TRATAMIENTO MASIVO DE DOCUMENTOS CON LENGUAJE NATURAL

MINECO. Ministerio de Economía y Competitividad
REUNIÓN DE TRABAJO INTERCAMBIO DE INFORMACIÓN DE I+D+I
Madrid, 15 de septiembre de 2015.
Resumen de Sixto Jansa sjansa@pas.uned.es de la ponencia de la Secretaría de Estado para la Sociedad Digital. Editado en junio de 2016.

Utilidades de los sistemas de tratamiento de documentos basados en el lenguaje natural en el ámbito de las políticas de apoyo a la investigación. Utilidades del lenguaje natural para gestores de innovación

A. Áreas de interés

1. Se necesita tener la visión de conjunto del sector TIC que las estadísticas no pueden aportar debido al ritmo de cambio este sector.
2. Hay multitud de organismos relacionados con innovación, existe la dificultad de saber qué hacen y en qué coinciden
3. Hace posible el análisis de tópicos de los abstract de los proyectos de I+D+I (Convocatorias competitivas).
4. Interesa observar la evolución en el tiempo de un tema
5. Seleccionar documentos de un área concreta, señalando el grado de pertenencia de cada uno a un tópico
6. En patentes se necesita ayudar al evaluador (examinador) a identificar si hay novedad en una solicitud a partir de las citas
7. Permitiría ganar tiempo reduciendo el rango de búsqueda. Sería un sistema automático de ayuda al examinador procesando muchos documentos y estableciendo un ranking de documentos.
8. Buscan herramientas para hacer búsquedas y localizar contenidos anonimizando el documento ¿Razón para el anonimato?
9. Ver de qué trata un documento sin haberlo leído para clasificarlo.
10. Poder responder preguntas sobre si hay solicitudes en otros organismos
11. Permitirá determinar por donde están tirando los investigadores.
12. Establecer la relación entre los proyectos financiados y los resultados de patentes y publicaciones.

13. Establecer cómo asignar evaluadores (de forma automática por temas de especialización)
14. Mejorar la colaboración entre investigadores e innovadores detectando cuales trabajan en lo mismo. (Recurso para Vigilancia Tecnológica)
15. El sistema automáticamente puede dar la alarma ante copias.
16. En el futuro procesaría información de varios idiomas

B. Requerimientos de trabajo: (características)

1. Aplicable a grupos de varios cientos o miles de documentos
2. El programa hace una bolsa de palabras por documento, realiza análisis de tópicos,
3. Permite encontrar proyectos duplicados
4. Identifica los temas tratados en cada documento
5. No se trata de una clasificación automática
6. Establece una huella textual de cada documento (Ej. Patente)
7. Detecta documentos anormalmente parecidos: copias, reescrituras, similares.
8. Facilita la búsqueda textual
9. Agrupaciones por metadatos
10. Hace un proceso de clustering