

# David Soares Batista

## PERSONAL DATA

---

PLACE AND DATE OF BIRTH: Lisbon, January 4th, 1981  
EMAIL: dsbatista@gmail.com  
HOMEPAGE: <http://www.davidsbatista.net>

## 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 (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   | <b>Data Scientist @ COMTRAVO</b> <ul style="list-style-type: none"><li>• Enhancement of the NLP pipeline for processing incoming booking requests.</li><li>• Supervised sequence prediction for fine grained Named-Entity Recognition.</li><li>• Technologies: spaCy, scikit-learn, Dask, Keras, AWS.</li></ul>  |
| JAN. 2016 – JUN. 2017 | <b>Data Engineer @ HELLOFRESH</b> <ul style="list-style-type: none"><li>• Text analysis and classification of customer reviews with NLP and ML.</li><li>• Modelling Data Warehousing Star Schema: dimensions, fact tables.</li><li>• Design, build and maintain ETLs and the infrastructure for batch processing.</li><li>• Technologies: Python NLTK, scikit-learn, Spark, Impala, Airflow, AWS.</li></ul>  |
| JUN. 2011 – APR. 2014 | <b>Researcher and Developer @ INESC-ID</b> <p>Project name: REACTION - Computational Journalism</p> <p>Project description: <a href="http://dmir.inesc-id.pt/project/Reaction">http://dmir.inesc-id.pt/project/Reaction</a></p> <ul style="list-style-type: none"><li>• Named-Entity Linking over news articles to Wikipedia.</li><li>• Semantic Relationship extraction between named-entities.</li><li>• Topic Modeling (i.e., LDA) applied to news articles.</li></ul>                                      |
| OCT. 2009 – OCT. 2011 | <b>Researcher and Developer @ LASIGE</b> <p>Project name: GREASE - Geographic Reasoning for Search Engines</p> <p>Project description: <a href="http://xldb.di.fc.ul.pt/wiki/Grease">http://xldb.di.fc.ul.pt/wiki/Grease</a></p> <ul style="list-style-type: none"><li>• Language identification of web-crawled text using n-grams models.</li><li>• Disambiguation of geographic ambiguous names in Portuguese text.</li><li>• Alignment between geo-ontologies by linking geographical references.</li></ul> |
| OCT. 2007 – JUL. 2008 | <b>Software Developer @ NOKIA SIEMENS NETWORKS</b> <ul style="list-style-type: none"><li>• Development of data collection modules for a GSM monitoring system.</li><li>• Technologies: Java, CORBA Architecture, Oracle RDBMS</li></ul>  |

## COMPUTER SKILLS

---

Programming:	Python, Java, Shell Scripting
Machine Learning Frameworks:	scikit-learn, gensim, Keras
Natural Language Processing:	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.