UvA-DARE (Digital Academic Repository)

Significant Words Representations of Entities

Dehghani, M.

DOI

[10.1145/2911451.2911474](https://doi.org/10.1145/2911451.2911474)

Publication date

2016

Document Version

Final published version

Published in

SIGIR'16

[Link to publication](#)

Citation for published version (APA):

Dehghani, M. (2016). Significant Words Representations of Entities. In *SIGIR'16: the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval: Pisa, Italy, July 17-21, 2016* (pp. 1183). Association for Computing Machinery. <https://doi.org/10.1145/2911451.2911474>

General rights

It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations

If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: <https://uba.uva.nl/en/contact>, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

Significant Words Representations of Entities

Mostafa Dehghani
Institute for Logic, Language and Computation
University of Amsterdam, The Netherlands
dehghani@uva.nl

ABSTRACT

Transforming the data into a suitable representation is the first key step of data analysis, and the performance of any data oriented method is heavily depending on it. We study questions on how we can best learn representations for textual entities that are: 1) precise, 2) robust against noisy terms, 3) transferable over time, and 4) interpretable by human inspection. Inspired by the early work of Luhn [4], we propose *significant words language models* of a set of documents that capture all, and only, the significant shared terms from them. We adjust the weights of common terms that are already well explained by the document collection as well as the weight of incidental rare terms that are only explained by specific documents, which eventually results in having only the significant terms left in the model.

KEYWORDS: Significant Words; Language Model.

1. SIGNIFICANT WORDS MODELS

Transformation of raw data to a representation that can be effectively exploited is motivated by the fact that data oriented methods often require input that is convenient to process. In this research, we introduce *significant words language models* (SWLM) as a family of models aiming to learn representations for the set of documents that are not affected by neither general properties nor specific properties. The general idea of SWLM is inspired by the early work of Luhn [4], in which he argues that to extract **significant words** by avoiding both common observations and rare observations (Figure1).

In order to estimate SWLM, we assume that terms in the each document in the set are drawn from three models: 1. *General model*, representative of common observation, 2. *Specific model*, representative of partial observation, and 3. *Significant Words model* which is a latent model representing the significant characteristics of the whole set. Then, we try to extract the latent significant words model.

1.1 Applications of the SWLM

The proposed approach is generally applicable to any system that requires the estimation of an effective model representing the significant features of a group of objects. Until now, we have employed the model in three main applications:

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

SIGIR '16 July 17-21, 2016, Pisa, Italy

© 2016 Copyright held by the owner/author(s).

ACM ISBN 978-1-4503-4069-4/16/07.

DOI: <http://dx.doi.org/10.1145/2911451.2911474>

Figure 1: Establishing a set of “Significant Words” based on Luhn [4]

Group Profiling for Content Customization: We have proposed to use SWLM to extract the ‘abstract’ group level latent model from users group that captures all, and only, the essential features of the whole group. We employed the resulting models in the task of contextual suggestion [1, 3] and observed improvements in the performance of customization.

(Pseudo-)Relevance Feedback: We have presented a variant of SWLM, Regularized SWLM (RSWLM) for estimating a robust language model for a set of feedback documents by incorporating the information from the query. We have conducted extensive experiments on the effectiveness of RSWLM for (pseudo-)relevance feedback and demonstrated that it captures the essential terms representing the mutual notion of relevance.

Hierarchical Classification: We have also extended SWLM to be able to estimate proper models for hierarchical entities which take their position in the hierarchy into consideration [2]. We have employed SWLM on the task of hierarchical classification and observed that since estimated models of entities in the hierarchy are both horizontally and vertically separable, they are precise, robust and transferable over time.

2. REST OF THE JOURNEY

Besides applying the model in further applications, there are several other interesting directions to pursue in this research. These include using SWLM as an analytical tool to investigate and better understanding of the data, extending the method to be applicable to non-textual features, and employing the general idea of SWLM to the representation learning systems like embedding methods.

Acknowledgments This research is funded by the Netherlands Organization for Scientific Research (ExPoSe project, NWO CI # 314.99.108).

References

- [1] M. Dehghani, H. Azarbonyad, J. Kamps, and M. Marx. Generalized group profiling for content customization. In *CHIIR '16*, pages 245–248, 2016.
- [2] M. Dehghani, H. Azarbonyad, J. Kamps, and M. Marx. Two-way parsimonious classification models for evolving hierarchies. In *CLEF'16*, 2016.
- [3] S. H. Hashemi, M. Dehghani, and J. Kamps. Parsimonious user and group profiling in venue recommendation. In *TREC 2015*. NIST, 2015.
- [4] H. P. Luhn. The automatic creation of literature abstracts. *IBM J. Res. Dev.*, 2 (2):159–165, 1958.