Creado el 26/05/2014
Última versión: 26/05/2014
Paolo Rosso, Universitat Politècnica de Valencia
Comunidad moocTLH
Esta obra está bajo una licencia de Creative Commons Reconocimiento-CompartirIgual 3.0 Unported.

Aplicaciones de las TLH/PLAGIO Y AUTORÍA

Detección automática del plagio

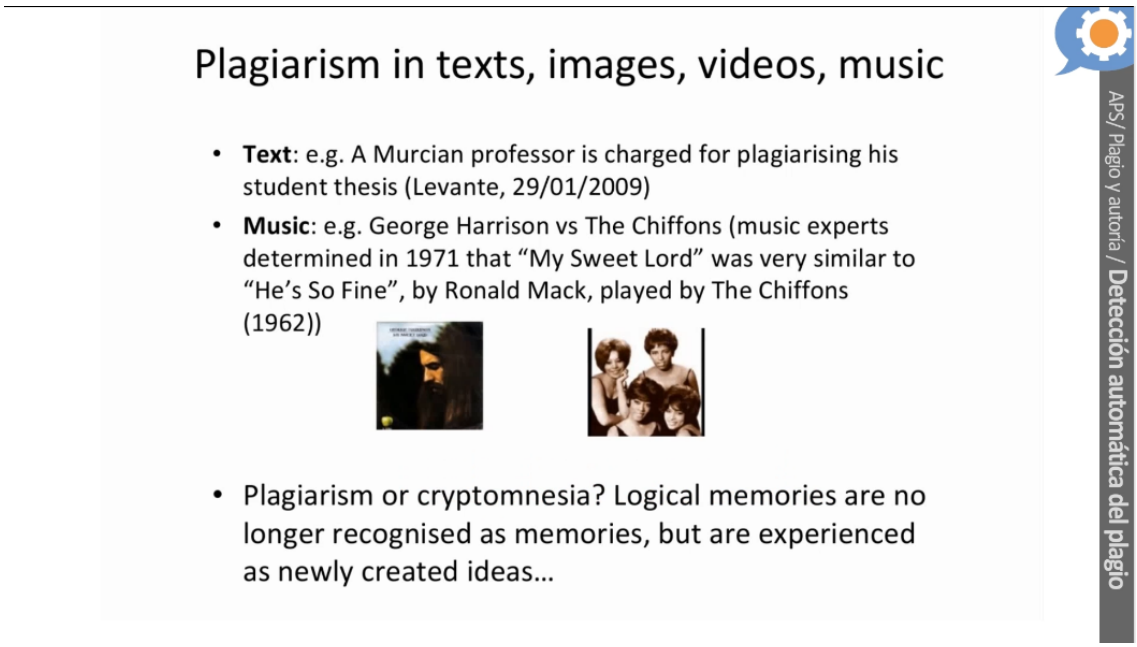
Hola. Soy Paolo Rosso de la Universitat Politècnica de Valencia y en estos 10 minutos introduciré el tema de la detección de plagio.

Primero qué entendemos por plagio. No sólo plagio cuando nos referimos a palabras sino también plagio de ideas, procesos, resultados, … y esto ocurre cuando en la fuente original y el autor no viene citado de la manera correcta.

Entonces hay que hacer una primera distinción entre el reutilización de texto y plagio. Un texto puede ser reutilizado de manera correcta siempre y cuando se cite la fuente y el autor. Si esto no ocurre es cuando entonces hablamos de plagio.

Plagio no sólo se refiere a plagio en texto y no sólo por parte de estudiantes, desgraciadamente hay casos de plagio por parte de profesores, investigadores, …

Y no sólo en texto sino que también en imágenes, vídeo y música.



A veces se habla de criptomnesia en lugar de plagio cuando uno tiene recuerdos de, por ejemplo, una melodía o una canción y asegura no haber cometido plagio.

Detección del plagio

Con respecto a los métodos de detección de plagio podemos distinguir los métodos externos y métodos intrínsecos.

Desde un punto de vista de proporcionar evidencia lingüística a un experto como un lingüista forense, lógicamente el lingüista computacional intenta proporcionarle el documento desde el cual ha sido cometido plagio. Es decir, en este caso hablamos de métodos de detección externa.

Hay técnicas que se basan en secuencias de caracteres, de palabras, …, o técnicas de *hashing* que intentan dado un documento, dado un fragmento de texto, proporcionar un valor de *hash.*

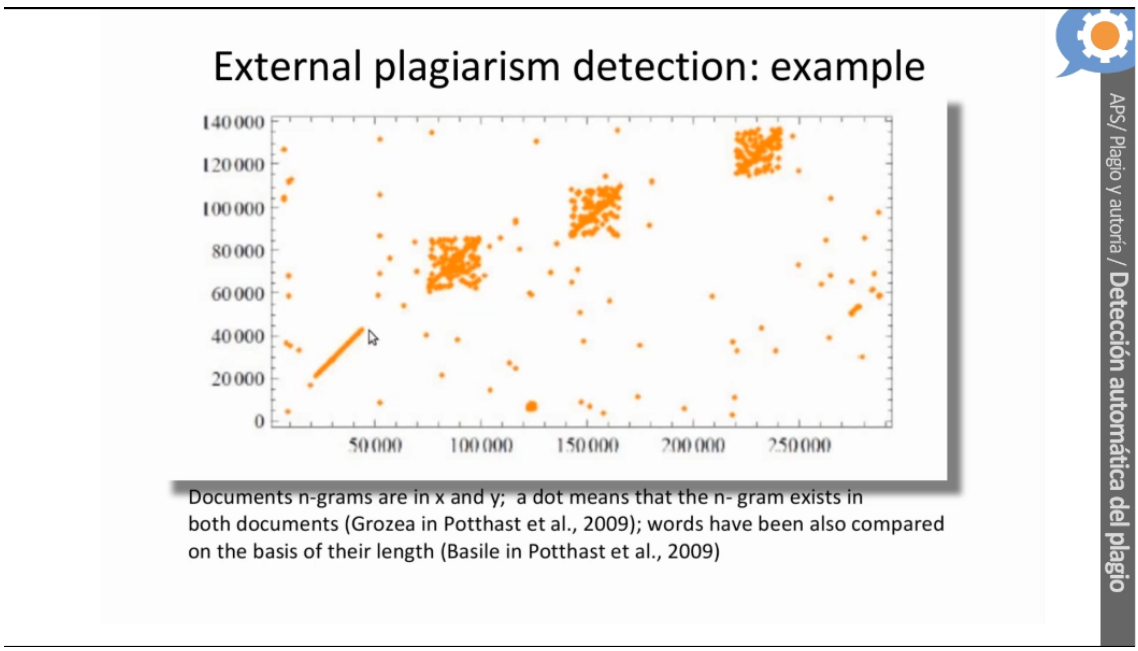
Cuando esto no es posible -no siempre el documento fuente está disponible y se puede recuperar- entonces se trata de detectar el plagio de manera intrínseca, basándose exclusivamente en el texto del documento que se está analizando.

En este caso hay que buscar irregularidades a nivel de cambio de estilo de escritura. Entonces él se utilizan técnicas como basarse en la riqueza del vocabulario, estadísticas básicas como la longitud de un promedio de frase, de palabras, o perfiles de secuencias de caracteres; o también el índice Gunning-Fog que determina la complejidad de un determinado fragmento de texto.

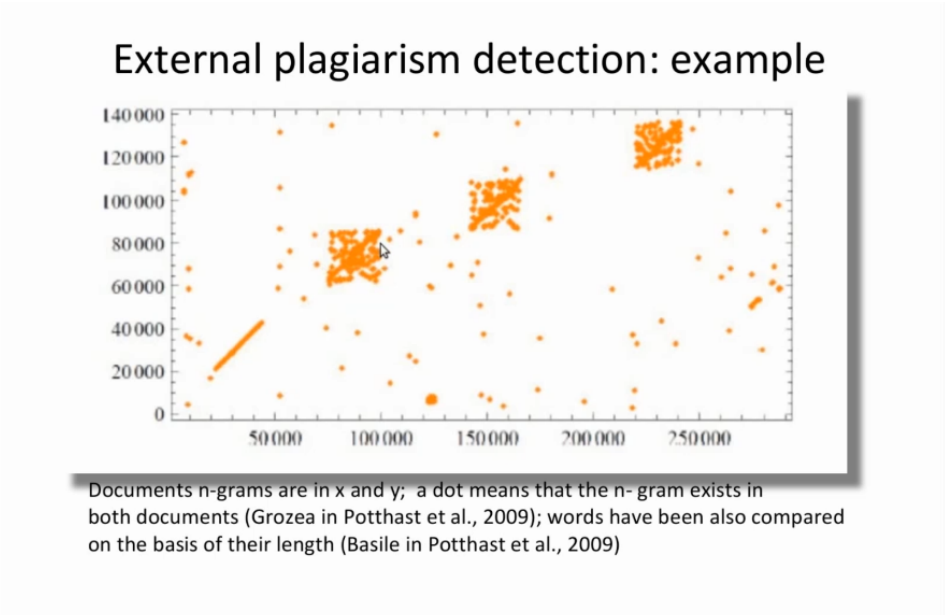
Comparación externa

En esta transparencia tenéis de manera gráfica la comparación de un documento con documentos sospechosos cuando se ha producido plagio de tipo “copiar pegar”: plagio *verbatim*.

Una copia *verbatim* se puede visualizar de esta manera , con la diagonal.



Cuando ha ocurrido un cambio de alguna de las palabra, entonces tenemos por una representación de este estilo.



Análisis intrínseco

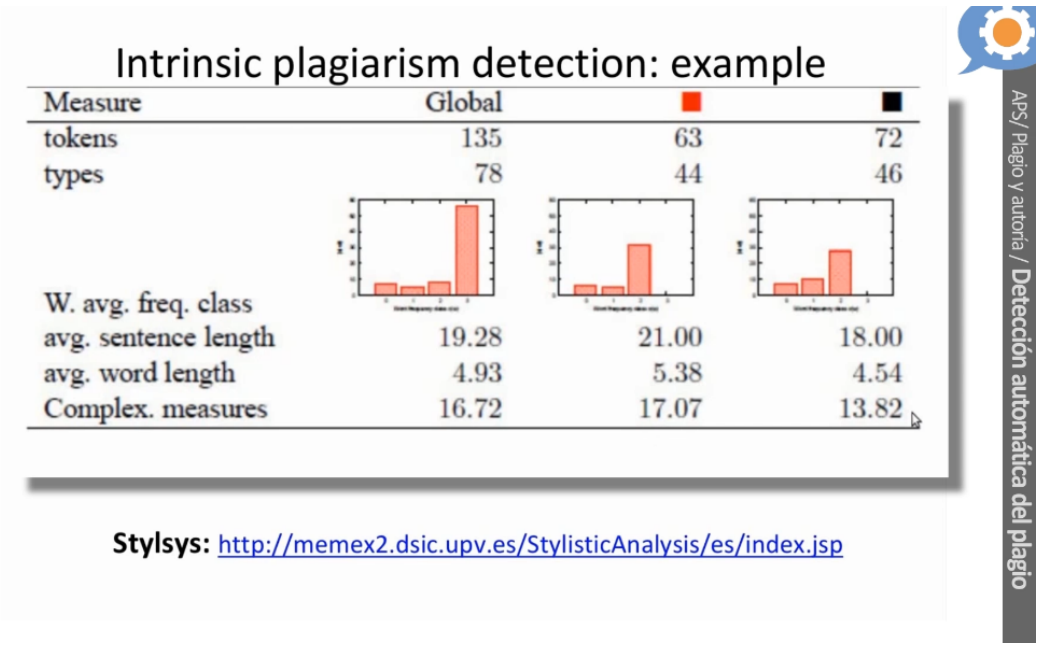
Con respecto a la detección intrínseca del plagio, como decía anteriormente, se trata de ver si hay un cambio de estilo. Podéis darle a la pausa y leer tranquilamente éste texto:

> En este trabajo, hemos hecho una investigación acerca de la influencia que tiene la cantidad de sales minerales en el humor de las personas. Para la investigación he trabajado con 5 personas que han tomado agua con distinta cantidad de sales minerales. Nuestra teoría es que entre más sales minerales haya en el agua, las personas son más volubles. […]
>
> Las sales minerales son moléculas inorgánicas de fácil ionización en presencia de agua y que en los seres vivos aparecen tanto precipitadas como disueltas. Las sales minerales disueltas en agua siempre están ionizadas. Estas sales tienen función estructural y funciones de regulación del pH, de la presión osmótica y de reacciones bioquímicas, en las que intervienen iones específicos. […]
>
> Me parece que los resultados son buenos. […]

Como habréis podido comprobar, en el primer párrafo y en el último hay expresiones más coloquiales: “hemos hecho”, he trabajado”, “nuestra teoría”, “me parece”,… Mientras que en el segundo párrafo el texto es algo más formal.

Si queremos verlo de una manera de cómo, por ejemplo, una herramienta como Stylsys, que hemos desarrollado aquí en la Universitat Politècnica de Valencia, esto es lo que se podría ver con respecto al texto anterior. La parte roja se refieren a la parte más formal. Como veis hay

una complejidad mayor con respecto al primer y tercer párrafo; hay una longitud mayor con respecto a las frases y a las palabras.



Measure	Global	■	■
tokens	135	63	72
types	78	44	46
W. avg. freq. class			
avg. sentence length	19.28	21.00	18.00
avg. word length	4.93	5.38	4.54
Complex. measures	16.72	17.07	13.82

Stylsys: http://memex2.dsic.upv.es/StylisticAnalysis/es/index.jsp

Podéis investigar más subiendo algunos textos o utilizando los textos de prueba que encontraréis en esta dirección: http://memex2.dsic.upv.es/StylisticAnalysis/es/index.jsp.