Curriculum Vitae
Kyle Williams

University Park, PA.
kwilliams@psu.edu
http://www.personal.psu.edu/kiw5209/
Concise version of CV: http://www.personal.psu.edu/kiw5209/kwilliams_cv_short.pdf
Updated: 16 April 2016

Personal Profile

I am Ph.D. candidate at The Pennsylvania State University advised by Dr. C. Lee Giles. My research interests include information retrieval, machine learning and digital libraries. Since beginning my Ph.D. I've interned at Microsoft and Oracle Labs; achieved the highest F_1 score among all participants in the source retrieval task for plagiarism detection at PAN 2013 and PAN 2014; placed 2nd and 3rd at the Penn State Graduate Exhibition in successive years; and presented my work at several conferences.

Education

- **The Pennsylvania State University (2012-present)**
  Ph.D. in Information Sciences and Technology, advised by Prof. Lee Giles.
- **University of Cape Town (2010-2012)**
  Masters in Computer Science by Dissertation, advised by A/Prof. Hussein Suleman.
  Degree awarded with distinction
- **University of Cape Town (2006-2009)**
  Bachelor of Business Science in Management Studies in the field of Computer Science.
  Degree awarded with second class division two honours

Employment History

- **Microsoft, Bellevue, WA (May 2015 - August 2015)**
  Position: Research Intern
  Research Project: Worked on identifying signals for detecting satisfaction in mobile search.
- **Oracle Labs, Burlington, MA (June 2014 - August 2014)**
  Position: Research Intern
  Research Project: Worked on relevance ranking and search result diversification for eCommerce search.
- **College of Information Sciences and Technology, Pennsylvania State University (August 2012 - Present)**
  Position: Graduate Research Assistant and Teaching Assistant
  Responsibilities: Research under the guidance of Prof. Lee Giles .
- **Digital Libraries Laboratory, Department of Computer Science, University of Cape Town (May 2012 - August 2012)**
  Position: Research Assistant
  Responsibilities: Writing of articles, assisting honours students, assisting with day-to-day operations in the lab.

- **Department of Computer Science, University of Cape Town (February 2010 - June 2011)**
  Position: Teaching assistant for CSC1015F, CSC2001F and CSC2020S
  Responsibilities: Setting test-cases for practical tests, supervising marking, managing class tutors, handling student queries and other administrative aspects of the course.

- **Digital Libraries Laboratory, Department of Computer Science, University of Cape Town (December 2009)**
  Position: Developer
  Accomplishments: Implemented an experimental framework for online and offline static collections of digital objects. A collection using this framework was launched at "The courage of ||kabbo and a century of Specimens of Bushman folklore" conference in Cape Town in August 2011 (http://kabbo.cmc-uct.co.za/)
  Programming Languages Used: JavaScript, XML+XSLT, Perl

- **amaAmbush Productions (Pty) Ltd (2004 - 2010)**
  Company Description: amaAmbush is involved in the development of African music on the marimbas. The organization provides marimba teachers to schools, manufactures and supplies marimba instruments, and operates a number of performing marimba groups.
  Position: Teacher, performer, leader.
  Responsibilities: Teaching high school students, meeting with clients and ensuring performances ran smoothly

- **Faranani Facilitation Services (Pty) Ltd (July 2008 - June 2009)**
  Company Description: Faranani Facilitation Services (PTY) provides consulting services and offers full qualifications and skills programmes through the Faranani Learning Academy.
  Position: IT support and accounting admin
  Responsibilities: Administrative accounting (Pastel), assisting staff with IT issues

Research, Experience and Skills

- **Research Areas:** Information retrieval, applied machine learning, digital libraries, cultural heritage preservation

- **Projects**
  - 2012-Present: SimSeerX
    SimSeerX is a similarity search engine where documents are submitted as queries and similar documents in a collection are retrieved based on arbitrary similarity functions.
  - 2012-Present: Source Retrieval for Plagiarism Detection
    This project involves investigating techniques for retrieving potential sources of plagiarism for a given suspicious document. The techniques developed as part of this project have gone on to achieve the highest F1 scores in the 2013 and 2014 PAN Source Retrieval Task.
  - 2012-Present: CiteSeerX
    CiteSeerX is a scientific digital library and search engine that makes use of autonomous citation indexing. I have been involved in the project since August 2012 and my involvement primarily revolves around the technical maintenance and development of CiteSeerX as well as the management of the code repositories.

  - 2009-2012: Digital Bleek and Lloyd Collection
    A project involving the application of various technologies for managing and interacting with this collection of historical document. My primary involvement was in the construction of an image-based search engine and investigating the automatic transcription of the texts that appear in this collection.

- **Teaching:** Teaching assistant for 1st, 2nd and 4th year University courses; 5 years of extra-curricular music teaching on the Marimba at high school level,
- **Programming Languages:**
  - **High Experience:** Java, Python
  - **Medium Experience:** PHP, JavaScript (incl. jQuery), Perl, C++
  - **Low Experience:** XML + XSLT, SQL, Scala
- **Core Technologies:** Solr, Lucene, scikit-learn. Experience with Hadoop and Pig

Awards and Leadership

2014

- Highest precision and F_1-score among all participants in the PAN 2014 Source Retrieval Task for Plagiarism Detection
- 3rd place winner in the Engineering category at the 2014 Penn State Graduate Exhibition
- SIGWEB DocEng 2014 Student Travel Award
- Best deployed application award at 2014 International Conference on Innovative Applications of Artificial Intelligence for *CiteSeerX: AI in a Digital Library Search Engine*
- Best paper award nomination at 2013 International Conference on Cloud Engineering for *Migrating a Digital Library to a Private Cloud*

2013

- Best Paper Award at 2013 South African Institute for Computer Scientists and Information Technologists Conference (SAICSIT) for *A Comparison of Machine Learning Techniques for Handwritten |Xam Word Recognition*
- Highest F_1-score among all participants in the PAN 2013 Source Retrieval Task for Plagiarism Detection
- 2nd place winner in the Engineering category at the 2013 Penn State Graduate Exhibition
- President of Graduate Students in Information Sciences and Technology Club

2012

- Jordan H. Rednor Graduate Fellowship
- Entelect UCT Computer Science Best Publication Award by an MSc (Computer Science) Student

- NRF Free-standing Scholarship

Before 2012

- **2011** - Honourable mention at 13th International Conference on Asia-Pacific Digital Libraries (ICADL) for *Creating a Handwriting Recognition Corpus for Bushman Languages*; NRF Free-Standing Scholarship; UCT Travel Grant; UCT Equity Scholarship.
- **2010** - UCT Equity Scholarship; NRF Prestigious/Equity Scholarship; UCT University Research Scholarship; SIGWEB JCDL 2010 Student Travel Award; UCT Michaelis Fine Art Research Scholarship
- **2008** - Certificate of achievement for achieving top mark for Project and Operations Management in the Academic Development Bachelor of Business Science Class
- **2007** - Certificate of achievement for third overall position in the 3rd Year Academic Development Bachelor of Business Science Class
- **2005** - Elected Head Prefect for 2005 at St Joseph's Marist College

Publications

Journals

2015

[1] Jian Wu, Kyle Williams, Hung-Hsuan Chen, Madian Khabsa, Cornelia Caragea, Suppawong Taurob, Alexander Ororbia, Douglas Jordan, Prasenjit Mitra, C. Lee Giles. 2015. CiteSeerX: AI in a Digital Library Search Engin. In: *Artificial Intelligence Magazine (AI Magazine) 36(3)*, pages 35-48.

Conferences/Workshops

2016

[2] Kyle Williams, Julia Kiseleva, Aidan C. Crook, Imed Zitouni, Ahmed Hassan Awadallah, Madian Khabsa. 2016. Detecting Good Abandonment in Mobile Search. To appear in: *Proceedings of the 2016 International World Wide Web Conference (WWW '16)*.

[3] Kyle Williams, Julia Kiseleva, Aidan C. Crook, Imed Zitouni, Ahmed Hassan Awadallah, Madian Khabsa. 2016. Is This Your Final Answer? Evaluating the Effect of Answers on Good Abandonment in Mobile Search. To appear in: *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '16)*.

[4] Kyle Williams, C. Lee Giles. 2016. Improving Similar Document Retrieval Using a Recursive Pseudo Relevance Feedback Strategy. To appear in: *Proceedings of the 2016 International Joint Conference on Digital Libraries (JCDL '16)*.

[5] Julia Kiseleva, Kyle Williams, Jiepu Jiang, Ahmed Hassan Awadallah, Imed Zitouni, Aidan C. Crook, Tasos Anastasakos. 2016. Predicting User Satisfaction with Intelligent Assistants. In: *Proceedings of the International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '16)*, pages 495-505.

[6] Kyle Williams, Jian Wu, Zhaohui Wu, C. Lee Giles. 2016. Information Extraction for Scholarly Digital Libraries. To appear in: *Proceedings of the 2016 International Joint Conference on Digital Libraries (JCDL '16)* **(Tutorial)**.

[7] Julia Kiseleva, Kyle Williams, Jiepu Jiang, Ahmed Hassan Awadallah, Imed Zitouni, Aidan C. Crook, Tasos Anastasakos. 2016. Understanding User Satisfaction with Intelligent Assistants. In: *ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '16)*, pages 121-130.

[8] Chen Liang, Shuting Wang, Zhaohui Wu, Kyle Williams, Bart Pursel, Benjamin Brautigam, Sherwyn Saul, Hannah Williams, Kyle Bowen, C. Lee Giles. 2016. BBookX: Building Online Open Books for Personalized Learning. In: *Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI '16).*

2015

[9] Kyle Williams, C. Lee Giles. 2015. On the Use of Similarity Search to Detect Fake Scientific Papers. In: *Proceedings of the 2015 International Conference on Similarity Search and Applications (SISAP '15)*, pages 1-7.

[10] Chen Liang, Shuting Wang, Zhaohui Wu, Kyle Williams, Bart Pursel, C. Lee Giles. 2015. BBookX: An Automatic Book Creation Framework. In: *Proceedings of the 2014 ACM Symposium on Document Engineering (DocEng '15)*, pages 121-123.

[11] Alexander Ororbia, Jian Wu, Madian Khabsa, Kyle Williams, C. Lee Giles. 2015. Big Scholarly Data in CiteSeerX: Information Extraction from the Web. In: *BigScholar, The Second WWW Workshop on Big Scholarly Data*, pages 597-602.

[12] Jian Wu, Jason Killian, Huaiyu Yang, Kyle Williams, Sagnik Ray Choudhury, Suppawong Taurob, C. Lee Giles. 2015. PDFMEF: A Multi-Entity Knowledge Extraction Framework for Scholarly Documents and Semantic Search. In: *Proceedings of the 8th International Conference on Knowledge Capture (K-Cap '15)* **(Best Paper Nomination)**.

[13] Shuting Wang, Chen Liang, Zhaohui Wu, Kyle Williams, Bart Pursel, C. Lee Giles. 2015. Concept Hierarchy Extraction from Textbooks. In: *Proceedings of the 2014 ACM Symposium on Document Engineering (DocEng '15).*

2014

[14] Kyle Williams, Jian Wu, C. Lee Giles. 2014. SimSeerX: A Similar Document Search Engine. In: *Proceedings of the 2014 ACM Symposium on Document Engineering (DocEng '14)*, pages 143-146.

[15] Kyle Williams, Hung-Hsuan Chen, C. Lee Giles. 2014. Classifying and Ranking Search Engine Results as Potential Sources of Plagiarism. In: *Proceedings of the 2014 ACM Symposium on Document Engineering (DocEng '14)*, pages 97-106.

[16] Kyle Williams, Lichi Li, Madian Khabsa, Jian Wu, Patrick C. Shih, C. Lee Giles. 2014. A Web Service for Scholarly Big Data Information Extraction. In: *21st IEEE International Conference on Web Services.*

[17] Kyle Williams, Jian Wu, Sagnik Ray Choudhury, Madian Khabsa, C. Lee Giles. 2014. Scholarly Big Data Information Extraction and Integration in the CiteSeerX Digital Library. In: *10th International Workshop on Information Integration on the Web*, pages 68-73.

[18] Jian Wu, Kyle Williams, Hung-Hsuan Chen, Madian Khabsa, Cornelia Caragea, Alexander Ororbia, Douglas Jordan, C. Lee Giles. 2014. CiteSeerX: AI in a Digital Library Search Engine. In: *Twenty sixth Annual Conference on Innovative Applications of Artificial Intelligence*, pages 2930-2937.

[19] Kyle Williams, Hung-Hsuan Chen, C. Lee Giles. 2014. Supervised Ranking for Plagiarism Source Retrieval - Notebook for PAN at CLEF 2014. In: *CLEF 2014 Evaluation Labs and Workshop Working Notes Papers* **(Highest F1-score in Source Retrieval task at PAN 2014)**.

[20] Jian Wu, Pradeep Teregowda, Kyle Williams, Madian Khabsa, Douglas Jordan, Eric Tree, Zhaohui Wu, C. Lee Giles. 2014. Migrating a Digital Library to a Private Cloud. In: *IEEE International Conference on Cloud Engineering.*

[21] Cornelia Caragea, Jian Wu, Kyle Williams, Sujatha G. Das, Madian Khabsa, Pradeep Teregowda, C. Lee Giles. 2014. Automatic Identification of Research Articles from Crawled Documents. In: *Web-Scale Classification: Classifying Big Data from the Web (Workshop at WSDM 2014).*

[22] Jian Wu, Alexander Ororbia, Kyle Williams, Madian Khabsa, Zhaohui Wu, C. Lee Giles. 2014. Utility-Based Control Feedback in a Digital Library Search Engine: Cases in CiteSeerX. In: *9th International Workshop on Feedback Computing.*

[23] Jian Wu, Kyle Williams, Madian Khabsa, C. Lee Giles. 2014. The Impact of User Corrections To Crawl-Based Digital Libraries: A CiteSeerX Perspective. In: *10th IEEE International Conference on Collaborative Computing.*

[24] Zhaohui Wu, Jian Wu, Madian Khabsa, Kyle Williams, Hung-Hsuan Chen, Wenyi Huang, Suppawong Taurob, Sagnik Ray Choudhury, Alexander Ororbia, Prasenjit Mitra, C. Lee Giles. 2014. Towards Building a Scholarly Big Data Platform: Challenges, Lessons and Opportunities. In: *International Conference on Digital Libraries.*

[25] Cornelia Caragea, Jian Wu, Alina Ciobanu, Kyle Williams, Juan Fernandez-Ramirez, Hung-Hsuan Chen, Zhaohui Wu, C. Lee Giles. 2014. CiteSeerX: A Scholarly Big Dataset. In: *36th European Conference on Information Retrieval*, pages 311-322.

2013

[26] Kyle Williams, C. Lee Giles. 2013. Near Duplicate Detection in an Academic Digital Library. In: *Proceedings of the 2013 ACM Symposium on Document Engineering (DocEng '13)*, pages 91-94, ACM, New York, NY, USA.

[27] Kyle Williams, Hung-Hsuan Chen, Sagnik Ray Choudhury, C. Lee Giles. 2013. Unsupervised Ranking for Plagiarism Source Retrieval - Notebook for PAN at CLEF 2013. In: *CLEF 2013 Evaluation Labs and Workshop Working Notes Papers* **(Highest F1-score in Source Retrieval task at PAN 2013)**.

[28] Kyle Williams, Jorgina Paihama, Hussein Suleman. 2013. A Comparison of Machine Learning Techniques for Handwritten —Xam Word Recognition. In: *Proceedings of the South African Institute for Computer Scientists and Information Technologists Conference (SAICSIT '13)*, pages 37-46, ACM, New York, NY, USA **(Best Paper Award)**.

2012

[29] Jorgina Paihama, Kyle Williams, Hussein Suleman. 2012. Assessing the Design of Web Interoperability Protocols. In: *Proceedings of the South African Institute for Computer Scientists and Information Technologists Conference (SAICSIT '12)*, pages 353-362, ACM, New York, NY, USA.

[30] Marius Nel, Kyle Williams, Hussein Suleman. 2012. Simple Large Image Support in DSpace. In: *Proceedings of 14th International Conference on Asia-Pacific Digital Libraries (ICADL '12), Volume 7634 of Lecture Notes in Computer Science*, pages 140-143, Springer Berlin / Heidelberg.

[31] Tresor Mvumbi, Flora Kundaeli, Zafika Manzi, Kyle Williams, Hussein Suleman. 2012. An Online Meeting Tool for Low Bandwidth Environments. In: *Proceedings of the South African Institute for Computer Scientists and Information Technologists Conference (SAICSIT '12)*, pages 226-235, ACM, New York, NY, USA.

[32] Lighton Phiri, Kyle Williams, Miles Robinson, Stuart Hammar, Hussein Suleman. 2012. Bonolo: A General Digital Library System for File-based Collections. In: *Proceedings of 14th International Conference on Asia-Pacific Digital Libraries (ICADL '12), Volume 7634 of Lecture Notes in Computer Science*, pages 49-58, Springer Berlin / Heidelberg.

[33] Marwan Nour, Kyle Williams, Hussein Suleman. 2012. ORchiD: Evaluating Simple Repository Deposit for Open Educational Resources. In: *Proceedings of 14th International Conference on Asia-Pacific Digital Libraries (ICADL '12), Volume 7634 of Lecture Notes in Computer Science*, pages 289-298, Springer Berlin / Heidelberg.

[34] Michelle Havenga, Kyle Williams, Hussein Suleman. 2012. Motivating Users to Build Heritage Collections Using Games on Social Networks. In: *Proceedings of 14th International Conference on Asia-Pacific Digital Libraries (ICADL '12), Volume 7634 of Lecture Notes in Computer Science*, pages 279-288, Springer Berlin / Heidelberg.

2011

[35] Kyle Williams, Hussein Suleman. 2011. Creating a Handwriting Recognition Corpus for Bushman Languages. In: *Proceedings of 13th International Conference on Asia-Pacific Digital Libraries (ICADL '12), Volume 7008 of Lecture Notes in Computer Science*, pages 222-231, Springer Berlin / Heidelberg **(Honorable Mention)**.

[36] Kyle Williams, Hussein Suleman. 2011. Using a Hidden Markov model to Transcribe Handwritten Bushman Texts. In: *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries (JCDL '11)*, pages 445-446, ACM, New York, NY, USA.

2010

[37] Rizmari Versfeldi, Spencer lee, Edward A. Fox, Hussein Suleman, Kyle Williams. 2010. Digital Library in a 3D Virtual World: The Digital Bleek and Lloyd Collection in Second Life. In: *Proceedings of the 14th European Conference on Research and Advanced Technology for Digital Libraries (ECDL'10), Volume 6273 of Lecture Notes in Computer Science*, pages 550-553, Springer Berlin / Heidelberg.

[38] Kyle Williams, Sanvir Manilal, Lebogang Molwantoa, Hussein Suleman. 2010. A Visual Dictionary for an Extinct Language. In: *Proceedings of 12th International Conference on Asia-Pacific Digital Libraries (ICADL '10), Volume 6102 of Lecture Notes in Computer Science*, pages 1-4, Springer Berlin / Heidelberg.

[39] Kyle Williams, Hussein Suleman. 2010. Translating handwritten bushman texts. In: *Proceedings of the 10th Annual International ACM/IEEE Joint Conference on Digital Libraries (JCDL '10)*, pages 109-118, ACM, New York, NY, USA.

Technical Reports

[40] Christopher Parker, Kyle Williams, Hussein Suleman. 2012. A Lightweight Interface to Local Grid Scheduling Systems. In: *Technical Report CS12-05-00, Department of Computer Science, University of Cape Town.*

[41] Kyle Williams. 2010. Feasibility of Automatic Transcription of Neatly Rewritten Bushman Texts. In: *Technical Report CS12-06-00, Department of Computer Science, University of Cape Town.*

Other Academic Output

[42] Kyle Williams, C. Lee Giles. 2015. Classifying Search Engine Results as Potential Sources of Plagiarism. In: *The Pennsylvania State University Annual Graduate Exhibition.*

[43] Kyle Williams, C. Lee Giles. 2014. Using Documents to Search for Documents. In: *The Pennsylvania State University Annual Graduate Exhibition* **(Third place winner in the Engineering category)**.

[44] Kyle Williams, C. Lee Giles. 2013. Automatic Document Collection Management: The Case of Duplicates. In: *The Pennsylvania State University Annual Graduate Exhibition* **(Second place winner in the Engineering category)**.

[45] Kyle Williams, Hussein Suleman. 2010. Learning to Read Bushman. In: *SAICSIT 2010 Postgraduate Symposium.*

Theses

[46] Kyle Williams. 2012. Learning to Read Bushman: Automatic Handwriting Recognition for Bushman Languages. MSc Thesis. In: *Department of Computer Science, University of Cape Town.*

Professional Presentations

- The Impact of User Corrections to Crawl-based Digital Libraries, at *Collaborative Computing*, Miami, FL, USA, October 2014.
- Scholarly Big Data: Information Extraction, Data Mining and Semantics (with Madian Khabsa and C. Lee Giles), Keynote at *Semantic Analysis of Documents Workshop*, Fort Collins, CO, USA, September 2014.
- Classifying and Ranking Search Engine Results as Potential Sources of Plagiarism, at *ACM Symposium on Document Engineering*, Fort Collins, CO, USA, September 2014.
- SimSeerX: A Similar Document Search Engine, at *ACM Symposium on Document Engineering*, Fort Collins, CO, USA, September 2014.
- Scholarly Big Data Information Extraction and Integration in the CiteSeerX Digital Library, at *10th International Workshop on Information Integration on the Web*, Chicago, IL, USA, March 2014.

- Generic Similar Document Search, guest presentation for *IST 441: Information Retrieval and Search Engines*, Pennsylvania State University, February 2014.
- Unsupervised Ranking for Plagiarism Source Retrieval, at *SIGComp Seminar at Penn State IST*, Pennsylvania State University, November 2013.
- The Bleek and Lloyd Collection: Clicks, Computation and the Digital Humanities, at *Penn State African Studies Program, Brown Bag Series*, Pennsylvania State University, September 2013.
- Information and Communication Technology for Development in South Africa, guest presentation for *IST 402: Information and Communication Technologies for Development*, Pennsylvania State University, November 2012
- Learning to Read Bushman: Automatic Handwriting Recognition for Bushman Texts, at *University of Cape Town-University of Stellenbosch Industry Seminar, Centre for Broadband Networks*, University of Cape Town, July 2012
- Creating a Handwriting Recognition Corpus for Bushman Languages, at *13th International Conference on Asia-Pacific Digital Libraries*, Beijing, China, October 2011
- The Digital Bleek and Lloyd Collection, at *Courage of ||Kabbo Conference*, University of Cape Town, August 2011
- Handwriting Recognition for Bushman Languages: Clicks, Problems and Preliminary Results, at *University of Cape Town-University of Stellenbosch Industry Seminar, Centre for Broadband Networks*, University of Cape Town, 2011
- Learning to Read Bushman, at *South African Institute for Computer Scientists and Information Technologists Conference Postgraduate Symposium*, Bela-Bela, South Africa, October 2010
- Translating Handwritten Bushman Text, at *10th Annual International ACM/IEEE Joint Conference on Digital Libraries*, Gold Coast, Australia, June 2010.

Professional Activities

- **Reviewer:** IEEE Computer, South African Computer Journal (SACJ), iConference
- **Sub-Reviewer:** WWW, TPDL, ASONAM, DL, DSAA
- **Assisted with Reviewing:** SIGIR, WWW, CIKM, JCDL, ICADL, SAICSIT
- Judge in Penn State 2014 and 2015 Undergraduate Exhibitions