

Explotación de la Información

Módulo 2. Ejercicio 2

De los siguientes documentos, calcular el conjunto de términos que superarían la fase de filtrado:

- En función de las palabras tal y como aparecen en el texto.
- En función de los lemas.
- En función de los stems del lema.
- Con listas de palabras de parada.
- Con listas de palabras de parada según categorías morfológicas (pronombres, adverbios, preposiciones, determinantes, conjunciones, signos de puntuación).
- Comparar los diferentes resultados: ¿cuál sería la idónea?

Texto 10:

*TEXT 027 01/04/63 PAGE 024

EGYPT UP THE REBELS IT WAS VICTORY DAY IN PORT SAID, SIX YEARS EXACTLY SINCE THE LAST BRITISH SOLDIER LEFT SUEZ . THERE THE REBELS CELEBRATE THE OCCASION WAS EGYPT'S PRESIDENT GAMAL ABDEL NASSER .

Texto 1:

*TEXT 017 01/04/63 PAGE 020

THE REBELS AFTER NASSAU IN DECEMBER 1960, THE U.S . FIRST PROPOSED TO HELP NATO DEVELOP ITS OWN NUCLEAR STRIKE FORCE . BUT EUROPE MADE NO ATTEMPT TO DEVISE A PLAN .

Lista de palabras de parada:

a about above across actually after again against ago all almost along already also although always among an and another any anyone around as at b back bad because before behind best better between big biggest both but by c cent complete d day down during e each early eight enough entire ep etc even ever every everything f face fact far fell few finally first five for found four from g good got h he held her here him himself his hour hours how however i idea if in including instead into it its itself j k l lack last later least led less little long longer lot m man many matter may me men miles million moment month months more morning most much my n near nearly necessary never night no nor not note nothing now o of off often on once one only or other others our out outside over own p page part past per perhaps place point proved q qm question r really recent recently reported round s same sec second section sense seven she short should showed since single six small so some soon still such t ten text than that the their them themselves then there these they thing things third this those though thought thousands three through thus time tiny to today together too took toward two u under until up upon us v very w warning way we week weeks well went were what when where whether which while who whom whose why will with without word words would x y year years yet you your z

Texto 10 con formato “PalabraFrase EtiqMorfológica Lema StemDelLema”

*TEXT VB *text text	PORT NN port port	THERE EX there there
027 CD 027 027	SAID VBD say sai	THE DT the the
01/04/63 CD 01/04/63 010463	, , , ,	REBELS NNS rebel rebel
PAGE NN page page	SIX CD six six	CELEBRATE VB celebrate celebr
024 CD 024 024	YEARS NNS year year	THE DT the the
EGYPT NP egypt egypt	EXACTLY RB exactly exactli	OCCASION NN occasion occasion
UP IN up up	SINCE IN since sinc	WAS VBDBE be be
THE DT the the	THE DT the the	egypt's NP egypt's egypt
REBELS NNS rebel rebel	LAST JJ last last	PRESIDENT NN president presid
IT PPH1 it it	BRITISH JJ british british	GAMAL NP gamal gamal
WAS VBDBE be be	SOLDIER NN soldier soldier	ABDEL NP abdel abdel
VICTORY NN victory victori	LEFT VBN leave leav	NASSER NP nasser nasser
DAY NN day dai	SUEZ NP suiez suiez	. SENT . .
IN IN in in	. SENT . .	

Texto 1 con formato “PalabraFrase EtiqMorfológica Lema StemDelLema”

*TEXT VB *text text

017 CD 017 017

01/04/63 CD 01/04/63 010463

PAGE NN page page

020 CD 020 020

THE NP the the

REBELS NNS rebel rebel

AFTER IN after after

NASSAU NP nassau nassau

IN IN in in

DECEMBER NP december decemb

1960 CD 1960 1960

’,’,’,’

THE DT the the

U.S NP u.s u

. SENT . .

FIRST RB first first

PROPOSED VBN propose propos

TO TO to to

HELP VB help help

NATO NP nato nato

DEVELOP VB develop develop

ITS POS3N its it

OWN JJ own own

NUCLEAR JJ nuclear nuclear

STRIKE NN strike strike

FORCE NN force forc

. SENT . .

BUT CC but but

EUROPE NP europe europ

MADE VBD make make

NO DT no no

ATTEMPT NN attempt attempt

TO TO to to

DEVISE VB devise devis

A DT a a

PLAN NN plan plan

. SENT . .

Significado de las etiquetas morfológicas:

tipoWBB('CC', TPal) :- TPal=[conj].	% Coordinating conjunction
tipoWBB('CD', TPal) :- TPal=[adjSimple(____,num)].	% Cardinal number
tipoWBB('DT', TPal) :- TPal=[art(____,_)].	% Determiner
tipoWBB('EX', TPal) :- TPal=[pronEnglish(____,terc,ind,____,_)].	% antes era "verbo(____,auxiliary,____)" % Existential "there"
tipoWBB('FW', TPal) :- TPal=[sust(____,comun)].	% Foreign word
tipoWBB('IN', TPal) :- TPal=[prepSimple(____,resto)].	% MOD: Preposition or subordinating conjunction
tipoWBB('TO', TPal) :- TPal=[prepSimple(____,to)].	% MOD: Preposition TO
tipoWBB('OF', TPal) :- TPal=[prepSimple(____,of)].	% MOD: Preposition OF
tipoWBB('JJ', TPal) :- TPal=[adjSimple(____,cal)].	% Adjective
tipoWBB('JJR', TPal) :- TPal=[adjSimple(____,cal)].	% Adjective, comparative
tipoWBB('JJS', TPal) :- TPal=[adjSimple(____,cal)].	% Adjective, superlative
tipoWBB('LS', TPal) :- TPal=[marker].	% List item marker
tipoWBB('MD', TPal) :- TPal=[verbo(____,auxiliary)].	% Modal
tipoWBB('NN', TPal) :- TPal=[sust(sing,____,comun)].	% Noun, singular or mass
tipoWBB('NNP', TPal) :- TPal=[sust(sing,____,propio)].	% Proper noun, singular
tipoWBB('NNPS', TPal) :- TPal=[sust(pl,____,propio)].	% Proper noun, plural
tipoWBB('NNS', TPal) :- TPal=[sust(pl,____,comun)].	% Noun, plural
tipoWBB('NP', TPal) :- TPal=[sust(sing,____,propio)].	% Proper noun, singular
tipoWBB('NPS', TPal) :- TPal=[sust(pl,____,propio)].	% Proper noun, plural
tipoWBB('PDT', TPal) :- TPal=[adjSimple(____,ind),pronEnglish(____,____,ind,____,_)].	% Predeterminer
tipoWBB('POS', TPal) :- TPal=[genitive].	% Possessive ending
tipoWBB('PP', TPal) :- TPal=[pronEnglish(____,____,____,____,_)].	% Personal pronoun
tipoWBB('PR', TPal) :- TPal=[pronEnglish(____,____,____,____,_)].	% Pronoun
tipoWBB('PRP', TPal) :- TPal=[pronEnglish(____,____,____,____,_)].	% Pronoun
tipoWBB('PRP\$', TPal) :- TPal=[pronEnglish(____,____,____,____,_)].	% Pronoun, plural
tipoWBB('PPX1', TPal) :- TPal=[pronEnglish(sing,____,persRefl,____,_)].	% singular reflexive personal pronoun
(yourself, itself)	
tipoWBB('PPX2', TPal) :- TPal=[pronEnglish(pl,____,persRefl,____,_)].	% plural reflexive personal pronoun
(yourselves, ourselves)	
tipoWBB('PP\$', TPal) :- TPal=[pronEnglish(____,terc,pos,____,_)].	% Pron. posesivo
tipoWBB('RBNOT', TPal) :- TPal=[adv(neg)].	% Adverbio negacion NOT
tipoWBB('NOT', TPal) :- TPal=[adv(neg)].	% Adverbio negacion NOT
tipoWBB('RB', TPal) :- TPal=[adv(general)].	% Adverb
tipoWBB('RBR', TPal) :- TPal=[adv(general)].	% Adverb, comparative
tipoWBB('RBS', TPal) :- TPal=[adv(general)].	% Adverb, superlative
tipoWBB('RP', TPal) :- TPal=[particle].	% Particle
% tipoWBB('SYM', TPal) :- TPal=[sust(____,comun)].	% Symbol
tipoWBB('UH', TPal) :- TPal=[interjection].	% Interjection
%% VERBOS -----	
tipoWBB('VB', TPal) :- TPal=[verbo(____,infinitivo,resto)].	% Verb, base form
tipoWBB('VBD', TPal) :- TPal=[verbo(____,pasado,resto)].	% Verb, past tense
tipoWBB('VBG', TPal) :- TPal=[verbo(____,gerundio,resto)].	% Verb, gerund or present participle
tipoWBB('VBN', TPal) :- TPal=[verbo(____,participio,resto)].	% Verb, past participle
tipoWBB('VBP', TPal) :- TPal=[verbo(____,presente,resto)].	% Verb, non-3rd person singular present
tipoWBB('VBZ', TPal) :- TPal=[verbo(sing,terc,presente,resto)].	% Verb, 3rd person singular present
%% MOD: Verbo BE -----	
tipoWBB('VBBE', TPal) :- TPal=[verbo(____,infinitivo,be)].	% Verb, base form
tipoWBB('VBDBE', TPal) :- TPal=[verbo(____,pasado,be)].	% Verb, past tense
tipoWBB('VBGBE', TPal) :- TPal=[verbo(____,gerundio,be)].	% Verb, gerund or present participle
tipoWBB('VBNBE', TPal) :- TPal=[verbo(____,participio,be)].	% Verb, past participle

tipoWBB('VBPBE', TPal) :- TPal=[verbo(__,__ ,presente,be)].	% Verb, non-3rd person singular
present	
tipoWBB('VBZBE', TPal) :- TPal=[verbo(sing,terc,presente,be)].	% Verb, 3rd person singular
present	
%% MOD: Verbo HAVE -----	
tipoWBB('VBHV', TPal) :- TPal=[verbo(__,__ ,infinitivo,have)].	% Verb, base form
tipoWBB('VBDHV', TPal) :- TPal=[verbo(__,__ ,pasado,have)].	% Verb, past tense
tipoWBB('VBGHV', TPal) :- TPal=[verbo(__,__ ,gerundio,have)].	% Verb, gerund or
present participle	
tipoWBB('VBNHV', TPal) :- TPal=[verbo(__,__ ,participio,have)].	% Verb, past participle
tipoWBB('VBPHV', TPal) :- TPal=[verbo(__,__ ,presente,have)].	% Verb, non-3rd person
singular present	
tipoWBB('VBZHV', TPal) :- TPal=[verbo(sing,terc,presente,have)].	% Verb, 3rd person singular
present	
%% MOD: Verbo DO -----	
tipoWBB('VBDO', TPal) :- TPal=[verbo(__,__ ,infinitivo,do)].	% Verb, base form
tipoWBB('VBDDO', TPal) :- TPal=[verbo(__,__ ,pasado,do)].	% Verb, past tense
tipoWBB('VBGDO', TPal) :- TPal=[verbo(__,__ ,gerundio,do)].	% Verb, gerund or
present participle	
tipoWBB('VBNDO', TPal) :- TPal=[verbo(__,__ ,participio,do)].	% Verb, past participle
tipoWBB('VBPDO', TPal) :- TPal=[verbo(__,__ ,presente,do)].	% Verb, non-3rd person singular
present	
tipoWBB('VBZDO', TPal) :- TPal=[verbo(sing,terc,presente,do)].	% Verb, 3rd person singular
present	
tipoWBB('WDT', TPal) :- TPal=[pronEnglish(__,__ ,rel,__,__)].	% Wh-determiner
tipoWBB('WP', TPal) :- TPal=[pronEnglish(__,__ ,rel,__,__)].	% Wh-pronoun
tipoWBB('WP\$', TPal) :- TPal=[pronEnglish(__,__ ,rel,__,__)].	% Possessive wh-pronoun
tipoWBB('WRB', TPal) :- TPal=[adv(general)].	% Wh-adverb