UMBERTO SELVA

Linguistics PhD, NLP / AI Engineer
Python, Deep Learning, Data Science

u.o.selva@gmail.com +39 333 5218014

linkedin.com/in/umberto-selva

leidenuni.academia.edu/UmbertoSelva

github.com/umbertoselva

EXPERIENCE

- **AI Engineer**

  Pharmazie.com 2023–Today

  Pharmazie.com is a web platform that gathers data from over 25 databases to provide reliable information on pharmaceutical products to healthcare professionals.

  My task is to oversee the implementation of Artificial Intelligence and NLP projects. I work at designing LLM-powered conversational assistants (chatbots), from data ingestion, pre-processing, embedding, retrieval QA pipeline design, testing and prompt engineering, all the way to server management and production deployment.

  Skills: Python, LLMs, LangChain, OpenAI, Huggingface, Prompt Engineering, Retrieval QA, Back-end Web Development, AWS Cloud

- **NLP / AI / ML / Data Science Consultant**

  Freelance 2020–Today

  Between 2020 and 2022 I transitioned from academia to the private sector, training in NLP, AI, Machine Learning and Data Science, working on various personal projects, as well as offering freelancing services.

  Skills: Python, NLP, Machine Learning, Data Science, Statistics, LLMs, Prompt Engineering, Back-end Web/API Development, SQL, bash/zsh, Linux

- **Academic Researcher and Lecturer (Linguistics)**

  LUCL / IIAS, Leiden 2014–2020

  Between 2014 and 2020, I was research assistant, PhD researcher, post-doctoral researcher and lecturer at the Leiden University Centre for Linguistics (2014-2019) and at the International Institute for Asian Studies (2019-2020) in Leiden (NL).

  Skills: Linguistics (General, Historical, Computational), Academic Research, Academic Writing, Teaching, Research Project Management

PORTFOLIO

github.com/umbertoselva

🏛️🤖 GPT-3.5 Museum FAQ Chatbot [Web app | Github Repo]

A chatbot built with Langchain, OpenAI GPT-3.5 and a Streamlit UI that can answer questions based on the Stedelijk Museum Amsterdam's FAQ list. This project is a follow-up to my previous Museum FAQ Chatbot built with Sentence Transformers and ElasticSearch.

Skills: Python, LangChain, OpenAI, Streamlit, Prompt Engineering

📞🤖 SMS Guidelines Assistant [Web app | Github Repo]

An AI chatbot built with LangChain, OpenAI GPT-3.5 and a Streamlit UI that can assist customers with the guidelines for sending SMS to each country with a popular telecommunication service.
I devoted special attention to creating a dedicated pipeline to pre-process the original data and significantly improved the chatbot performance.

Skills: Python, LangChain, OpenAI, Streamlit, Prompt Engineering

🔪⚖️ GPT-3 Unbiased News App [Web app | Github Repo]

A Streamlit app built with OpenAI GPT-3 that rewrites news articles by removing bias and highlights logical fallacies.

Skills: Python, OpenAI, Streamlit, Prompt Engineering

🎥 Fine-tuning BERT for NER-based Sentiment Analysis [Notebooks]

A Data Science project in which I combine NER-extraction with spaCy and fine-tuning a BERT model with TensorFlow to perform sentiment analysis on a real-world movie review dataset extracted from the popular "I Just Watched" subreddit with the goal of finding out which actors are most popular among the subreddit users (spoiler: it was Tom Hanks!).

Skills: Python, TensorFlow, Keras, spaCy, Pandas, Huggingface

📊 Exploratory Data Analysis and House Price Prediction [Notebook | Repo]

A step-by-step example of how to apply Data Science to a classic regression task. By employing a number of custom Python utils, I conduct exploratory data analysis, data preprocessing, feature selection, model design, training, evaluation and model explainability.

Skills: Python, Pandas, NumPy, Sci-kit Learn, TensorFlow, Keras, Scipy.stats, Statsmodels, Matplotlib, Seaborn, Lime

🎵 The Sanskrit Meterscan [Website | Github Repo]

A digital humanities web app, built with Python's back-end framework Flask, that automates the metrical scansion of Sanskrit and Vedic texts. Used by academic researchers in the fields of Linguistics and Philology

Skills: Python, Flask, HTML, CSS, JavaScript

CERTIFICATIONS

- **Natural Language Processing Specialization**

  Deeplearning.ai January 2022 (four-month program)

  In depth theoretical introduction to probabilistic, sequence and deep learning models and techniques for classification, sentiment analysis, POS tagging, NER, autocomplete, next-word prediction, etc. Logistic regression, naive bayes, Markov models, N-grams, vector space, word embedding techniques (Word2Vec, CBOW, Glove, etc.), Neural Networks, RecurrentNNs, LSTM, Attention, Transformer models, etc.

- **NLP with Transformers in Python**

  Udemy - by J. Briggs July 2022

  Transformer (Encoder-Decoder with Attention) architecture and research history, pre-training, fine-tuning, applied tasks (text classification, sentiment analysis, named entity recognition (NER), reader-retriever QA, etc.), NLP preprocessing, metrics, etc., using Huggingface, TensorFlow, SpaCy, FAISS, etc.

- **Machine Learning**

  Stanford Online August 2020

  Andrew Ng's classic introduction to ML covering linear and logistic regression, cost functions, gradient descent, backpropagation, optimization and regularization techniques, neural networks, supervised and unsupervised learning, principal component analysis (PCA), reinforcement learning, recommender systems, data augmentation, etc.

- **Deep Learning Specialization**

  Deeplearning.ai November 2022

  Neural Networks architectures, activation functions, hyperparameter tuning, advanced regularization and optimization techniques, transfer learning, metrics, error analysis, best practices for structuring ML projects, etc.

- **Statistics with Python Specialization**

  IBM October 2022

  Descriptive statistics, data visualization, probability distributions, hypothesis testing, regression analysis, etc.

- **Practical Data Science on the AWS Cloud**

  AWS & Deeplearning.ai December 2022

  Build, train and deploy ML pipelines using BERT with Amazon SageMaker.

EDUCATION

- **PhD Linguistics**

  Leiden University, Netherlands 2019

- **Research MA Linguistics**

  Leiden University, Netherlands 2014

- **BA Classics / Linguistics**

  Università degli Studi di Torino, Italy 2011

SKILLS

- **Coding languages**

  Python, SQL (MySQL), Prompt Engineering

- **NLP / AI**

  NLTK, SpaCy, Huggingface, OpenAI, LangChain
  RNNs, LSTM, Transformers, BERT, GPT, LLMs

- **Machine Learning**

  Scikit-Learn, Tensorflow, Keras

- **Data Science**

  Pandas, NumPy, Matplotlib, Seaborn

- **Web app / API / Back-end development**

  Flask, FastAPI, Gradio, Streamlit

- **Various**

  Git, Github, Docker, Amazon SageMaker, Jupyter Notebook, Google Colab, ElasticSearch, AWS Cloud, Nginx, Linux (Ubuntu), Linux servers, bash/zsh

LANGUAGES

Native proficiency: English, Italian

Grammatical knowledge of various modern and classical/ancient languages

Last update: August 2023