Daniel Campos

707 S Birch Street, Urbana, Il (415)-272-9964 —dcampos3@illinois.edu— https://spacemanidol.github.io/

EDUCATION

University of Illinois Urbana-Champaign(UIUC)-*PhD Computer Science* 06/2020- *3.88/4.0*
University of Washington-*MS Computational Linguistics* 06/2018-12/2020 *3.68/4.0*
Rensselaer Polytechnic Institute-*BS Computer Science* 08/2011-05/2015 *3.43/4.0*

Publications

2021

- Significant Improvements over the State of the Art? A Case Study of the MS MARCO Document Ranking Leaderboard - Jimmy Lin, **Daniel Campos**, Nick Craswell, Bhaskar Mitra, Emine Yilmaz
- Overview of the TREC 2020 deep learning track - Nick Craswell, Bhaskar Mitra, Emine Yilmaz, **Daniel Campos** - TREC 2020

2020

- Explorations In Curriculum Learning Methods For Training Language Models - University of Washington Computational Linguistics Master's Thesi
- ORCAS: 18 Million Clicked Query-Document Pairs for Analyzing Search - Nick Craswell, **Daniel Campos**, Bhaskar Mitra, Emine Yilmaz, Bodo Billerbeck- CIKM 2020
- XGLUE: A New Benchmark Dataset for Cross-lingual Pre-training, Understanding and Generation - Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, Xiaodong Fan, Ruofei Zhang, Rahul Agrawal, Edward Cui, Sining Wei, Taroon Bharti, Ying Qiao, Jiun-Hung Chen, Winnie Wu, Shuguang Liu, Fan Yang, **Daniel Campos**, Rangan Majumder, Ming Zhou - EMNLP 2020
- On the Reliability of Test Collections to Evaluating Systems of Different Types - Emine Yilmaz, Nick Craswell, Bhaskar Mitra and **Daniel Campos** - SIGIR
- Overview of the TREC 2019 deep learning track Nick Craswell, Bhaskar Mitra, Emine Yilmaz, **Daniel Campos**, Ellen M. Voorhees, - TREC 2019
- Leading Conversational Search by Suggesting Useful Corbin Rosset, Chenyan Xiong, Xia Song, **Daniel Campos**, Nick Craswell, Saurabh Tiwary and Paul Bennett - WWW 2020

2019

- Open Domain Web Keyphrase Extraction Beyond Language Modeling - Lee Xiong, Chuan Hu, Chenyan Xiong, **Daniel Campos**, Arnold Overwijk and Xiayu Huang - EMNLP 2019
- Overview of the TREC 2019 deep learning track Nick Craswell, Bhaskar Mitra, Emine Yilmaz, **Daniel Campos**, Ellen M. Voorhees, - TREC 2019

2018

- MS MARCO: A Human Generated MAchine Reading COmprehension Dataset Payal Bajaj, ¡strong¿Daniel Campos¡/strong¿, Nick Craswell, Li Deng, Jianfeng

Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, Tong Wang

Awards and Research Activities

Fellowships

- Gene Golub FELLOWSHIP at UIUC - 2020-2021
- Summer Predoctoral Institute Fellow at UIUC - 2020

Patents

- Using a Multi-Task-Trained Neural Network to Guide Interaction with a Query-Processing System via Useful Suggestions- 408364-US-NP - Filed 4/16/2020
- Keyphrase Extraction Beyond Language Modeling - U.S. Appln. No. 16/460,853 - Filed July 2nd, 2019

Academic Activity

- NIST TREC 2021 Deep Learning-Track Coordinator
- NIST TREC 2020 Deep Learning-Track Coordinator
- NIST TREC 2019 Deep Learning-Track Coordinator
- ACM SIGIR/SIGKDD Africa Summer School on Machine Learning for Data Mining and Search 2019-Invited Lecturer for Deep Learning in Search
- AACM SIGIR/SIGKDD Africa Summer School on Machine Learning for Data Mining and Search 2020-Invited Lecturer for Deep Learning in Search
- LatinX in AI Research at Neurips-Reviewer
- CIKM 2020 Reviewer
- CoLing 2020 Reviewer

WORK EXPERIENCE

AI Research - Neural Magic 011/2020-Present

- Researching how to best prune language models like BERT and building the sparseml library family.

Research Assistant- UIUC 06/2020-Present

- Researching the cross point of information extraction and information retrieval and applying it to Molecular synthesis and farming as part of Molecule Maker Lab Institute under Prof. Heng Ji.

Senior Product Manager-Microsoft Research & AI, Bing 11/2017-10/2020

- Architected relevance metrics stack into a real time streaming system through new human labeling tasks, novel data sampling methods, and metrics aggregation. Labeling scaled to $7m/year across 16 markets generating more than 1,000,000 judgments a week. The pipeline went from measuring user experience years old to seconds old. To scale the pipeline created a novel multi class-stratified sampling method using the Horvitz-Thompson estimator, work under review.
- Built family of MSMARCO Datasets and baselines including QnA, Passage Ranking, and Keyphrase Extraction, which collectively have been used by over 3000 Researchers. Papers Cited over 250 times and competitive datasets with over 200 research submissions. Dataset creation using Pandas and baselines models built in Pytorch and Tensorflow.
- Research and analyzed search relevance and user experience for the executive team across various verticals, languages, and metrics using language embeddings,

sampling methodologies, and classification methods. Analysis and visualization done through tSNE, Pandas, Numpy, ANNOY, and Plotly

Product Manager 2-Microsoft Cloud + AI, Global Services 05/2016-11/2017

- Designed, built, and deployed Neural Machine Translation(NMT) into a software localization pipeline all Cloud and Enterprise Microsoft products. Using the NMT models, supplier contracts were renegotiated, saving 20% $3m/year over the next three years.
- Optimized and scaled the NLP system for customer feedback translation, filtering, and categorization, decreasing team effort by 80% using NLTK, Tensorflow, and sci-kit-learn.
- Drove end to end localization and internationalization for Azure and Visual Studio products, including automation, bug triage, budgeting, vendor, and release management for $4m/year worth of translations.

Product Manager-Microsoft, Azure RemoteApp Summer 2014 & 08/2015-05/2015

- Designed scalable multi-cloud network architecture for cloud-based Remote Desktop Service and drove large scale deployment to alpha customers and Microsoft.
- Created and scaled a predictive user demand modeling pipeline which, optimized VM usage by 20%.

NLP Researcher-Basis Technology 01/2014-05/2014

- Grew the Rosette Entity Resolution pipeline to decease document analysis time, increase accuracy, and increase throughput while expanding language support to Spanish, Russian, French, Farsi, and Chinese.
- Re-designed the systems machine-learned architecture to state of the art neural methods, which improved multilingual accuracy to English parity.
- Formalized data creation, data ingestion, retrieval, formatting, and analysis pipeline using Bash, Java, and Python primarily. Optimized models were ANN, KNN, SVMs, LSTMs, and Decision Trees.

Software Engineer-Cisco Systems Summer 2013

- Developed JavaScript tooling for collecting user feedback, bug tracking, and HTML5 based screenshot used with alpha customers during initial product launches.
- Led a team of 12 other interns to develop a novel food ordering system for cafeterias that used containerized Node.js and MongoDB applications allowing efficient scaling.

Co-Founder/Software Engineer- Gapelia 08/2013-05/2014

- Built an end to end digital multi modal publishing platform which allowed non-technical users to create and publish beautiful digital magazines in a matter of minutes.
- Created REST API in Java, front end in Express, and deployed in AWS with Apache Tomcat, Maven, and MongoDB.
- Won the RPI 2014 Business Model Competition 1st place, 2014 Why not change the world grant, Harvard iLab cultural entrepreneurship challenge, and was member of Harvard's Innovation Lab 01/2014-05/2014.

Software Engineer-GE Capital Summer 2012

- Developed a reusable web system in Boostrap, Javascript and PHP which allowed GE employees to create websites that matched corporate guidelines around colors, fonts, icons, etc.