



Becas colaboración curso 2019/2020

Fecha: 07 Junio 2019

Vicerrectorado de Investigación, Innovación y Transferencia

Subcomisión de I+D+i

Propuesta del departamento *SISTEMAS INFORMATICOS Y COMPUTACION*

Núm Proyecto: 2019/32/00012

Responsable

Sánchez Peiró, Joan Andreu

E-mail

jandreu@prhlt.upv.es

Ext.

77358

Responsable

Benedí Ruiz, José Miguel

E-mail

jmbenedi@prhlt.upv.es

Ext

79722

Título proyecto

Desarrollo de un corpus anotado para entrenar modelos de indexación de expresiones matemáticas en colecciones masivas de documentos impresos.

Valoración proyecto

4

Descripción proyecto

Para abordar el problema de la indexación y búsqueda de expresiones matemáticas (EM) en grandes colecciones de imágenes de documentos impresos, es necesario disponer de una gran cantidad de datos de entrenamiento. En el caso de EM, estos datos están compuestos de pares, región de la imagen (caja de inclusión) y transcripción (expresión latex). Habitualmente, la preparación de estos datos es una tarea costosa que se realiza manualmente. En este proyecto se estudiarán técnicas semi-automáticas de generación de un gran corpus que permita entrenar los correspondientes modelos de indexación.

Actividades a realizar por el alumno

1. Familiarizarse con el problema del reconocimiento de expresiones matemáticas.
2. Construir un sistema que permita detectar las expresiones matemáticas en un documento (pdf), teniendo como entrada tanto el fichero (pdf) como los fuentes latex.
3. Explotar este sistema para la obtención de un gran corpus formado por regiones que contengan expresiones matemáticas (caja de inclusión) y su transcripción (expresión latex).
4. Redacción de la memoria y documentación del corpus generado.

Horario



Becas colaboración curso 2019/2020

Fecha: 07 Junio 2019

15 horas semanales distribuidas de lunes a viernes, en horario que no afecte al horario regular del alumno para asistir a clase.