Uma estratégia para identificação de gênero em repositórios de dados abertos utilizando um modelo de rede neural artificial

Sérgio José de Sousa
Mestrando em Modelagem Matemática e Computacional pelo Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG) – Belo Horizonte, MG, Brasil. Graduação em Sistemas de Informação pela Faculdade Pitágoras de Divinópolis (FAP) - Brasil.
http://lattes.cnpq.br/1639967799540564
E-mail: sergio7sjs@gmail.com

Monique de Oliveira Santiago
Mestranda em Modelagem Matemática e Computacional pelo Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG) – Belo Horizonte, MG, Brasil. Graduação em Engenharia da Computação pela Universidade do Estado de Minas Gerais - Unidade Divinópolis (UEMG Divinópolis) - Brasil.
http://lattes.cnpq.br/3530976051984613
E-mail: moniqueosantiago@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-MG) – Belo Horizonte, MG, Brasil.
http://lattes.cnpq.br/4687858846001290
E-mail: thiagomagela@gmail.com

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

Diversos são os estudos que tentam identificar e relacionar o gênero a uma quantidade de produção, em especial, na área acadêmica existe um esforço em responder à questão: há diferenças entre a produtividade científica feminina e masculina? Responder esse tipo de questão é de suma importância para identificar polaridades e desigualdades. Mas muitas vezes para realizar este tipo de análise não temos essa informação disponível, propomos aqui uma estratégia para obtenção dessa informação com base no nome completo. Este trabalho utiliza dados dos grupos de pesquisa do censo de 2016 que estão disponíveis na Plataforma Lattes. Totalizando 622.383 nomes, sendo 352.804 femininos e 269.581 masculinos. O pré-processamento inclui inverter a ordem dos nomes: “Alcides da Silva Diniz” se torna “Diniz Silva da Alcides”. Dessa maneira, as informações mais importantes para a classificação são posicionadas no fim, dando uma indicação mais forte ao modelo, proporcionando que ela aprenda mais rapidamente e de uma maneira mais eficiente. A seguir, os nomes são transformados em tensores cujas letras são convertidas em valores inteiros sequencialmente. O modelo consiste em uma camada Embedding com 97 neurônios de entrada e 4 de saída, uma camada LSTM com entrada de 4 neurônios e 256 na saída e, por fim, a camada de saída com entrada de 256 e saída com 2 neurônios, já classificando o nome com a função Softmax. Treinado por 16 épocas e utilizando estratégia de validação cruzada, o modelo proposto atingiu uma acurácia média de 99,549% no treino e 98,995% no teste.

Palavras-chave: Rede Neural. Nome Completo. Identificação de gênero. Deep Learning.

A strategy for gender identification in open data repositories using an artificial neural network model

ABSTRACT

There are several studies that attempt to identify and relate gender to a quantity of production, especially in the academic area, there is an effort to answer the question: are there differences between female and male scientific productivity? Answering such questions is of utmost importance in identifying polarities and inequalities. But often, to perform this type of analysis, we do not have this information available. Therefore, we propose a strategy to obtain information based on the full name. This document uses data from the 2016 census research groups that are available on the Lattes Platform. Totaling 622,383 names, 352,804 women and 269,581 men. Preprocessing includes reversing the order of names: "Alcides da Silva Diniz" becomes "Diniz Silva da Alcides". In this way, the most important information for classification is placed at the end, which gives the model a stronger indication, allowing it to learn faster and more efficiently. Then, the names are transformed into tensors whose letters become integer values sequentially. The model consists of an inlay layer with 97 input and 4 output neurons, an LSTM layer with 4 input and 256 output neurons and, finally, the output layer with 256 inputs and 2 output neurons, which is already Sorting the name with the Softmax function. Trained for 16 seasons and using a cross-validation strategy, the proposed model achieved an average accuracy of 99.549% in training and 98.995% in the test.

Keywords: Neural network. Full name. Gender identification. Deep Learning.

Una estrategia para la identificación de género en repositorios de datos abiertos utilizando un modelo de red neuronal artificial

RESUMEN

Hay varios estudios que intentan identificar y relacionar el género con una cantidad de producción, especialmente en el área académica, hay un esfuerzo por responder la pregunta: ¿existen diferencias entre la productividad científica femenina y masculina? Responder tales preguntas es de suma importancia en la identificación de polaridades y desigualdades. Pero a menudo, para realizar este tipo de análisis, no tenemos esta información disponible. Por lo tanto, proponemos una estrategia para obtener información basada en el nombre completo. Este documento utiliza datos de los grupos de investigación del censo de 2016 que están disponibles en la Plataforma Lattes. Totalizando 622,383 nombres, siendo 352,804 mujeres y 269,581 hombres. El preprocesamiento incluye invertir el orden de los nombres: "Alcides da Silva Diniz" se convierte en "Diniz Silva da Alcides". De esta manera, la información más importante para la clasificación se coloca al final, lo que le da al modelo una indicación más fuerte, lo que le permite aprender más rápido y más eficientemente. Después, los nombres se transforman en tensores cuyas letras se convierten en valores enteros secuencialmente. El modelo consiste en una capa de incrustación con 97 neuronas de entrada y 4 de salida, una capa LSTM con 4 neuronas de entrada y 256 de salida y, finalmente, la capa de salida con 256 entradas y 2 neuronas de salida, que ya está clasificando El nombre con la función Softmax. Entrenado durante 16 temporadas y utilizando una estrategia de validación cruzada, el modelo propuesto logró una precisión promedio de 99.549% en el entrenamiento y 98.995% en la prueba.

Palabras clave: *Red neuronal. Nombre completo. Identificación de género. Deep Learning.*