Universidade do Minho

Mestrado Integrado em Engenharia Informática

LEI - Laboratório Engenharia Informática

Sistema de apoio à ingestão e catalogação de posts e comentários

André Salgueiro (A77617)

Bruno Carvalho(A76987)

Fábio Araújo (A78508)

15 de Abril de 2019

Resumo

Este documento diz respeito ao projeto de LEI, Laboratórios de Engenharia Informática, relativo ao perfil de PLC, Processamento de Linguagens e Conhecimento, do quarto ano de Engenharia Informática.

De uma forma sucinta, este projeto encontra-se inserido num projeto internacional, denominado NetLang, cujo principal objetivo é a análise de comentários de teor discriminatório em plataformas online, como são exemplo, as redes sociais.

Conteúdo

Lista de Figuras

1 Introdução

Este documento refere-se ao projeto de LEI, Laboratórios de Engenharia Informática, relativo ao perfil de Processamento de Linguagens e Conhecimento, do quarto ano de Engenharia Informática.

O projeto encontra-se inserido num projeto internacional, denominado NetLang, cujo objetivo é o estudo de comentários ofensivos e discriminatórios em plataformas online, como as redes sociais.

Ao longo deste documento iremos explanar o desenvolvimento deste projeto. Numa primeira fase procura-se apresentar o projeto, indicando no que consiste e o seu propósito. Numa segunda fase procura-se apresentar a abordagem tomada pelo grupo de trabalho ao longo deste projeto de maneira a atingir os objetivos anteriormente definidos.

Numa fase final deste documento serão expostos os resultados até agora conseguidos, com a apresentação de exemplos concretos, e a projeção do trabalho futuro.

2 Enquadramento

Este projeto surge no âmbito dos projetos de Laboratórios de Engenharia Informática, sendo estes à escolha dos alunos que possuam o perfil que abrange o projeto em específico. Este projeto encontra-se inserido no perfil de Processamento de Linguagens e Conhecimento, e é suportado no aprendizado deste perfil até então fornecido que irá decorrer o trabalho desenvolvido sobre o projeto.

Este projeto, é um projeto internacional de investigação suportado pela FCT, Fundação para a Ciência e a Tecnologia, denominado NetLang. O NetLang tem como objetivo a criação de corpora anotado que permita estudar formas de insulto e discriminação presente nos painéis de comentários das redes sociais e sites de jornais.

A corpora anteriormente referida, representa meta-informação que advém da análise dos *posts* pelos comentários surgentes. Um dos principais pontos da meta-informação é inserir cada um dos *posts* numa área muito abrangente, que é caracterizada pelo tipo de preconceito. A ideia geral é que através de blocos de informação previamente concebidos, com a identificação de palavras-chave seja possível associá-las ao tipo de preconceito.

3 Motivação

É importante perceber que a pretensão do projeto NetLang visa a compilação de uma vasta coleção de *posts* e os seus respetivos comentários, pelo que a obtenção da meta-informação por meios manuais representa uma tarefa quase impossível, senão impossível, impraticável quando associado o tempo de processamento de tal tarefa.

Por este motivo, a criação de um método computacional que possibilite a obtenção dos mesmos resultados mas que elimine a arduidade deste processo é estritamente necessária.

A ferramenta a criar, dentro das suas capacidades, deverá possibilitar a recolha de dados estatísticos, o processamento semi-automático da meta-informação e a representação dos *posts* e comentários em formato TEI.

4 Objetivos

Relembrando aquilo que foi referido anteriormente, com a introdução das temáticas relativas aos comentários e às palavras-chave *(keywords)*, e principalmente a matéria de trabalho (ficheiros em análise) é possível traçar com clareza os principais objetivos deste projeto.

Primeiramente, dado que a informação relativa às palavras-chave encontra-se em contexto de tabela, é importante que a mesma seja representada num formato estruturado que permita um fácil manuseio, como é o formato JSON. Na mesma linha de pensamento, uma vez que a informação relativa aos comentários possa encontrar-se no formato *TAB* é necessário desenvolver um programa responsável pela transformação dessa informação para o formato JSON.

Seguidamente, depois de normalizada toda a informação a tratar, por forma a possibilitar a associação dos *posts* aos tipos de preconceito, é necessária a criação de um motor de procura das palavras-chave em cada um dos comentários.

Posteriormente, em resultado do trabalho realizado pelo motor de procura, é solicitada a criação de uma análise estatística, baseada na ocorrência das palavras-chave, com apresentação em formato à escolha, e por fim a criação de uma interface híbrida que para além de acolher os resultados do motor para a construção da meta-informação, dá oportunidade ao utilizador de a complementar.

5 Arquitetura do Sistema

Nesta secção é apresentada a figura 1, que representa a arquitetura do sistema, ou seja, o trabalho a desenvolver ao longo do projeto, em forma de esquema, o que proporciona uma visão mais ampla e facilitada do projeto.

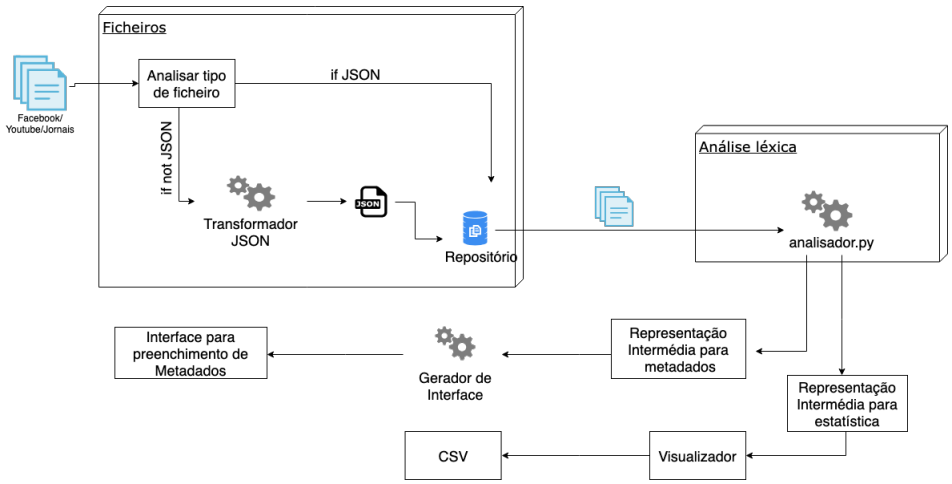


Figura 1: Arquitetura do Sistema

6 Resultados

Tendo em conta a definição dos objetivos para este projeto, presentes na secção anterior, passemos agora a apresentar os resultados obtidos, até então.

- Transcrição de documentos em análise para formato JSON
  - ◇ Coletânea de palavras-chave *(keywords)* relativa a cada preconceito

    As seguintes imagens são exemplo da transcrição efetuada, neste caso de um excerto associado ao preconceito *'Sexismo'*.

TYPES OF PREJUDICE	SOCIOLINGUISTIC VARIABLES	KEYWORDS (English)	KEYWORDS (Portuguese)
SEXISM	Gender	- Male chauvinism /chauvinist. - Gender. Sex- (Sexual, Sexism…). Misogyn-(-y, -ous, -nist, -e). Patriarchy. - Woman. Chick. Dame. Cock tease. Dumb blonde. Becky. Make me a sandwich. - (Old) Hag. Crone. Witch. Minger (UK). - B*tch. Promiscuous. Sl*t. Tart. Wh*re. - Feminazi.	- Machismo. Machista. - Género. Sex- (Sexual, Sexismo). Misogin- (Misógino, Misoginia). Patriarcado. - Mulher. Gaja. Tipa. Garina. Chavala. Miúda. Burra loura/loira. Boazuda. - Bruxa. Feiosa. Mostrenga. Monstra. Velha. Camafeu. - Badalh*ca. C*bra. C*dela. Galdéria. Promíscua. P*ta. Quenga. Mula. Ninfomaníca. Rameira. Safada. Serigaita. Vadia. Vaca. Ordinária. Mulher de vida fácil. -Brasil (Piranha, Ganda, Marmita, Biscaite, Ridícula, Cachorra, Preparada, Popozuda, Vagabunda, Égua, Potranca, Boca de Veludo, Boqueteira, Prostituta, Maria Gasolina, Maria Chuteira, Machadeira, Seringueira, Chupona, Boca de Pêlo).

Figura 2: Excerto do documento com o conjunto das *keywords* associadas a ca preconceito.

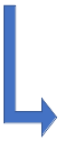
```
{
    "type_prejudice": "Sexism",
    "Sociolinguistic variables" : {
        "Gender": [/**Keywords**/]
    }
},
...
```

Figura 3: Transcrição do excerto para JSON

◇ Comentários

A seguinte figura representa a transcrição de um comentário em formato tabular para o formato JSON.

position	post_id	post_by	post_text	post_published	comment_id	comment_by	is_reply	comment_message	...
61_0	469974510474_10161491403995475	c992ae2630a5223d06b076adb83368922 5214a4e	" Eu não quero ser mártir. Eu quero viver afirmou Jean Wyllys em entrevista"	2019-01-25T16:15:00+0000	101614914039 95475_101614 92275085475	da39a3ee5e6b4b0d3255bfef95601890afd80709	1	"Ele já chegou. Está a viver em casa do Mamadou."	...




```
{ "position": "61_0",
  "post_id": "469974510474_10161491403995475",
  "post_by": "c992ae2630a5223d06b076adb833689225214a4e",
  "post_text": "Eu não quero ser mártir. Eu quero viver  afirmou Jean Wyllys em entrevista",
  "post_published": "2019-01-25T16:15:00+0000",
  "comment_id": "10161491403995475_10161492275085475",
  "comment_by": "da39a3ee5e6b4b0d3255bfef95601890afd80709",
  "is_reply": 1,
  "comment_message": "Ele já chegou. Está a viver em casa do Mamadou.",
  "comment_published": "2019-01-25T17:57:52+0000",
  "comment_like_count": 0,
  "attachment_type": "",
  "attachment_url": ""
 }
```


Figura 4: Transcrição de comentário

- Motor de procura

Relativamente ao motor de procura, o que poderá ser apresentado, é a sua mecânica, consistindo na procura das palavras reservadas nos documentos em análise. Esta procura, é sempre categorizada segundo o preconceito em análise. Para tal, a transcrição dos documentos para um formato mais adequado, como o JSON, permite proceder a esta categorização de uma forma mais facilitada.

Para o desenvolvimento do motor, foi utilizada a linguagem *Python*, uma linguagem que possui imensas bibliotecas de leitura de ficheiros, nomeadamente de ficheiros JSON, e possui imensas outras qualidades para o desempenho de tarefas de *scripting*.

As seguintes imagens apresentam, de momento, de inicio a fim, o funcionamento do motor.

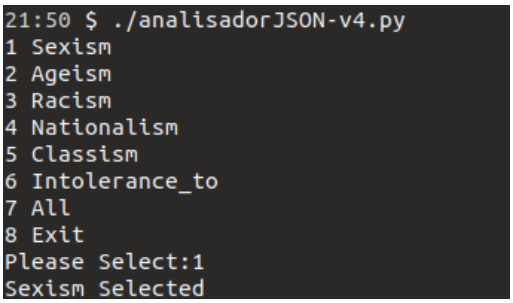


Figura 5: Motor - Fase de seleção do tipo de preconceito.

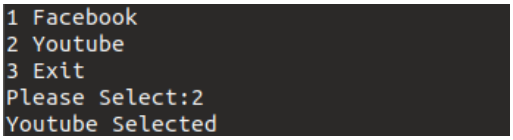
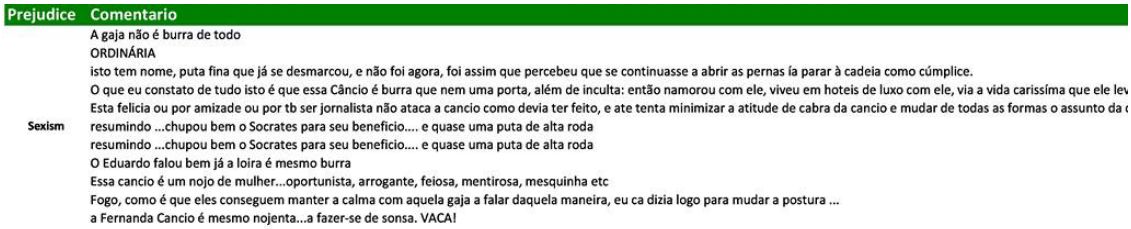


Figura 6: Motor - Fase de seleção do tipo de documentos em análise .

- Representação estatística

  A partir do motor de motor de procura é possível a obtenção de resultados estatísticos como os que se encontram abaixo.

  A primeira parte da tabela representa os comentários onde foram encontradas correspondências com as palavras reservadas *(keywords)*



Prejudice	Comentario
	A gaja não é burra de todo
	ORDINÁRIA
	isto tem nome, puta fina que já se desmarcou, e não foi agora, foi assim que percebeu que se continuasse a abrir as pernas ía parar à cadeia como cúmplice.
	O que eu constato de tudo isto é que essa Câncio é burra que nem uma porta, além de inculta: então namorou com ele, viveu em hoteis de luxo com ele, via a vida carissíma que ele le
	Esta felicia ou por amizade ou por tb ser jornalista não ataca a cancio como devia ter feito, e ate tenta minimizar a atitude de cabra da cancio e mudar de todas as formas o assunto da
Sexism	resumindo ...chupou bem o Socrates para seu beneficio.... e quase uma puta de alta roda
	resumindo ...chupou bem o Socrates para seu beneficio.... e quase uma puta de alta roda
	O Eduardo falou bem já a loira é mesmo burra
	Essa cancio é um nojo de mulher...oportunista, arrogante, feiosa, mentirosa, mesquinha etc
	Fogo, como é que eles conseguem manter a calma com aquela gaja a falar daquela maneira, eu ca dizia logo para mudar a postura ...
	a Fernanda Cancio é mesmo nojenta...a fazer-se de sonsa. VACA!

Figura 7: Representação estatística - Parte I da tabela

A segunda parte da tabela representa as frequências relativas e absolutas de cada palavra reservada, em cada comentário, representadas como um triplo, em que a primeira posição representa a palavra reservada, a segunda a contabilização de ocorrências e a última posição o total de palavras de cada comentário. Esta parte da tabela também apresenta o total de comentários preconceituosos, e o total de palavras reservadas por *post*.

Frequencia	Total	Ocorrencias
C6A5z1RTUk-y2A94AaABAg [('Gaja', 1, 7), ('Burra', 1, 7)]		Gaja ----> 2
:NZEky7rl1k02_0t4AaABAg [('Ordinária', 1, 1)]		Loira ----> 1
\cg_wbFCVXy0_qd4AaABAg [('P*ta', 1, 30)]		Cabra ----> 2
!Y04s840dmR8uYt94AaABAg [('Burra', 1, 86)]		Burra ----> 3
Gg597h03Ze-U-4l4AaABAg [('Cabra', 2, 56), ('Vaca', 1, 56)]		Vaca ----> 2
'LVoPyjwDRBac_4J4AaABAg [('P*ta', 1, 15)]	11/42	Puta ----> 3
16d-Zi5AFab6xNZ54AaABAg [('P*ta', 1, 15)]		Feiosa ----> 1
i41Jskp2s708xLD94AaABAg [('Burra', 1, 10), ('Loira', 1, 10)]		Ordinária ----> 1
JlqI0oCXRWSMRWx4AaABA [('Feiosa', 1, 12)]		
14c60BZB-p1LBhpd4AaABAg [('Gaja', 1, 25)]		
YGPEb2OJKqH4brt4AaABAg [('Vaca', 1, 10)]		

Figura 8: Representação estatística - Parte II da tabela

7 Conclusão

Em suma é possível retirar que o estado atual do projeto se encontra avançado, sendo que já é possível a visualização de resultados bastante interessantes para a análise dos *posts* e respetivos comentários.

No entanto, ainda falta uma parte do projeto bastante importante, que se trata do desenvolvimento da *interface* que permita a construção semi-automática da meta-informação.

Dado isto, é relegado para trabalho futuro o melhoramento do motor, dentro do possível, como a procura mais eficaz das palavras reservadas nos documentos em análise e principalmente o desenvolvimento da *interface*.

Referências

[1]

8 Anexos