David Soares **Batista**

dsbatista@gmail.com | www.davidsbatista.net | davidsbatista | dsbatista

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 production ready software, not just results.

Professional Experience

Comtravo GmbH — *Berlin*

LEAD NLP ENGINEER — *Aug. 2017 - PRESENT*

- I joined the company as an NLP Engineer on August 2017, was promoted to Senior NLP Engineer on July 2019, and in June 2021 to Lead NLP Engineer.
- Leading a team of 3 developers + 4 annotators, working on the system that automatically answers incoming email travel requests and assists travel-agents in handling them. Coordinating development tasks based on system performance and feature requests.
- Developed several modularised Python components making use of type-annotations, code linting and test coverage above 95%.
- Trained, evaluated and improved different 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 into unique Knowledge Base identifiers, e.g: airports, train stations, hotels, geographic locations.
- Built airport and train stations Knowledge Bases based on open resources: Wikidata, GeoNames, DB open-data and in-house operational data.
- Defined quantifiable measures in collaboration with the Data Engineering team, resulting in performance reports and monitoring dashboards.

HelloFresh SE — *Berlin*

DATA ENGINEER — *Jan. 2016 - Jun. 2017*

- Built and maintained several ETLs using PySpark (Apache Spark) and Hive.
- Developed a 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 scikit-learn linear models to identify customer reviews mentions to different types of issues with the meal kits.

INESC-ID Research & Development Institute — *Lisbon*

RESEARCHER AND DEVELOPER: REACTION - COMPUTATIONAL JOURNALISM — *Jun. 2011 - Apr. 2014*

- 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 a graph based on LDA topic models extracted from news articles and persons co-occurrences, allowing to explore which topics connected two personalities.

LaSIGE - Department of Informatics Research Unit — *Lisbon*

RESEARCHER AND DEVELOPER: GREASE - GEOGRAPHIC REASONING FOR SEARCH ENGINES — *Sep. 2009 - Jul. 208*

- 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.

Nokia-Siemens Networks — *Lisbon*

SOFTWARE DEVELOPER — *Sep. 2007 - Sep. 2011*

- 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
Natural Language Processing Libraries	spaCy, NLTK, HuggingFace Transformers, gensim
Machine Learning Frameworks	scikit-learn, PyTorch, Keras
Information Retrieval	ElasticSearch, Kibana, 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

Doctor of Philosophy (Ph.D.) - "Large-Scale Relationship Extraction"

Lisbon

Instituto Superior Técnico, University Lisbon — *2011 - 2015*

- I researched different methods to perform semantic relationship extraction between named-entities and proposed 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*.

Master of Science (M.Sc.) - "Geographic Text Mining"

Lisbon

Faculty of Sciences, University of Lisbon — *2007 - 2009*

- Developed a geographic information extraction system based on Conditional Random Fields and a geographic ontology, to generate geographic summaries. The summary lists all the geographic entities found in a document mapped to unique concepts in the geographic ontology.
- The system was then applied to a crawl of the Portuguese Web (25GB raw text) using a Hadoop cluster, generating summaries for hundreds of thousands of documents.

Bachelor's in Informatics Engineering

Lisbon | Karlsruhe

Faculty of Sciences, University of Lisbon — *2003 - 2007*

- 2005-2006, Karlsruhe Universität (TH), Germany - exchange student
- 2004 November - 2005 July - IT support (part-time) - troubleshooting network infrastructure and preventative maintenance.
- 2003 November - 2004 March - Sys Admin (part-time) - administration and configuration of the IT department network and services.

Personal Projects

Extraction and analysis of support and opposition political relationships

Politiquices.PT - https://www.politiquices.pt — *Oct. 2020 - PRESENT*

- A tool to explore political interactions of support and opposition, based on news articles from the past 25 years, awarded the **2nd place on the "Arquivo.pt Awards 2021"** organised by the Portuguese Web Archive.
- Using an API, I crawled news headlines mentioning politicians from archived newspapers websites, annotated a sample in order to build supervised models to detect relationships between politicians and to link them to Wikidata.
- By applying the trained models to the crawled data, I was able to generated a semantic graph connecting Wikidata politicians entries through a support or opposition relationship sustained by archived news articles.
- Using a SPARQL endpoint it's possible to ask: *Which politicians affiliated with party X opposed/supported politicians from party Y?*
- The project gain the interest of journalists, political scientists and social humanities researchers.

Relevant Publications

- **Semi-Supervised Bootstrapping of Relationship Extractors with Distributional Semantics**. David S Batista, Bruno Martins, and Mário J Silva. In *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*
- **POWER - Politics Ontology for Web Entity Retrieval** Silvio Moreira, David Batista, Paula Carvalho, Francisco M Couto, and Mário J Silva. In *Advanced Information Systems Engineering Workshops. Springer, 2011*
- **A Statistical Study of the WPT05 Crawl of the Portuguese Web** David Batista and Mário J Silva. In *FALA 2010 VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop, Vigo, Spain*
- **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*