plataforma de arquivo digital, recorrendo a *software open source* já existente.

No que toca ao objectivo da dissertação em si, ou seja, o reconhecimento de entidades nomeadas, a metodologia utilizada contempla a aplicação de técnicas de aprendizagem automática para criar modelos de suporte à extracção de informação, utilizando vários algoritmos e avaliando os resultados obtidos, através da ferramenta Minorthird [Coh04]. Assim sendo, estes modelos podem ser utilizados futuramente, aplicando-os em novos documentos para a recuperação de informação importante sobre o AHS.

1.3 Estrutura da dissertação

Aqui é descrito um pouco a organização desta dissertação, para melhor entendimento dos leitores.

No primeiro capítulo, fala-se sobre os objectivos desta dissertação e metodologias nela utilizadas. O capítulo 2 explica o conceito de EI, técnicas abordadas no trabalho efectuado, métodos de avaliação dos resultados e ferramentas utilizadas. Após explicação do estado da arte, contempla-se o capítulo 3, que demonstra os casos tidos em conta para instalação e utilização, bem como soluções encontradas para obstáculos que surgiram ao longo da execução deste trabalho. O capítulo 4 demonstra a realização dos testes e comparação de resultados na forma de gráficos. Por fim, o capítulo 5, fala sobre as conclusões obtidas através desta dissertação e do trabalho futuro pretendido.

2. Trabalho Relacionado

Neste capítulo, é feita uma breve descrição sobre alguns conceitos do interesse para este trabalho. Inicialmente é apresentada uma definição de EI. De seguida, são descritas as principais abordagens, métodos de classificação, assim como métricas normalmente utilizadas na avaliação em sistemas de EI.

2.1 Introdução à extracção de informação

O aumento significativo da Internet e a sua popularidade, criou uma quantidade enorme de fontes de informação. Contudo, devido à falta de estrutura e às naturezas diferentes das fontes de informação na Web, o acesso a esta grande colecção de informação tem estado limitado à pesquisa e navegação, o que torna difícil a tarefa de extrair dados relevantes [CKGS06].

A EI tem a finalidade de detectar informação relevante em grandes quantidades de documentos, e apresentar essa mesma informação num formato adequado aos sistemas computacionais. O objectivo é isolar fragmentos de texto do documento, extrair informação relevante dos fragmentos sobre tipos de eventos, entidades, ou relações, e em seguida, reunir a informação específica para posterior processamento [KM05].

Os dados de entrada podem ser documentos não estruturados como texto escrito em linguagem natural ou documentos semi-estruturados que são predominantes na Web, tais como tabelas ou listas [CKGS06].

2.2 Abordagens

Existem duas abordagens muito utilizadas em técnicas de EI: baseada em regras ou aprendizagem automática.

2.2.1 Baseada em regras

Esta abordagem é baseada na definição de regras para extrair, a partir de texto, informação relevante. O programador deve estar familiarizado com o domínio e com a actividade que o sistema executará. Nalguns textos onde há um certo nível de estrutura, é comum especificar regras de extracção, utilizando expressões regulares. Apesar desta técnica ser simples de se usar e de interpretar, apenas consegue extrair padrões muito simples.

Outra abordagem comum baseada em regras é a utilização de técnicas de Processamento de Linguagem Natural. Estas técnicas são normalmente aplicadas para extracção de informação em textos de Língua Natural, desde que tenham a capacidade de lidar com as irregularidades da Língua Natural [AI99]. Neste caso, a ideia é usar informação sintáctica (ex., estrutura de uma frase) e/ou semântica (ex., representação lógica de uma frase) para efectuar a extracção.

Segundo Applet e Israel [AI99], uma das vantagens de sistemas baseados em regras é serem conceptualmente simples de se criar. Outra vantagem é que os sistemas de EI baseados em regras obtém melhores resultados quando são aplicadas as métricas descritas na secção 2.5 pois conseguem usar propriedades específicas de cada fonte de texto.

Contudo, o processo de desenvolvimento de um sistema de EI baseado em regras pode precisar de um grande número de iterações. Se as regras são demasiado específicas para o domínio, alterações na especificação devido a modificações de domínio podem ser difíceis de introduzir. Por vezes, torna-se mais fácil desenvolver um novo sistema de raíz. Outro problema desta abordagem é o facto de necessitar de recursos que podem não estar disponíveis, como dicionários ou gramáticas.

2.2.2 Baseado em aprendizagem automática

Para tentar superar algumas dificuldades provenientes da mudança de domínio em sistemas de EI baseados em regras, algumas abordagens utilizam algoritmos de aprendizagem automática para EI. Aprendizagem automática permite uma adaptação eficiente de um sistema de EI para novos domínios de extracção [Sod99].

A ideia por detrás desta abordagem é usar um *corpus* anotado para treinar o sistema utilizando um algoritmo de aprendizagem automática. O *corpus* é normalmente pro-

duzido por um especialista com conhecimento suficiente sobre o domínio da extracção e os resultados necessários da actividade de EI.

Pode-se dividir os sistemas de EI baseados em aprendizagem automática nos seguintes grupos [SBP05]:

- Baseado em autómatos finitos: visa a aprendizagem das regras de extracção sob a forma de autómatos finitos. O *input* do autómato é constituído por fragmentos de texto. Durante o procedimento de EI, alguns fragmentos de texto são aceites (levando a um estado de transição) e todos os outros são ignorados. Os fragmentos aceites irão preencher uma estrutura de dados que será o *output* do processo de EI.

- Baseado na correspondência de padrões: o sistema aprende regras de extracção na forma de expressões regulares.

- Baseado em classificadores: o texto de *input* é dividido em fragmentos. Estes fragmentos são candidatos para preencher um campo na estrutura de dados de *output*. Métodos estatísticos são aplicados a cada fragmento para determinar que campo na estrutura de *output* irá preencher.

Os dois primeiros tipos de sistema têm a vantagem de maior compreensibilidade, desde que as regras sejam descritas utilizando uma linguagem simbólica. Contudo, não são muito adequados para textos com grande variação na estrutura. Sistemas baseados em classificadores usam várias características dos fragmentos, conseguindo assim bons resultados para o reconhecimento de entidades a partir do texto. Ainda assim, têm limitações ao classificar através das características de cada termo individualmente e perdem alguma informação sobre a relação entre os fragmentos. Devido a essas limitações, outras técnicas começaram por surgir. Actualmente, as técnicas baseadas nos *Markov Models* (Modelos Markov) são utilizadas em grande escala.

Com a evolução de técnicas de aprendizagem automática, tem havido alguns esforços no desenvolvimento de soluções, para se reduzir a necessidade de utilizar grandes *corpus* anotados para treino.

2.3 Reconhecimento de Entidades Mencionadas

Do inglês Named Entity Recognition, o Reconhecimento de Entidades Mencionadas (REM) é uma tarefa de EI, cujo objectivo, como em qualquer sistema de extracção

de informação, é a identificação e classificação de elementos atómicos em texto, pertencentes a categorias tais como nomes de pessoas, organizações, localizações, expressões de tempo, quantidades, valores monetários, percentagens, etc [KG05].

Esta tarefa, bastante utilizada e relevante para várias áreas de processamento de linguagem natural (PLN), como a resposta automática a perguntas, a extracção de informação e a tradução automática, entre outras, foi introduzida em 1996, na sexta Message Understanding Conference [GS96], tendo sido esta o primeiro evento de avaliação a introduzir uma tarefa independente de avaliação em REM. A tarefa de REM da Message Understanding Conferences (MUC) consistiu em marcar as anotações que se dividem em três categorias:

- `ENAMEX`: categoria nominal composta pelos tipos `PERSON` (pessoa), `ORGANIZATION` (organização) e `LOCATION` (local).
- `NUMEX`: categoria numérica composta pelos tipos `MONEY` (moeda) e `PERCENT` (percentagem).
- `TIMEX`: categoria temporal composta pelos tipos `TIME` (hora) e `DATE` (data).

As MUC ocorreram no período de 1987 até 1998. Nessa altura, as MUC concentravam-se nas tarefas de EI, onde informação estruturada de actividades de companhias de defesa militar era extraída a partir de texto não estruturado, tal como artigos de jornais [RBC+98].

Como as MUC, outras competições e congressos foram realizados para melhorar o desenvolvimento de sistemas de avaliação para o reconhecimento de entidades mencionadas, como por exemplo Conference on Computational Natural Language Learning (CoNLL) e Avaliação de Reconhecimento de Entidades Mencionadas (HAREM).

A CoNLL foi uma conferência de avaliação que teve como objectivo promover a avaliação em diversas áreas de Processamento de Linguagem Natural. O primeiro evento remonta a 1999. Os eventos de 2002 e 2003 focaram a tarefa de REM, encorajando a investigação em sistemas independentes da língua, produzindo conjuntos de dados padrões e resultados nos quais um trabalho significativo em linguística computacional é baseado actualmente [BM06].

No evento CoNLL de 2002 utilizou-se colecções de textos em Espanhol e Holandês, e no evento CoNLL de 2003, foi oferecido aos participantes colecções de treino e teste em Inglês e Alemão.

A metodologia de avaliação abordada nestas conferências não difere muito em relação às MUC, apresentando quatro categorias de classificação semântica: `LOC` (local), `ORG`

(organização), `PER` (pessoa) e `MISC` (diversos) [SM03].

Em Portugal também se trabalha nestas áreas. O HAREM é um evento de avaliação conjunta em reconhecimento de entidades mencionadas para o português criado e organizado pela Linguateca [1]. Até ao momento decorreram três eventos HAREM: Primeiro HAREM (2005), MiniHAREM (2006), Segundo HAREM (2008).

Segundo o modelo de avaliação conjunta em REM, vários grupos comparam entre si, os resultados de avaliação obtidos, utilizando um conjunto de recursos comum, e uma métrica acordada entre todos [SC07].

O HAREM teve inspiração directa do MUC, e devido ao interesse na abstracção em português e para o português, fez duvidar das iniciativas multilingues. Visa avaliar o sucesso na identificação e classificação automática dos nomes próprios na língua portuguesa [SC07]. A metodologia do HAREM inclui a definição das directivas de etiquetagem dos textos, a especificação das tarefas de avaliação e o processo de criação das colecções de texto, suportando a vagueza das Entidades Mencionadas (EM) na anotação manual das colecções de texto, pois no caso de REM existem casos de EM vagas que não permitem desambiguar o seu significado semântico. Assim o sistema possibilita a atribuição de várias categorias semânticas para cada EM.

O modelo semântico do HAREM distingue-se de outros modelos utilizados na avaliação de REM, devido a dois factores: a identificação e classificação de uma EM depende apenas do seu uso em contexto, não estando restringida lexicalmente a nenhum dos atributos a que possa estar associada noutros recursos linguísticos, como dicionários, almanaques, ontologias; a possibilidade de atribuir mais do que uma classificação a uma mesma EM, se o contexto não conseguir distinguir apenas uma classificação, demonstrando que o HAREM suporta a vagueza de nomes próprios, que podem ter mais de uma interpretação associada [San07].

No evento foram entregues aos participantes, subconjuntos da colecção HAREM, denominados de **Colecções Douradas**, para que anotassem automaticamente e avaliassem o resultado segundo as directivas fornecidas. Com esta iniciativa a Linguateca atribuía aos participantes uma função no desenvolvimento do HAREM. Após as anotações, os subconjuntos foram revistos por anotadores independentes e outros participantes, para que depois o resultado desses subconjuntos, fosse englobado numa colecção maior, a colecção HAREM, que era distribuída aos participantes na própria avaliação conjunta [har].

As directivas da etiquetagem do HAREM são seguidas pelos participantes no desenvolvimento dos sistemas, e são usadas na anotação manual da colecção de textos. A

[1] http://www.linguateca.pt

categorização é composta por uma hierarquia de dois níveis, denominados categorias e tipos. Existem dez categorias principais: `VALOR`, `VARIADO` que no segundo evento viu o seu nome alterado para `OUTRO`, `PESSOA`, `ORGANIZACAO`, `LOCAL`, `ACONTECIMENTO`, `OBRA`, `ABSTRACCAO` e `COISA` [SCFO08]. As categorias representam as classes semânticas principais das entidades mencionadas e são compostas por vários tipos, que são especializações de cada categoria [CS07].

Os eventos tiveram a participação de vários grupos de investigação, a maioria portugueses. De seguida, alguns dos sistemas participantes, são explicados:

- **SIEMÊS**

  Luís Sarmento, Luís Cabral e Ana Sofia Pinto do Pólo do Porto da Linguateca, participaram no primeiro HAREM e Mini-HAREM com o sistema SIEMÊS - Sistema de Identificação de Entidades Mencionadas com Estratégia Siamesa. Utilizando uma abordagem que evita o uso de regras e heurísticas de reconhecimento de EM, a estratégia do SIEMÊS passa pela utilização de uma camada de classificação, possuindo uma cadeia de geração de hipóteses com cinco componentes, que procura semelhanças lexicais entre as EM no texto e as EM contidas no seu almanaque, o REPENTINO. O REPENTINO armazena 450.000 exemplos de nomes de entidades, recolhidos através grandes corpora, distribuídos por 11 classes e 103 subclasses [Sar07].

- **RENA**

  O RENA foi um protótipo de sistema de extracção de entidades mencionadas construído por Edgar Alves, participante no primeiro evento HAREM, sob a supervisão de José Almeida, no âmbito do projecto IKF (*Information + Knowledge + Fusion*). O sistema RENA marca e extrai as EM a partir do texto dependendo de um conjunto de ficheiros e de regras configuráveis, que simbolizam o conhecimento geral e regras de contexto usados na extracção.

  Assim sendo o RENA é um sistema de REM constituído por uma biblioteca Perl, baseada num conjunto de configurações, tendo em vista extrair a lista das entidades ou então marcá-las através de um conjunto de textos.

  Com estes elementos de configuração, o RENA possibilita a adaptação da utilização por parte dos utilizadores[dA07].

- **CaGE**

Capturing Geographic Entities (CaGE) foi desenvolvido no âmbito de um trabalho de doutoramento que aborda o problema do reconhecimento e desambiguação de nomes de locais, argumentando que este é um trabalho indispensável na geocodificação de documentos textuais, ou seja, informação não estruturada, diferenciando-se dos sistemas de informação geográfica tradicionais, que lidam com dados estruturados [Mar08]. O sistema também já foi utilizado em vários projectos relacionados com a recuperação de informação geograficamente contextualizada, falamos então do GREASE e no DIGMAP. No primeiro HAREM e Mini-HAREM, o CaGE focou-se na tarefa de identificação e de classificação de EM de categoria `LOCAL`, utilizando selectivamente o universo de EM com informação geográfica, tais como locais ou códigos postais [MSC07]. No segundo evento, o sistema CaGE abrangeu mais categorias na avaliação, correspondendo ao reconhecimento de entidades das categorias `PESSOA`, `ORGANIZACAO` e `TEMPO`, e ao reconhecimento e classificação em tipos e subtipos de entidades da categoria `LOCAL` [Mar08].

- **Priberam**

  A Priberam[2] é uma empresa que desenvolveu um sistema usado como módulo independente em vários produtos relacionados com a extracção de informação em motores de pesquisa como os utilizados nos sítios TSF[3] e do JN[4] e no IncogniX, ferramenta utilizada no Supremo Tribunal da Justiça. Participou pela primeira vez no segundo HAREM, e em todas as categorias existentes devido à adaptação para reconhecimento das categorias, tipos e subtipos propostos, surgindo algumas dificuldades na adaptação à categoria `TEMPO`, devido a critérios para criação de regras no sistema Priberam, a nível de detecção, construção e classificação das EM [AFM+08].

2.4 Classificação

Nesta secção é feita uma breve descrição sobre os algoritmos de classificação onde se obteve melhores resultados, nos testes realizados neste trabalho.

2.4.1 Support Vector Machines

SVM [CV95], como classificador, é conhecido pelo seu desempenho e capacidade de tratamento de dados de grande dimensão [HGMZ03]. Resultados demonstram que exige

[2] http://www.priberam.pt/
[3] http://www.tsf.pt/
[4] http://www.jn.pt/

menos dados de treino, para atingir desempenhos iguais ou superiores aos de outros classificadores [WC00].

Este classificador pertence a um grupo de algoritmos de aprendizagem kernel, provenientes da área da teoria de aprendizagem estatística e são baseados no princípio de minimização do risco estrutural [GQ03].

SVM assume que é possível mapear segmentos de texto num espaço vectorial, segundo propriedades linguísticas (ex., informação lexical) ou gráficas (ex., posição ou estilo no texto) de segmentos e palavras próximas. Depois, a ideia é separar elementos positivos (elementos que pertencem à classe) de elementos negativos (elementos que não pertencem à classe) de uma classe dada por um hiperplano [HGMZ03] [Joa98].

SVM é um classificador binário linear supervisionado, logo, falha em apresentar uma solução quando a fronteira entre duas classes não é linear. Neste caso, a solução recai-se na projecção do espaço de entrada X num novo espaço F e tentar definir uma separação linear entre as duas classes em F [GQ04].

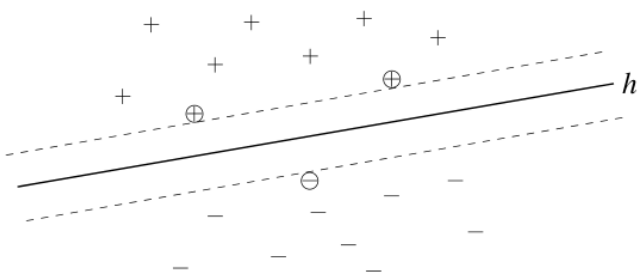

Figura 2.1: Hiperplano h descoberto por SVM, o qual separa os exemplos de treino positivos e negativos. Os exemplos mais próximos do hiperplano são chamados de *Support Vectors* (marcados com círculos)

A figura 2.1 mostra um caso linearmente separável. O hiperplano de decisão separa exemplos positivos e negativos pela maior margem. A linha sólida indica o hiperplano de decisão e duas linhas tracejadas paralelas indicam a margem entre exemplos positivos e negativos.

Sendo o SVM um classificador, tem uma fase de treino e uma de teste. Na fase de treino o classificador auxilia-se de um conjunto de documentos/dados com entidades, para construir um modelo. Ao construir o modelo, tem como objectivo descobrir os hiperplanos que alcançam melhores distinções entre exemplos positivos e negativos. Já na fase de teste, o modelo criado pela fase de treino é utilizado para classificar os dados

de teste, ou seja, a classificação é efectuada ao verificar de que lado do hiperplano os dados de *input* se encontram, por outras palavras, se é um elemento positivo ou negativo de uma dada classe.

2.4.2 Conditional Random Fields

CRF [LMP01] é uma metodologia que constrói modelos probabilísticos de modo a classificar conjuntos de dados. Estes tipos de modelos são aplicados em problemas de etiquetagem de sequências estruturadas, tal como texto em linguagem natural. CRF é baseado no conceito de maximização da entropia, que permite estimar a probabilidade de distribuição de um conjunto de dados de treino. O conceito de maximização da entropia afirma que a probabilidade de distribuição construída a partir de informação incompleta, tal como um número finito de dados de treino, é a que tem submetida entropia máxima para um conjunto de restrições que representam a informação disponível. Está provado que uma probabilidade que maximiza a entropia deve ser o mais uniforme possível. Qualquer outra distribuição implicará pressupostos não justificados [Wal04].

CRF oferece várias vantagens além das já existentes nos modelos hidden Markov, uma delas, é a capacidade de atenuar fortes pressupostos de independências, necessários nesses modelos. Também evita uma limitação fundamental dos Maximum entropy Markov models (MEMM) e outros modelos Markov característicos baseados em modelos gráficos, que podem tender para estados com poucos estados sucessores [LMP01].

CRF foi desenvolvido como solução que oferece todas as vantagens dos MEMM mas ultrapassa o problema de *label bias*. A diferença entre o CRF e o MEMM é a utilização de um modelo exponencial de probabilidades condicionais por parte do modelo MEMM para cada estado, enquanto que o CRF apenas usa um único modelo exponencial para a probabilidade conjunta da sequência de etiquetas, dada a sequência de observação. Com este modelo é possível normalizar as probabilidades a nível global.

Normalmente CRF ultrapassa tanto HMM e MEMM em termos de desempenho [LMP01] mas o treino de um modelo CRF é muito dispendioso, tornando-o difícil de usar se actualizarmos o nosso modelo ao longo do tempo (forçando o treino em cada iteração).

2.5 Avaliação

Entre 1987 e 1998 ocorreram algumas conferências onde sistemas de EI foram avaliados. Estas conferências foram apelidadas de MUC fundadas pela DARPA [5] para encorajar

[5] www.darpa.mil

o desenvolvimento de novas e melhores técnicas de EI. Além do desenvolvimento que estas conferências trouxeram a esta área, as MUC foram importantes porque criaram um consenso de como os sistemas de EI deveriam ser avaliados. Inicialmente, utilizavam-se medidas provenientes da área da Recuperação de Informação: cobertura e precisão. Embora os nomes mantenham-se os mesmos, o método de calcular as medidas foi alterado a fim de considerar os casos gerais de EI. Esta secção explica essas medidas [GS96] [RBC+98].

2.5.1 Cobertura

Cobertura dá-nos a relação entre a quantidade de informação correctamente extraída dos textos e a informação disponível dos textos. Portanto, cobertura mede a quantidade de informação extraída relevante e é dada pela equação:

$$cobertura = \frac{N_{correct}}{N_{key}} \tag{2.1}$$

onde Ncorrect representa o número de entidades correctamente extraídas enquanto que Nkey representa o número total de entidades que deveriam ser extraídas.

A desvantagem desta medida é o facto de retornar valores altos quando é extraída toda a informação correcta e incorrecta do texto.

2.5.2 Precisão

Precisão é a relação entre a quantidade de informação correctamente extraída dos textos e toda a informação extraída. Então, a precisão é a medida de confiança sobre a informação extraída e é dada pela equação:

$$precisao = \frac{N_{correct}}{N_{response}} \tag{2.2}$$

onde Ncorrect representa o número de entidades correctamente extraídas e Nresponse representa as entidades extraídas, ou seja, entidades extraídas correcta e incorrectamente.

A desvantagem desta medida é que podemos obter resultados altos extraindo apenas informação do qual temos a certeza que está correcta, ignorando informação que está no texto e que pode ser relevante.

2.5.3 Medida-F

A relevância de precisão e cobertura pode variar dependendo do contexto da sua utilização. Para se obter uma avaliação entre estes dois valores, criou-se uma medida designada por medida-F:

$$F = \frac{(\beta^2 + 1) \times P \times R}{\beta^2 \times P + R} \tag{2.3}$$

onde P = precisão, R = cobertura e β é o factor que quantifica a preferência de cobertura sobre a precisão. Quando β é igual a 1, isto é, cobertura e precisão têm o mesmo peso, é uma média harmónica e designa-se F1 [GS96] [Moe06].

2.6 Ferramentas

Nesta secção fala-se um pouco das aplicações/ferramentas utilizadas ou ponderadas no âmbito deste trabalho. Nalguns casos houve a necessidade de se instalar e testar programas para o mesmo fim, com o intuito de verificar qual o que melhor cumpria as necessidades impostas. De seguida é feita uma breve descrição de cada uma das ferramentas.

2.6.1 Minorthird

Minorthird [Coh04] é uma colecção de classes Java, que permite classificação documental e classificação sequencial utilizando algoritmos de aprendizagem automática. Minorthird oferece uma vasta variedade de algoritmos para classificação tais como CRF, SVM (ver 2.4 para uma explicação mais detalhada sobre estes dois algoritmos) , HMM e MEMM.

Minorthird consegue carregar vários tipos de formatos de texto, mas o formato recomendado para introdução de dados é texto simples com etiquetas XML referenciando determinados segmentos importantes. Este programa não permite ao utilizador definir como o texto deve de ser dividido. Um texto é sempre dividido por espaços e sinais de pontuação.

2.6.2 Archon

Archon [6] é uma ferramenta para gestão de informação descritiva sobre material arquivístico e para publicação na web. Permite facilmente a criação e publicação de pesquisas para todos os tipos de material arquivístico. Inclui métodos de gestão de

[6] www.archon.org

sessões e uma aplicação de biblioteca digital integrada. O sistema permite a introdução de registos descritivos que são compatíveis com os padrões de arquivamento.

Os scripts criados pelo Archon, produzem automaticamente um website para navegação e pesquisas, onde os utilizadores podem procurar e aceder a colecções de resumos de descrições, conteúdos digitais (incluindo vista de miniaturas) e termos de assuntos. Na figura 2.2 é possível visualizar o aspecto de uma colecção, através da vista do utilizador.

Figura 2.2: Demonstração da visualização de uma colecção por parte do utilizador

No Archon, as colecções podem ser agrupadas em registos de grupos e subgrupos ou outras classificações definidas nos repositórios. Os dados podem ser importados para o sistema a partir de ficheiros CSV (comma separated value), EAD, e MARC. Os repositórios podem construir websites customizados utilizando temas e modelos, facilmente modificáveis.

A interface administrativa (ver figura 2.3) para criar, editar e aceder a registos descritivos, necessita apenas de um web browser padrão que consiga correr javascript.

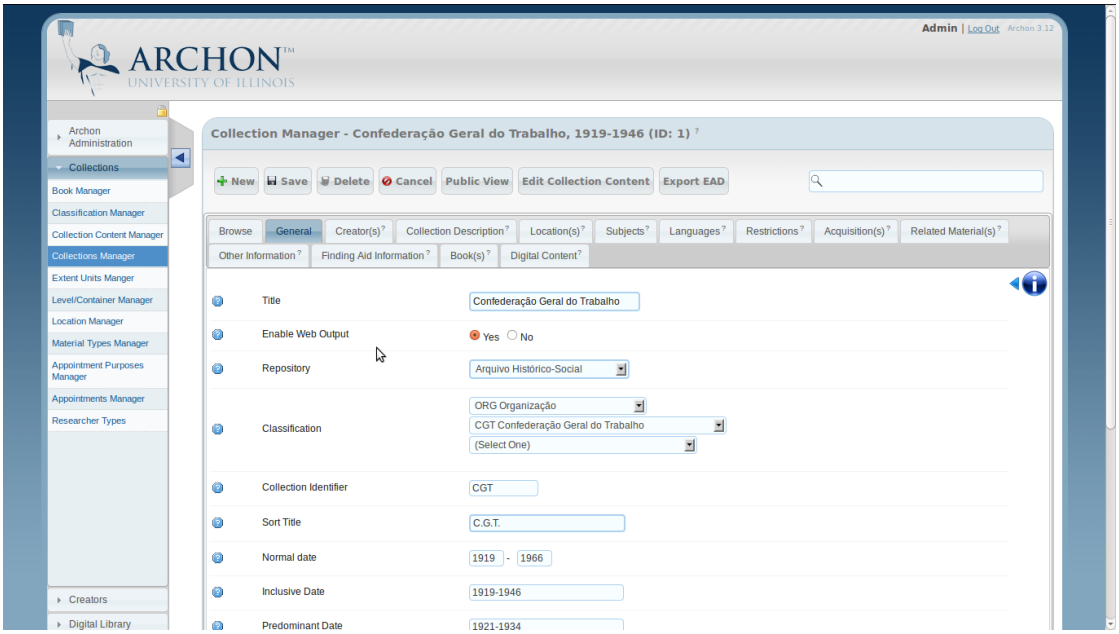

Figura 2.3: Exemplo da área administrativa do Archon

2.6.3 DSpace

DSpace [7] é o *software open source* de escolha das organizações académicas, sem fins lucrativos e comerciais que oferece as ferramentas de gestão para implementar repositórios digitais e um enorme apoio de uma comunidade de utilizadores, em constante crescimento.

Suporta uma grande variedade de dados, tal como livros, dissertações, fotografias, ficheiros vídeo, relatórios de investigação e outras formas de conteúdo.

A maneira como os dados são organizados no DSpace destina-se a reflectir a estrutura da organização que utiliza o seu sistema. Cada website DSpace está dividido em comunidades, que podem vir a ser divididas em sub-comunidades, como colégio, departamento, centro de investigação, ou laboratório, que espelham a estrutura típica de uma universidade.

As comunidades contém colecções, que são agrupamentos de conteúdos relacionados, e as colecções são compostas por itens, que são os elementos básicos de arquivo. Cada item é propriedade de uma colecção. Itens são ainda subdividos em pacotes de *bitstreams* (fluxos de dados), que são ficheiros normais.

[7] www.dspace.org

Figura 2.4: Exemplo da listagem de colecções da comunidade do Departamento de Informática do Repositório Digital de Publicações Científicas da Universidade de Évora

Na figura 2.4, pode-se verificar que a pesquisa de conteúdos é efectuada através do título, autor, palavras-chave, data, etc.

2.6.4 Archivists' Toolkit

O Archivists' Toolkit, é uma aplicação de base de dados *open source*, que oferece um suporte amplo e integrado para gestão de arquivos. Os principais objectivos do Archivists' Toolkit são o suporte ao processamento arquivístico e a produção de opções de acesso, promover a criação de um modelo de dados padrão, promover a eficiência e baixar os custos do processamento e treino de dados.

Suporta a gestão e descrição de conteúdos de arquivo, incluindo o acesso, registo de fontes de informação para recursos arquivísticos, fornecendo tópicos e nomeando pontos de acesso para assuntos e autores de recursos, podendo também produzir vários relatórios administrativos. A figura 2.5, demonstra como estão organizadas as várias áreas da aplicação.

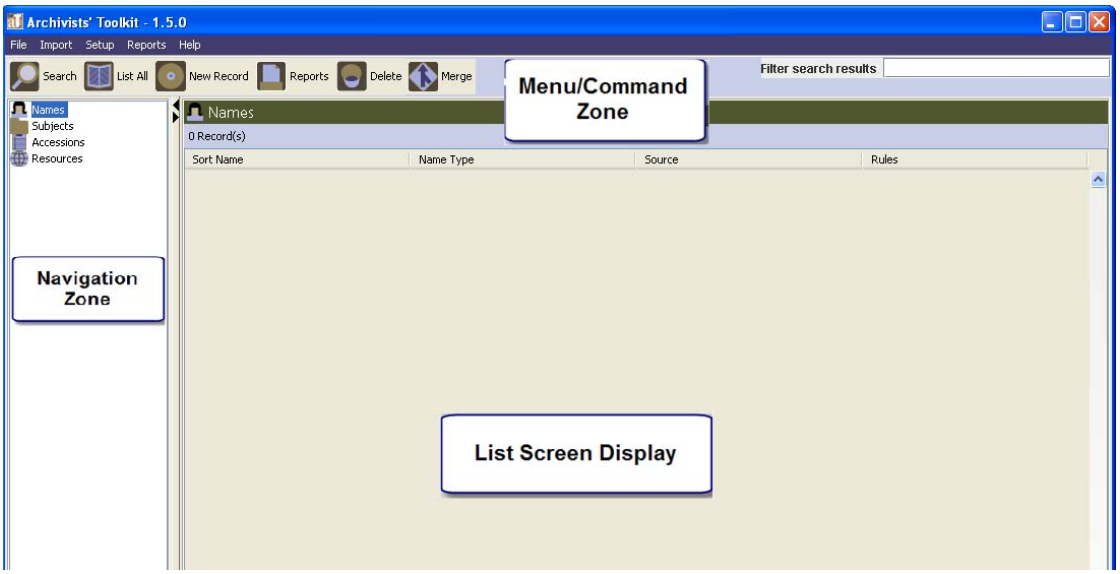

Figura 2.5: Ecrã da aplicação após inicialização, demonstrando as várias áreas de navegação e pesquisa

Neste capítulo, é apresentado todo o processo de criação do arquivo digital e do portal, desde as configurações efectuadas, ferramentas utilizadas para alojamento dos documentos do AHS, etc.

2.7 Criação do arquivo digital

Numa primeira fase, era necessário, a escolha da ferramenta adequada para alojar a informação dos documentos do AHS. No leque de opções, as ferramentas que melhor se adaptam a este tipo de arquivos históricos, como já foi referido no capítulo anterior em 2.6, são: Archon [8] ideal para gestão de informação descritiva sobre material arquivístico e para publicação via web; Archivists' Toolkit [9], um sistema de gestão de dados arquivísticos *open source*; DSpace [10] um software que permite o acesso fácil e aberto a todos os tipos de conteúdo digital incluindo texto, imagens, vídeos.

Figura 2.6: Página inicial do Archon, após instalação

Das três ferramentas que foram mencionadas e testadas acima, apenas uma foi utilizada, o Archon, ver figura 2.6. Ofereceu muitas funcionalidades direccionadas para o tipo de documentos do AHS. O facto de ser programado sob a base de php e mysql, possibilitava em muito a integração de algumas funcionalidades para o portal que seria construído no futuro. O Archivists' Toolkit tem pouca documentação, está confuso, não tem muitos exemplos/tutoriais, de uma integração um pouco mais complexa num ambi-

[8] www.archon.org
[9] www.archiviststoolkit.org
[10] www.dspace.org

ente web, aparenta estar num nível de desenvolvimento menor e com uma comunidade reduzida de utilizadores, no que toca ao esclarecimento de dúvidas e apresentação de exemplos. Já com o DSpace, a possibilidade de se criar os modelos de metadados à nossa medida, pareceu ser trabalho desnecessário, visto o Archon já oferecer todos os campos/modelos dos metadados por defeito, moldados às nossas necessidades.

Após a instalação, configuração e população do suporte arquivístico, foi necessário encontrar/desenvolver uma plataforma web que alojasse a informação e as funcionalidades pretendidas do AHS dentro do Archon, permitindo também a interacção de outros utilizadores com o projecto, submetendo novidades, desenvolvimentos do projecto, etc. A escolha recaiu-se então pelo Joomla, visto ser um Content Management System (CMS) *open source* e grátis, ver figura 2.7. Um dos factores mais apelativos é o facto deste *software* estar assente sobre php, javascript e mysql. Como o Archon, também foi concebido em php, javascript e mysql, existe a possibilidade de o implementar dentro do portal Joomla, tornando a navegação mais amigável, centrando toda a informação num único sítio. Assim os utilizadores, não sentiriam o desconforto de navegação e de pesquisa de conteúdos entre páginas web diferentes.

Figura 2.7: Página inicial do portal do projecto em Joomla

Tendo já ambas as ferramentas/plataformas instaladas e configuradas, iniciou-se o processo de adaptação das principais funcionalidades e características do Archon para o portal Joomla, funcionalidades estas que podemos dividir nos seguintes pacotes:

- Colecções: Nesta parte estão identificadas todas as colecções do AHS, onde estão todas as informações sobre os documentos pertencentes e administrativas.

- Conteúdos Digitais : Aqui pode-se visualizar ou pesquisar miniaturas das imagens dos documentos inseridos no sistema. Também se pode ver cada objecto conteúdo digital e toda a informação associada, tal como título, data, descrição, etc.

- Assuntos: Na secção sobre assuntos, pesquisa-se todos os registos dentro do sistema, sejam colecções, imagens, dando como argumento o assunto a que pertencem.

- Produtores: Como o nome indica, são os produtores das obras inseridas no AHS, demonstrando todos os registos associados a cada produtor.

- Classificação: Árvore de estrutura da descrição dos documentos do AHS.

- Cesto de Documentos: Ferramenta que possibilita aos utilizadores adicionar as colecções do seu agrado, para posterior download/leitura.

Na adaptação, foi necessária a programação em php e javascript, para se integrar dentro do portal as funções e classes essenciais ao funcionamento do Archon. Assim sendo, dispomos da informação da base de dados do Archon e também do seu aspecto, facilitando a pesquisa de informação, sem requerer à navegação entre dois sites distintos, ver figura 2.8.

Figura 2.8: Lista de colecções do Archon dentro do portal

Outro factor importante, era o de obter o sistema Archon na língua portuguesa, visto só existir as línguas inglesa e espanhola nas opções de instalação. Próxima tarefa seria então a de traduzir os ficheiros *XML* de cada pacote de funções do Archon. Aproveitando

os termos em inglês, procedeu-se à tradução dos mesmos para português, trabalho este que teve de ser totalmente feito à mão, não podendo utilizar ferramentas de programação para auxílio, tal como a *API* do tradutor da *Google* [11], pois os termos arquivísticos são demasiado específico e a tradução não era adequada.

O Archon oferece um sistema de ficheiros *XML* para instalação das linguagens pretendidas, ou seja, inglês e espanhol, como já tinha referido anteriormente. Criando novos ficheiros na mesma estrutura mas na língua portuguesa, oferecíamos então a possibilidade de uma instalação em português.

Estando este trabalho efectuado, entrou-se em contacto com a equipa de desenvolvimento do Archon, no caso de estarem interessados na oferta da tradução, para que em futuros lançamentos de versões do Archon, já existisse a língua portuguesa nos pacotes de instalação.

Com o desenrolar dos trabalhos, notou-se que o pacote de "Conteúdos Digitais"do Archon, era insuficiente e não muito apelativo à visualização dos documentos. Depois de alguma deliberação sobre o assunto, chegou-se à conclusão que seria uma mais valia a criação de um visualizador de documentos personalizado. Para tal, utilizou-se a linguagem de programação java, pois oferece bibliotecas para carregamento de imagens, que suportam os tipos de ficheiros utilizados nas digitalizações dos documentos, sendo estes de extensão TIFF. Assim, temos uma estante virtual, ferramenta que se adaptava às exigências dos utilizadores, que pretendem pesquisar e ler, tal e qual como o fazem numa biblioteca real, pois no final o objectivo é recriar o método de leitura de documentos como o fazem numa biblioteca, ver figura 2.9.

Agora, será preciso fundir a estante virtual com o Archon, fornecendo as informações sobre os documentos e respectivas imagens. Esta função é executada aquando da pesquisa por parte dos utilizadores. Estes podem adicionar ao "Cesto de Documentos", as colecções que vão procurando dentro do Archon, estando finalmente disponíveis na estante virtual.

Para a concepção da estante virtual, foi necessária a biblioteca *jai_imageio*, que possibilita o carregamento de imagens de extensão TIFF, extensão esta utilizada nos documentos. Não se considerou a transformação dos documentos para uma extensão padrão, como por exemplo JPG, pois iria ser um processo moroso e diminuiria a qualidade da imagem do documento. O funcionamento do programa java começa quando recebe a informação sobre as colecções que o utilizador deseja ler. Ao receber essa informação, verifica, através de um *web service*, quais os ficheiros a carregar, pertencentes às colecções. Depois é feito o carregamento na aplicação de leitura, para que o leitor

[11] http://translate.google.com/

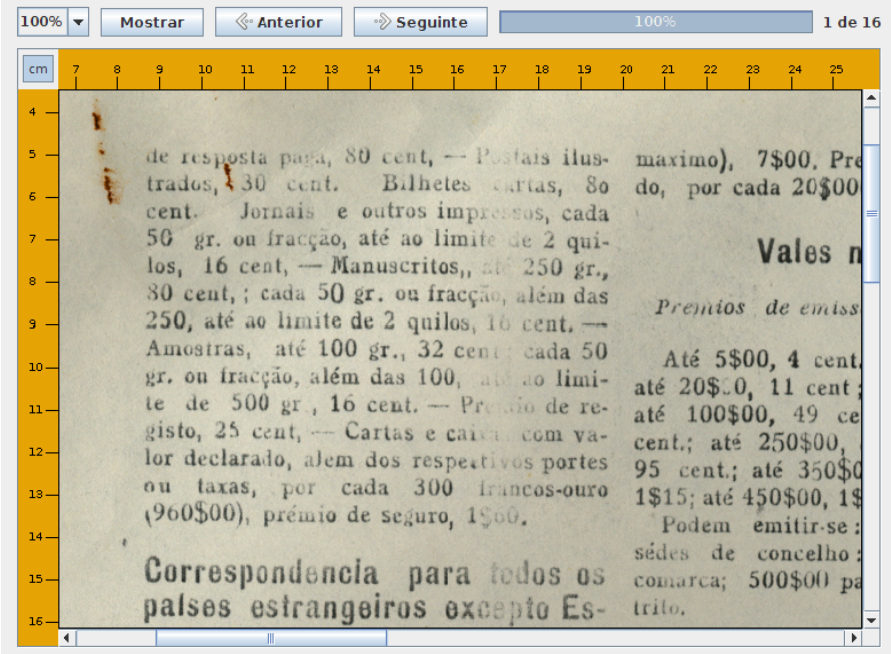

Figura 2.9: Estante virtual de leitura dos documentos

possa navegar através dos documentos, como de um livro se tratasse. O portal encontra-se disponível em http://http://mosca-servidor.xdi.uevora.pt/projecto/.

2.8 Exemplo de utilização

Nesta secção é demonstrado um exemplo de utilização normal, de como se poderá usufruir do arquivo digital que foi construído, para pesquisar e visualizar conteúdos.

Na página principal do portal, estamos perante uma secção central com notícias e novidades colocadas ao longo do projecto, e a zona de menus do lado esquerdo, dos quais destaco o menu **Archon**.

Na figura 2.10, contempla-se a página inicial do arquivo digital Archon, com os menus **Colecções**, **Conteúdos Digitais**, **Assuntos**, **Produtores**, **Classificação** e uma caixa de texto para pesquisa livre. Efectuando uma pesquisa pelo termo "Confederação", obtém-se a página Resultados da Pesquisa com os resultados obtidos, indicando em que secções se obteve a correspondência. No caso da figura seguinte 2.11, visualiza-se a correspondência para Manuscritos/Colecções, indicando qual a Classificação do objecto encontrado.

Ao clicar no objecto, é apresentada toda a informação disponível da colecção, visível

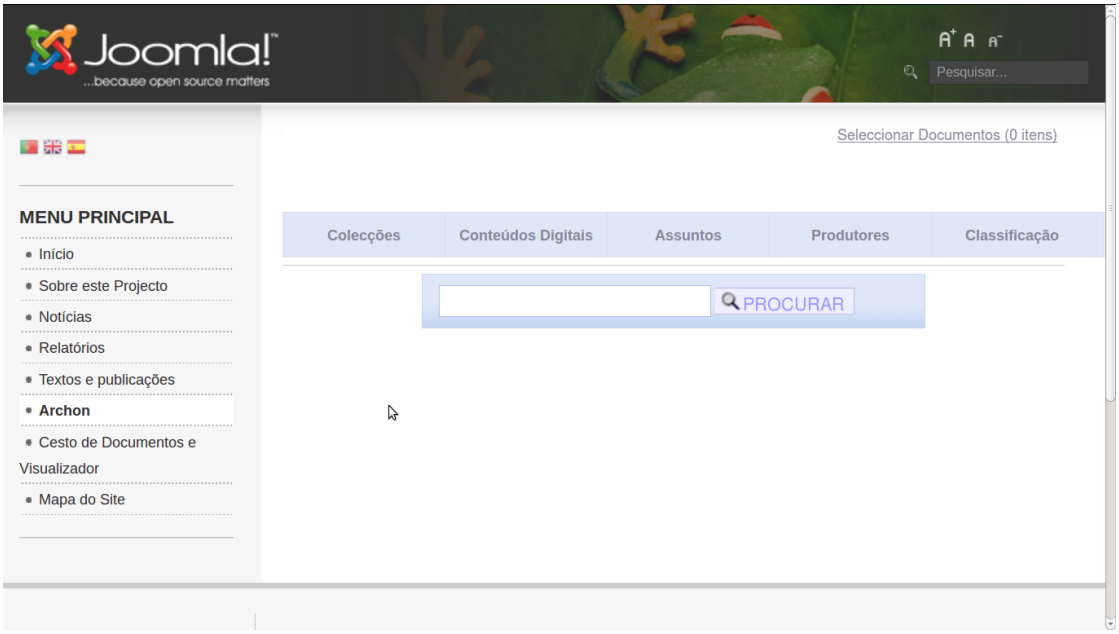

Figura 2.10: Página inicial do Archon no portal

na figura 2.12, tal como "Título", documentos digitais associados em , "Organização", "Criado por", que demonstra quem foi o autor da colecção, etc. É possível visualizar as ligações que a colecção tem, por exemplo, com a página do seu produtor. De notar que ao longo de toda a navegação, o utilizador é sempre alertado sobre a secção onde se encontra, com a indicação da mesma nos menus horizontais.

Figura 2.11: Página Resultados da Pesquisa para o termo Confederação

Na figura 2.13 é demonstrada a página do Produtor "Confederação Geral do Trabalho", ou seja, a colecção "Confederação Geral do Trabalho, 1919-1946"foi criada pela própria Confederação Geral do Trabalho. Aqui também é visível toda a informação associada ao produtor, tal como "Nome", produtores relacionados, colecções criadas pelo produtor, conteúdos digitais criados pelo produtor, etc.

Os conteúdos digitais criados pelo produtor, fazem a ligação ao mesmo objecto que também se encontrava disponível na figura 2.12 em Imagens. Ao clicarmos no objecto "Organizações aderentes à C.G.T.", é-nos apresentado as miniaturas das imagens presentes e informações sobre o objecto, como se pode observar na figura 2.14.

Por fim, para fazermos a ligação entre as colecções que desejamos ver e a estante virtual, existe um botão na forma de cesto de compras, com um símbolo verde, que serve para esse mesmo efeito. Assim que são adicionadas as colecções desejadas ao Cesto de Documentos, basta clicar em Cesto de Documentos e Visualizador para sermos direccionados para a estante virtual.

Figura 2.12: Página de apresentação da colecção Confederação Geral do Trabalho, 1919-1946

Figura 2.13: Página de apresentação do produtor Confederação Geral do Trabalho do AHS

Figura 2.14: Conteúdo digital das Organizações aderentes à C.G.T.

3. Reconhecimento de Entidades Mencionadas

3.1 Etiquetação documental

Aqui são descritos, os métodos utilizados para identificar supostos elementos pertencentes às categorias pretendidas, tais como: entidades, datas, lugares e pessoas. Numa primeira fase, teve que se escolher os documentos dactilografados com melhor qualidade, para se conseguir passar do formato imagem para formato texto. Com o auxílio do programa tesseract-ocr[1], obteve-se resultados satisfatórios, facilitando em muito esta tarefa. Para que o programa Minorthird funcionasse correctamente, teve que se proceder à eliminação de todos os caracteres especiais dos documentos, pois o programa é designado para a língua inglesa e não reconhece acentos nem cedilhados. Por fim, para marcar cada uma das categorias, existentes nos documentos do AHS, foram utilizadas *tags XML* (`<variável> ... </variável>`), neste caso propriamente dito, foram utilizadas `<e> ... </e>` para entidades, `<d> ... </d>` para datas, `<l> ... </l>` lugares e finalmente `<p> ... </p>` para pessoas. Um exemplo da aplicação das tags no texto é apresentado na figura 3.1:

Na figura 3.1, "Diario do Governo"é classificado como entidade e "7 de Marco"é classificado como data.

Os documentos pertencentes ao corpus, são provenientes do AHS, foram escolhidos apenas os documentos dactilografados com maior qualidade, para permitir a utilização de um programa OCR e assim passar os documentos do formato imagem para texto.

O AHS tem por vocação reunir, conservar e tratar tudo o que diga respeito à memória

[1] http://code.google.com/p/tesseract-ocr/

```
CAMARADAS: Acaba de ser publicado no <e>≪Diario do Governo≫</e> n.º1
13- Serie de 16 do corrente, o Decreto-Lei n.º21.238 que aprova o
Regulamento Provisorio da Caixa de Auxilio aos Desempregados que foi
creada pelo decreto com forca de lei n.º20.984 de <d>7 de Marco</d>
do corrente.  O Regulamento Provisorio da Caixa de Auxilio aos
Desempregados publicado em vesperas do <d>29 de Fevereiro</d>, que
serviu de bomba de choque do movimento que ia desencadear-se nao
nos serve, pelo espirito acentuadamente reaccionario que presidiu a
sua confeccao.  A burguesia pretende que a classe operaria pague as
consequencias da sua pessima administracao pelo que estabelece um
imposto de 2% nos nossos ja minguados salarios.
```

Figura 3.1: Excerto de documento do AHS etiquetado

do antigo Movimento Operário e do Anarquismo em Portugal. Os assuntos incluídos nos documentos dizem respeito, não apenas às questões do trabalho e do sindicalismo e da filosofia política libertária, mas igualmente aos campos por eles proximamente influenciados ou criticados: religião, guerra, mulher, educação, cultura, cooperativismo, socialismo, república, política e outros problemas sociais.

Os materiais integrados no AHS são, especificamente: publicações unitárias (livros, brochuras); publicações periódicas (jornais, revistas, números únicos); iconografia (cartazes, gravuras, fotografias, panfletos, objectos); manuscritos (documentos internos, correspondência, papéis administrativos); e registos orais.

A delimitação temporal do antigo Movimento Operário e do Anarquismo em Portugal é estabelecida, no que respeita ao seu início, pelo surgimento das primeiras formas de associativismo operário e das ideias socialistas em Portugal, particularmente depois de 1870. O final é definido pelo desaparecimento progressivo das suas expressões organizativas, reivindicativas e propagandísticas, sob o Estado Novo, que ilegalizou e perseguiu aqueles Movimentos. Dentro destas referências, o AHS privilegiará particularmente o período compreendido entre 1900 e 1940.

3.2 Treino e Classificação

Depois de o *corpus* de documentos ter sido devidamente etiquetado, está pronto para a próxima fase, ou seja, executar a tarefa de EI, através do *software* Minorthird [Coh04], descrito em 2.6.1.

Na primeira parte, temos o treino, que dado um conjunto de documentos etiquetados é treinado segundo a entidade e algoritmo fornecidos, sendo depois classificado pelo

programa. Na fase de teste, é feita a comparação entre o resultado do treino e os documentos a testar. No treino, todos os algoritmos foram testados para cada etiqueta, retirando daí os dois melhores, para utilização posterior na fase de teste. De salientar que foram utilizadas todas as opções padrão nos algoritmos disponíveis no Minorthird. Nos testes foi utilizada a validação cruzada com 5 pastas como método de análise de resultados obtidos. Pode-se verificar os outputs dos testes realizados nos Conteúdos em Anexo A.

Será demonstrado, na forma de gráficos, os valores da avaliação obtidos pelo programa Minorthird, tendo como argumento 98 documentos pertencentes ao AHS.

Documentos	98
Palavras	42416
Caracteres	272451
Etiquetas Entidade	761
Etiquetas Data	109
Etiquetas Lugar	284
Etiquetas Pessoa	76
Total etiquetas	1230

Tabela 3.1: Estatística sobre o corpus de treino utilizado

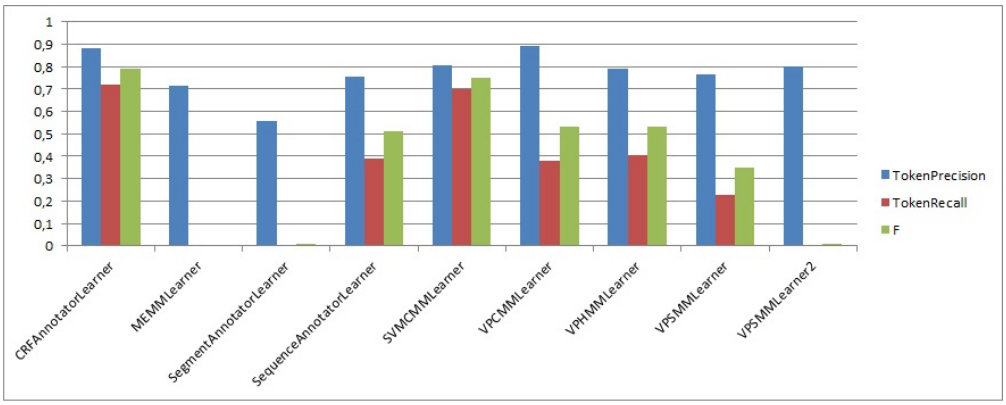

Figura 3.2: Resultados dos vários algoritmos do Minorthird para a tag entidade

O gráfico, representado na figura 3.2, demonstra que para a tag entidade, os algoritmos CRF e SVM destacam-se claramente. Para o algoritmo CRF, obteve-se 0,8807 de precisão, o que significa que 88,07% das tags extraídas, eram de facto tags entidades; 0,7214 de cobertura, indicando que 72,14% das tags entidades foram classificadas correctamente em relação às tags entidades existentes; e 0,7931 de medida-f.No algoritmo SVM, temos valores 0,8061 de precisão, 0,7016 de cobertura e 0,7503 de medida-f, ligeiramente inferiores aos do algoritmo CRF.

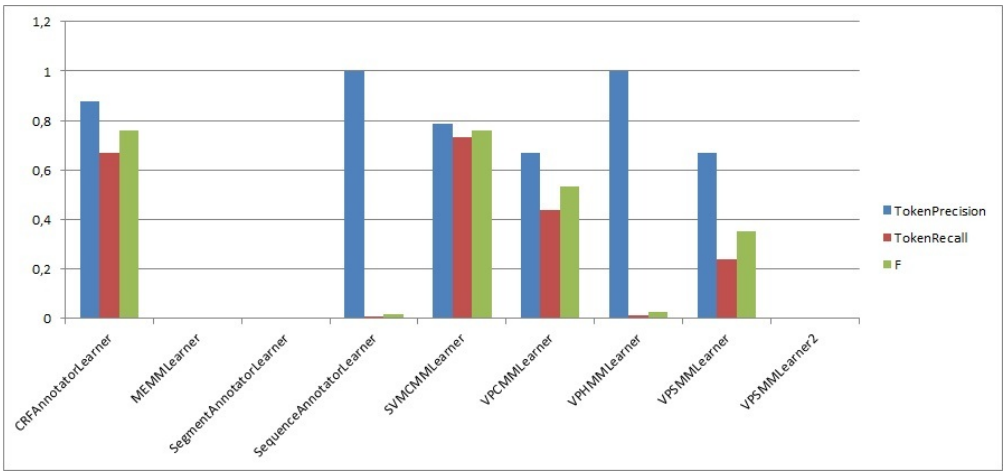

Figura 3.3: Resultados dos vários algoritmos do Minorthird para a tag data

Nos testes para a tag data (ver figura 3.3), a tendência mantém-se nos algoritmos CRF e SVM, ambos com resultados bastante próximos e bem distintos em relação aos restantes algoritmos testados. No CRF, verificou-se o valor de precisão 0,9094, indicando que 90,94% das tags extraídas, eram datas; 0,6077 de cobertura, 60,77% das tags datas foram bem classificadas em relação ao universo de tags data nos documentos; e 0,7286 de medida-f. No caso do SVM, observa-se um resultado de precisão de 0,7755, 0,7361 de cobertura e 0,7553 de medida-f.

Observando o gráfico, da figura 3.4, pode-se constatar que os valores não são tão bons como os dos testes anteriores. Talvez, mesmo tendo mais ocorrências desta tag nos documentos em comparação, por exemplo, com a tag data, os resultados sejam piores, porque na tag data ocorrem números, o que possibilita um reconhecimento mais fácil. Ainda assim, verificam-se valores de precisão de 0,8636 para o CRF e de 0,8519 para o SVM, não existindo quase diferença entre eles. No cobertura e medida-f, já existe alguma diferença, favorecendo o SVM, tendo valores de cobertura de 0,3828 para o CRF e 0,4634 para o SVM; finalmente para o medida-f temos 0,5305 para o CRF e 0,6002

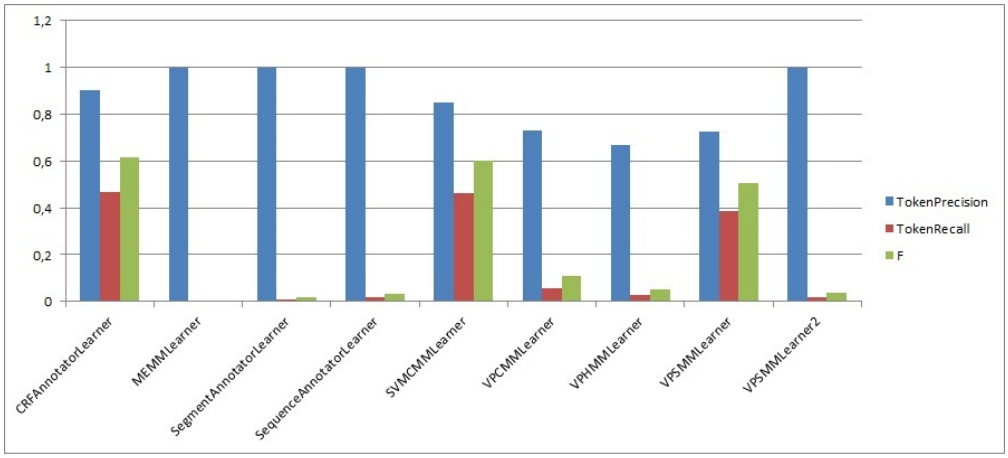

Figura 3.4: Resultados dos vários algoritmos do Minorthird para a tag lugar

para o SVM.

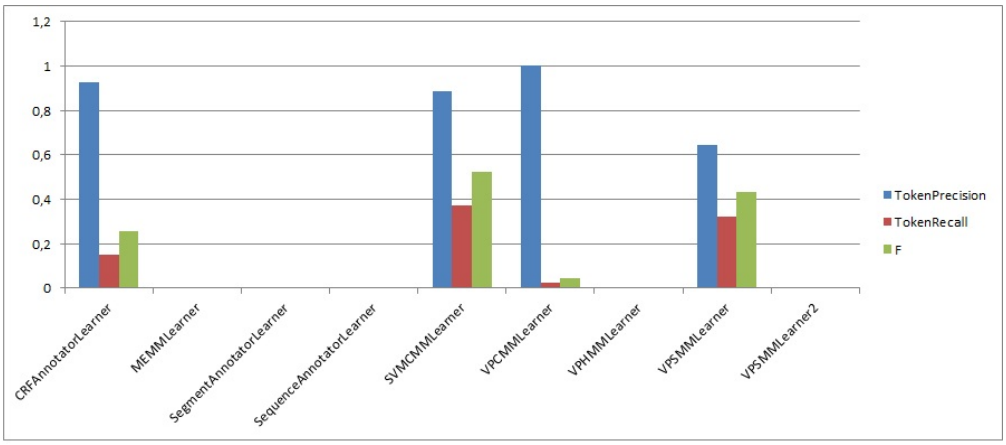

Figura 3.5: Resultados dos vários algoritmos do Minorthird para a tag pessoa

Na tag pessoa (ver figura 3.5), é onde se obtém os piores resultados, devido a esta tag, ser a menos ocorrente de todas nos documentos. No algoritmo CRF, o valor de precisão é de 1 indicando que todos os elementos extraídos, são elementos "pessoa"; o valor de 0,1453 de cobertura indica que apenas 14,53% das tags pessoa foram classificadas em relação a todo o universo de tags pessoa existentes nos documentos; medida-f de 0,2537. No algoritmo SVM verificam-se valores próximos de precisão 0,9153, mas é nos valores de cobertura e medida-f que há melhorias significativas, 0,3017 e 0,4538 respectivamente.

Em todos os gráficos acima, observamos que os algoritmos com melhor desempenho são o CRF e o SVM. Das quatro classes, segundo os valores de precisão e cobertura de cada gráfico, a que demonstra melhores resultados é a das entidades, enquanto que a classe pessoa é a que apresenta os valores mais baixos. Isto deve-se ao facto de que a quantidade total de elementos "pessoa", dentro do *corpus* de documentos, ser muito inferior em relação às outras classes. Os valores para a classe pessoa seriam melhorados caso o número de documentos no *corpus* fosse aumentado, possibilitando o incremento de ocorrências de "pessoas"nos documentos.

Com os resultados obtidos dos gráficos anteriores, criaram-se dois novos gráficos (figuras 3.6 e 3.7), com a comparação das quatro tags para o algoritmo CRF e SVM, pois foram estes que melhores valores alcançaram.

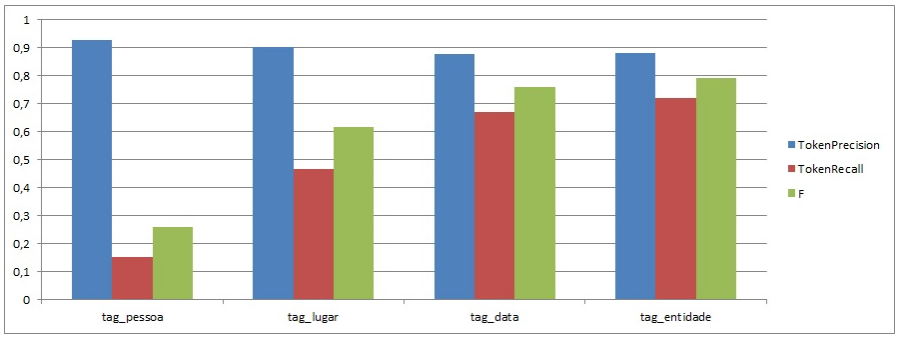

Figura 3.6: Resultados do algoritmo CRF para as várias tags

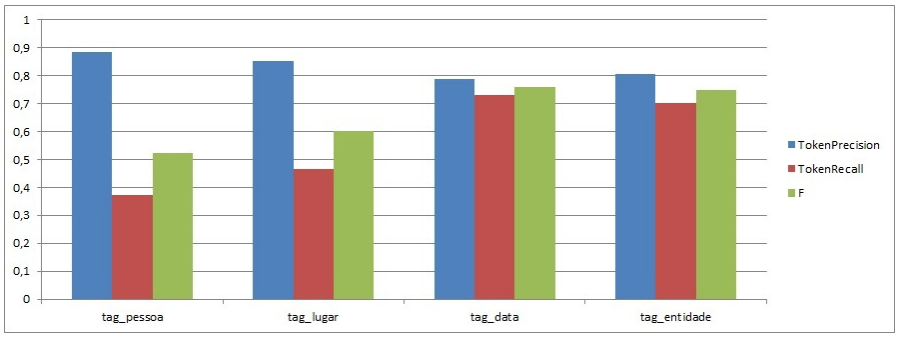

Figura 3.7: Resultados do algoritmo SVM para as várias tags

Tanto num algoritmo, como no outro, as tags entidade e data são as que apresentam valores mais satisfatórios. Ambas as tags situam-se na casa dos 70% nos dois classificadores, tendo apenas um melhor registo no valor de precisão do algoritmo CRF, que se

situa nos 88%. Já no caso da tag pessoa, existe uma diferença significativa entre os dois algoritmos. Os valores de precisão são semelhantes, mas para o cobertura e medida-f há um melhoramento de aproximadamente 20% no SVM.

Os três gráficos seguintes, 3.8, 3.9 e 3.10, representam a comparação dos resultados de precisão, cobertura e medida-f nas quatro etiquetas, para os dois melhores algoritmos, SVM e CRF.

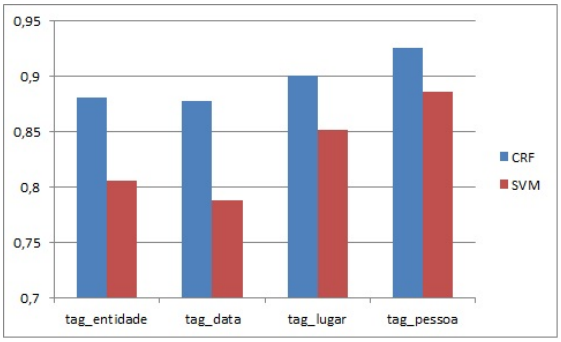

Figura 3.8: Valores de precisão

Figura 3.9: Valores de cobertura

Figura 3.10: Valores de Medida-F

No geral, os valores de cobertura e medida-f, são muito idênticos para os dois algoritmos, menos para a etiqueta pessoa, na qual o SVM se destaca com melhores resultados. Já no caso do gráfico de precisão, o CRF obtém os melhores valores, em todas as etiquetas.

3.3 Como aplicar a classificação a novos documentos

Nesta parte é explicado como se deve proceder para se obter a classificação de entidades de novos documentos.

Primeira tarefa será a criação do ficheiro de treino. Para tal é necessário o programa Minorthird, e o seguinte comando, em que `ALGORITMO` indica qual o algoritmo a ser utilizado, `DIR_TAGS` a directoria com o conjunto de documentos etiquetados e `FICHEIRO_TREINO` o ficheiro de treino necessário para a fase de teste:

```
java edu.cmu.minorthird.ui.TrainExtractor -learner ALGORITMO -output TAG
-labels DIR_TAGS -saveAs FICHEIRO_TREINO -spanType TAG
```

Terminada a fase de treino, desse processo resultará um ficheiro com nome `FICHEIRO_TREINO` que servirá de argumento para a próxima fase.

Tendo sido o ficheiro de treino criado na fase de treino, para executar uma tarefa de teste, tem que executar o comando em baixo, em que `DIR_SEM_TAGS` representa a directoria onde estão localizados os documentos que serão classificados:

```
java edu.cmu.minorthird.ui.ApplyAnnotator -format xml -labels DIR_SEM_TAGS
-loadFrom FICHEIRO_TREINO -saveAs DIR
```

Assim obterá uma nova pasta com nome `DIR`, onde se encontrarão os novos documentos já classificados com as tags encontradas.

3.3.1 Exemplo

Nesta secção é demonstrado o antes e o depois, da aplicação do comando do programa Minorthird para reconhecimento das entidades mencionadas, que foi descrito na secção 3.3. Na figura 3.11 é apresentado um texto sem o reconhecimento de entidades.

```
A Comissao Inter-Sindical de Lisboa Pro-defesa do Horario de
Trabalho, eleita em 6 de Marco p.  p., em reuniao a que assistiram
os organismos operarios:  Associacao dos Maquinistas Fluviais, Asso-
ciacao Fraternal de Classe dos Operarios Alfaiates, Sindicato do
Pessoal da Exploracao do Porto de Lisboa, Sindicato dos Empregados
no Comercio e Industria de Lisboa, Sindicato dos Empregados de Ou-
rivesaria de Lisboa, Sindicato do Pessoal do Arsenal do Exercito,
Sindicato Ferroviario, Sindicato dos Profissionais Culinarios,
Associacao de Classe dos Caixeiros de Lisboa, ...
```

Figura 3.11: Excerto de documento do AHS antes do teste

Considerando que o treino foi efectuado anteriormente, e que utilizamos os ficheiros de treino como argumento no comando 3.3, vamos obter os textos já com o reconhecimento das entidades mencionadas. Para obtermos todas as entidades mencionadas, que poderão vir a estar no documento, tem que se correr o programa com os diferentes ficheiros de treino, respectivos a cada entidade mencionada, obtendo o resultado que está na figura 3.12.

```
A <e>Comissao Inter-Sindical de Lisboa Pro-defesa do Horario de
Trabalho</e>, eleita em <d>6 de Marco</d> p.  p., em reuniao a que
assistiram os organismos operarios:  <e>Associacao dos Maquinistas
Fluviais</e>, <e>Asso- ciacao Fraternal de Classe dos Operarios
Alfaiates</e>, <e>Sindicato do Pessoal da Exploracao do Porto de
Lisboa</e>, <e>Sindicato dos Empregados no Comercio e Industria
de Lisboa</e>, <e>Sindicato dos Empregados de Ou- rivesaria de
Lisboa</e>, <e>Sindicato do Pessoal do Arsenal do Exercito</e>,
<e>Sindicato Ferroviario</e>, <e>Sindicato dos Profissionais
Culinarios</e>, <e>Associacao de Classe dos Caixeiros de Lisboa</e>,
...
```

Figura 3.12: O mesmo documento do AHS após o teste e reconhecimento das entidades

3.4 Pesquisa através de Entidades Mencionadas

Uma das potencialidades da extracção das EM, é a possibilidade de efectuar pesquisas sobre pessoas, locais, entidades e datas, referidas nos documentos do AHS. Para isso existe uma ligação ao sistema de pesquisa do portal, que procurará pelos termos introduzidos, indicando a que EM e documento pertencem. Desta forma, facilita ao utilizador a pesquisa de textos do seu interesse, para posterior leitura na estante virtual. Graças ao REM esta tarefa torna-se muito mais simples, visto o trabalho de identificação das EM não ser efectuado na altura da pesquisa, mas sim anteriormente, sendo esses dados guardados numa base de dados.

4. Conclusões e trabalho futuro

4.1 Conclusões

O trabalho desenvolvido nesta dissertação teve como principal tema o reconhecimento de entidades mencionadas em textos provenientes do AHS, estando também enquadrado no projecto de investigação PTDC/CPJ-CPO/098500/2008 da FCT. Relativamente ao objectivo principal, o reconhecimento de entidades mencionadas, utilizou-se a ferramenta Minorthird, que permitiu o teste e classificação das etiquetas propostas para identificação nos documentos históricos, possibilitando assim a comparação entre os vários algoritmos aplicados a cada uma das etiquetas.

Verificou-se que para algumas das etiquetas, os resultados eram pouco satisfatórios, isto explicado pelo número insuficiente de documentos com qualidade suficiente para se retirar o texto automaticamente, através de programa de OCR (corpus constituído por 98 textos). Com este trabalho, pretende-se que no futuro se apliquem estas técnicas para extracção automática de entidades, datas, lugares e pessoas a textos novos, introduzidos no arquivo digital do AHS, possibilitando assim meios para efectuar pesquisas personalizadas, recuperando apenas a informação desejada.

É no suporte digital que se enquadra o projecto de investigação da FCT. Para o projecto foi necessário o desenvolvimento de um arquivo digital, que correspondesse às necessidades do AHS, tendo sido necessário a instalação e configuração de um servidor, dotado de capacidade para armazenamento das digitalizações de alta resolução dos documentos. Ainda no servidor, foram instaladas ferramentas para a gestão do arquivo digital (Archon) e uma outra para gestão de conteúdos (Joomla) num portal, para acesso a utilizadores. No portal, houve a necessidade de introduzir os conteúdos do Archon, para que permitisse a navegação do AHS dentro do próprio portal, com suporte à leitura

dos documentos através da criação, em java, da estante de leitura virtual.

No final, conseguiu-se implementar uma plataforma que possibilita a pesquisa e leitura de vários documentos, através de ferramentas vocacionadas para o efeito e também métodos para a extracção e identificação de entidades importantes nos textos, que ajudem a classificação de documentos.

4.2 Trabalho futuro

Como trabalho futuro pretende-se efectuar a ligação a um sistema geo-referenciado, tipo *Google maps*[1], onde se poderiam seleccionar documentos, onde fossem referidos locais de determinadas zonas do mapa. Exemplo disso seria efectuar uma pesquisa na zona de Évora, resultando todos os documentos que contenham Évora, como local. Mais complexo seria a pesquisa de documentos, tendo como epicentro Évora novamente, estabelecer um raio de X quilómetros como parâmetro e retornar todos os documentos que contenham, como locais, as localidades dentro dessa circunferência.

[1] http://maps.google.pt/

A. Conteúdos em Anexo

Outputs do programa Minorthird, com os resultados dos testes para as várias etiquetas e algoritmos.

- Etiqueta data:
  - **CRFAnnotatorLearner**
    Overall performance:
    TokenPrecision: 0,8780 TokenRecall: 0,6710 F: 0,7606
    SpanPrecision: 0,7397 SpanRecall: 0,4954 F: 0,5934
    Total time for task: 1452.192 sec
  - **MEMMLearner**
    Overall performance:
    TokenPrecision: 0,0000 TokenRecall: 0,0000 F: 0,0000
    SpanPrecision: 0,0000 SpanRecall: 0,0000 F: 0,0000
    Total time for task: 2899.223 sec
  - **SegmentAnnotatorLearner**
    Overall performance:
    TokenPrecision: 0,0000 TokenRecall: 0,0000 F: 0,0000
    SpanPrecision: 0,0000 SpanRecall: 0,0000 F: 0,0000
    Total time for task: 610.586 sec
  - **SequenceAnnotatorLearner**
    Overall performance:

TokenPrecision: 1,0000 TokenRecall: 0,0078 F: 0,0154

SpanPrecision: 0,0000 SpanRecall: 0,0000 F: 0,0000

Total time for task: 126.97 sec

- **SVMCMMLearner**

  Overall performance:

  TokenPrecision: 0,7877 TokenRecall: 0,7306 F: 0,7581

  SpanPrecision: 0,6111 SpanRecall: 0,5046 F: 0,5528

  Total time for task: 144.455 sec

- **VPCMMLearner**

  Overall performance:

  TokenPrecision: 0,6693 TokenRecall: 0,4404 F: 0,5313

  SpanPrecision: 0,2653 SpanRecall: 0,1193 F: 0,1646

  Total time for task: 323.881 sec

- **VPHMMLearner**

  Overall performance:

  TokenPrecision: 1,0000 TokenRecall: 0,0130 F: 0,0256

  SpanPrecision: 0,0000 SpanRecall: 0,0000 F: 0,0000

  Total time for task: 43.537 sec

- **VPSMMLearner**

  Overall performance:

  TokenPrecision: 0,6715 TokenRecall: 0,2383 F: 0,3518

  SpanPrecision: 0,3390 SpanRecall: 0,1835 F: 0,2381

  Total time for task: 3247.726 sec

- **VPSMMLearner2**

  Overall performance:

  TokenPrecision: 0,0000 TokenRecall: 0,0000 F: 0,0000

  SpanPrecision: 0,0000 SpanRecall: 0,0000 F: 0,0000

  Total time for task: 1931.529 sec

- Etiqueta entidade:

  - **CRFAnnotatorLearner**

    Overall performance:

TokenPrecision: 0,8807 TokenRecall: 0,7214 F: 0,7931

SpanPrecision: 0,7492 SpanRecall: 0,6037 F: 0,6686

Total time for task: 1755.296 sec

- **MEMMLearner**

  Overall performance:

  TokenPrecision: 0,0000 TokenRecall: 0,0000 F: 0,0000

  SpanPrecision: 0,0000 SpanRecall: 0,0000 F: 0,0000

  Total time for task: 2609.405 sec

- **SegmentAnnotatorLearner**

  Overall performance:

  TokenPrecision: 0,5556 TokenRecall: 0,0049 F: 0,0098

  SpanPrecision: 0,2727 SpanRecall: 0,0040 F: 0,0078

  Total time for task: 534.82 sec

- **SequenceAnnotatorLearner**

  Overall performance:

  TokenPrecision: 0,7542 TokenRecall: 0,3875 F: 0,5120

  SpanPrecision: 0,1299 SpanRecall: 0,0964 F: 0,1107

  Total time for task: 143.073 sec

- **SVMCMMLearner**

  Overall performance:

  TokenPrecision: 0,8061 TokenRecall: 0,7016 F: 0,7503

  SpanPrecision: 0,6787 SpanRecall: 0,5694 F: 0,6193

  Total time for task: 367.89 sec

- **VPCMMLearner**

  Overall performance:

  TokenPrecision: 0,8906 TokenRecall: 0,3776 F: 0,5304

  SpanPrecision: 0,2385 SpanRecall: 0,0819 F: 0,1219

  Total time for task: 999.043 sec

- **VPHMMLearner**

  Overall performance:

  TokenPrecision: 0,7887 TokenRecall: 0,4026 F: 0,5331

  SpanPrecision: 0,1577 SpanRecall: 0,1004 F: 0,1227

  Total time for task: 56.89 sec

  - **VPSMMLearner**

    Overall performance:

    TokenPrecision: 0,7675 TokenRecall: 0,2270 F: 0,3503

    SpanPrecision: 0,6491 SpanRecall: 0,3421 F: 0,4481

    Total time for task: 3161.017 sec

  - **VPSMMLearner2**

    Overall performance:

    TokenPrecision: 0,8000 TokenRecall: 0,0039 F: 0,0079

    SpanPrecision: 0,2857 SpanRecall: 0,0026 F: 0,0052

    Total time for task: 2415.745 sec

- Etiqueta lugar:

  - **CRFAnnotatorLearner**

    Overall performance:

    TokenPrecision: 0,9012 TokenRecall: 0,4670 F: 0,6152

    SpanPrecision: 0,8492 SpanRecall: 0,5409 F: 0,6609

    Total time for task: 1757.515 sec

  - **MEMMLearner**

    Overall performance:

    TokenPrecision: 1,0000 TokenRecall: 0,0018 F: 0,0037

    SpanPrecision: 1,0000 SpanRecall: 0,0036 F: 0,0071

    Total time for task: 2961.792 sec

  - **SegmentAnnotatorLearner**

    Overall performance:

    TokenPrecision: 1,0000 TokenRecall: 0,0085 F: 0,0169

    SpanPrecision: 1,0000 SpanRecall: 0,0142 F: 0,0281

    Total time for task: 501.307 sec

  - **SequenceAnnotatorLearner**

    Overall performance:

    TokenPrecision: 1,0000 TokenRecall: 0,0171 F: 0,0335

    SpanPrecision: 1,0000 SpanRecall: 0,0285 F: 0,0554

    Total time for task: 124.666 sec

  - **SVMCMMLearner**

    Overall performance:

    TokenPrecision: 0,8516 TokenRecall: 0,4648 F: 0,6014

    SpanPrecision: 0,8342 SpanRecall: 0,5552 F: 0,6667

    Total time for task: 244.317 sec

  - **VPCMMLearner**

    Overall performance:

    TokenPrecision: 0,7297 TokenRecall: 0,0576 F: 0,1067

    SpanPrecision: 0,8065 SpanRecall: 0,0890 F: 0,1603

    Total time for task: 552.835 sec

  - **VPHMMLearner**

    Overall performance:

    TokenPrecision: 0,6667 TokenRecall: 0,0256 F: 0,0493

    SpanPrecision: 0,5882 SpanRecall: 0,0356 F: 0,0671

    Total time for task: 45.168 sec

  - **VPSMMLearner**

    Overall performance:

    TokenPrecision: 0,7269 TokenRecall: 0,3859 F: 0,5042

    SpanPrecision: 0,7430 SpanRecall: 0,5658 F: 0,6424

    Total time for task: 3045.856 sec

  - **VPSMMLearner2**

    Overall performance:

    TokenPrecision: 1,0000 TokenRecall: 0,0192 F: 0,0377

    SpanPrecision: 1,0000 SpanRecall: 0,0320 F: 0,0621

    Total time for task: 582.675 sec

- Etiqueta pessoa:

  - **CRFAnnotatorLearner**

    Overall performance:

    TokenPrecision: 0,9259 TokenRecall: 0,1497 F: 0,2577

    SpanPrecision: 0,9000 SpanRecall: 0,1184 F: 0,2093

    Total time for task: 1396.939 sec

- **MEMMLearner**

  Overall performance:

  TokenPrecision: 0,0000 TokenRecall: 0,0000 F: 0,0000

  SpanPrecision: 0,0000 SpanRecall: 0,0000 F: 0,0000

  Total time for task: 2876.285 sec

- **SegmentAnnotatorLearner**

  Overall performance:

  TokenPrecision: 0,0000 TokenRecall: 0,0000 F: 0,0000

  SpanPrecision: 0,0000 SpanRecall: 0,0000 F: 0,0000

  Total time for task: 500.186 sec

- **SequenceAnnotatorLearner**

  Overall performance:

  TokenPrecision: 0,0000 TokenRecall: 0,0000 F: 0,0000

  SpanPrecision: 0,0000 SpanRecall: 0,0000 F: 0,0000

  Total time for task: 120.933 sec

- **SVMCMMLearner**

  Overall performance:

  TokenPrecision: 0,8516 TokenRecall: 0,4648 F: 0,6014

  SpanPrecision: 0,8342 SpanRecall: 0,5552 F: 0,6667

  Total time for task: 244.317 sec

- **VPCMMLearner**

  Overall performance:

  TokenPrecision: 1,0000 TokenRecall: 0,0240 F: 0,0468

  SpanPrecision: 1,0000 SpanRecall: 0,0132 F: 0,0260

  Total time for task: 317.963 sec

- **VPHMMLearner**

  Overall performance:

  TokenPrecision: 0,0000 TokenRecall: 0,0000 F: 0,0000

  SpanPrecision: 0,0000 SpanRecall: 0,0000 F: 0,0000

  Total time for task: 42.257 sec

- **VPSMMLearner**

  Overall performance:

TokenPrecision: 0,6429 TokenRecall: 0,3234 F: 0,4303

SpanPrecision: 0,5385 SpanRecall: 0,2763 F: 0,3652

Total time for task: 3081.309 sec

– **VPSMMLearner2**

Overall performance:

TokenPrecision: 0,0000 TokenRecall: 0,0000 F: 0,0000

SpanPrecision: 0,0000 SpanRecall: 0,0000 F: 0,0000

Total time for task: 570.139 sec

Bibliografia

[AFM+08] Carlos Amaral, Helena Figueira, Afonso Mendes, Pedro Mendes, Cláudia Pinto, and Tiago Veiga. Adaptação do sistema de reconhecimento de entidades mencionadas da priberam ao harem. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 2008.

[AI99] Douglas E. Appelt and David J. Israel. Introduction to information extraction technology. *AI Communications*, 1999.

[BM06] Sabine Buchholz and Erwin Marsi. Conll-x shared task on multilingual dependency parsing. *Proceedings of the Tenth Conference on Computational Natural Language Learning*, 2006.

[CKGS06] Chia-Hui Chang, Mohammed Kayed, Moheb Ramzy Girgis, and Khaled Shaalan. A survey of web information extraction systems. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, 2006.

[Coh04] William W. Cohen. Minorthird: Methods for identifying names and ontological relations in text using heuristics for inducing regularities from data. 2004.

[CS07] Nuno Cardoso and Diana Santos. Directivas para a identificação e classificação semântica na colecção dourada do harem. *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*, 2007.

[CV95] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 1995.

[dA07] José João Dias de Almeida. Rena - reconhecedor de entidades. *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*, 2007.

[GQ03] Teresa Gonçalves and Paulo Quaresma. A preliminary approach to the multilabel classification problem of portuguese juridical documents. *Lecture Notes in Computer Science*, 2003.

[GQ04] Teresa Gonçalves and Paulo Quaresma. Using ir techniques to improve automated text classification. *Natural Language Processing and Information Systems*, 2004.

[GS96] Ralph Grishman and Beth Sundheim. Message understanding conference-6: a brief history. *COLING '96 Proceedings of the 16th conference on Computational linguistics*, Volume 1, 1996.

[har] Harem - metodologia. *http://www.linguateca.pt/primeiroHAREM/harem.html*.

[HGMZ03] Hui Han, C. Lee Giles, Eren Manavoglu, and Hongyuan Zha. Automatic document metadata extraction using support vector machines. 2003.

[Joa98] Thorsten Joachims. Text categorization with support vector machines: Learning with many relevant features. 1998.

[KG05] Vijay Krishnan and Vignesh Ganapathy. Named entity recognition. 2005.

[KM05] Katharina Kaiser and Silvia Miksch. Information extraction: A survey. 2005.

[LMP01] John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

[Mar08] Bruno Martins. O sistema cage no segundo harem. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 2008.

[Moe06] Marie-Francine Moens. Information extraction: Algorithms and prospects in a retrieval context. 2006.

[MSC07] Bruno Martins, Mário J. Silva, and Marcirio Silveira Chaves. O sistema cage no harem - reconhecimento de entidades geográficas em textos em língua

portuguesa. *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*, 2007.

[RBC+98] Patricia Robinson, Erica Brown, Nancy Chinchor, John Burger, Aaron Douthat, Lisa Ferro, and Lynette Hirschman. Overview: Information extraction from broadcast news. *In Proceedings of DARPA Broadcast News Workshop*, 1998.

[San07] Diana Santos. O modelo semântico usado no primeiro harem. *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM*, 2007.

[Sar07] Luís Sarmento. O siemês e a sua participação no harem e no mini-harem. *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*, 2007.

[SBP05] Eduardo F.A. Silva, Flávia A. Barros, and Ricardo B. C. Prudêncio. Uma abordagem de aprendizagem híbrida para extração de informação em textos semi-estruturados. *XXV Congresso da Sociedade Brasileira de Computação*, 2005.

[SC07] Diana Santos and Nuno Cardoso. Breve introdução ao harem. *Reconhecimento de entidades mencionadas em português: Documentação e actas do HAREM, a primeira avaliação conjunta na área*, 2007.

[SCFO08] Diana Santos, Paula Carvalho, Cláudia Freitas, and Hugo Gonçalo Oliveira. Segundo harem: Directivas de anotação. *Desafios na avaliação conjunta do reconhecimento de entidades mencionadas: O Segundo HAREM*, 2008.

[SM03] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the conll-2003 shared task: language-independent named entity recognition. *Proceedings of the seventh conference on Natural language learning*, 2003.

[Sod99] Stephen Soderland. Learning information extraction rules for semi-structured and free text. *Machine Learning*, 1999.

[Wal04] Hanna M. Wallach. Conditional random fields: An introduction. 2004.

[WC00] Vincent Wan and William M. Campbell. Support vector machines for speaker verification and identification. *IEEE Proceeding*, 2000.