

David Soares Batista

PERSONAL DATA

PLACE OF BIRTH:	Lisbon, Portugal	HOME PAGE:	www.davidsbatista.net
DATE OF BIRTH:	January 4th, 1981	LINKEDIN	www.linkedin.com/in/dsbatista
EMAIL:	dsbatista@gmail.com	GITHUB	www.github.com/davidsbatista
		PUBLICATIONS	www.goo.gl/uihrcx

SUMMARY

I like to explore and extract knowledge from large volumes of data, and transform natural language into structured data. I'm experienced in Machine Learning (ML), Natural Language Processing (NLP) 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

AUG. 2017 – PRESENT	NLP Researcher/Engineer @ COMTRAVO Enhancement of the NLP pipeline for processing incoming booking requests: <ul style="list-style-type: none">• Supervised sequence prediction for fine grained Named-Entity Recognition.• Named-Entity evaluation metrics at a full entity level instead of token level.• Slot Filling to extract pre-defined attributes from incoming booking requests.• Relationship-Extraction between named-entities• Participate in hiring processes, evaluating and interviewing candidates.• Technologies: spaCy, Keras, scikit-learn, Docker, AWS, Dask.
JAN. 2016 – JUN. 2017	Data Engineer @ HELLOFRESH <ul style="list-style-type: none">• Text analysis and classification of customer reviews with NLP and ML.• Modelling Data Warehousing Star Schema: dimensions, fact tables.• Design, build and maintain ETLs and the infrastructure for batch processing.• Designed a supervised model for market attribution.• Technologies: Python NLTK, scikit-learn, Spark, Impala, Airflow, AWS.
JUN. 2011 – APR. 2014	Researcher and Developer @ INESC-ID Project name: REACTION - Computational Journalism Project description: http://dmir.inesc-id.pt/project/Reaction <ul style="list-style-type: none">• Named-Entity Linking over news articles to Wikipedia.• Semantic Relationship extraction between named-entities.• Topic Modeling (i.e., LDA) applied to news articles.
OCT. 2009 – OCT. 2011	Researcher and Developer @ LASIGE Project name: GREASE - Geographic Reasoning for Search Engines Project description: http://xldb.di.fc.ul.pt/wiki/Grease <ul style="list-style-type: none">• Language identification of web-crawled text 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 <ul style="list-style-type: none">• Development of data collection modules for a GSM monitoring system.• Technologies: Java, CORBA Architecture, Oracle RDBMS

COMPUTER SKILLS

Programming:	Python, Java, SQL, Bash Shell Script
Machine Learning Frameworks:	scikit-learn, Keras, gensim
Natural Language Processing Libraries:	spaCy, Python NLTK, Stanford CoreNLP, LingPipe
Distributed Computing:	Apache PySpark, Hadoop, Hive, Impala
Information Retrieval:	Apache Lucene/Solr
Databases/NoSQL:	MySQL, PostgreSQL, REDIS

EDUCATION

- 2011-2015 | **Doctor of Philosophy (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.)** - Faculty of Sciences, University Lisbon.
Informatics Engineering
- 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, IMAP, Apache HTTP Server, IPTables, crontab scheduling, backups;

LANGUAGES

PORTUGUESE:	Spoken: <i>Native</i>	Written: <i>Native</i>	
ENGLISH:	Spoken: <i>Fluent</i>	Written: <i>Fluent</i>	
GERMAN:	Spoken: <i>Fluent</i>	Written: <i>Fair</i>	Certification: <i>Goethe-Zertifikat B.1</i>
SPANISH:	<i>Conversational</i>		

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.

Where in the Wikipedia is that answer? The XLDB at the GikiCLEF 2009 task. Nuno Cardoso, David Batista, Francisco J Lopez-Pellicer, and Mário J Silva. In *Multilingual Information Access Evaluation I. Text Retrieval Experiments*. Springer, 2010.