Anca Dumitrache
ℹ: http://ancad.ro | ✉: anca.dmtrch@gmail.com | ☎: +31-614-622-712

INTERESTS

Data Science, Natural Language Processing, Human – Computer Interaction, Machine Learning, Crowdsourcing, Linked Open Data.

PROFESSIONAL EXPERIENCE

Talpa Network, Amsterdam, Netherlands **02/2020 to present**

Data scientist

Developed recommender systems for radio and TV programs, using collaborative filtering and matrix factorization. Designed an offline evaluation to select the best model architecture and hyperparameters. Implemented in PySpark, with a data pipeline in Amazon Web Services.

FD Mediagroep, Amsterdam, Netherlands **01/2019 to 12/2019**

Data scientist

Developed a content-based recommender system for news in the financial domain, using gradient boosting decision trees with features from the user profile and article representation. Optimized the system in production with alpha and A/B tests. Implemented in Python, using a data pipeline in Amazon Web Services. Resulting publication:

- Lu, Dumitrache, Graus: *Beyond Optimizing for Clicks: Incorporating Editorial Values in News Recommendation*. UMAP 2020.

Center for Advanced Studies (CAS), IBM, Amsterdam, Netherlands **02/2013 to 11/2018**

Research scientist

Implemented the CrowdTruth project (collecting gold standard data for the training, evaluation of machine learning systems) in the context of IBM Watson powered business solutions. Trained and evaluated question answering models from the IBM Bluemix stack for applications in the medical field, cultural heritage and open domain. Part-time position. Resulting publications:

- Dumitrache, Aroyo, Welty: *A Crowdsourced Frame Disambiguation Corpus with Ambiguity*. NAACL 2019.
- Dumitrache et al.: *Empirical Methodology for Crowdsourcing Ground Truth*. Semantic Web Journal 2019.

Google AI, New York, USA **06/2016 to 09/2016**

Software engineering intern

Developed a model for relation classification from sentences in the open domain, using a convolutional neural network with word embeddings as features. Implemented in Python with Tensorflow. Resulting publications:

- Dumitrache, Aroyo, Welty: *Crowdsourcing Semantic Label Propagation in Relation Classification*. FeVeR Workshop at EMNLP 2018.
- Dumitrache, Aroyo, Welty: *False Positive and Cross-relation Signals in Distant Supervision*. AKBC Workshop at NeurIPS 2017.

Watson Research Group, IBM, New York, USA **01/2014 to 06/2014**

Research intern

Trained and evaluated a model for relation extraction from sentences in the medical domain, showing that models trained with crowdsourcing annotations perform as well as those trained with expert annotations. Implemented in Java. Resulting publications:

- Dumitrache, Aroyo, Welty: *Crowdsourcing Ground Truth for Medical Relation Extraction*. ACM TiiS, 8(2), 12.
- Dumitrache, Aroyo, Welty: *CrowdTruth Measures for Language Ambiguity*. LD4IE Workshop at ISWC 2015. Best paper award.

EDUCATION

Vrije Universiteit Amsterdam, Netherlands **11/2013 to 11/2018**

PhD, User-Centric Data Science research group

Thesis: Truth in Disagreement - Crowdsourcing Labeled Data for Natural Language Processing

Vrije Universiteit Amsterdam, Netherlands **09/2011 to 08/2013**

MSc, Artificial Intelligence (cum laude)

Thesis: Combining Gamification Techniques and Crowdsourcing to Create a Gold Standard for Medical Text

Jacobs University Bremen, Germany **09/2008 to 06/2011**

BSc, Computer Science

TECHNICAL SKILLS

Programming languages:
- Good: Python (PySpark, Tensorflow, Xgboost, SpaCy, Gensim, NLTK), R (tidyverse)
- Moderate: Java, PHP, Clojure, C/C++

Amazon Web Services: SageMaker, EMR, ECR, Lambda

Software: Jupyter, RStudio, Git, LaTeX, Gephi

Databases: MySQL, DynamoDB

WORKSHOPS & INVITED TALKS

Tutorial on *Crowdsourcing Inclusivity with CrowdTruth* at the Web Conference (WebConf). San Francisco, CA, USA. May, 2019.

Tutorial on the *CrowdTruth Methodology for Crowdsourcing Ground Truth* at the International Semantic Web Conference (ISWC). Monterey, CA, USA. October, 2018.

Organized *Subjectivity, Ambiguity and Disagreement (SAD) Workshop* at the Human Computation (HCOMP) conference. Zürich, Switzerland. July 2018.

Keynote talk on *Harnessing the Diversity in Human Annotation with CrowdTruth* at the ESSENCE Network *Conference on Computational Approaches to Diversity in Interaction and Meaning*. Venice, Italy. October 2017.

Talk on *Watson & Natural Language Processing* at the *Watson Experience MeetUp*. Utrecht, Netherlands. March 2016.

AWARDS & HONORS

IBM PhD Fellowship awards program, IBM, Netherlands, 2013 – 2016.

Best Poster in the Human and the Machine track, *2nd Prize* for best poster of the conference, ICT Open, Netherlands, 2016.

VU Fellowship Programme (VUFP) scholarship for academic excellence, Vrije Universiteit Amsterdam, Netherlands, 2011 – 2013.

President's List distinction for academic excellence, Jacobs University Bremen, Germany, 2010 – 2011 and 2008 – 2009.

SELECTED COURSES

Transylvanian Machine Learning Summer School (TMLSS) on machine & reinforcement learning. Cluj Napoca, Romania. July 2018.

Deep Learn summer school on deep learning. Bilbao, Spain. July 2017.

CBS Data Camp & Advanced Course on Managing Big Data and working with Spark. Enschede, Netherlands. December 2016.

Lisbon Machine Learning Summer School (LxMLS) on natural language processing. Lisbon, Portugal. July 2015.

Data Science in Society summer school. Southampton, UK. July 2014.

LANGUAGES

Romanian: native

English: proficient (C2)

French: intermediate (B2)

Dutch: basic (A2)