David Soares Batista

Personal Data

Place and Date of Birth:	Lisboa, January 4th, 1981
Phone:	+351 93 178 31 64 (mobile)
email:	dsbatista@inesc-id.pt
Homepage:	http://davidsbatista.github.io

Summary

Experienced in Natural Language Processing, Machine Learning and Information Retrieval. I like to explore and extract knowledge from large datasets, either structured or unstructured.

Currently, I am a researcher at INESC-ID and a Ph.D. candidate at Instituto Superior Técnico, the engineering school from the University of Lisbon. My Ph.D. proposal relates to extracting relationships from textual documents with an on-line classifier based on min-hash and locality sensitive hashing, while generating initial training data for this classifier by a bootstrapping approach relying on distributional word representations (i.e. neural word embeddings).

Research Experience

June 2011 – Present	Researcher @ INESC-ID (www.inesc-id.pt) Within the context of my Ph.D. thesis and research, I developed techniques to analyse large collections of documents: • Named Entity Recognition and Disambiguation • Analysis of Network from Entities Co-Occurrences • Extract Semantic Relationships • Topic Modelling
Sep. 2008 – May 2011	Researcher @ LaSIGE (www.lasige.di.fc.ul.pt) Developed an information extraction system to generate geographic summaries. Each summary lists geographic entities found in a document and linked (i.e., disambiguated) to a geographic ontology. The system was applied to a crawl of the Portuguese Web using a Hadoop cluster.

Education

2011 - Sept. 2015 (expected)	Doctor of Philosphy (Ph.D.) candidate in Informatics, **Instituto Superior Técnico**, University Lisbon. Thesis: "Large-Scale Semantic Relationship Extraction"
2007- 2009	Master's Degree (M.Sc.) in Information Extraction, **Faculty of Sciences**, University of Lisbon. Thesis: "Geographic Text Mining"
2003-2007	Bachelor of Science (B.Sc.), Informatics Engineering, **Faculty of Sciences**, University Lisbon.
	2005-2006, **Karlsruhe Universität (TH)**, Germany Two semesters abroad as an exchange student.

Professional Experience

Oct. 2007 – Jul. 2008	Software Developer @ Nokia Siemens Networks Responsible for different modules within a mobile network monitoring system. Tasks included the implementation of new features and code maintenance. Technologies: Java, CORBA, Oracle
Nov. 2004 - Jul. 2005	IT support @ Informatics Centre, University of Lisbon • Linux network administration • Troubleshooting network problems • Technical support regarding software and hardware problems
Nov. 2003 - Mar. 2004	Sysadmin @ Faculty of Sciences, University of Lisbon • Administration of a mixed Linux/Windows environment • Maintenance of services: HTTP, SMTP/IMAP, DNS, LDAP, IPTables

Languages

Portuguese:	*Native Speaker*		
English:	Spoken: *Fluent*	Written: *Fluent*	
German:	Spoken: *Good*	Written: *Fair*	Certification: *Goethe-Zertifikat B.1*

Computer Skills

Programming Languages:	Java, Python, Unix Shell Scripting
Databases/NoSQL:	MySQL, PostgreSQL, REDIS
Full-Text Index/Retrieval:	Apache Lucene/Solr
Machine Learning Libraries:	Python Scikit-Learn, Gensim, SVM^{light}
Natural Language Processing Libraries:	Python NLTK, Stanford CoreNLP, LingPipe
Distributed Computing:	Apache Hadoop

Selected Publications

Semi-Supervised Bootstrapping of Relationship Extractors with Distributional Semantics. David S Batista, Bruno Martins, and Mário J Silva. In *Empirical Methods in Natural Language Processing-EMNLP'15*. ACL, 2015. (Honorable Mention for Best Short Paper)

A Minwise Hashing Method for Addressing Relationship Extraction from Text. David S Batista, Rui Silva, Bruno Martins, and Mário J Silva. In *Web Information Systems Engineering-WISE'13*. Springer, 2013.

Exploring DBpedia and Wikipedia for Portuguese Semantic Relationship Extraction David Soares Batista, David Forte, Rui Silva, Bruno Martins, and Mário J. Silva. *Linguamática,* 5(1), 2013. Linguamática, 2013

Toponym Disambiguation using Ontology-based Semantic Similarity David S Batista, João D Ferreira, Francisco M Couto, and Mário J Silva. In *Computational Processing of the Portuguese Language 2012*. Springer, 2012

A Statistical Study of the WPT05 Crawl of the Portuguese Web David Batista and Mário J. Silva. In *In FALA 2010 VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop* Universidade de Vigo, 2010.

Geographic Signatures for Semantic Retrieval David S Batista, Mário J Silva, Francisco M Couto, and Bibek Behera. In *Proceedings of the 6th Workshop on Geographic Information Retrieval* ACM, 2010.