Informationssysteme
Prof. Dr. Felix Naumann
Masterprojekt Sommer 2019

What's in a Life? Representing Vitae in 1000 Dimensions

Changing perspective on changing people

Life leads people through many different stations, usually gathered and organized in a CV. When people are particularly important or famous, reports of their lives might even fill several pages of text on Wikipedia, written by a variety of people in collaboration over many years. Such online, collaborative editing raises some interesting questions: How does the world's perspective on these prominent people change over time? When are new stages of life added to Wikipedia? Do articles from a person's surroundings change at the same time? Can we find a machine representation of their vitae?

Project Outline

Wikipedia is a collaborative encyclopedia that was launched 18 years ago, in 2001. Since then, countless authors have written and curated more and more information to reach today's almost 6 million articles (English Wikipedia). In this master-project, we want to analyze the evolution of this knowledge base. We focus the analysis on the over 890 thousand articles about persons and suppose that changes to these articles correspond with their biographies "in real time" over the past 18 years.

By processing the revisions of written free text and infobox-facts over time, we will generate a structured dataset for further analysis and exploration [4,5,6]. We will use this data to visualize the evolution of specific articles [1,2] and identify groups of people. We will not only use linguistic measures and structured metadata, but also train high-dimensional representations of the unstructured data to fulfil these tasks. These machine-representations could be used to query the data [3] or find arithmetic analogies as shown by doc2vec [7].

There are no explicit prerequisites for this project but an interest in one or more of the following research areas or topics might be helpful:

- Machine learning
- Text mining
- Working with very large datasets
- Building scalable applications

Contact

We offer this project for up to five students who will be supervised closely by Prof. Felix Naumann, Tobias Bleifuß (tobias.bleifuss@hpi.de), and Tim Repke (tim.repke@hpi.de). If you have any questions, please do not hesitate to contact us.

Informationssysteme
Prof. Dr. Felix Naumann
Masterprojekt Sommer 2019

Related Work

[1] Perez-Messina, et al, Organic Visualization of Document Evolution, IUI 2019

[2] Szafir et al, TextDNA: Visualizing Word Usage with Configurable Colorfields, EuroVis 2016

[3] Risch et al, Book Recommendation Beyond the Usual Suspects: Embedding Book Plots Together with Place and Time Information, ICADL 2018

[4] Flöck et al., WikiWho: precise and efficient attribution of authorship of revisioned content, WWW 2014

[5] Yao et al., Dynamic Word Embeddings for Evolving Semantic Discovery, WSDM 2018

[6] Bykau et al., “Tell me more" using Ladders in Wikipedia, WebDB 2017

[7] Mikolov et al., Distributed representations of words and phrases and their compositionality, NIPS 2013