FACULTAD DE CIENCIAS E INGENIERÍA

ANEXOS

Tesis para optar el Título de **Ingeniero Informático**, que presenta el bachiller:

Melissa Zevallos Franco

ASESOR: Fernando Alva Manchego

Lima, abril del 2015

ÍNDICE GENERAL

LISTADO DE TABLAS

En este anexo se muestran los resultados, sin ninguna modificación, después de evaluar el Ejemplo 4-A con cada una de las herramientas. En el encabezado de cada tabla se puede apreciar el nombre de la herramienta, la página en la se encuentra el demo online y la fecha correspondiente a la consulta. A continuación se muestra la relación de tablas correspondiente a cada herramienta:

a) NLTK →Tabla 9-B
b) NLP STANFORD → Tabla 9-C
c) FREELING → Tabla 9-A

Tabla A: Resultado - Evaluación de FreeLing

FREELING				
http://nlp.lsi.upc.edu/freeling/demo/demo.php				
1 de Junio 2014				
Pero	**Se**	**Hizo**	**más**	**famoso**
CC	P00CN000	VMIS3S0	RG	AQ0MS0
todavía	**un**	**poco**	**después**	**,**
RG	DI0MS0	PI0MS000	RG	Fc
apostando	**una**	**carrera**	**a**	**el**
VMG0000	DI0FS0	NCFS000	SPS00	DA0MS0
amanecer	**,**	**Desde**	**la**	**Plaza_San_Martín**
NCMS000	Fc	SPS00	DA0FS0	NP00G00
hasta	**el**	**Parque_Salazar**	**,**	**con**
SPS00	DA0MS0	NP00G00	Fc	SPS00
Quique_Ganoza	**,**	**éste**	**por**	**la**
NP00SP0	Fc	PD0MS000	SPS00	DA0FS0
buena	**pista**	**,**	**Pichulita**	**contra**
AQ0FS0	NCFS000	Fc	NP00SP0	SPS00
el	**tráfico**	**.**	**Los**	**patrulleros**
DA0MS0	NCMS000	Fp	DA0MP0	NCMP000
lo	**Persiguieron**	**desde**	**Javier_Prado**	**,**
PP3CNA00	VMIS3P0	SPS00	NP00SP0	Fc
sólo	**lo**	**alcanzaron**	**en**	**Dos_de_Mayo**
RG	PP3CNA00	VMIS3P0	SPS00	NP00G00
,	**cómo**	**correría**	**.**	
Fc	PT000000	VMIC3S0	Fp	

Tabla B: Resultado - Evaluación de NLTK

NLTK http://text-processing.com/demo/tag/ 15 de Mayo 2014
Phrases and Named Entities **PER:** m\xc3/-None- ¡/-None- s/-None- famoso/AQ todavĂ/-None- /-None- **PER:** despuĂ/-None- ©/-None- s/-None- **LOC:** amanecer/VMN **ORG:** Plaza/NC San/AQ MartĂ/-None- /-None- n/NC **ORG:** Parque/NC Salazar/AQ **PER:** Quique/NC Ganoza/NC **PER:** Ã/-None- ©/-None- ste/DD **MISC:** trĂ/-None- ¡/-None- fico/AQ **PER:** Javier/NC Prado/NC **LOC:** Mayo/NC **PER:** cómo/-None- correrĂ/-None- /-None-

Tabla C: Resultado - Evaluación de NLP STANDFORD

NLP STANFORD http://nlp.stanford.edu:8080/corenlp/process 20 de Mayo 2014
1 Pero se hizo más famoso todavía un poco después, apostando una carrera al amanecer, desde la Plaza San Martín (Location) hasta el Parque Salazar (Person), con Quique Ganoza (Person), éste por la buena pista, Pichulita (Person) contra el tráfico. 2 Los patrulleros lo persiguieron desde Javier Prado (Person), sólo lo alcanzaron en Dos de Mayo (Person), cómo correría.

En este anexo se presenta en la tabla 9-D la información sobre los textos literarios recolectados. Se puede apreciar la columna Autor, Titulo, Editorial, Nro. Página inicial y final (representa el intervalo escogido para cada obra), nombre del archivo correspondiente y finalmente la cantidad de palabras que contiene cada archivo para resaltar que todos son de tamaños muy semejantes.

Tabla D: Información sobre los textos literarios recolectados

Autor	Título Obra	Editorial Año	Nro. Página		Nombre Archivo	Cantidad Palabras
			Inicio	Fin		
Fernando Ampuero	¨Miraflores Melody¨		30	32	input1	904
			34	36	input2	989
			62	64	input3	840
Alfredo Bryce Echenique	¨No me esperen en Abril¨	El Comercio 2006	29	33	input4	1140
			43	45	input5	1422
			53	55	input6	1424
Julio Ramón Riveyro	¨Los Geniecillos Dominicales¨	El Comercio 2007	3	6	input7	1506
			26	28	input8	1172
			29	33	input9	1311
			36	37	input10	1257
			39	41	input11	1048
Mario Vargas LLosa	¨La ciudad y los perros¨	Peisa 2006	81	83	input12	991
			94	97	input13	1035
			105	107	input14	772
			245	248	input15	1294
			252	255	input16	1019
Mario Vargas LLosa	¨Los Jefes y Los cachorros¨	Peisa 1997	72	74	input17	762
			118	120	input18	968
			127	129	input19	1722
			130	133	input20	1217
Mario Vargas LLosa	¨Travesuras de la niña mala¨	Alfaguara 2011	10	13	input21	1116
			14	18	input22	1271
			20	23	input23	1147

ANEXO 5 B: Información por cada lugar obtenido en un archivo recolectado

En este anexo se presenta en la Tabla 9-E las características por cada lugar mencionado para el archivo input10.txt. Se puede apreciar que en total hay 8 lugares identificados los cuales 7 son reales y 3 son completos.

Tabla E: Información por cada lugar obtenido en un archivo recolectado.

No.	Nombre Lugar	Atributo	Completa	Real
1	Agua Dulce			✔
2	Club Regatas			✔
3	Margarita	pasaje	✔	
4	La Parada			✔
5	Paracas			✔
6	La Parada			✔
7	Pardo	avenida	✔	✔
8	Dos de Mayo	calle	✔	✔

En este anexo se presenta en la Tabla 9-F los resultados de cada archivo de texto después de evaluarlos con el módulo de NER nativo de la herramienta FreeLing. Al final de la tabla se presentan las cantidades totales y se puede comprobar que se han identificado 95% de las entidades que se deseaban recuperar (266/279).

Tabla F: Resultados - Evaluación módulo de NER nativo

Nombre Archivo	Entidades	NER Nativo
input1	8	7
input2	11	11
input3	14	13
input4	11	10
input5	11	4
input6	17	17
input7	3	3
input8	7	7
input9	4	4
input10	8	7
input11	9	9
input12	7	6
input13	4	4
input14	12	12
input15	36	35
input16	7	7
input17	5	5
input18	8	8
input19	17	17
input20	12	12
input21	23	23
input22	37	38
input23	8	7
	279	266

En este anexo se presenta en la Tabla 9-G los resultados de cada archivo de texto después de evaluarlos con el módulo de NEC nativo de la herramienta FreeLing. Al final de la tabla se presentan las cantidades totales y se puede comprobar, utilizando la fórmula en 9.9, que el nivel de calidad es de 53% (2*(0.61+0.47)/(0.61*0.47)).

Tabla G: Resultados - Evaluación módulo de NEC nativo

Nombre Archivo	Precisión		Recall	
	P1	P2	R1	R2
input1	4	7	4	8
input2	7	8	7	11
input3	7	9	7	14
input4	5	15	5	11
input5	1	4	1	11
input6	7	17	7	17
input7	2	4	2	3
input8	3	3	3	7
input9	2	5	2	4
input10	5	10	5	8
input11	5	19	5	9
input12	1	2	1	7
input13	0	1	0	4
input14	6	9	6	12
input15	14	16	14	36
input16	3	7	3	7
input17	3	6	3	5
input18	2	3	2	8
input19	8	12	8	17
input20	4	7	4	12
input21	15	16	15	23
input22	23	28	23	37
input23	5	6	5	8
	132	214	132	279
	Precisión total: 0,617		Recall total: 0,473	

ANEXO 5 G: Detalle de resultados: Evaluación los corpus de prueba

En este anexo se presentan los resultados de cada archivo de texto después de evaluarlos con el módulo de NEC especializado con cada uno de los corpus. Al final de cada tabla se presentan los valores totales de precisión y recall. Con estos datos se puede comprobar, utilizando la fórmula en 9.9, el F-Measure de cada uno. Cabe resaltar que los archivos utilizados para generar el corpus el entrenamiento (4, 8, 10, 14, 15) y prueba (6, 18 y 21) no han sido considerados en ninguna de las 3 evaluaciones anteriores. A continuación se muestra la relación de tablas correspondiente a cada corpus:

a) Resultados del CORPUS A→Tabla 7-I
b) Resultados del CORPUS B → Tabla 7-J
c) Resultados del CORPUS C → Tabla 7-K

Tabla H: Resultados – Evaluación del CORPUS A

Nombre Archivo	Precisión		Recall	
	P1	P2	R1	R2
input1	8	10	8	8
input2	8	12	8	11
input3	10	11	10	14
input5	2	3	2	11
input7	2	6	2	3
input9	4	11	4	4
input11	7	13	7	9
input12	6	8	6	7
input13	4	5	4	4
input16	6	8	6	7
input17	5	7	5	5
input19	16	26	16	17
input20	11	17	11	12
input22	26	31	26	37
input23	6	10	6	8
	121	178	121	157
	Precisión total: 0,680		Recall total: 0,771	

Tabla I: Resultados – Evaluación del CORPUS B

Nombre Archivo	Precisión		Recall	
	P1	P2	R1	R2
input1	4	9	4	8
input2	7	9	7	11
input3	6	7	6	14
input5	2	3	2	11
input7	2	5	2	3
input9	2	4	2	4
input11	5	7	5	9
input12	4	6	4	7
input13	4	6	4	4
input16	6	7	6	7
input17	5	6	5	5
input19	12	15	12	17
input20	7	9	7	12
input22	29	31	29	37
input23	6	10	6	8
	101	134	101	157
	Precisión total: 0,754		Recall Total: 0,643	

Tabla J: Resultados – Evaluación del CORPUS C

Nombre Archivo	Precisión		Recall	
	P1	P2	R1	R2
input1	7	10	7	8
input2	5	12	5	11
input3	14	19	14	14
input5	4	7	4	11
input7	3	16	3	3
input9	4	18	4	4
input11	6	32	6	9
input12	6	9	6	7
input13	4	7	4	4
input16	6	18	6	7
input17	5	7	5	5
input19	18	34	18	17
input20	10	27	10	12
input22	35	50	35	37
input23	6	31	6	8
	133	297	133	157
	Precisión total: 0,448		Recall total: 0,847	

En este anexo se presentan los resultados después de cada actividad y evaluación para el archivo “input 11.txt”. A continuación se muestra la relación de tablas correspondiente:

a) Información por cada lugar obtenido →Tabla 9-K
b) Resultados de la evaluación de los módulos NER y NEC nativo → Tabla 9-L
c) Resultados de la evaluación de cada corpus prueba → Tabla 9-M
d) Detalle de la evaluación del módulo NEC nativo → Tabla 9-N
e) Detalle de la evaluación de cada corpus prueba→ Tabla 9-O

Tabla K: Información por cada lugar obtenido

No.	Nombre Lugar	Atributo	Completa	Real
1	Diagonal			✔
2	Diagonal			✔
3	D'Onofrio			✔
4	Porta	Calle	✔	✔
5	Terrazas			✔
6	Terrazas			✔
7	Terrazas			✔
8	Las Delicias			

Tabla L: Resultados de la evaluación de los módulos NER y NEC nativo

Resultados - NER Nativo		Resultados - NEC Nativo	
Entidad	Correcta	Entidad	Correcta
D'_Onofrio	✔	Las Delicias	✔
Diagonal	✔	Porta	✔
Diagonal	✔	San Nicolás	
Las_Delicias	✔		
Porta	✔		
Terrazas	✔		
Terrazas	✔		
Terrazas	✔		

Tabla M: Resultados de la evaluación de cada corpus prueba

Resultados - CORPUS A		Resultados - CORPUS B		Resultados - CORPUS C	
Entidad	Correcta	Entidad	Correcta	Entidad	Correcta
Diagonal	✔	Porta	✔	Diagonal	✔
Las_Delicias	✔	San_Nicolás		Las_Delicias	✔
Diagonal	✔	Terrazas	✔	D'_Onofrio	✔
Violín_Gitano		Chino		Diagonal	✔
Porta	✔	Terrazas	✔	Violín_Gitano	
San_Nicolás		Terrazas	✔	Porta	✔
Terrazas	✔			San_Nicolás	
Chino				Terrazas	✔
Terrazas	✔			Chino	
Terrazas	✔			Terrazas	✔
Lalo				Terrazas	✔
				Chispas	
				Lucio	
				A	
				A	
				Lalo	

Tabla N: Detalle de la evaluación del módulo NEC nativo

		Indicador de calidad – NEC				
Entidades		Precisión		Recall		F-measure
Totales	NER	P1	P2	R1	R2	
8	8	2	3	2	8	0,364
100%		Presión: 0,667		Recall: 0,250		

Tabla O: Detalle de la evaluación de cada corpus prueba

Indicador de calidad - CORPUS A					Indicador de calidad - CORPUS B					Indicador de calidad - CORPUS C				
Precisión		Recall		F-measure	Precisión		Recall		F-measure	Precisión		Recall		F-measure
P1	P2	R1	R2		P1	P2	R1	R2		P1	P2	R1	R2	
7	11	7	8	0,737	4	7	4	8	0,533	8	16	8	8	0,667
Precisión: 0,636		Recall:0,875			Precisión: 0,571		Recall:0,5			Precisión: 0,5		Recall: 1		

ANEXO 6 A: Detalle de la cantidad de entidades para los archivos recolectados

En este anexo se presenta el detalle de la cantidad de entidades para los archivos recolectados según su tipo de entidad.

Tabla P: Detalle de la cantidad de entidades para los archivos recolectados

Nombre Archivo	Entidades		
	Lugares	Completas	Reales
input1	8	2	4
input2	11	1	4
input3	14	2	12
input4	11	2	10
input5	11	0	6
input6	17	3	12
input7	3	1	3
input8	7	0	5
input9	4	1	4
input10	8	3	7
input11	9	2	7
input12	7	3	6
input13	4	0	3
input14	12	3	10
input15	36	13	33
input16	7	1	7
input17	5	4	5
input18	8	1	7
input19	17	5	14
input20	12	2	10
input21	23	2	19
input22	37	6	26
input23	8	0	7
	279	57	221