David Soares Batista

Personal Data

Place and Date of Birth:	Lisbon, January 4th, 1981
Phone:	+49 (0)1523 850 98 89 (mobile)
email:	dsbatista@inesc-id.pt
Homepage:	http://davidsbatista.github.io

Summary

I like to explore and extract knowledge from large volumes of data, be it structured or unstructured, numbers or text. I'm experienced in Machine Learning, Natural Language Processing and Big-Data.

In the past I've worked in projects, tackling problems with a strong text-mining and text-analysis components, involving tasks such as information extraction, classification, clustering and information retrieval.

In 2016, I successfully defended my Ph.D., proposing new methods to perform semantic relationship extraction from large collections of documents.

Professional Experience

Jan. 2016 – present | Data Engineer @ HelloFresh AG (http://www.hellofresh.de)
- Text analysis of customer NPS comments with NLP and ML.
- Data Warehousing and processes automation for data analysis.
- Design, build and maintain ETLs and infrastructure for batch processing.
- Technologies: Python NLTK, scikit-learn, Spark, Hive, Impala, Airflow.

Jun. 2011 – Apr. 2014 | Researcher and Developer @ INESC-ID (http://www.inesc-id.pt)
Project: REACTION - Computational Journalism
- Named Entity Recognition and Disambiguation over news articles.
- Semantic Relationship Extraction between named-entities.
- Topic Modeling (i.e., LDA) techniques applied to on-line news.
- Analysis of large graphs connecting entities and topics.

Oct. 2009 – Oct. 2011 | Researcher and Developer @ LaSIGE (http://www.lasige.di.fc.ul.pt)
Project: GREASE - Geographic Reasoning for Search Engines
- Language identification on a web-crawl using n-grams models.
- Disambiguation of geographic ambiguous names in Portuguese text.
- Alignment between geo-ontologies by linking geographical references.

Oct. 2007 – Jul. 2008 | Software Developer @ Nokia Siemens Networks
- Development of data collection modules for a GSM monitoring system.
- Technologies: Java, CORBA Architecture, Oracle RDBMS

Computer Skills

Programming:	Python, Java, Shell Scripting
Machine Learning:	scikit-learn, gensim, MinorThird
Natural Language Processing:	Python NLTK, Stanford CoreNLP, LingPipe
Distributed Computing:	Hadoop, Spark, Hive, Impala
Information Retrieval:	Apache Lucene/Solr
Databases/NoSQL:	MySQL, PostgreSQL, REDIS

Education

2011-2015	Doctor of Philosphy (Ph.D.) **Instituto Superior Técnico**, University Lisbon. **"Large-Scale Semantic Relationship Extraction"** To achieve scalable relationship extraction, I proposed using an on-line classifier, based on the idea of nearest neighbor classification, and leveraging min-hash and locality sensitive hashing for efficient similarity search. To obtain training data, for the classifier, I proposed a bootstrapping technique relying on distributional word representations.
2007-2009	Master's Degree (M.Sc.) - **Faculty of Sciences**, University of Lisbon. **"Geographic Text Mining"** I developed an information extraction system based on Conditional Random Fields to generate geographic summaries. The summary lists the geographic entities found in a document and mapped into geographic concepts in a geographic ontology. The system was applied to a crawl of the Portuguese Web (25GB raw text) using a Hadoop cluster.
2003-2007	Bachelor of Science (B.Sc.), Informatics Engineering, **Faculty of Sciences**, University Lisbon. 2005-2006, **Karlsruhe Universität (TH)**, Germany • Erasmus exchange student for two semesters 2004 November - 2005 July - IT support (part-time) @ University of Lisbon • Troubleshooting network connections and services, preventative maintenance, helping and educating new users. 2003 November - 2004 March - SysAdmin (part-time) @ University of Lisbon • Administration and configuration of networking software and services (e.g., samba, backups, quotas, mail, web, crontab, IPTables).

Languages

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

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.* - **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.*

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.

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.*

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.