sensors

MDPI

Article

Modeling Topics in DFA-Based Lemmatized Gujarati Text

Uttam Chauhan [1], Shrusti Shah [1], Dharati Shiroya [1], Dipti Solanki [1], Zeel Patel [1], Jitendra Bhatia [2,*], Sudeep Tanwar [2], Ravi Sharma [3], Verdes Marina [4,*] and Maria Simona Raboaca [5,6]

[1] Department of Computer Engineering, Vishwakarma Government Engineering College, Chandkheda, Ahmedabad 382424, India
[2] Department of Computer Science and Engineering, Institute of Technology, Nirma University, Ahmedabad 382481, India
[3] Ravi Sharma, Centre for Inter-Disciplinary Research and Innovation, University of Petroleum and Energy Studies, Dehradun 248001, India
[4] Faculty of Civil Engineering and Building Services, Department of Building Services, Technical University of Gheorghe Asachi, 700050 Iasi, Romania
[5] Doctoral School, University Politehnica of Bucharest, Splaiul Independentei Street No. 313, 060042 Bucharest, Romania
[6] National Research and Development Institute for Cryogenic and Isotopic Technologies—ICSI Rm. Vâlcea, Uzinei Street, No. 4, P.O. Box 7 Râureni, 240050 Râmnicu Vâlcea, Romania
* Correspondence: jitendra.bhatia@nirmauni.ac.in (J.B.); marina.verdes@academic.tuiasi.ro (V.M.)

Abstract: Topic modeling is a machine learning algorithm based on statistics that follows unsupervised machine learning techniques for mapping a high-dimensional corpus to a low-dimensional topical subspace, but it could be better. A topic model's topic is expected to be interpretable as a concept, i.e., correspond to human understanding of a topic occurring in texts. While discovering corpus themes, inference constantly uses vocabulary that impacts topic quality due to its size. Inflectional forms are in the corpus. Since words frequently appear in the same sentence and are likely to have a latent topic, practically all topic models rely on co-occurrence signals between various terms in the corpus. The topics get weaker because of the abundance of distinct tokens in languages with extensive inflectional morphology. Lemmatization is often used to preempt this problem. Gujarati is one of the morphologically rich languages, as a word may have several inflectional forms. This paper proposes a deterministic finite automaton (DFA) based lemmatization technique for the Gujarati language to transform lemmas into their root words. The set of topics is then inferred from this lemmatized corpus of Gujarati text. We employ statistical divergence measurements to identify semantically less coherent (overly general) topics. The result shows that the lemmatized Gujarati corpus learns more interpretable and meaningful subjects than unlemmatized text. Finally, results show that lemmatization curtails the size of vocabulary decreases by 16% and the semantic coherence for all three measurements—Log Conditional Probability, Pointwise Mutual Information, and Normalized Pointwise Mutual Information—from −9.39 to −7.49, −6.79 to −5.18, and −0.23 to −0.17, respectively.

Keywords: topic models; Gujarati text lemmatization; Latent Dirichlet Allocation; poor quality topics; overly general topics

Citation: Chauhan, U.; Shah, S.; Shiroya, D.; Solanki, D.; Patel, Z.; Bhatia, J.; Tanwar, S.; Sharma, R.; Marina, V.; Raboaca, M.S. Modeling Topics in DFA-Based Lemmatized Gujarati Text. *Sensors* **2023**, *23*, 2708. https://doi.org/10.3390/s23052708

Academic Editors: Chang Choi, Kiho Lim and Gyuho Choi

Received: 5 February 2023
Revised: 18 February 2023
Accepted: 24 February 2023
Published: 1 March 2023

1. Introduction

Topic modeling is statistical modeling for uncovering abstract "topics" hidden in a massive text collection. For example, Latent Dirichlet Allocation (LDA) infers the topics in a text collection [1]. Linguistic field researchers have shown great interest in techniques for discovering a smaller set of word clusters (known as topics) that represents the whole corpus without losing its significance. The set of techniques for modeling topics in domains are Latent Semantic Analysis (LSA) [2], probabilistic Latent Semantic Analysis (pLSA) [3], followed by LDA.