2006 Deidentification and Smoking Challenge

**NLP Data Set #1B:**889 de-identified discharge summaries with de-identification challenge annotations, training and test sets and ground truth.

 Please cite as:

* Uzuner Ö., Juo Y, Szolovits P. "Evaluating the state-of-the-art in automatic de-identification".  J Am Med Inform Assoc. 2007, 14(5):550-63. <http://www.jamia.org/cgi/content/abstract/14/5/550>.

Other related publications:

* Uzuner Ö. , Sibanda T, Luo Y, Szolovits P. "A De-identifier for Medical Discharge Summaries".  International Journal Artificial Intelligence in Medicine. 2008; 42(1): 13-35. [www.aiimjournal.com/article/SO933-3657(07)00132-7/pdf](http://www.aiimjournal.com/article/SO933-3657(07)00132-7/pdf).
* Hara K. "Applying a SVM based chunker and a text classifier to the deid challenge". Online only at <http://jamia.bmj.com/content/suppl/2009/11/20/14.5.550.DC1/Hara_M2443R1.pdf>.
* Wellner B., Huyck M, Mardis S, Aberdeen J, Morgan M, Peshkin L, Yeh A, Hitzeman J, Hirschman L. "Rapidly retargetable approaches to de-identification in medical records".  J Am Med Inform Assoc. 2007; 12(5):564-73. <http://www.jamia.org/cgi/content/abstract/14/5/564>.
* Szarvas Gy, Farkas R, Busa-Fekete R. "State-of-the-art anonymisation of medical records using an iterative machine learning framework".  J Am Med Inform Assoc.  2007; 14(5):574-80. <http://jamia.bmj.com/content/14/5/574.full.pdf>.

Source: <https://www.i2b2.org/NLP/DataSets/Main.php>

*"Deidentified clinical records used in this research were provided by the i2b2 National Center for Biomedical Computing funded by U54LM008748 and were originally prepared for the Shared Tasks for Challenges in NLP for Clinical Data organized by Dr. Ozlem Uzuner, i2b2 and SUNY."*

++++++++++++++++++++++++++++++++++

NLP Data Set #1:

* **NLP Data Set #1A:**  889 unannotated, de-identified discharge summaries

Please cite as:

* Uzuner Ö., Juo Y, Szolovits P. "Evaluating the state-of-the-art in automatic de-identification".  J Am Med Inform Assoc. 2007, 14(5):550-63. <http://www.jamia.org/cgi/content/abstract/14/5/550>.
* **NLP Data Set #1C:**  A subset of the above 889 (N = 502) de-identified discharge summaries with smoking challenge annotations, training and test sets and ground truth.

Please cite as: 

* Uzuner Ö., Goldstein I, Luo Y, Kohane I. "Identifying patient smoking status from medical discharge records".  *J Am Med Inform Assoc*.  2008; 15(1)15-24. [www.jamia.org/cgi/content/short/15/1/14](http://www.jamia.org/cgi/content/short/15/1/14).

Other related publications:

* Savova G., Ogren P, Duffy P, Buntrock J, and Chute C. "Mayo Clinic NLP System for patient smoking status identification".  *J Am Med Inform Assoc.*  2008; 15(1):25-28. <http://jamia.bmj.com/content/15/1/25.full.pdf>.
* Wicentowski R and Sydes M. "Using implicit information to identify smoking status in smoke-blind medical discharge summaries".  *J Am Med Inform Assoc*.  2008; 15(1):29-31. <http://jamia.bmj.com/content/15/1/29.full.pdf>.
* Cohen A. "Five-way smoking status classification using text hot-spot identification and error-correcting output codes".  *J Am Med Inform Assoc.*  2008; 15(1):32-35. <http://jamia.bmj.com/content/15/1/32.full.pdf>.
* Clark C., Good K, Jezierny L, Macpherson M, Wilson B and Chajewska U. "Identifying smokers with a medical extraction system".  *J Am Med Inform Assoc.*  2008; 15(1):40-43. <http://jamia.bmj.com/content/15/1/36.full.pdf>.
* Heinze D., Morsch M, Potter B and Sheffer, Jr R. "Medical i2b2 NLP smoking challenge: The A-Life System Architecture and Methodology".  *J Am Med Inform Assoc.*  2008; 15(1):40-43. <http://jamia.bmj.com/content/15/1/40.full.pdf>.