

Lista de reserva del producto

ToNgueLP, editor de corpus de textos para tareas NLP.

Producto: ToNgueLP Corpus Tools

Versión: 1.0

Autor: Ing. Abel Meneses Abad

Reglas de Confidencialidad

Clasificación: Software particular para la Universidad de Camagüey, factible de exportar por las partes desarrolladoras, de carácter nacional a evaluar en la convocatoria del CITMA en el 2014.

Este documento contiene información propietaria de **Abel Meneses Abad**, y es emitido confidencialmente para orientar al equipo de desarrollo de ToNgueLP.

El que recibe el documento asume la custodia y control, comprometiéndose a no reproducir, divulgar, difundir o de cualquier manera hacer de conocimientos público su contenido, excepto para cumplir el propósito para el cual se ha generado.

Estas reglas son aplicables a las **10** páginas de este documento.

Índice de contenido

Lista de reserva del producto.....	1
Leyenda:.....	1
Funcionalidades de los Módulos o Vistas.....	2
[M1] – Módulo Vista Principal (/ToNgueLP/modules/Linguist).....	2
[M2] – Módulo Vista de Comparación (/ToNgueLP/modules/Matching).....	2
[M6] – Módulo Vista de Diccionarios (/ToNgueLP/modules/Dicts).....	3
[M5] – Componente Editor de Texto (/ToNgueLP/modules/M5).....	3
Funcionalidades de la Capa de Control o Negocio.....	5
Funcionalidades de la Capa de Datos.....	6
[M4] – Módulo Corpus (/ToNgueLP/data).....	6
Requisitos de Configuración.....	7
[M3] – Módulo de Documentación (/ToNgueLP/html/index.html).....	7
Requisitos No Funcionales.....	8
Módulos Futuros.....	9
[M7] – Módulo Edit XML Structure (/ToNgueLP/modules/M7).....	9

Leyenda:

#.naranja4	Funcionalidades que pertenecen al requisito #.
Vista ...	Vista del módulo X donde está ubicada, pero corresponde al módulo de las vistas HTML.

Funcionalidades de los Módulos o Vistas

[M1] – Módulo Vista Principal (/ToNgueLP/modules/Linguist)

Asignado a	Ítem	Descripción	Estimación	Estimado por
Prioridad Muy Alta				
	1.5	Diseño simple y muy usable de la vista principal para personas con pocos conocimientos de computación.	1	Abel
	1.1	Mostrar datos generales del corpus: o sea leer los metadatos del ToNgueLPYYY-plag-cases-corpus.xml. Opción de Menú “About Corpus” y UI pequeña independiente.	0.2	Abel
	1.2	Botón Verificar y generar las etiquetas de metadatos del ToNgueLP-plag-cases-corpus.xml en función de los la información que contenga el corpus. Ej: # de docs que contiene, # de casos, etc.		
Prioridad Alta				
Prioridad Media				
	1.3	Insertar un caso de plagio desde 0. Significa inscribir el texto src y el texto susp.		
	1.3.1	Sugerir texto donde insertar.		
	1.3.2	Mostrar pos inicio y final donde se va a insertar.		
	1.3.3	Al confirmar la adición regenerar el metadata-corpus.xml		
Prioridad Baja				

Otras funcionalidades:

- Abrir varios corpus y ver en pestañas laterales o en una columna(estilo tabs derecho de Konqueror).
- Ver lista de casos del corpus en el foco.
- Abrir diferentes casos para estudiar o editar en pestañas horizontales (estilo tabs Firefox).
- Pestaña de un caso:
 - Visualizar fragmento de texto de src-doc en el 1er componente de texto.
 - Visualizar datos del fragmento src debajo de este.
 - Visualizar fragmento de texto de susp-doc en el 2do componente de texto.
 - Visualizar datos del fragmento susp debajo de este.
 - Operaciones a realizar:
 - Validar por humano (debe estar logueado, y debe ser un usuario autorizado) Esto debería conectarse por webservice para tener una BD online que pueda recibir la retroalimentación.
 - Modificar el límite de los casos.
 - Editar los textos.
 - Clasificar el fragmento por su clase y tipo de plagio, proyección.
 - Adicionar un nuevo caso.
 - El sistema debe sugerir, en una ventana emergente, los 3 documentos donde podría hacerse la inserción. Al escoger escribir los datos en el XML del corpus, en la sección de este caso, y actualizar la barra de datos del fragmento con el nombre del documento.
 - Calcular datos del caso y visualizar en la barra de datos del fragmento sospechoso o fuente según sea el caso. Escribir estos datos en el XML.
 - Mostrar las etiquetas (este botón debe clickearse cada vez para ver una anotación diferente

ejemplo para cambiar entre SRL y POS).

- Relativo al menú corpus:
 - Adicionar un nuevo corpus
 - Adicionar un nuevo caso a un corpus abierto.
 - Ver los datos del corpus. (incluso actualizado si se ha insertado un caso recientemente)
- Relativo a las vistas (Esto es como el menú de “Ventanas” en LibreOffice):
 - Abrir o Poner en el Foco la Vista Principal.
 - Abrir o Poner en el Foco la Vista de Comparación.
 - Abrir o Poner en el Foco la Vista de Diccionarios.
- Relativo a Help
 - About
 - Ayuda para Lingüistas.
 - Ayuda para Investigadores NLP.
 - Ayuda para Desarrolladores.
- Relativo al menú “Herramientas” (similar al “Herramientas” de LibreOffice):
 - Normalizar TXT
 - Convertir documentos TXT o PDF a ZZZdoc.XML
 - Generar XML estandar a partir de la carpeta OUT de la salida de un algoritmo de detección.
 - Comparación del XML estandar con XML del corpus cargado.
 - Actualizar ciertos tags del XML de datos del algoritmo y actualizar ciertos tags del XML reporte de casos detectados.
- Relativo a Archivo
 - Cargar TXT, que se desea analizar para incorporar a la colección.
 - Introducir TXT al Corpus, debe seleccionar la carpeta: src o susp. (debe estar autenticado)
 -

[M2] – Módulo Vista de Comparación (/ToNgueLP/modules/Matching)

Módulo de detección de plagio

Asignado a	Ítem	Descripción	Estimación	Estimado por
		Prioridad	Muy Alta	
	2.1	Mostrar en el lateral derecho casos del corpus presentes en ToNgueLP-plag-cases-corpus.xml.	1	Abel
	2.2	Mostrar en el lateral izquierdo casos detectados por el algoritmo que se está analizando, y que aparecen en el suspXXX-plag-report.xml	1	Abel
		Prioridad	Alta	
	2.3	Algoritmo para machear casos detectados versus casos en el reporte.	1	Abel
	2.4		3	Abel
		Prioridad	Media	
		Prioridad	Baja	

- ToNgueLP debe analizar, en la vista step-by-step una sola carpeta de ficheros “out” (salida que dan los algoritmos) pues los nombres de los reportes podrían coincidir entre varias carpetas “out”.

[M6] – Módulo Vista de Diccionarios (/ToNgueLP/modules/Dicts)

Asignado a	Ítem	Descripción	Estimación	Estimado por
		Prioridad	Muy Alta	
		Conectar con los diccionarios de FreeLing.	1	
	6.1	Ver el diccionario de Español de FreeLing al estilo Artha.(también llamado por Abel estilo Wordnet) Nota: Usar Artha como prototipo, y su código sí es posible.	8	Abel
		Prioridad	Alta	
		Diseñar un espacio donde se vean las palabras que hay en el corpus que no están en el diccionario.	0.2	
		Prioridad	Media	
		Mostrar las operaciones que se pueden hacer con palabras detectadas que no están en el diccionario. Ej: Adicionar, No Incluir, Conjugación del Verbo X, etc.	1	
		Asociar palabra o frase a un acrónimo. Ej: spanish y english → 'Spanglish'	0.2	
		Adicionar lexicón o glosses(wordnet) Ej: a la palabra 'quijote' → la oración 'Su quijotesca figura rompía con el paisaje abundante'	0.2	
		Asociar conjunto de una frase con una sigla. Ej: IGSW → Ingeniería de Software.	0.2	
		Editar la BD o usar el algoritmo de FreeLing para los nombres-proprios.	1	
		Sugerir palabras detectadas como nombres propios. Permitir adicionar nombre-propio detectado a la BD.	0.2	
		Sugerir contracciones detectadas con sus	0.2	

[Link al Índice \(+usabilidad\)](#)

Elaborado por el Ing Abel Meneses Abad, 30/07/13
Basado en las planillas originales de la UCI, Copyleft Creative Common
Share Alike Non-Comercial Use

		correspondientes miembros. Permitir adicionar contracciones detectadas a la BD.		
		Identificar collocations. Sugerir lista de collocations no contenidas en la BD de collocations. Permitir adicionar collocations o desestimar.	0.2	
		Prioridad	Baja	

[M5] – Componente Editor de Texto (/ToNgueLP/modules/M5)

Componente de las vistas Principal y side-by-side que permite a los lingüistas modificar las fronteras de los casos de plagio, permite señalar con colores los textos de manera que se pueda reconocer el tipo de plagio que se está visualizando así como sus límites, señalará POS , Entity Labeling, Semantic Role Labeling y Aceptación Labeling.

Asignado a	Ítem	Descripción	Estimación	Estimado por
		Prioridad	Muy Alta	
	5.1			
		Prioridad	Alta	
		Prioridad	Media	
		Prioridad	Baja	

Funcionalidades de la Capa de Control o Negocio

Asignado a	Ítem	Descripción	Estimación	Estimado por
		Prioridad Muy Alta		
	8.1	Parser de lectura escritura para leer ToNgueLP-plag-cases-corpus.xml	1	Abel
	8.2	Flujo para Verificar y generar las etiquetas de metadatos del ToNgueLP-plag-cases-corpus.xml en función de los la información que contenga el corpus. Ej: # de docs que contiene, # de casos, etc.		
		Prioridad Alta		
	8.3	Reconocer automáticamente la estructura del plag-report que se está probando a partir de los dtd que soporte ToNgueLP en sus diferentes momentos.		
		Prioridad Media		
	8.4	Flujo para calcular Sugerencia de texto susp y src donde insertar. Esto podría ser un algoritmo de clasificación basado en bag of words del fragmento, o de cálculo temático del discurso para insertar en el lugar adecuado.		
	8.5	Flujo Al confirmar la adición de un caso regenerar el metadata-corpus.xml		
		Prioridad Baja		

Funcionalidades de la Capa de Datos

[M4] – Módulo Corpus (/ToNgueLP/data)

Funcionalidades que corresponden a la definición del corpus de ToNgueLP. Dígase por ejemplo la cantidad de textos que contendrá la versión, casos definidos de plagio, etc.

Asignado a	Ítem	Descripción	Estimación	Estimado por
		Prioridad <i>Muy Alta</i>		
	4.1	XML del corpus de ToNgueLP → ToNgueLP-plag-cases-corpus.xml	1	Abel
Abel	4.1.1	Agregar un caso básico de plagio, para probar las vistas de ToNgueLP, y el funcionamiento correcto de los parsers y los XML diseñados en la arquitectura de datos.	0.1	Abel
		Prioridad <i>Alta</i>		
	4.1.2	Agregar 10 casos de plagio literal.	0.2	Abel
	4.1.3	Agregar 10 casos de plagio del tipo paráfrasis "low ofuscation".	2	Abel
	4.2	XML de casos detectados por algoritmoXXX , o conocido como el "Reporte de Plagio" → algorithmXXX-plag-report.xml		
		Prioridad <i>Media</i>		
	4.3	XML del algoritmoXXX que contenga los datos del mismo, como las técnicas NLP, fechas de creado, autores, etc → algorithmXXX-data-report.xml	1	Abel
		Prioridad <i>Baja</i>		
	4.4	XML de los documentos → ZZZdoc.xml	2	Abel

Definición: Los ejemplos contenidos en el corpus(4.1) deberían tener las siguientes características:

- 1/3 del tamaño de una página.
- 1/3 del tamaño de un artículo (9 páginas promedio)
- 1/3 del tamaño de una tesis (80 páginas promedio)
- 0% del tamaño de un libro (200 – 600 páginas)
- 0% del tamaño de un super-libro (+ 600 páginas)
- 0% del tamaño de un párrafo.
- 0% del tamaño de una oración.
- 0 documentos vacíos

- Agregar a los XMLs de ToNgueLP los xmlns para identificarlos de forma única.

Requisitos de Configuración

- Configurar repositorio para versionar utilizando git. Configurando el .gitignore para tomar solo la parte del árbol necesaria de configurar. 2h.

[M3] – Módulo de Documentación (/ToNgueLP/html/index.html)

Módulo de documentación. Por lo general los proyectos no trabajan en esta parte, ello hace que mueran muchos proyectos. ToNgueLP contiene la documentación dentro de cada carpeta de la arquitectura del proyecto. Y además contiene una carpeta /ToNgueLP/doc que contiene la mayoría de los ficheros que se refieren en la documentación contenida en la URL (~html/index.html). En esta sección se documentan las necesidades básicas que consideran los desarrolladores y no los clientes (porque ellos no han solicitado nada al respecto, pero como visión de los desarrolladores para que el proyecto sea escalable, etc, etc.)

Asignado a	Ítem	Descripción	Estimación	Estimado por
		Prioridad	Muy Alta	
	3.1	Configurar carpeta para trabajo con Sphinx	0.4	Abel
	3.2	Diseñar árbol de carpetas del proyecto para la documentación.	0.2	Abel
	3.3	Diseñar index.rst de ToNgueLP. Árbol inicial de la documentación.	0.4	Abel
	3.4	Documentar el Overview de ToNgueLP	02	Abel
		Prioridad	Alta	
	3.5	Documentar carpeta "modules" de ToNgueLP	0.2	Abel
		Prioridad	Media	
	3.6	Documentar carpeta "modules" de ToNgueLP	0.2	Abel
	3.7	Documentar carpeta "lib" de ToNgueLP	0.2	Abel
		Prioridad	Baja	
	3.8	Documentar carpeta "lib" de ToNgueLP		
RNF (Requisitos No Funcionales de la documentación)				
Compatibilidad	NF1	Debe poder leerse en cualquier sistema operativo.		

- Documentar página del parser dentro del nivel de "Control Level" una subsección para los Parsers.
- Documentar un ejemplo para alguien que desea hacer un parser nuevo o reutilizar el XML.

Requisitos No Funcionales

-
-

RNF (Requisitos No Funcionales)				
Usabilidad	21	El sistema podrá ser usado por cualquier persona con conocimientos básicos de informática y navegación en internet.		
Software	22	Las estaciones de trabajo deberán contar con soporte para python y Qt		
Hardware	23	Para la aplicación servidora es necesaria una PC con microprocesador Pentium 4 3.0GHz, 512MB de RAM y una capacidad de 10GB en disco duro.		
Soporte	24	Debe poder ser mantenido por el equipo creador.		

Módulos Futuros.

[M7] – Módulo Edit XML Structure (/ToNgueLP/modules/M7)