Project Description

Our project interrogates and responds to intersecting manifestations of Anglocentrism—in technological developments like artificial intelligence and in digital humanities methods and practices—by seeking to fine-tune OpenAI’s GPT-2 natural language model in Spanish—and in Spanish literature. The aim is that these models will serve as engines for literary and archival discovery, allowing students and researchers to explore well-known authors and their works in novel and creative ways.

Following a series of dead ends and failed attempts, we successfully built datasets needed to be available in the public domain—published in 1925 or before—and digitized (with character recognition). Given the historical exclusivity (in terms of gender and race) of the literary canon, we were limited in our ability to include diverse voices. We attempted to reconcile this to some degree by selecting two women authors.

Our initial aims were largely experimental and exploratory—to test the limits of the open-access AI and to test our own technical capabilities. We hope to continue learning and developing this project, ideally through the creation of an interactive web application that will allow users to test out our text-generating author models.

Central Problems & Guiding Theories

Programmed and algorithmic biases plague artificial intelligence and a wide range of technologies, a problem charted extensively by media and science scholars like Adam Hadhazy, Safiya Umoja Noble, and Sara Wachtter-Boettcher.

In GPT-2’s case, we observe that biases from the primarily Anglocentric—when prompted in Spanish, it does not generate coherent texts.

As Simon Mahony and Jin Gao, Roopika Risam, Thea Pitman and Claire Taylor, and others note, linguistic legacies of implicit and systemic biases—theorized to be rooted in the coloniality of knowledge—are pervasive in digital humanities tools and practices.

Tools & Processes

GPT-2: a large-scale, open-source, transformer-based language model programmed by OpenAI. Trained on a very large text corpus—8 million web pages—this language model is designed to predictively generate the next “token” (word) in a given sequence of text. GPT-2 stands for “Generative Pre-Trained Transformer, version 2.”

The original GPT-2 model is trained primarily and most effectively to generate texts in English. However, we found a fine-tuned version for Spanish: “GPT-2—small-spanish,” trained by the Datificate research group. They processed 3GB data of the Spanish language Wikipedia and applied transfer learning and fine-tuning techniques from the Hugging Face libraries to obtain the resulting model. According to them, the training took around 70 hours with four GPU NVIDIA GTX 1080-Ti with 11GB of DDR5.

We used their model as a baseline to train our own. To fine-tune each author model, we prepared datasets ranging from 0.5 to 6GB containing a corpus of the texts of each author. We fine-tuned three separate GPT-2 models using the Hugging Face transformer library. Our software platform was a Google Colab notebook which executes code using a cloud-based GPU environment.

Results: Generated Texts

El acto de la escritura es el elemento clave en toda explicación de la existencia humana. // The act of writing is the essential element of every explanation of human existence.

---GPT-Milana Pardo Bazán

La escritura es cosa muy curiosa y muy rara, y así los unos en la lengua inglesa, la otra en el árabe; pero . . . no hay artificio más extraordinario que el arteficio que el español usa. // Literature is a curious and odd thing, and so some [do it] in the English language, another in Arabic; but . . . there is no craft more extraordinary than that which Spanish employs.

---Cide GPT-Mete Benengeli (Miguel de Cervantes GPT Saavedra)

Mi escritura es del alma—por ser ella la deidad—sin haber sido sólo de los sentidos. // My writing is from my soul—as she is the deity—without having been only of the senses.

---Sor Juana Inés de la Cruz

Lingering Issues & Future Plans

While the ostensible aim of our project is to respond to—and help to begin to remedy—linguistic biases in AI technology and digital humanities, certain limitations of the tools at our disposal inherently—and inevitably—reproduce implicit biases related not only to language but also to race and gender.

Wikipedia training dataset: Because the “Small Spanish” model of GPT-2 is fine-tuned with a Wikipedia training set, our fine-tuned model likely contains hidden biases based on the source data’s issues of selection and representation (Wikipedia has been scrutinized both for its exclusion of certain perspectives and for its questionable factuality).

Public domain, digitized training dataset: The texts we used in our fine-tuning datasets needed to be available in the public domain—published in 1925 or before—and digitized (with character recognition). Given the historical exclusivity (in terms of gender and race) of the literary canon, we were limited in our ability to include diverse voices. We attempted to reconcile this to some degree by selecting two women authors.

Our initial aims were largely experimental and exploratory—to test the limits of the open-access AI and to test our own technical capabilities. We hope to continue learning and developing this project, ideally through the creation of an interactive web application that will allow users to test out our text-generating author models.

Selected References & Works Cited


For more resources, visit our project website: https://tbt.gy/kkuhd

Acknowledgements

This work was funded by an Andrew W. Mellon Fellowship for the Digital Humanities. We are grateful for the collegiality, comments, and constructive criticism of our fellow 2021 Mellon DH Fellows. (Gracias a todes)