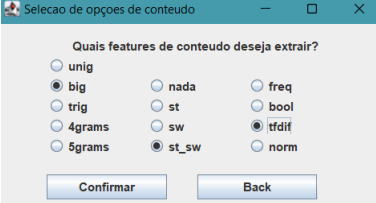

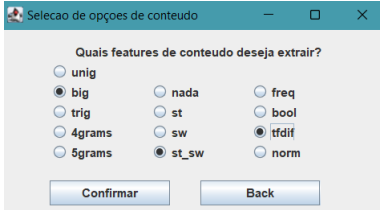


Comentários ao Feature Extractor v01

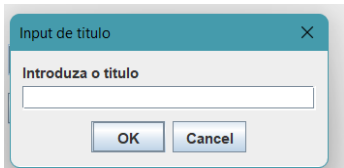



Em geral lê a partir da pasta 'Origem' E salva na 'Output'.

Content-Based Features (CBF)	<p>CBF Features (número variável de features)</p> <p>Está a criar corretamente os ficheiros ex: 'CBF_unig_nada_freq.csv' mas na raiz do projeto em vez de ser na pasta 'src/Output'.</p>  <p>À partida em CBF basta extrair de unig, big e trig, depois sw e st_sw e depois freq (eventualmente também o tfidf)</p> <p>Atenção que o tfidf não está a funcionar. Não está a gerar o ficheiro csv.</p>
<p>Stanford POS Tagger</p>  <p>Eventualmente tiraria a opção do 'Stanford POS Tagger' do menu CBF e colocaria no menu principal (em cima) com o nome 'Representação da letra como POSTags'.</p> <p>Assim o menu CBF ficava com apenas uma opção e sendo assim podia passar diretamente para a vista</p> 	<p>Esta opção não corresponde a nenhuma feature. O que faz é atribuir a cada palavra da letra na pasta 'Origem' a sua classe gramatical.</p> <p>Ex:</p> <p>Want/JJ to/TO find/VB out/RP where/WRB the/DT moon/NN goes/VBZ When/WRB it/PRP leaves/VBZ the/DT western/JJ sky/NN And/CC night/NN dissolves/VBZ again/RB to/TO morning/NN Azure/NNP turns/VBZ to/TO gold/NN Azure/NNP turns/VBZ to/TO gold/NN</p> <p>Está a gravar na pasta 'Output' mas com um nome incorreto ex: CapitalLetters_M45L001-141.csv para a letra L001-141.csv. Devia fazer com que o nome fosse algo como L001-141_with_POSTags.csv</p>
Stylistic Based Features	

Comentários ao Feature Extractor v01

Slang (1 feature)	<p>Está a extrair bem (na pasta Output) mas atribui um nome errado ao ficheiro 'WordsDictionary.csv'. Deveria ser algo como 'slang.csv'.</p> <p>Está a extrair também um ficheiro auxiliar chamado 'Combined_Features_outputDetails.csv' que eventualmente pode evitar que seja produzido (não trás nada de novo e pode confundir).</p>
Capital Letters (2 features: ACL e FCL)	<p>Está a extrair bem para a pasta correta, mas o nome do ficheiro produzido é 'CapitalLetters_M45.csv'. Pode tirar o '_M45'.</p>
POS_Frequency (36 features) Indica o número de adjetivos, nomes, determinantes, etc (num total de 36 classes gramaticais) na letra.	<p>Passos</p> <p>1º Para cada letra criar um ficheiro com as POS Tags ex: JJ TO VB RP WRB DT NN VBZ WRB PRP VBZ DT JJ NN CC NN VBZ RB TO NN NNP VBZ TO NN NNP VBZ TO NN</p> <p>Eventualmente guardar estas letras txt numa pasta chamada 'Origem_POS'. Se a pasta 'Origem' tiver 180 letras a pasta 'Origem_POS' fica com 180 ficheiros.</p> <p>2º Aplicar as CBF com 'unigramas', 'nada', 'freq' que vai criar um csv com 36 features.</p> <p>NOTA: No file 'StanfordPosTagger/SPT_Initial' colocar o parâmetro =2 como em:</p> <pre>34 35 public class SPT_Initial { 36 static final String LocTagger = "src/AuxiliarFiles/bidirectional-distsim-wsj-0-18.tagger"; 37 String sourceFolder; // pasta onde estão as líricas 38 // a processar 39 static final int withTags = 1; 40 static final int onlyTags = 2; 41 42 int option; 43 44 //output folder 45 String outputFolder = "src/Output/"; 46 47 String outputFile; 48 49 public static void main(String[] args) throws ClassNotFoundException, 50 IOException { 51 SPT_Initial spt_initial = new SPT_Initial(true,"src/Origem/L001-141.txt","teste.txt",2); 52 } 53 54 public SPT_Initial(boolean onlyOneFile, String input, String outputFile, int option) throws Clas 55 if(input != null && !input.isEmpty()) { 56 this.sourceFolder = input; 57 } 58 else { 59 this.sourceFolder = "src/Origem/"; 60 } 61 62 if(outputFile != null && !outputFile.isEmpty()) { 63 this.outputFile = outputFile; 64 } 65 }</pre> <p>Problems Javadoc Declaration Console x terminated> SPT_Initial [Java Application] C:\Users\vicar\p2\pool\plugins\org.eclipse.justi.openjdk.hotspot.jre.full.win32.x86_64_17.0.0.v202110 WRB PRP VBZ DT JJ NN CC NN VBZ RB TO NN DT NN VBZ IN PRP\$ NNS DT NN VBZ IN PRP\$ NNS DT NN VBZ IN PRP\$ NNS DT NN VBZ IN PRP\$ NNS</p> <p>Falta fazer esta chamada no GUI e guardar os ficheiros em pasta 'OrigemPOS' e em seguida criar a partir dessa pasta o ficheiro unig_nada_freq_POS.csv</p>
Structural Based Features	

Comentários ao Feature Extractor v01

#Title (1 feature)	<p>Difícil de extrair as features porque tem que ser uma a uma.</p>  <p>O ideal era ter um ficheiro onde fosse ler os títulos e depois confrontasse com o respetivo ficheiro de letra.</p>
#Refrão (1 feature)	ESTÁ A DAR ERRO.
Semantic Based Features	
General Inquirer	Está correto mas o nome do ficheiro não faz muito sentido. Está 'GI_Features-1180.csv'. Eventualmente deveria sair o '-1180'.
Synesketh	Está correto mas como n  terior o nome do ficheiro gerado não faz muito sentido.
DAL_ANEW	Ok mas gera um ficheiro txt a ma 
LIWC	À parte. 
Gazetteers (14 Features)	<p>Só extrai 2 features 'AvgValence e AvgArousal' a partir do ficheiro 'Gazetteers.txt'.</p> <p>Devia ter o 'AvgValence e AvgArousal' para os Gazetteers 'GazQ1-dal.txt, GazQ2-dal.txt, GazQ3-dal.txt, GazQ4-dal.txt e GazQ1Q2Q3Q4_dal.txt'. Isto dá um total de 10 features. Depois extrair mais 4 features que correspondem à frequência de cada um dos quadrantes nas letras i.e., features '#GAZQ1, #GAZQ2, #GAZQ3 e #GAZQ4'.</p>