Assignment 1 – Data Science

Decolonial and Data centric approaches on improving Wikipedia entries

Crysern Smith

The Wikipedia entry chosen is that of Jan Smuts from Lucy Panesar's excel spreadsheet, toppletheracists.org, an interactive map featuring statues honouring and commemorating the achievements of those with racial history. The rationale for this decision was due to the legislation Smuts drafted and executed, which eventually became the foundational laws[i] that governed South Africa throughout apartheid. There was no mention of the legislation on his Wikipedia entry page, except that he supported segregation.

Information and data collection must be collected from reliable sources to provide accurate facts regarding the topic mentioned above; this includes the library and video archives collected and stored by academics that specialise in the field, as well as any recorded and archived material from Smuts himself, who was a well-known academic and author of several books.

Jan Smuts groundwork on segregation in his homeland of South Africa was most likely overlooked because he was considered a war hero, having meaningfully contributed to the winning ally campaign in World War 2 and being a close collaborator of Winston Churchill. This might be why his influence on South Africa's segregation regime and its contribution to creating the Apartheid system has been left out of his Wikipedia page (and most other online available sources).

One of the more challenging aspects of gathering historical data is that the narrative differs significantly across sources. Each scholar, author, and academic may have approached the subject from a different perspective, which could have led to inaccuracies and bias. Nonetheless, most of the data on segregation and Jan Smuts have been well presented; hence the reasoning behind omittance could be separated into categories of either personal bias or language barrier.

The available historical information is in the native tongue of Jan Smuts, Afrikaans, which might be challenging to comprehend and translate given the lack of resources to accurately translate a niche language such as Afrikaans, especially if the researcher is not a native speaker.

Other areas of challenge could be the location of resources; Jan Smuts was a prominent figure but not a famous one. His documented history can be found in South Africa and the United Kingdom, and this is a potential challenge if a researcher is not living in either country. If there is no demand for digitising records, they will remain in the libraries and archives in their respective countries. This is just one of the more severe challenges that could be faced by anyone looking to further delve into the history of Jan Smuts.

These are some of the steps regarding South African history, unfortunately not at the rate it would take to ensure that all Wikipedia entries are accurate. Archival digitisation is an excellent form of preserving the past, making data available to all, and developing language translation. This way of documenting history is time-consuming and costly, which are two things that South Africa cannot give away freely.

We think the best thing to do is to ensure that the relevant data used has been sourced through different publicly available mediums and ones that are not as readily available. Subsequently, the collected information can be digitised and archived for future use.

Some advantages of using a data-centric approach are uploading the relevant documentation, OCR technology has made leaps and bounds. A simple OCR scan on historical documents can now be a useable dataset. Furthermore, using NLP techniques, we can use tokenisation to pull information given a timeline and keyword; this can allow for historical data to be categorised and cross-referenced against other sources of information to see patterns and correlations within the text. Another advantage is that the information then becomes available on a global scale.

The disadvantages of data-centric approaches are the need for human interaction and engagement, and an unbiased representative would need to upload all relevant documentation. However, even with that, there could be documents that are media propaganda or articles that have been lost in translation and then interpreted to make it seem that the same law Jan Smuts wrote was the backbone on which apartheid laws were governed.

The best we can do is to ensure that a dataset is as extensive and as diverse as possible, and when we see a correlation of timeline and facts, we do our due diligence to cross-reference and then accurately report on topics.

---

[i] (Feinberg, 1993; Beinart and Delius, 2014)