Personal Data

Name:	David Soares Batista	Homepage:	www.davidsbatista.net
Place of Birth:	Lisbon, Portugal, 1981	GitHub:	www.github.com/davidsbatista
email:	dsbatista@gmail.com	Publications	http://goo.gl/uihrcx

Summary

Experienced in both research and industry I enjoy working on solutions from concept to production and transform natural language text into structured data. In the past I've tackled problems with strong Machine Learning and Natural Language Processing components, involving tasks like: information extraction, classification, clustering and information retrieval. I considered myself a practical problem solver and like to deliver software, not just results.

Professional Experience

2017-08 – Present | Comtravo GmbH

2021-06 – Present **Lead NLP Engineer**
2019-08 – 2021-05 **Senior NLP Engineer**
2017-08 – 2019-07 **NLP Engineer**

- Leading the Automation team, 3 developers plus 4 annotators, working on the system that automatically answers incoming email travel requests and assists travel-agents in handling them. Coordinating tasks based on system performance and feature requests.
- Developed several modularised Python components with type-annotations, linting and test coverage above 95%.
- Trained and evaluated models for text classification and fine-grained NER, increasing the performance of identifying specific booking requests and information extraction.
- Developed algorithms to map input text to corresponding unique identifiers in a target knowledge base, e.g: airports, train stations, hotels, geographic locations.
- Defined performance and monitoring metrics in collaboration with the Data Engineering team, resulting in detailed performance reports and system monitoring dashboards.

2016-01 – 2017-06 | **Data Engineer** @ HelloFresh SE

- Built and maintained several ETLs using PySpark (Apache Spark) and Hive.
- Developed the first prototype to manage ETLs pipelines based on Airflow operators which later went into production and was used by the team.
- Built a classifier using NLTK and linear models from scikit-learn, to identify customer reviews mentions to different types of issues with the meal kits.

2011-06 – 2014-04 | **Researcher and Developer** @ INESC-ID - REACTION - Computational Journalism

- Explored and implemented methods for entity-relationship extraction, based on: hand-built patterns, supervised linear classifiers with linguistic features and semi-supervised methods based on seed relationships and large amounts of unannotated text.
- Developed a method to link personalities in news articles to Wikipedia entries based on textual similarities and Wikipedia graph structure.
- Built graph based on LDA topic models extracted from news articles and persons co-occurrences, allowing to explore which topics connected two personalities.

2009-09 – 2011-09 | **Researcher and Developer** @ LaSIGE - Geographic Reasoning for Search Engines

- Developed a method to disambiguate toponyms based on textual context, information content and topological similarity measures
- Built an alignment method between two geo-ontologies resulting in a single linked-data ontology, allowing the inclusion of features from both ontologies in SPARQL queries.

2007-09 – 2008-07 | **Software Developer** @ Nokia Siemens Networks

- Developed data collection modules for a GSM monitoring system.
- Technologies: Java, CORBA Architecture, Oracle RDBMS

Skills

Programming:	Python (`pytest, pylint, mypy`), Java, SQL, Shell Script
NLP Libraries:	spaCy, NLTK, HuggingFace Transformers, gensim
Machine Learning Frameworks:	scikit-learn, PyTorch, Keras
Information Retrieval:	ElasticSearch, Apache Lucene/Solr
Semantic Web/Linked Data:	Apache Jena, SPARQL
Infrastructure:	Docker, Flask, FastAPI, AWS: EC2, S3, Batch
Databases/Document stores:	MySQL, PostgreSQL, MongoDB

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 a classifier based on the idea of nearest neighbour 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 which was awarded an **Honorable Mention for Best Short Paper at EMNLP'15**

2007-2009 | **Master's Degree (M.Sc.)** - Faculty of Sciences, University of Lisbon.
"Geographic Text Mining"
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's in Informatics Engineering** - Faculty of Sciences, University Lisbon
- 2005-2006, Karlsruhe Universität (TH), Germany - exchange student
- 2004 November - 2005 July - IT support (part-time)
- 2003 November - 2004 March - System Administrator (part-time)

Personal Projects

Oct. 2020 – present | **Politiquices.PT - https://www.politiquices.pt**

A semantic graph allowing to explore political interactions of support and opposition based on news articles from the past 25 years. The project gain the interest of journalists, political scientists and social humanities researchers and was awarded **2nd place on the "Arquivo.pt Awards 2021"** organised by the Portuguese Web Archive research institution.

I crawled, using an API, news headlines mentioning politicians from archived newspapers websites, and then annotated a sample in order to build supervised models to detect relationships between politicians and to link them to the respective Wikidata entry. By applying the trained models to the crawled data, I was able to generated a semantic graph connecting politicians Wikidata entries through a support or opposition relationship sustained by news articles mentioning the politicians.

Then, I indexed the graph in a SPARQL endpoint, allowing to formulate queries such as: *Which politicians affiliated with party X opposed/supported politicians from party Y?*

Technologies: spaCy, Apache Jena, ElasticSearch, Neo4J, Docker

Relevant 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*.

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.

Languages

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