Corpus Interface Plan

First Interface: Data Analysis and Keyword Search

Purpose: Enable users to analyze data and search for keywords efficiently.

Features:

- **Dataframe Display:** Includes text, likes, and other features from the original dataset.
- **Category Labeling:** Drop-down menu for labeling categories.
- **Text Output Limit:**
  - Display up to **100 tokens** per text entry for concise results.
  - Users can click **"See More"** to view full content.
- **Search Functionality:**
  - Enables users to explore the corpus and retrieve relevant results easily.
- **Visualization Options:**
  - Helps users analyze data patterns with **frequency distributions** and **counts**.
- **Export Function:**
  - Users can download or export search results in **CSV** or **JSON** format.

Second Interface: Interactive Text Chunking with BERT

Purpose: Provide an interactive environment for text chunking using BERT or other ML models.

Features:

- **User Input:** Interactive text box for user-provided text.
- **Model Output:** Predicts and displays labels for the input text chunk.
- **Real-time Interaction:** Ensures quick model responses for an efficient user experience.

Technical Stack

Framework:

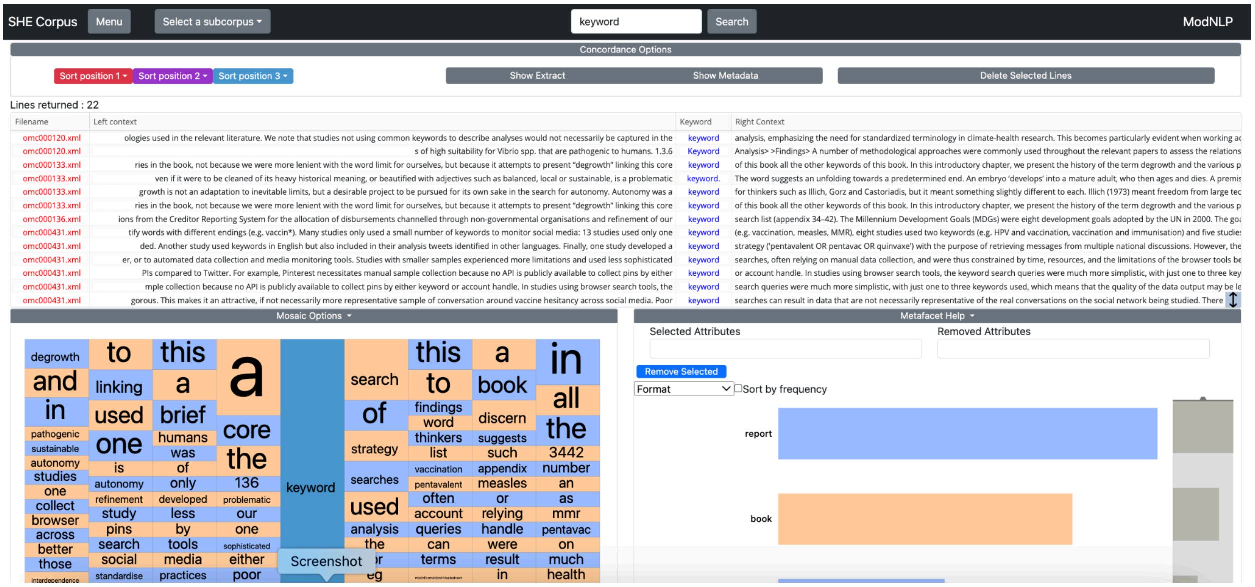
- **Streamlit** – For building interactive web interfaces.

Packages:

- **Whoosh** – Full-text indexing and search.
- **Pandas** – Data manipulation and analysis.

Interfaces for Reference:

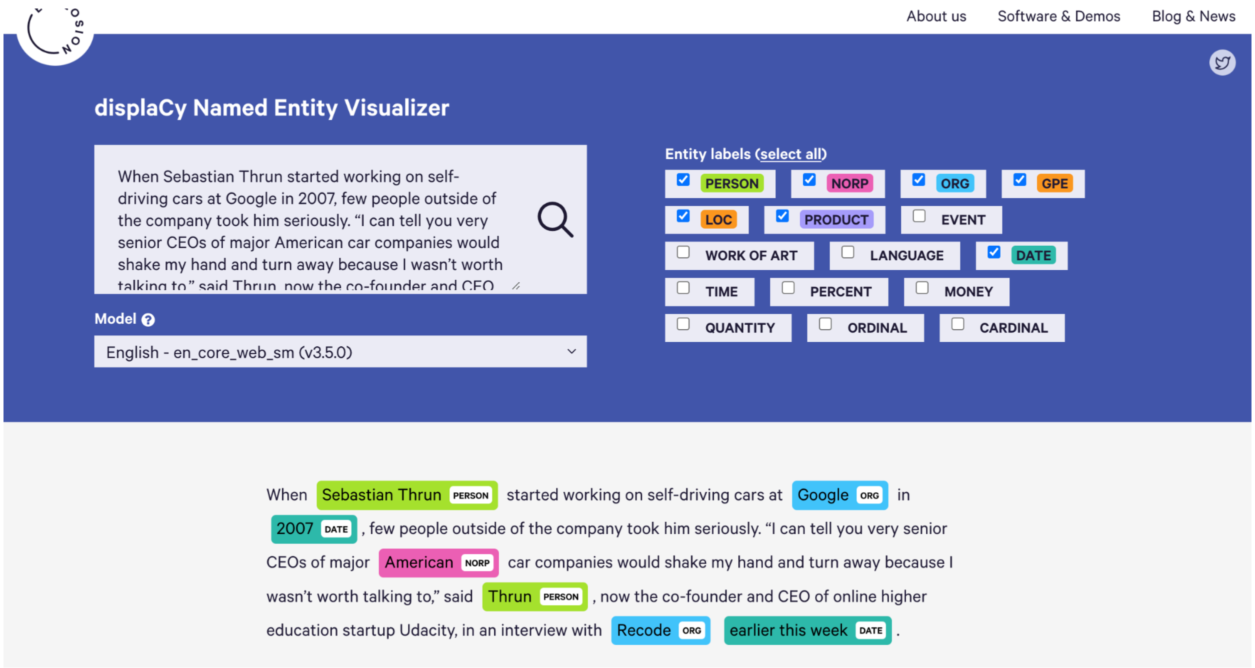
SHE Corpus: https://genealogies.mvm.ed.ac.uk/webcli/



Lines returned : 22

Filename	Left context	Keyword	Right Context
omc000120.xml	ologies used in the relevant literature. We note that studies not using common keywords to describe analyses would not necessarily be captured in the	keyword	analysis, emphasizing the need for standardized terminology in climate-health research. This becomes particularly evident when working ac
omc000120.xml	s of high suitability for Vibrio spp. that are pathogenic to humans. 1.3.6	Keyword	Analysis> >Findings> A number of methodological approaches were commonly used throughout the relevant papers to assess the relations
omc000133.xml	ries in the book, not because we were more lenient with the word limit for ourselves, but because it attempts to present "degrowth" linking this core	keyword	of this book all the other keywords of this book. In this introductory chapter, we present the history of the term degrowth and the various p
omc000133.xml	ven if it were to be cleaned of its heavy historical meaning, or beautified with adjectives such as balanced, local or sustainable, is a problematic	keyword.	The word suggests an unfolding towards a predetermined end. An embryo 'develops' into a mature adult, who then ages and dies. A premis
omc000133.xml	growth is not an adaptation to inevitable limits, but a desirable project to be pursued for its own sake in the search for autonomy. Autonomy was a	keyword	for thinkers such as Illich, Gorz and Castoriadis, but it meant something slightly different to each. Illich (1973) meant freedom from large tec
omc000133.xml	ries in the book, not because we were more lenient with the word limit for ourselves, but because it attempts to present "degrowth" linking this core	keyword	of this book all the other keywords of this book. In this introductory chapter, we present the history of the term degrowth and the various p
omc000136.xml	ions from the Creditor Reporting System for the allocation of disbursements channelled through non-governmental organisations and refinement of our	keyword	search list (appendix 34–42). The Millennium Development Goals (MDGs) were eight development goals adopted by the UN in 2000. The goa
omc000431.xml	tify words with different endings (e.g. vaccin*). Many studies only used a small number of keywords to monitor social media: 13 studies used only one	keyword	(e.g. vaccination, measles, MMR), eight studies used two keywords (e.g. HPV and vaccination, vaccination and immunisation) and five studies
omc000431.xml	ded. Another study used keywords in English but also included in their analysis tweets identified in other languages. Finally, one study developed a	keyword	strategy ('pentavalent OR pentavac OR quinvaxe') with the purpose of retrieving messages from multiple national discussions. However, the
omc000431.xml	er, or to automated data collection and media monitoring tools. Studies with smaller samples experienced more limitations and used less sophisticated	keyword	searches, often relying on manual data collection, and were thus constrained by time, resources, and the limitations of the browser tools be
omc000431.xml	PIs compared to Twitter. For example, Pinterest necessitates manual sample collection because no API is publicly available to collect pins by either	keyword	or account handle. In studies using browser search tools, the keyword search queries were much more simplistic, with just one to three key
omc000431.xml	mple collection because no API is publicly available to collect pins by either keyword or account handle. In studies using browser search tools, the	keyword	search queries were much more simplistic, with just one to three keywords used, which means that the quality of the data output may be le
omc000431.xml	gorous. This makes it an attractive, if not necessarily more representative sample of conversation around vaccine hesitancy across social media. Poor	keyword	searches can result in data that are not necessarily representative of the real conversations on the social network being studied. There

displaCy Named Entity VIsualizer: https://demos.explosion.ai/displacy-ent



displaCy Named Entity Visualizer

When Sebastian Thrun started working on self-driving cars at Google in 2007, few people outside of the company took him seriously. "I can tell you very senior CEOs of major American car companies would shake my hand and turn away because I wasn't worth talking to," said Thrun, now the co-founder and CEO

Model English - en_core_web_sm (v3.5.0)

Entity labels (select all)

☑ PERSON ☑ NORP ☑ ORG ☑ GPE
☑ LOC ☑ PRODUCT ☐ EVENT
☐ WORK OF ART ☐ LANGUAGE ☑ DATE
☐ TIME ☐ PERCENT ☐ MONEY
☐ QUANTITY ☐ ORDINAL ☐ CARDINAL

When Sebastian Thrun PERSON started working on self-driving cars at Google ORG in 2007 DATE, few people outside of the company took him seriously. "I can tell you very senior CEOs of major American NORP car companies would shake my hand and turn away because I wasn't worth talking to," said Thrun PERSON, now the co-founder and CEO of online higher education startup Udacity, in an interview with Recode ORG earlier this week DATE.