Newspapers to Numbers

Introduction and Summary

I sought to **quantify how language has changed over time**. I also wanted to see how **historical trends** appeared in this data. Using the **BERT AI model**, scanned **newspapers** from the **Library of Congress**, and more than 16 hours of model training, I have made large models that can produce **vectors that represent the idea** of a series of words. And by graphing the **changes in similarity** between different **ideas over time**, we can see **correlations** with **historical trends**.

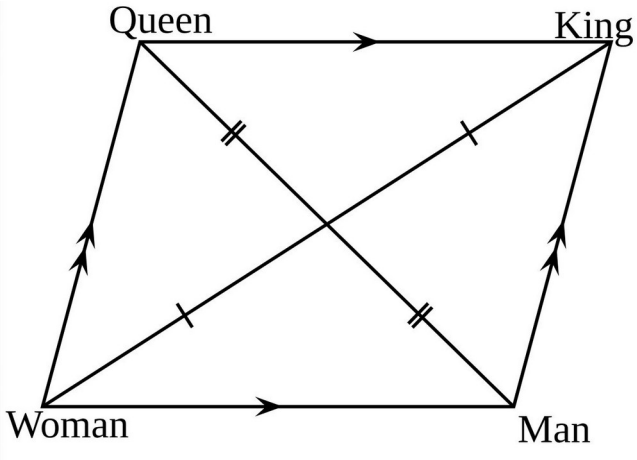
Background

Many methods in AI allow for the **quantification of meaning**.

This is commonly done by constructing a **space of words.**

The **locations** of these words in the space **correspond** to their **definition and meaning.**

The **BERT** architecture, created by Google in 2019, makes use of these ideas to **model language.**



Sometimes word relationships can appear in these word-spaces - something similar to the above parallelogram is formed within many AI systems [Efficient Estimation of Word Representations in Vector Space, 2013]

Methods

1) I downloaded pre-scanned **American newspapers** from the Library of Congress
2) I **filtered** the articles and only used those from **1850-1950**
3) I then **split each article** into 20 **excerpts** to make training easier
4) While training, the BERT model learned to **associate the year** to newspaper **content** from the year

5) I took the **word-vectors** the **BERT** created from **pairs of phrases**. An example phrase would be *"[year1903]:Empire"* to which the the model would give a **representative vector** for the meaning of the phrase

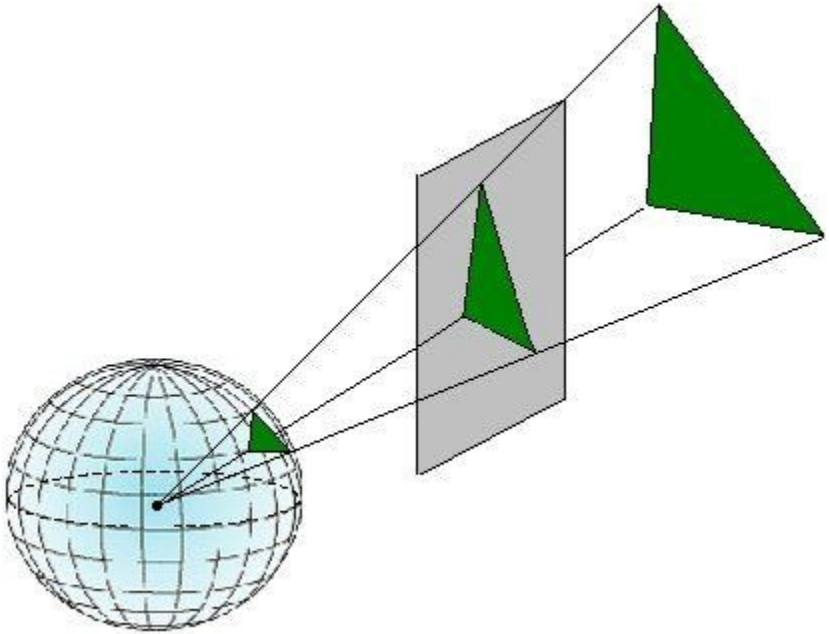
6) I then **repeated this** for phrases relating to various countries, such as the **United States of America**, **Great Britain** and **Spain**.

7) I then measured and **graphed** the **similarity** between the **ideas** of these **pairs of phrases** in the BERT's "understanding", for each year. This was done by finding the **distance between** the **representative vectors** for each phrase.

Technical notes for the methods

To calculate the the **similarity between ideas** of phrases, I mapped the **word-space to a globe** and took the **cosine of the angle** between the phrases on the globe. More **similar ideas** have **cosines closer to one.**

This method is used for similarity because other methods for traditionally finding distance in such a large space perform worse as the dimension increases.



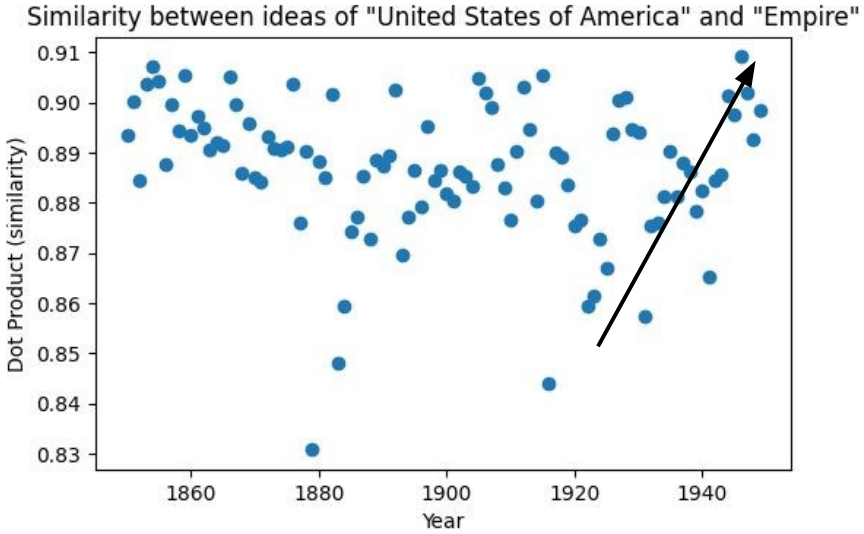
https://www.researchgate.net/figure/Projection-of-the-3D-and-2D-points-on-the-sphere_fig3_236027560

Results and discussion

To the left we can see the **similarity between** the **changing ideas** of **“United States of America”** and **“Empire”**.

We also see an **increase** in the period during and following **World War II**, and is often considered when America became a global superpower.

This could show that while the **idea of** what **America** was **stayed constant**, the idea of what an **empire** was **became** more **“American”.**

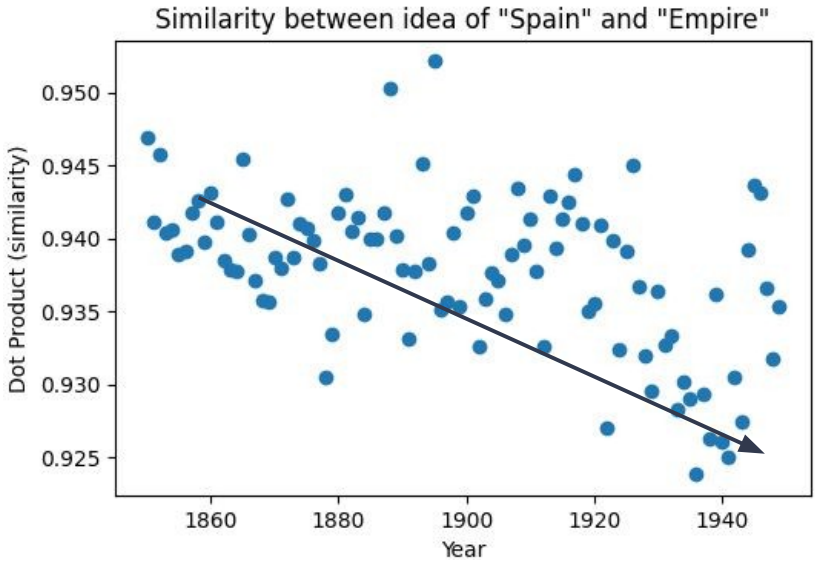


Results and discussion

To the left we see a gradual **decrease** in the **similarity between** the **changing ideas** of **“Spain”** and **“Empire”**.

The largest downward shift occurred during **World War II**, during which Spain was in the midst of a **civil war**.

This downward shift may have also occurred because the **idea** of **“Empire”** was becoming more **“American”**.

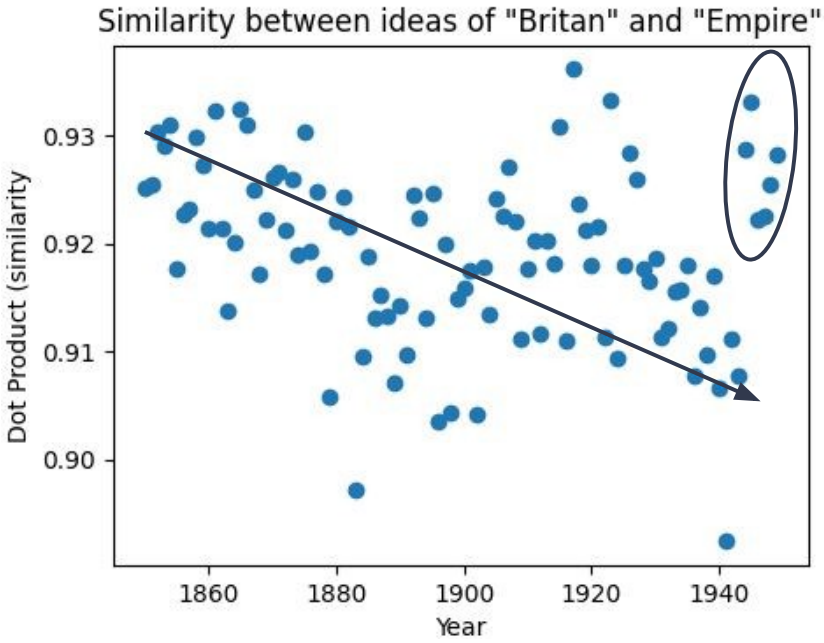


Results and discussion

To the left we see **several trends** relating to the **similarity between** the **changing ideas** of **“Britain”** and **“Empire”** over time.

We see a **gradual decrease** in similarity, and this **correlates** with the **decline of** the **empire**.

We also see a **quick spike** in similarity during **World War II**, which may have been caused by **American newspapers describing** Britain’s **fighting** in the war.



Discussion

This model allows us to see the **correlation** between the **language use** in newspapers and **history**. It also allows us to see how **words change over time**, and how they **change** when surrounded by additional **context**.

It's important to remember that this model was trained on **American** newspapers and therefore **embodies the biases** these **newspapers** have.

For example, American newspapers in the **1890s** may have put extra **emphasis** on the **decline of the Spanish Empire**. This probably would have been **caused by** yellow journalism's **sensational** and **nationalistic reporting** of the **Spanish-American War.**

Acknowledgments

I want to thank Mr. [redacted], for guiding me through this project and suggesting directions to explore.

I want to thank my parents, who encouraged me to learn Linear Algebra.

I also want to thank my friends, [redacted] and [redacted], for helping me build the computer which I trained my models on.

Citations

Devlin, J. et al. (2019) Bert: Pre-training of deep bidirectional Transformers for language understanding, arXiv.org. Available at: https://arxiv.org/abs/1810.04805 (Accessed: 10 February 2026).

Humanities, N.E. for the (no date) Chronicling America: Library of Congress, News about Chronicling America RSS. Available at: https://chroniclingamerica.loc.gov/ocr/ (Accessed: 10 February 2026).

Historically interesting excerpt of an article

Excerpt of an article from September 24th 1904

…Secretary of War Taft heads the list of those members of the jury of awards who will judge the exhibits of the Philippines at the World's fair. A French army officer and a private who attempted to take a photograph of the Forbidden City at Pek<n were severely beaten by Chinese soldiers. Dr. Anita McGee, in charge of Red Cross work with the Japanese army, denies that the Japanese soldiers are hypocritical in their kind treatment of Russian prisoners….