Seminario del LabTL:

Comparativa de los métodos participantes en la primera competencia internacional de detección automático de plagio (en el marco del PAN´09)

Fernando Sánchez Vega

El plan de la plática

- -Cómo es la competencia
  - *Corpus*
  - Medidas de evaluación
- -Enfoques generales
- -Los primeros lugares
- -El resto
- 20-25 min.

El contexto

- No existe una marco claro de referencia
- Algunos intentos :
  - Corpus METER
  - Corpus WEBER (31 casos, 19 sistemas) irrepetibilidad

Los enfoques de la tarea

sospechoso

conjunto
de
fuentes

- Detección extrínseca: encontrar los pasajes plagiados de un conjunto de fuentes .

- Detección intrínseca: encontrar los pasajes plagiados de un documento particular.

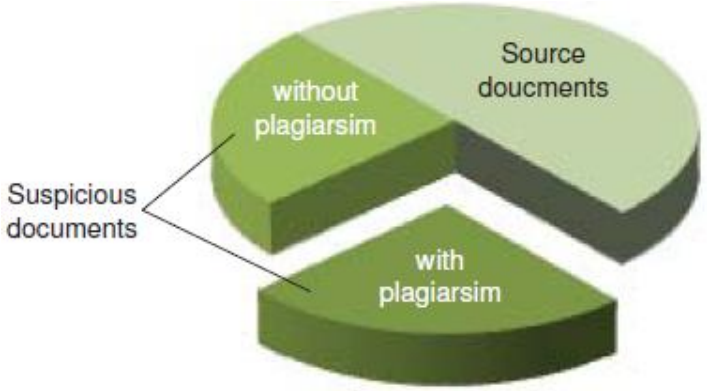
Detección intrínseca (las únicas palabras dedicadas)

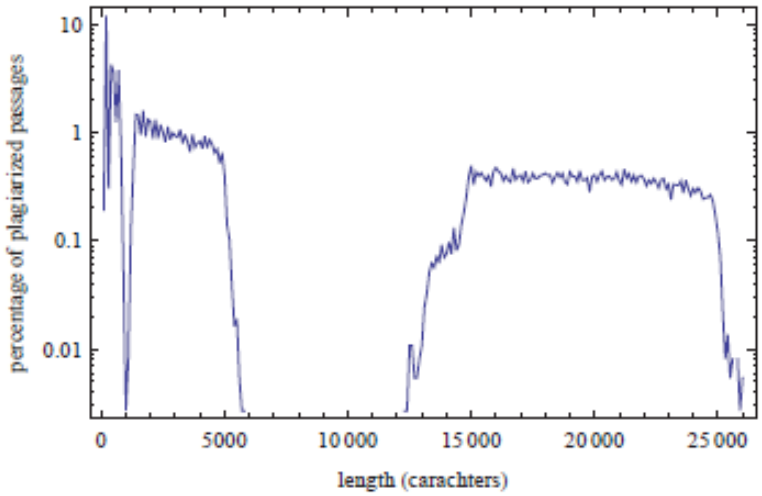
- Procedimientos de atribución de autoría.
- Resultados muy malos

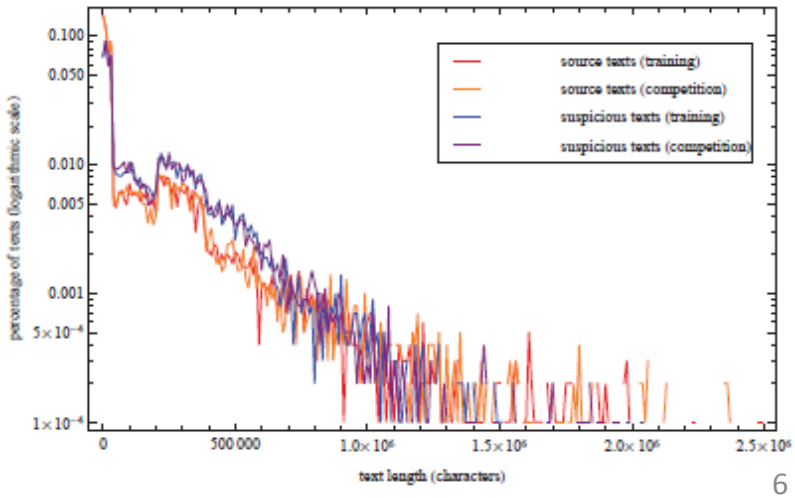
Intrinsic Detection Quality							
Rank	Overall	F	Precision	Recall	Granularity	Participant	
1	0.2462	0.3086	0.2321	0.4607	1.3839	Stamatatos (2009)	
2	0.1955	0.1956	0.1091	0.9437	1.0007	Hagbi and Koppel (2009)	(Baseline)
3	0.1766	0.2286	0.1968	0.2724	1.4524	Muhr et al. (2009)	
4	0.1219	0.1750	0.1036	0.5630	1.7049	Seaward and Matwin (2009)	

Itinerario de la competencia

- Corpus de entrenamiento :
  - 7 214 sospechosos;
  - 7 214 fuentes;
  - 37 046 pasajes plagiados;
  - 50 % de sospechosos sin plagio;
  - 25% de fuentes no usadas;
- Corpus de entrenamiento.

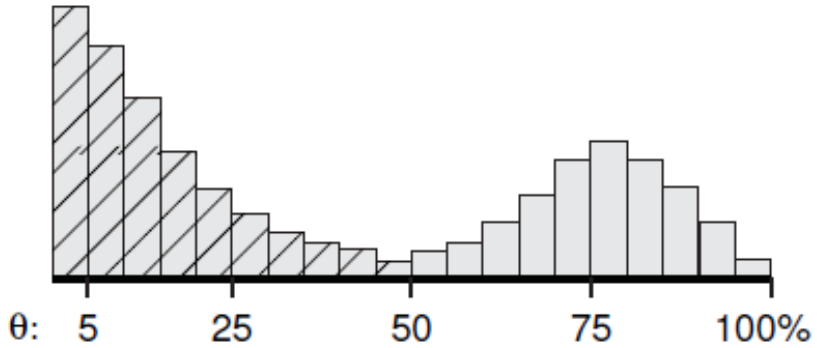






Corpus del PAN´09:

- Composición:
  - 41 223 documentos;
  - 94 202 casos de plagio artificiales;
  - Basados en 22 874 documentos (libros) del Project Gutenberg;
- Longitud:
  - 50% 1-10 páginas;
  - 35% 10-100;
  - 15% 100-1000;
- Plagio tras lingue: 10%;
- Longitud de plagio: 50-5000;
- Porcentaje del plagio del sospechoso (autenticidad):
  - 0-100%



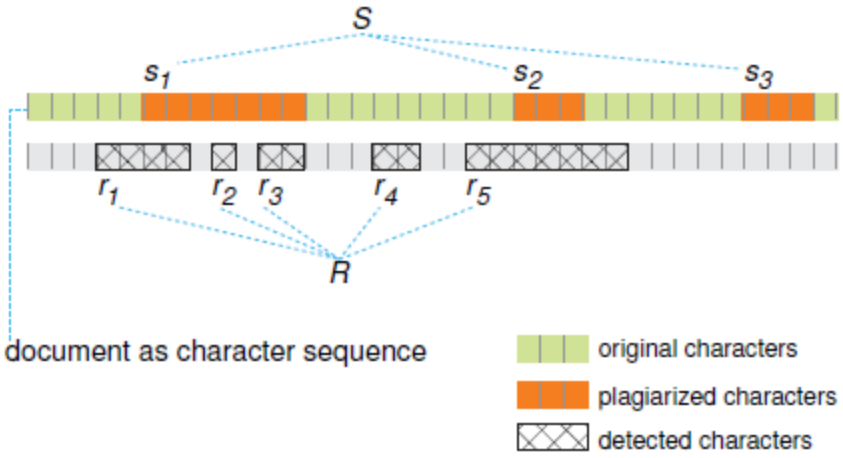
Corpus del PAN´09:

- Ofuscación o reescritura:
  - Nula a alta;
  - Se prefiere la ofuscación simple;
  - Operadores:
    - Textuales aleatorios: inserción, remoción o intercambio.
    - Semánticos: sinónimos, hiperónimos, hipónimos…
    - Barajeo manteniendo POS.

Criticas al corpus

- La rescritura no es humanamente-legible;
  - Existen oraciones a normales;
  - El problema se aleja del PLN;
- La construcción no es repetible (web);

Medidas de calidad.



- Recuerdo

$$rec_{PDA}(S,R) = \frac{1}{|S|}\sum_{s\in S}\frac{|s \sqcap \bigcup_{r\in R} r|}{|s|},$$

- Granularidad

$$gran_{PDA}(S,R) = \frac{1}{|S_R|}\sum_{s\in S_R}|C_s|,$$

$$S_R = \{s \mid s \in S \wedge \exists r \in R : s \cap r \neq \emptyset\}$$
$$C_s = \{r \mid r \in R \wedge s \cap r \neq \emptyset\}$$

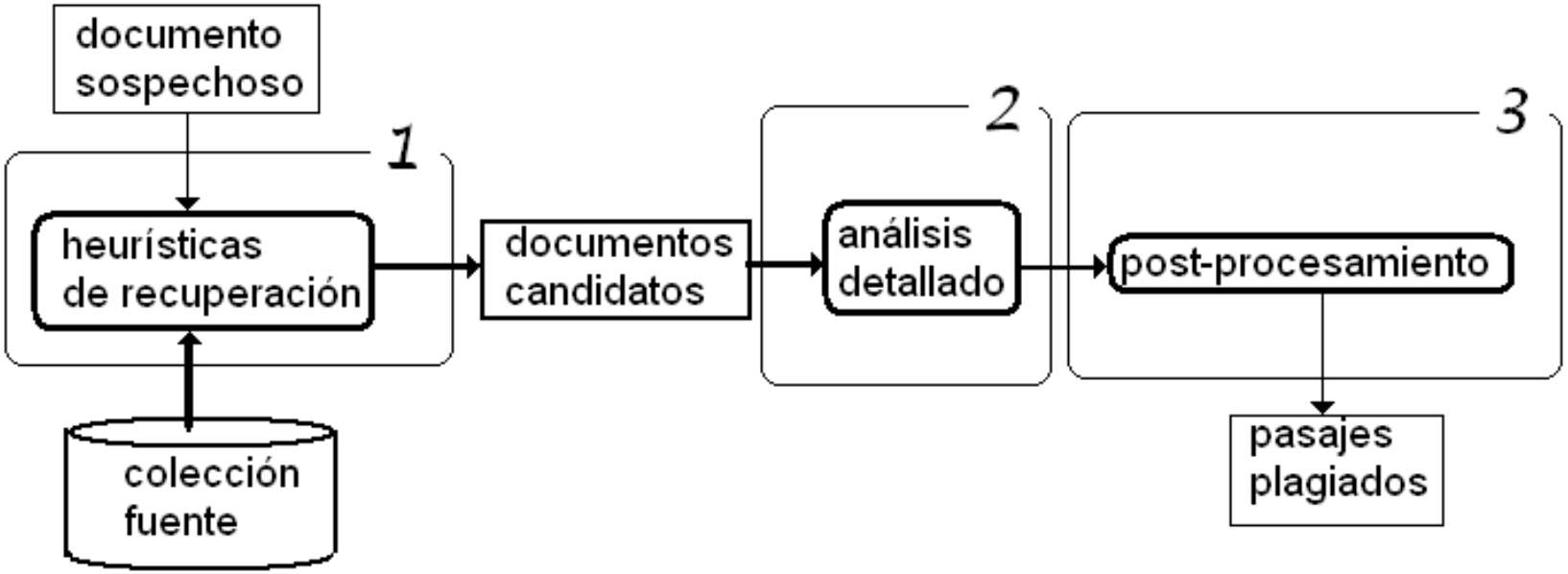
- Precisión

$$prec_{PDA}(S,R) = \frac{1}{|R|}\sum_{r\in R}\frac{|r \sqcap \bigcup_{s\in S} s|}{|r|},$$

- General

$$overall_{PDA}(S,R) = \frac{F}{\log_2(1+gran_{PDA})},$$

Esquema genérico (3 fases)



1. Similitud temática a nivel documento
2. Verificación textual a nivel pasajes
3. Refinamiento (agrupamientos)

Fase 1 (doc. probables)	Fase 2 (comparación exhaustiva)	Fase 3 (refinamiento)	inconveniente	Unidad - análisis	T. ejec.
Dist. coseno tf 16-gram (partiendo el conjunto) Ventaneo 2-3 palabras Podando a 51 (--10%) .	Comparación de los 16-gramas: listas ordenadas comparando en orden para ahorrar tiempo	Une los n-gramas separados por 3 o menos palabras	Los plagios deben de estar en orden para ser unidos	16-gram Carácter	12 h
5-*chunk* (traslape) @ hash 32 bits Umbral de +20 hash comunes	Se aglutinan los *chuck* en intervalos válidos: Mínima coincidencia, máxima separación	Se eliminan pasajes traslapados	Los pasajes válidos mantiene la cohesión en la fuente y el sospechoso - El incremento introduce ruido	*5-Chunks* de Palabras	1h 14 min
De A.A: Reducción por tamaño 1-9, 8-gram Distancia tf´ 10 primeros	Compresión T9 Cadenas largas por umbral ni sub preexistentes. Se mantiene en *suffix tree*	Mapeo gráfico “líneas y cuadros” Se unen los cuadros al ir expandiendo.	Confeccionado a medida no genérico	8-gram Palabra-Carácter	2.3 días

Fase 1 (doc. probables)	Fase 2 (comparación exhaustiva)	Fase 3 (refinamiento)	inconveniente	Unidad - análisis	T. ejec.
Divide en pasajes (500). Aplica BOW tf. Selecciona con distancia los centroides (2000) cercanos.	Compara los clúster con los vectores de los pasajes.	Une los pasajes adyacentes que sean identificados como plagio.	No captura nada acerca del orden .	palabras	

Resultados

		External Detection Quality				
Rank	Overall	F	Precision	Recall	Granularity	Participant
1	0.6957	0.6976	0.7418	0.6585	1.0038	Grozea, Gehl, and Popescu (2009)
2	0.6093	0.6192	0.5573	0.6967	1.0228	Kasprzak, Brandejs, and Křipač (2009)
3	0.6041	0.6491	0.6727	0.6272	1.1060	Basile et al. (2009)
4	0.3045	0.5286	0.6689	0.4370	2.3317	Palkovskii, Belov, and Muzika (2009)
5	0.1885	0.4603	0.6051	0.3714	4.4354	Muhr et al. (2009)
6	0.1422	0.6190	0.7473	0.5284	19.4327	Scherbinin and Butakov (2009)
7	0.0649	0.1736	0.6552	0.1001	5.3966	Pereira, Moreira, and Galante (2009)
8	0.0264	0.0265	0.0136	0.4586	1.0068	Vallés Balaguer (2009)
9	0.0187	0.0553	0.0290	0.6048	6.7780	Malcolm and Lane (2009)
10	0.0117	0.0226	0.3684	0.0116	2.8256	Allen (2009)

- Software desarrollado anteriormente
- Utiliza búsquedas a una base de datos

Gracias

¿Preguntas?

¿Plagios?

Apéndices:

Solo para el caso de preguntas sobre detalles específicos.

Segundo lugar

- Función Hash MD5 o similar
- Umbral 24 palabras
- Algoritmo de intervalos válidos:
  A. La primera y última *chuck* del intervalos son comunes.
  B. Los intervalos debían de contener al menos 20 *chucks* que puedan tener traslape.
  C. Entre cada par de *chuck* comunes no puede haber más de 49 *chucks* no comunes.

- la entrada del algoritmo es todo *chuck* que pertenece a un par de chucks comunes en la fuente-sospechoso, uno por cada par.

- I. profundidad =0
- II. ordenar la lista de pares de chucks ID en D1
- III. dividir la lista en los chucks pertenecientes a los intervalos validos en D1
- IV. si la lista de entrada es cubierta por un único intervalo válido entonces profundidad se incrementa
- V. si la profundidad es igual a 2 se regresa todo el rango de chunck ID como el pasaje plagiado.
- VI. para cada intervalo válido:
- (a)generar una nueva lista con los pares chunck ID en D1 con los del intervalo válido actual.
- (b)profundidad =1
- (c)regresar al paso =2
- (lo que hace básicamente es dividir los pasajes para analizar cada uno de los pedazos para encontrar los pasajes válidos de las listas)

3 er lugar

- Distancia basada en frecuencia:
- tercer lugar, primera etapa:

$$d_n(x,y) := \frac{1}{|D_n(x)| + |D_n(y)|} \sum_{\omega \in D_n(x) \cup D_n(y)} \left(\frac{f_y(\omega) - f_x(\omega)}{f_y(\omega) + f_x(\omega)} \right)$$

•Umbral diferenciado:
  - •15 para documentos pequeños (menos de 500000 caracteres)
  - •25 para documentos grandes.

•Algoritmo de unión de cuadros:
  A. los coincidencias deben de ser subsecuentes en x (sospechoso)
  B. la distancia entre las proyecciones en x está en el rango [0,d_x . l_x) con l_x la mayor de las longitudes de las dos secuencias (se esperaría que hasta ese momento) y d_x es un factor para escalar.
  C. el mismo caso que b pero para las proyecciones en y.

- d_x y d_y iniciales = 3
- d_x y d_y de escalabilidad =0.5.