Summary.

With 10+ years of experience in research and industry, I excel in driving solutions from concept to production, particularly focusing on transforming unstructured text into structured data. My expertise lies in applying Machine Learning and Natural Language Processing techniques for tasks like information extraction, classification, clustering, and information retrieval. I have thrived in diverse environments, from academic research to freelancing and successful start-ups. Collaborative by nature, I enjoy working with dedicated individuals to achieve collective solutions. I take pride in my strong software engineering skills, delivering production-ready software solutions rather than just reporting results.

Professional Experience ____

TripActionsBerlin, Germany

LEAD NLP ENGINEER May 2022 - Feb 2023

- · Joined through the acquisition of Comtravo by TripActions, tasked with developing a chatbot's NLP/NLU component.
- Defined, together with the team, an intent and entities annotation schema and set up the infrastructure for annotation of a training corpus for intent and entity recognition.
- Fine-tuned Transformers models to perform intent detection and entity recognition, and threshold tuning for deployment in production.

Comtravo GmbH Berlin, Germany

LEAD NLP ENGINEER

Aug. 2017 - Apr. 2022

- Joined as an NLP Engineer in August 2017, was promoted to Senior NLP Engineer in July 2019, and in June 2021 to Lead NLP Engineer.
- Led a team of 3 developers + 4 annotators, working on the system that automatically answers incoming email travel requests and assists travelagents 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 performing information extraction to automatically fulfill booking requests.
- 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, Germany

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 become the ETL management platform in production.
- 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, Portugal

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 topic models extracted from news articles and person co-occurrences, allowing to explore topics connecting persons.

LaSIGE - Department of Informatics Research Unit

Lisbon, Portugal

RESEARCHER AND DEVELOPER: GREASE - GEOGRAPHIC REASONING FOR SEARCH ENGINES

Sep. 2008 - Oct. 2010

- 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, Portugal

SOFTWARE DEVELOPER Oct. 2007 - Jul. 2008

• Developed data collection modules for a GSM monitoring system using Java, CORBA Architecture and 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, EMR (PySpak)

Databases/Document stores MySQL, PostgreSQL, MongoDB

May 19, 2023 David S. Batista · Résumé



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

Lisbon, Portugal

INSTITUTO SUPERIOR TÉCNICO, UNIVERSITY LISBON

Feb. 2011 - Feb. 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, Portugal

FACULTY OF SCIENCES, UNIVERSITY OF LISBON

Sep. 2007 - Jul. 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, Portugal Karlsruhe, Germany

Sep. 2003 - Jul. 2007

FACULTY OF SCIENCES, UNIVERSITY OF LISBON

- 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

Politiquices.PT - https://www.politiquices.pt

Analysis of support and opposition political relationships

- A website which allows to explore political interactions of support and opposition based on news articles, which was awarded the 2nd place
 on the "Arquivo.pt Awards 2021" organised by the Portuguese Web Archive.
- I've built and tuned supervised models to detect relationships between politicians and link them to Wikidata, based on news headlines. The trained models were applied to archived news articles, generating an RDF semantic graph connecting politicians through a support or opposition relationship sustained by news articles.
- Using a SPARQL endpoint it's possible to issue queries like: Which politicians affiliated with party X opposed/supported politicians from party Y?
- The project gained the interest of journalists, political scientists and social humanities researchers.

BREDS - https://pypi.org/project/breds

BOOTSTRAPPING SEMANTIC RELATIONSHIPS WITH DISTRIBUTIONAL SEMANTICS

• A Python package implementation based on results from my Ph.D. thesis. BREDS is an approach to extract named-entity relationships without labelled data by relying instead on an initial set of seeds, i.e. pairs of named-entities representing relationship type to be extracted. The algorithm uses the seeds to learn extraction patterns and expands the initial set of seeds using distributional semantics to generalize the relationship while limiting the semantic drift.

nervaluate - https://pypi.org/project/nervaluate

NAMED-ENTITY RECOGNITION CONSIDERING PARTIAL MATCHING

An open-source software package to evaluate named-entity recognition systems considering partial entity matching. Originally started with
a blog post I wrote about the subject which attracted the interest of several people and converged into a Python package which is currently
maintained by myself and other contributors.

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.