Clasificación semántica y visual de documentos digitales

Classificació semàntica i visual de documents digitals

Marçal Rusiñol

Item: revista de biblioteconomia i documentació, ISSN 0214-0349, n. 65-66, juliol-desembre 2018 / gener- juny 2019, p. 74-87

Se analizan los sistemas de procesamiento automático que trabajan sobre documentos digitalizados con el objetivo de describir los contenidos. De esta forma contribuyen a facilitar el acceso, permitir la indización automática y hacer accesibles los documentos a los motores de búsqueda. El objetivo de estas tecnologías es poder entrenar modelos computacionales que sean capaces de clasificar, agrupar o realizar búsquedas sobre documentos digitales. Así, se describen las tareas de clasificación, agrupamiento y búsqueda. Cuando utilizamos tecnologías de inteligencia artificial en los sistemas de clasificación esperamos que la herramienta nos devuelva etiquetas semánticas; en sistemas de agrupamiento que nos devuelva documentos agrupados en clusters significativos; y en sistemas de búsqueda esperamos que dada una consulta, nos devuelva una lista ordenada de documentos en función de la relevancia. A continuación se da una visión de conjunto de los métodos que nos permiten describir los documentos digitales, tanto de manera visual (cuál es su apariencia), como a partir de sus contenidos semánticos (de qué hablan). En cuanto a la descripción visual de documentos se aborda el estado de la cuestión de las representaciones numéricas de documentos digitalizados tanto por métodos clásicos como por métodos basados en el aprendizaje profundo (deep learning). Respecto de la descripción semántica de los contenidos se analizan técnicas como el reconocimiento óptico de caracteres (OCR); el cálculo de estadísticas básicas sobre la aparición de las diferentes palabras en un texto (bag-of-words model); y los métodos basados en aprendizaje profundo como el método word2vec, basado en una red neuronal que, dadas unas cuantas palabras de un texto, debe predecir cuál será la siguiente palabra. Desde el campo de las ingenierías se están transfiriendo conocimientos que se han integrado en productos o servicios en los ámbitos de la archivística, la biblioteconomía, la documentación y las plataformas de gran consumo, sin embargo los algoritmos deben ser lo suficientemente eficientes no sólo para el reconocimiento y transcripción literal sino también para la capacidad de interpretación de los contenidos.

Resumen elaborado por María Osuna González

El nacimiento de un encabezamiento de materia

Anna M. Ferris

Library Resources & Technical Services, ISSN 0024-2527, Vol. 62, n. 1, 2018, p. 16-27

Un aspecto esencial de la clasificación es la creación del encabezamiento de materia, por el que todos los materiales incluidos en el catálogo referentes a ese tema se regirán. Esto tiene dos claros beneficios: facilitar que todos los nuevos ítems que atañan a este tema se sitúen bajo un único término descriptivo y ayudar en las búsquedas de los usuarios. Las propuestas de nuevos encabezamientos de materia en la Library of Congres (LC) se producen a través de SACO. Este proceso puede ser intimidante. El artículo explica cómo funciona. En 1994 la LC creó el Program for Cooperative Cataloging (PCC), para estandarizar los procesos de la catalogación compartida. Un año después se creó SACO, centrada en la creación de Program for Cooperative Cataloging (PCC). Cualquier catalogador, sin tener que estar especializado, puede enviar su propuesta a SACO en línea. Cada propuesta es analizada y si se aprueba pasa a formar parte del archivo de autoridades de materia. En 2001 se publicó un manual sobre políticas y procedimientos, y en 2007 se editó un curso en línea. El sistema de propuestas de SACO se actualiza y mejora continuamente. SACO también organiza talleres para enseñar cómo se crea un encabezamiento de materia. Las tres principales características que debe cumplir para ser admitido son: tiene que ser un concepto nuevo, tiene que haber obras que traten este tema y tiene que haber información de calidad que apoye su establecimiento. Además, hay varios que requisitos que se deben cumplir: tiene que ser un encabezamiento uniforme, único y específico. A continuación el autor detalla el proceso que siguieron sus propuesta de “libido” y “negacionistas del Holocausto” hasta ser admitidas por la LC.

Resumen elaborado por Antonio Rodríguez Vela

My Life as a “Like-Minded Misfit,” or, Experiences in Zine Librarianship

Heidy Berthoud

Serials Review, ISSN: 1879-095X, Vol. 44, n.1, 2018, p. 4-12

Los fanzines son publicaciones ajenas al mundo académico. Tienen muchas formas y tamaños, son difíciles de catalogar para las bibliotecas porque puede no haber autoridades o ser difíciles de identificar, las materias son difusas, la adquisición es complicada y los campos del registro a rellenar son confusos. Además, ¿son monografías o publicaciones periódicas? Pueden ser una fuente primaria de historia social y cultural. Suelen ser de dos tipos, personales y políticas. No hay una norma de suscripciones aplicable a ellos. Para adquirirlos hay algunos distribuidores, pero también es necesario el uso de redes sociales o los contactos personales. Tampoco hay calendarios fijos de entrega. Clasificarlos como monografías o publicaciones periódicas es una cuestión personal del catalogador. Hay muchas maneras aceptables de proporcionar acceso a un título. Siempre hay que comprobar si hay una copia en OCLC y de ser así seguir su modelo. Los títulos pueden ser muy largos y extravagantes, se recomienda el uso literal de RDA para su representación en el registro MARC. A la hora de crear autoridades lo mejor es utilizar punteros y aplicarlos de manera consistente. Para catalogar fanzines divididos, con dos títulos, también es recomendable seguir las RDA, con un título colectivo. La información sobre copyright también suele ser muy variable, se puede utilizar el campo 542 para anotar las peculiaridades. En cuanto a la proveniencia, se utiliza el campo 541. Los sumarios suelen ser más fáciles, al estar muchas veces casi incluidos en los títulos. El vocabulario de la Library of Congress es inadecuado para los fanzines. La autora prefiere la utilización de Anchor Archive Thesaurus. También se pueden crear materias propias. Los géneros suelen estar bien delimitados. A menudo es conveniente avisar sobre contenido potencialmente perturbador o inadecuado. Para la clasificación se usa el sistema Cutter. Si se basa en el autor el número es seguido por la primera letra del título. Si se trata de una publicación periódica, el número Cutter es seguido por el número de la publicación.

Resumen elaborado por Antonio Rodríguez Vela