Modelagem semântica de dados abertos: a viabilidade de aplicação de word embeddings sobre o currículo lattes

Felipe de Paula Oliveira
Mestrando em Modelagem Matemática e Computacional pelo Centro Federal de Educação Tecnológica de Minas Gerais (CEFET/MG) – MG - Brasil. Especialização em Segurança da Informação pelo Centro Universitário Estácio Ribeirão Preto (Estácio) – Ribeirão Preto, SP - Brasil. Servidor Público da Prefeitura Municipal de Divinópolis - Brasil.
http://lattes.cnpq.br/6651319979088361
E-mail: engcomp.felipedepaula@gmail.com

Thiago Magela Rodrigues Dias
Doutor em Modelagem Matemática e Computacional pelo Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG) - Belo Horizonte, MG - Brasil. Professor do Centro Federal de Educação Tecnológica de Minas Gerais (CEFET) - Divinópolis, MG - Brasil.
http://lattes.cnpq.br/4687858846001290
E-mail: thiagomagela@gmail.com

Adilson Luiz Pinto
Pós-Doutorado pelo Institut de Recherche en Sciences de l´Information et de la Communication (IRSIC) - França. Doutor em Documentación pela Universidad Carlos III de Madrid (UC3M) - Espanha. Professor da Universidade Federal de Santa Catarina,(UFSC) - Florianópolis, SC - Brasil.
http://lattes.cnpq.br/4767432940301118
E-mail: adilson@cin.ufsc.br

Data de submissão: 13/09/2019. Data de aprovação no ConfOA: 11/06/2019. Data de publicação:

RESUMO

O ato de recuperar estudar e avaliar a atividade científica brasileira disponibilizada em dados abertos, como na Plataforma Lattes, é um processo extenso e complexo, porém de extrema necessidade para construir indicadores de produção e desempenho científico. Para isso, realizam-se diversos tipos de análises bibliométricas, podendo ser aplicados tanto métodos ou modelos tradicionais quanto alternativos para a avaliação da ciência. Este estudo analisa a viabilidade de aplicação de PLN (processamento de linguagem natural) por meio da similaridade semântica, a qual visa determinar se o contexto de um trecho implica outro. A inferência textual realiza a atribuição de uma pontuação de similaridade semântica ao par, sendo aplicada a ferramenta Word2Vec, utilizando o algoritmo PV-DBOW, tornando possível a realização de inferência dos termos. Como resultados, é possível obter índices de similaridade calculados pelo modelo entre palavras contidas em títulos de publicações. Espera-se que, a partir dessa verificação, obtenham-se futuramente resultados mais consideráveis com a modelagem semântica de outros elementos do currículo Lattes, visando por exemplo, a classificação automática de descrições de projetos de pesquisa.

Palavras-chave: Plataforma Lattes. Processamento de linguagem natural. Similaridade semântica.

Open data semantic modeling: the feasibility of applying word embeddings on the lattes curriculum

ABSTRACT

The act of retrieving studying and evaluating the Brazilian scientific activity available in open data, such as the Lattes Platform, is an extensive and complex process. However, it is extremely necessary to build indicators of production and scientific performance. For this, several types of bibliometric analyzes are performed, and can be applied either traditional or alternative methods or models for the evaluation of science. The present study tries to affirm about the viability of applying PLN (Natural Language Processing) through semantic similarity, which aims to determine if the context of one passage implies another. The textual inference performs the assignment of a semantic similarity score to the pair, the Word2Vec tool was applied using the PV-DBOW algorithm, making it possible to infer the terms. As a result it is possible to obtain similarity indices calculated by the model between words contained in titles of publications. It is hoped that from this verification, more results can be obtained in the future with semantic modeling of other elements of the Lattes curriculum, aiming, for example, the automatic classification of research project descriptions.

Keywords: *Lattes platform. Natural language processing. Semantic similarity.*

Modelado semántico de datos abiertos: la viabilidad de aplicar incrustaciones de palabras en el plan de estudios lattes

RESUMEN

El acto de recuperar el estudio y la evaluación de la actividad científica brasileña disponible en datos abiertos, como la Plataforma Lattes, es un proceso extenso y complejo. Sin embargo, es extremadamente necesario construir indicadores de producción y desempeño científico. Para esto, se realizan varios tipos de análisis bibliométricos, y pueden aplicarse métodos o modelos tradicionales o alternativos para la evaluación de la ciencia. El presente estudio trata de afirmar la viabilidad de aplicar PLN (Procesamiento del Lenguaje Natural) a través de la similitud semántica, cuyo objetivo es determinar si el contexto de un pasaje implica otro. La inferencia textual realiza la asignación de un puntaje de similitud semántica al par, se aplicó la herramienta Word2Vec, utilizando el algoritmo PV-DBOW, siendo posible la inferencia de los términos. Como resultado, es posible obtener índices de similitud calculados por el modelo entre palabras contenidas en títulos de publicaciones. Se espera que a partir de esta verificación, se puedan obtener más resultados en el futuro con el modelado semántico de otros elementos del plan de estudios Lattes, con el objetivo, por ejemplo, de la clasificación automática de descripciones de proyectos de investigación.

Palabras clave: *Plataforma Lattes. Procesamiento de lenguaje natural. Similitud semántica.*