UNIVERSITY OF MINNESOTA
Driven to Discover℠

Automated De-Identification of Distributional Semantics Models

Technology No. 20180072

De-Identification Algorithm Maintains Word Disambiguation Performance

An automated model de-identification algorithm applies aggressive de-identification to a word co-occurrence model without sacrificing performance for word sense disambiguation. While some very common words must be included in the model (i.e., names in some of their occurrences, like “white”), the de-identification process removes anything that is not part of the SPECIALIST Lexicon and any words in patient information databases (e.g., names and addresses). The one exception to this rule, critical to maintaining good word disambiguation performance, is that the 2,000 most common words in the patient database are included in the model to allow for homonyms like “white,” as mentioned above.

May Be HIPAA Compliant

In the medical domain, electronic health record (EHR) data contains protected health information with highly restricted access. The U.S. Health Insurance Portability and Accountability Act (HIPAA) specifies requirements for protecting confidentiality in EHR datasets used for non-clinical purposes by removing certain identifying strings, such as names and addresses. Performing this de-identification process manually can be prohibitively expensive, and while automated methods have been successful, healthcare institutions often remain hesitant to permit the release of automatically de-identified text. This alternative approach de-dentifies a word co-occurrence table rather than raw text. Co-occurrence statistics comprise many distributional semantic models, with many applications in biomedical natural language processing (NLP). These models do not preserve syntactic and phrasal information of their source text, dramatically reducing confidentiality risk even before de-identification. If stripped of identifiers, these models could be safely shared with other researchers to improve outcomes in NLP and information retrieval. This tool both effectively removes HIPAA identifiers from a model and preserves a de-identified model’s effectiveness in NLP tasks.

BENEFITS AND FEATURES:

- Aggressive de-identification to a word co-occurrence model
- Preserves performance for word sense disambiguation
- Effectively removes HIPAA identifiers from a model
- Preserves de-identified model's effectiveness in NLP tasks

APPLICATIONS:

- De-identification of confidential information
- Healthcare

Phase of Development – license available for non-profit research.

Researchers

Serguei Pakhomov, PhD
Professor, Department of Pharmaceutical Care and Health Systems, School of Pharmacy
External Link (www.pharmacy.umn.edu)
Genevieve Melton-Meaux, MD, PhD
Professor, Department of Surgery
External Link (healthinformatics.umn.edu)

Publications

Corpus domain effects on distributional semantic modeling of medical terms
Bioinformatics, Volume 32, Issue 23, 1 December 2016, Pages 3635–3644
Automated De-Identification of Distributional Semantic Models
AMIA 2016 Annual Symposium, Nov 12 – 16, 2016

External Links

Natural Language Processing / Information Extraction (NLP/IE) Program

Available for Licensing
The Model De-ID algorithm is available from github under the Apache License 2.0
The Data Dictionary may be licensed from the University of Minnesota by completing the online license. This is a word2vec binary file that can be used with software libraries like DL4J or Gensim.
Please contact us if you have questions.