Thomas Compton CV

Statement

Interdisciplinary NLP researcher with a unique foundation in social science, English, and computer science. Experienced in applying NLP methods to large-scale textual corpora, including topic modelling, OCR pipelines, and discourse analysis. Strong communicator with a proven track record in undergraduate teaching and outreach across Maths, English, and Social Policy, spanning primary to university level.

Experience

School for Business and Society Graduate Teaching Assistant & Research Assistant

May 2023 – Aug 2025 | University of York

- Organised and interviewed 10 practitioners to understand their approaches to study EDI policies.
- Created chatbot to allow staff to understand EDI policy changes, using staff survey data.
- Delivered seminars to undergraduate students receiving positive feedback in evaluations.

Many Flavours of Wellingborough CIC

Founder & Director

Sept 2023 – Sept 2024 | Wellingborough

- Founded, registered and managed Many Flavours of Wellingborough CIC, including developing safeguarding, data protection, EDI and many other policies.
- Advertised, managed volunteers of youth drama group with 11 regular students.
- Successfully secured funding from Made With Many (£750), Wellingborough Town Council (£1000), Asda (£586).

Projects

OCR ACCURACY EVALUATION FOR IMAGE-BASED TEXT EXTRACTION Project Lead

Jan 2025 – Present | University of York

- **Packages:** EasyOCR, Paddle, Tesseract, Google Gemini, Jiwer, Deepseek, Qwen
- Compare WER & CER of different OCR pipelines.
- Develop a batch approach to OCR with Gemini, avoiding rate limits and errors.
- GitHub: OCR-evaluation

BERTOPIC EVALUATION AND FINE-TUNING

Project Lead

Jan 2024 – Present | University of York

- Develop metrics for evaluating multiple BERTopic runs & exploring trade-offs.
- Develop batch approach to output the best topic model out of multiple runs.

Education

University of York, 2021 – Expected Aug 2025

- 1+3 ESRC Funded PhD in NLP approaches to large historical corpora
- Title: Historical (dis)Continuities in Community Unionism Between The Boot and Shoe Union and Unite Community

University of Birmingham, 2016 – 2020

- MSc Distinction International Relations
- BA 2:1 English and Creative Writing

Skills

Programming

Python 3+ years

Technology/Packages

Git/GitHub • AWS •
SpaCy • NLTK • FAISS • BERTopic • LDA
Google Gemini • Jiwer
PyTorch • Sentence Transformers

Open Source Projects

PDF-to-Speech Reader:GitHub

- Converts PDF documents into audio files using text-to-speech
- Skills: pypdf, pyttsx3, pydub, AWS

Hansard OCR Dataset: GitHub

- Downloading and preparing historical UK parliamentary debates (Hansard) for use in OCR training
- Skills: torch, TrOCR, BeautifulSoup, transformers

Samuel Smiles DA: GitHub

- Educational repository showing how to collect, embed, cluster, and analyse historical texts using sentence-level methods, topic modelling, and frequency analysis
- Skills: sentence-transformers, faiss-cpu, bertopic, nltk, spacy, scikit-learn, matplotlib, requests, beautifulsoup4

ngram-2-topic:GitHub

- n-gram frequency analysis with sentence-level embeddings to generate interpretable and semantically rich topics.
- Skills: sentence-transformers, spacy

Publications

Articles

Compton, T. (2025). Holistic evaluations of topic models. https://arxiv.org/abs/2507.23364

Conferences

- BUIRA Conference (accepted), 2025, *Community Unionism Reconsidered*
- BUIRA Conference (accepted), 2024, *Northampton Boot Shoe Then Now*
- BSA Work, Employment and Society Conference 2023, poster *Northampton Boot Shoe Then Now* (accepted)
- PSA Conference 2023, *Humanitarian Intervention*