Methodology of Google Trends studies

Soumadeep Ghosh

Kolkata, India

Abstract
In this paper, I describe the methodology of Google Trends studies.
The paper ends with "The End"

Introduction

In today's world, with ever-increasing data, traditional statistical methods are either **inefficient** or face **data deluge** that makes discovering insights akin to finding diamonds in the dust.
However, Data Analysis, a subset of Data Science, offers a few methodologies that are able to find such diamonds in the dust.
One such Data Analytic methodology is that of **Google Trends studies**.
This methodology isn't new. In fact, this methodology was in use long before Large Language Models (LLMs) like BERT, GPT-1 etc. were built.
The methodology depends only on obtaining two time-series data and the use of similarity measures and correlation coefficients.
In this paper, I describe the methodology of Google Trends studies.

Similarity measures

There are many similarity measures that one may use in this methodology. In fact, there's ongoing research to find newer and more efficient similarity measures all over the world. But for time-series that are not too long, **cosine similarity** still performs remarkably well.

Cosine similarity

Given two vectors of length n, $\mathbf{A}$ and $\mathbf{B}$, the cosine similarity

$$\cos\theta = \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

measures the similarity between the two vectors.

Correlation coefficients

Again, there are many correlation coefficients that one may use in this methodology. And again, in fact, there's ongoing research to find newer and more efficient correlation coefficients all over the world. But for time-series that are not too long, **Pearson's correlation coefficient** still performs remarkably well.

Pearson's correlation coefficient

Given two vectors of length n, $\mathbf{A}$ and $\mathbf{B}$, the Pearson's correlation coefficient

$$r_{(\mathbf{A},\mathbf{B})} = \frac{cov(\mathbf{A},\mathbf{B})}{\sigma_{\mathbf{A}}\sigma_{\mathbf{B}}} = \frac{\sum_{i=1}^{n}(A_i-\bar{A})(B_i-\bar{B})}{\sqrt{\sum_{i=1}^{n}(A_i-\bar{A})^2}\sqrt{\sum_{i=1}^{n}(B_i-\bar{B})^2}}$$

measures the correlation between the two vectors.

Methodology of Google Trends studies

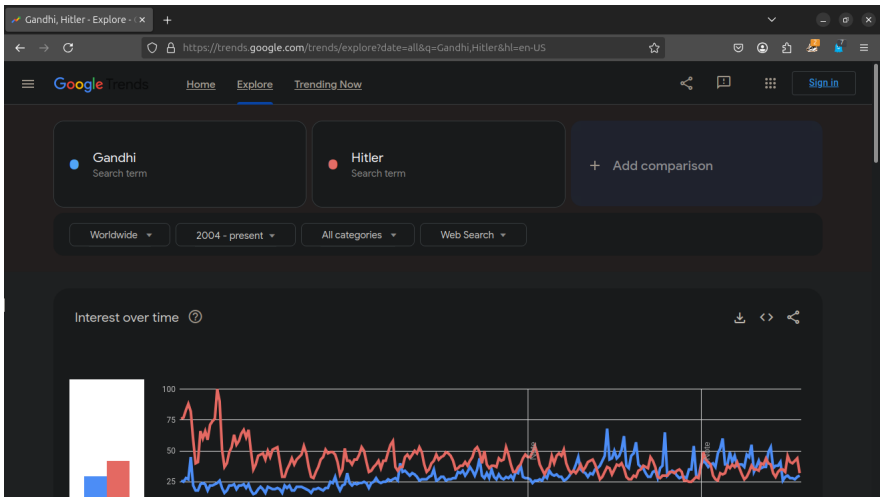
The methodology of Google Trends studies is simple:

1. Decide on two words or phrases that are real variables or serve as proxies of real variables.
2. Download simultaneous time-series data of equal length on the two words or phrases **as search terms** from `https://trends.google.com/trends/`.
3. Compute the similarity measure(s) and the correlation coefficient(s) on the two time-series.
4. Interpret the results using standard rules.

Best practices

1. The time-series data must be from the same interval of time.
2. To deter adversarial attacks, the sample size should be sufficient - usually around 25 to 35 observations.
3. Compute both the similarity measure(s) and the correlation coefficient(s) to ensure consistency, validity and veracity.
4. The standard rules of interpretation should be consistent across time, algorithms and regimes.
5. Leverage such studies in the **Exploratory Data Analysis** phase of the Data Science pipeline.

Exercise for the reader



The reader is encouraged to find the similarity measure(s) and the correlation coefficient(s) for the two Google Trends search terms 'Gandhi' and 'Hitler' to gain political insights.

References

1. Google Trends (`https://trends.google.com/trends/`)
2. Similarity measure (`https://en.wikipedia.org/wiki/Similarity_measure`)
3. Correlation coefficient `https://en.wikipedia.org/wiki/Correlation_coefficient`

The End