miércoles, septiembre 16, 2020

Resumo TFM

Identificación de TFM (para o profesorado):	SJOPFWD
Apelidos e Nome do Titor/a	Borrajo Diz, María Lourdes
Correo electrónico do Titor/a	lborrajo@uvigo.es
Apelidos e Nome do CoTitor/a (se procede)	Seara Vieira, Adrián
Apelidos e Nome do Alumno/a	Celard Pérez, Pedro
DNI do Alumno/a	
Título do TFM	Análisis y adaptación de modelos LDA en sistemas de recuperación de información textual

Resumo

Los sistemas de *information retrieval* [1] (recuperación de información) tratan de realizar búsquedas sobre datos no estructurados como
texto en base a una *query* (consulta).
Estos sistemas ordenan los resultados obtenidos por su relevancia en relación
con la búsqueda realizada. Entre ellos, una de las técnicas más utilizadas es
la Okapi BM25 [2] que utiliza la
representación de bolsa de palabras para calcular la relevancia de un documento
sobre la consulta dada.

A la hora de
realizar estas búsquedas, los modelos de *query expansion* [3] (expansión de
consultas) tratan de mejorar el rendimiento de los sistemas de recuperación de
información reformulando la consulta original.

Para poder realizar este proceso es necesario adaptar la representación de los documentos a un formato que los algoritmos de clasificación puedan comprender como entrada.
Actualmente la forma más común de representación es la denominada *bag-of-words* [4] (bolsa de palabras), en la que un documento pasa a ser un vector donde cada elemento indica la frecuencia de aparición de un término en el mismo. Sin embargo, esta representación utiliza vectores con una dimensionalidad alta, provocando serios problemas al tratar grandes conjuntos de datos.

Recientemente, los algoritmos de *Topic Modeling* [5] se están empleando para descubrir de forma no supervisada los temas inherentes a los textos. Entre estos algoritmos se encuentra el modelo basado en LDA (*Latent Dirichlet Allocation*) [6], que ofrece un vector de dimensionalidades reducidas para representar un documento a raíz una serie de *topics* (temas) encontrados en el corpus de documentos.

El trabajo principal consistirá en explorar y explotar las características del modelo LDA, analizando su aplicabilidad y eficiencia en los distintos ámbitos y fases de la recuperación de información documental. En concreto, los objetivos que se persiguen con este trabajo de tesis son:

I. Realizar un estado del arte en el que se revisen diversas técnicas de preprocesado, consulta y recuperación de documentos.

2. Desarrollo de un modelo de

expansión o retroalimentación de consultas basado en la aplicación de
algoritmos LDA para la recuperación de información.

3. Evaluación del modelo y
comparación con otros algoritmos de recuperación de información utilizando
corpus de datos biomédicos y otros utilizados ampliamente en el comentado
ámbito de estudio.

Nota: La memoria de este TFM será redactada en inglés.

[1] E. Lim, J. Liu y R. Lee, «Text Information Retrieval,» Intelligent Systems Reference Library, vol. 8, pp. 27-36, 2011.
[2] J. Whissell y C. Clarke, «Improving document clustering using Okapi BM25 feature weighting,» Information Retrieval, vol. 14, nº 5, pp. 466-487, 2011.
[3] D. Akila, S. Sathya y G. Suseendran, «Survey on query expansion techniques in word net application,» Journal of Advanced Research in Dynamical and Control Systems, vol. 10, nº 4, pp. 119-124, 2018.
[4] R. Baeza-Yates y B. Ribeiro-Neto, Modern Information Retrieval, Addison-Wesley Longman, 1999.
[5] D. Blei, «Probabilistic topic models,» Communications of the ACM, vol. 55, nº 4, pp. 77-84, 2012.
[6] D. Blei, A. Ng y M. Jordan, «Latent Dirichlet allocation,» Journal of Machine Learning Research, vol. 3, nº 4-5, pp. 993-1022, 2003.

(A empresa debe ter convenio asinado en vigor coa Universidade de Vigo. **Deberase entregar copia do nomeamento do/a titor/a pola empresa**)

*** O/A profesor/a recibirá copia desta solicitude, e deberá dar o Visto e Prace do mesmo dende o formulario online dispoñible na web da ESEI.**

Código do TFM (para o alumno): MEI 20/21-3

Introduce un correo electrónico válido (a continuación recibirá un código para a firma da solicitude)

*** Unha vez enviada a solicitude recibirá por correo electrónico copia da mesma**