Understanding Collective Memory Patterns by Analyzing View Patterns of People's Pages on the English Wikipedia

Final Project for the Course *The Art of Analyzing Big Data*

Netanel Madmoni

April 2024

1 Introduction

1.1 Wikipedia Statistics as Collective Memory Measures

There are many works on the concept of *collective memory*. Dating back to the 1920s, this concept of memory extends beyond individual recollection to encompass the shared experiences and representations of communities, cultures, and societies [1]. In 2009, Pentzold [2] proposed 'to interpret the web-based encyclopedia Wikipedia as a global memory place', as it lets locally disconnected members to combine their knowledge and perspectives into one collective memory in the cyberspace. Following this notion, many researchers often used Wikipedia statistics as a proxy for the collective memory [3] and attention [4] [5].

1.2 Time Series Pattern Extraction & Approximation

Pattern extraction from time series is a important aspect of time series analysis in all areas of time-series- or signal-processing-related research with many real-world applications. Many algorithms have been developed over the years for both local and global pattern extraction for a variety of different tasks, such as classification [7] [8] and clustering [9]. One of the useful tools used for time series analysis and pattern extraction is time series approximation. It is used not only for overcoming technical limitations such as computation time and storage space, but can also be used for representing time-series data in a meaningful way for specific tasks such as pattern extraction and features discovery. Symbolic ApproXimation (SAX) is a relatively simple and intuitive method for time series approximation proposed by Lin *et al.* [6]. The method reduces the dimensionality of a time series by dividing it into equal sized 'frames' and calculating the mean value for each frame, and then discretizing the reduced series into predetermined bins. See Figure 1 for an illustration of the algorithm.

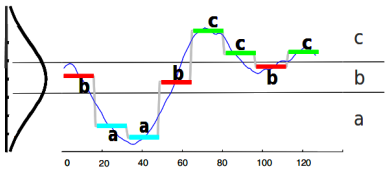

Figure 1: SAX algorithm illustration. The time series is discretized and mapped to specified breakpoints, resulting in the representation **baabccbc**. Source: [6]

1.3 Contribution

In this work I identify and investigate the commonness of six intuitive different page views patterns of around two million Wikipedia pages of persons over nine years (2015-2023), using publicly available page views information and pages data. My goal is to reveal properties of the collective memory in regards to people, and have a better intuitive understanding of different groups of pages by characterizing the patterns that are present in their views over time.

2 Method

2.1 Data Description & Acquisition

In this project I use three sources of data:

1. **English Wikipedia page view data** - from a public dataset hosted on Google BigQuery[1]. The dataset contains an hourly view count of all of the pages on Wikipedia over the years 2015-2023. The granularity of this dataset is at the hourly level, which is too high for the purpose of this project. I've taken a monthly sum of the views for each entity, resulting in a total of over 200 million data points across over 2 million rows (later filtered to around 1.5 million rows).

2. **Wikidata entity data** - a public dataset containing information about Wikidata entities. Also hosted on Google BigQuery[1].

3. **Database of notable people** - a cross-verified database published in 2022 by Laouenan *et al.* [10]. The database contains a verified list of people on Wikidata and their properties. Due to the importance and cleanliness of this information, I decided to only use the pages that are present in this database.[2]

The raw data was cleaned and transformed in preparation for the pattern extraction. For a detailed description of the data fields and cleaning process, refer to the accompanying code files, `notebook 1` in the Supplementary Material section.

[1] Google cloud bucket: `gs://cloud-samples-data/third-party/wikimedia/pageviews`.

[2] Also due to the Wikidata entity data being messy, incomplete and sometimes false, the verified database is the main source of the information in regards to entity properties.

2.2 Pattern Classification

In this section I describe the patterns I've selected to identify as well as the methods for classifying a pattern.

2.2.1 The Patterns

I decided to focus on the following patterns:

- *One-hit wonder pattern* - the term 'one-hit wonder' is taken from the music industry. Originally it refers to a band or an individual that became popular due to a single piece of work going viral. In this work I am borrowing this term to describe a page of a person that received one big 'peak' of views during the lifetime of the page.
- *Nostalgia pattern* - similar to *one-hit wonder*, but here the page views had two big "peaks" of views during its lifetime, emulating the "nostalgia" phenomenon.
- *Seasonal pattern* - where an entity has at least one strong seasonal components in the views its page receives.
- *Rising pattern* - where an entity has an upward trend in its views over time.
- *Fading pattern* - where an entity has an downward trend in its views over time.
- *Constant / no pattern* - where an entity has no particular pattern in its page views over time (the default).

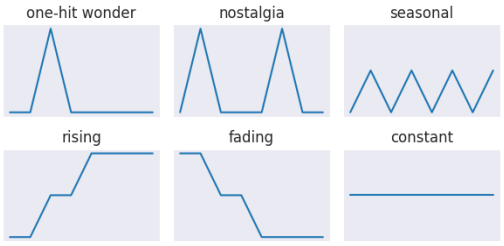

Figure 2: Illustrations of each of the patterns.

A simple illustration of each pattern is shown in Figure 2.

2.2.2 Assumptions

For simplicity, I decided to assume the following: 1) An entity can have more than one pattern, as long as the patterns don't contradict each other (for example, a person can have both *one-hit wonder* and *seasonal*, but not *one-hit wonder* and *nostalgia*, or *rising* and *fading*, as they are contradictory by definition). 2) The patterns are independent of the absolute volume of the series, meaning, popular pages (high sum of views over the years) as well as less popular can have the same pattern. In other words, it is the shape that I am interested in, regardless of its absolute values. 3) Each pattern spans across the entirety of the series, i.e., the subject of classification is the series as a whole, rather than just subsequences of it is (as opposed to, say, shapelets [8]). In future work it might be worth questioning all of these assumptions.

2.2.3 Determining the Pattern of Each Series

In order to determine the category into which each page views time-series fall, I first calculate the symbolic approximation (SAX) of each time series, classifying each observation into one of two bins: *high* and *low*. I then count the number of observations in the *high* bin. This is what I consider as the number of 'significant' peaks. If there is a singular peak, then the series is classified as *one-hit wonder*. If there are two peaks, then the series is classified as *nostalgia*. It is important to note that I've considered clusters of observations will to count as a singular peak if they are within distance d from each other, for a predetermined d (here I've chosen $d = 3$, meaning that peaks within the same quarter-year will count as a single peak).

Next, i test for the *seasonal* pattern. To determine if a series has a significant seasonal component, I use the autocorrelation function (ACF) for lags $d + 1 = 4$ to 21 and calculate the confidence interval for each autocorrelation one at significance level .95, using the Ljung–Box test [11]. If the confidence interval's lower bound of any autocorrelation is positive, then that series is considered to have a significant seasonal component and is classified as *seasonal*.

Then, I perform a simple linear regression on the series. If the regression's p-value is significant, then that series is classified as *rising* or *fading*, according to the slope of the regression.

Any series that hasn't been classified yet, will be classified as *constant*.

Figure 3 shows a few (good[3]) examples for various patterns.

[3] I admit that a lot of examples are not classified in an intuitive manner, see more on section Discussion & Future Work.

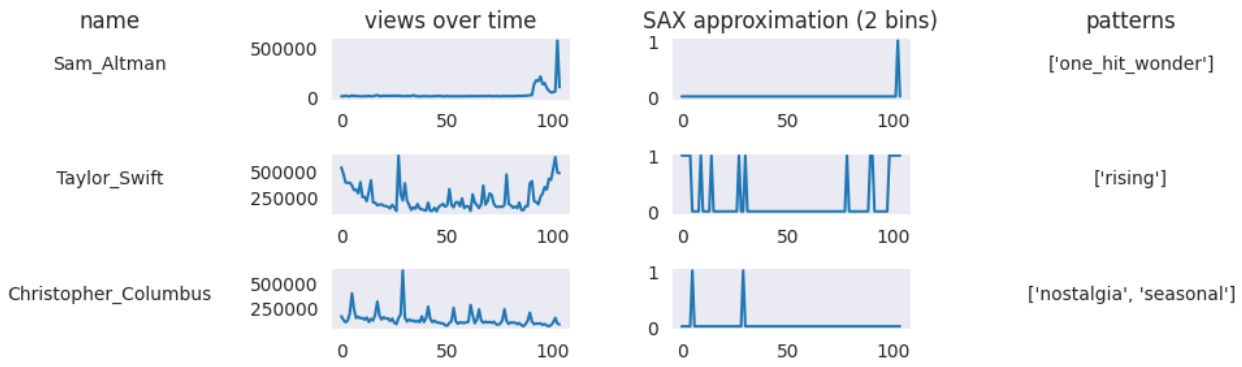

Figure 3: A few examples for various patterns.

A Note on Optimization

Since the data is very large in volume, I've employed different techniques for achieving satisfactory results in a reasonable amount of time and memory space. These include working with `parquet` files and using the `Polars` package [12] for data transformation, and more importantly: defining relatively simple, easy to calculate measures.

3 Results

For the dataset as a whole, the *one-hit wonder* pattern was the most common one (after the default *constant* pattern). A close second is the *seasonal* pattern, followed by the *nostalgia*, *fading* and *rising* patterns, in that order. Figure 4 shows the number of occurrences for each of the patterns in the dataset as a whole. I discuss these results in Section 4.

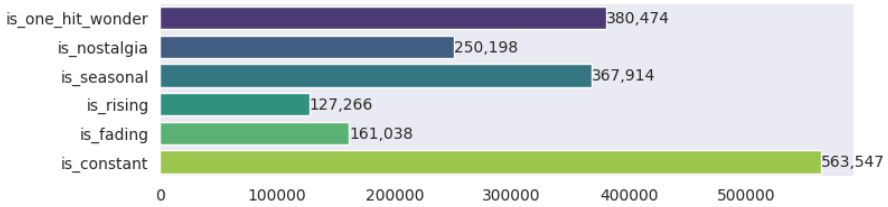

Figure 4: Count of patterns over the entire dataset.

Next are the pattern distributions by some individual properties[4]. Figure 5 shows the distributions by gender. Note that in all of the following figures, the counts have been normalized for each group (divided by the total number of patterns present in that group).

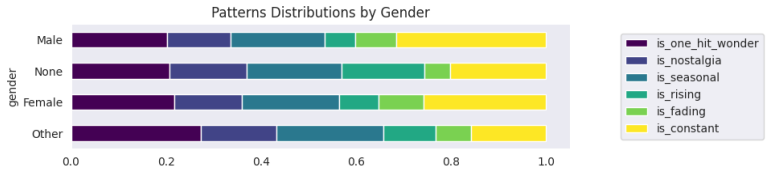

Figure 5: Count of patterns by gender.

Although they seem very similar, there is a statistically significant difference in proportions between the male and female group, i.e., the proportion of each pattern (except for the default pattern *constant*) within the male group was significantly smaller than that proportion in the female group ($p \ll .0001$).

Below (Figure 6 are the normalized pattern counts by occupation.

[4]I've chosen some of the interesting properties to display. Distributions of the patterns by other properties can be found in the code in the Supplementary Material, `notebook 2`.

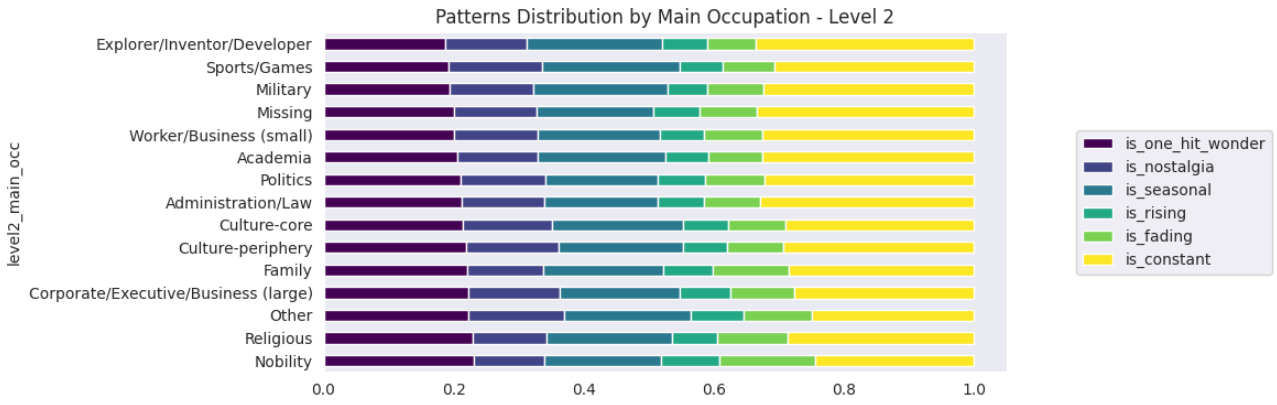

Figure 6: Count of patterns by occupation.

It seems that the proportions are similar between groups, however 2/3 of the combinations are significantly different between groups. It is interesting to see that the occupations *nobility* and *religion* have the most *one-hit wonder*-s, although being two groups with little members.

An interesting observation for both gender and occupation[5], while statistically different, the proportions within each group seem roughly similar to the global proportions presented in Figure 4. This is not the case for the distributions by birth period (Figure 7).

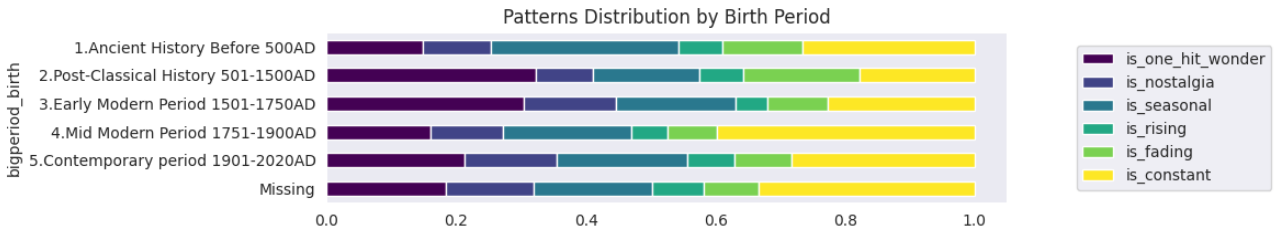

Figure 7: Count of patterns by birth period.

As it seems (and perhaps counter-intuitive), in years 501-1750 AD, the proportion[6] of *one-hit wonder*-s is much larger relative to the other eras. Also worth noting that the seasonal pattern is more prominent for people living before 500AD. I discuss this in Section 4.

Finally, we can look at the patterns distributions by Laouenan *et al.*'s [10] *notability rank* of a person. It is a measure comprised of the number of Wikipedia editions of each individual, the length in all available Wikidata biographies for the individual, the average number of biography views for them between 2015 and 2018 in all available language editions, the number of non-missing items retrieved from Wikipedia or Wikidata for birth date, gender and domain of influence. I have grouped the ranks into ten bins. The normalized pattern count for each bin is presented in Figure 8.

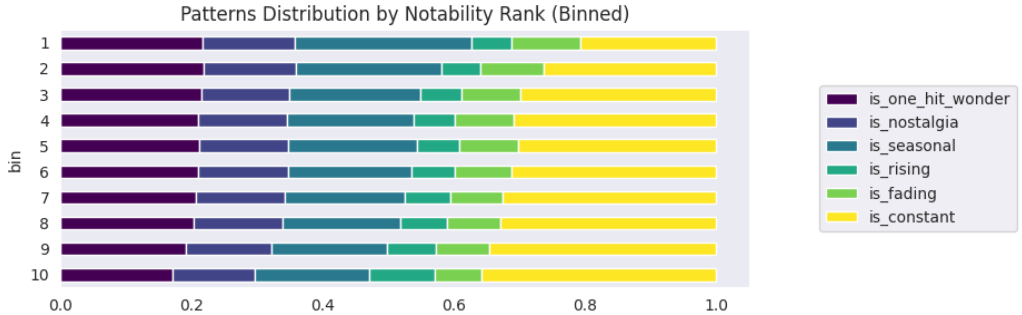

Figure 8: Count of patterns by binned notability rank. Lower bin number means more notable.

It seems that the percentage of the default non-pattern *constant* is correlated with notability rank (see

[5] As well as other properties present in the supplementary material, `notebook 2`.
[6] Remember: these are normalized proportions by group, and not absolute numbers.

Figure 9), meaning that less notable people have more constant patterns. By that logic, the more notable a person is, it is more likely to have one of the 'special' patterns.

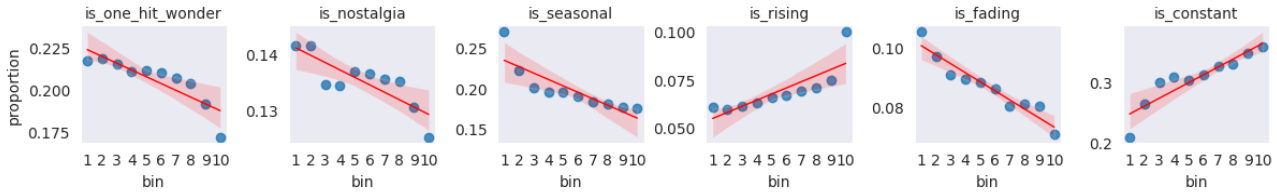

Figure 9: Regression plots describing the proportion of each pattern by binned notability rank. Lower bin number means more notable. Most of the patterns have decreasing proportion with rank, meaning the pattern is more present for more notable people.

4 Discussion & Future Work

In this work, in order to better understand the collective memory in regards to people, measured by Wikipedia page views, I presented a way to classify people's page views on the English Wikipedia into six simple patterns, using intuitive and statistical measures. I was able to classify a large portion of pages into one of five 'special' patterns: *one-hit wonder*, *nostalgia*, *seasonal*, *rising*, and *fading*. Surprisingly, out of those five, the largest group in the dataset as a whole was the *one-hit wonder* group, meaning the page had a high number of views in a single short period during the last nine years. This might reveal an important aspect of our 'collective memory': does the memory shifts globally from one subject to another in a short period of time?

For the seasonal pattern, the presence of seasonality in some people's Wikipedia articles might be due to a academic interest in these people that fluctuated over the academic year [13], which is a reasonable speculation for people from earlier historical eras, or real world-events happening seasonally for living people.

This work should be considered a preliminary investigation. There are many interesting ways in which it can be improved. Admittedly, for many samples, the chosen pattern was not very intuitive. I think this is might be due to the fact that page views, especially over nine years, have more complex patterns that can't be defined by a single 'shape' over their entire lifespan (local patterns might be more robust and intuitive than the global patterns). The selection of the minimum distance between peak clusters described in Section 2.2.3 also seems to have a big impact on the classification. The patterns themselves and the way of extracting them might need to be assessed as well. I have treated the patterns as independent of the absolute number of views in the series, an assumption worth questioning. Finally, it would be interesting to try to develop a machine learning model that can predict a views pattern from a person's properties such as occupation, gender, etc.

I believe this project can provide a solid foundation to a better understanding of the view patterns on Wikipedia and a window to the workings of our global memory.

Acknowledgements

I've used the assistance of generative AI (ChatGPT) in the writing of some of the code and this report.

Supplementary Material

- **Data.** I have uploaded both the page view data and the Wikidata entity data to Kaggle, and they are available in this dataset: https://www.kaggle.com/datasets/netanelmad/wikipedia-people-page-views-data. The verified people dataset by Laouenan *et al.* [10] is available at https://data.sciencespo.fr/dataset.xhtml?persistentId=doi:10.21410/7E4/RDAG3O.
- **Code.** My code is available at the GitHub repository: https://github.com/MNetanel/wikipedia-people-page-views-patterns.

References

[1] N. Igarashi, Y. Okada, H. Sayama, and Y. Sano, "A two-phase model of collective memory decay with a dynamical switching point," *Scientific Reports*, vol. 12, no. 1, p. 21484, Dec. 2022.

[2] C. Pentzold, "Fixing the floating gap: The online encyclopaedia Wikipedia as a global memory place," *Memory Studies*, vol. 2, no. 2, pp. 255–272, May 2009.

[3] T. Yasseri, P. Gildersleve, and L. David, "Collective memory in the digital age," in *Progress in Brain Research*. Elsevier, 2022, vol. 274, pp. 203–226.

[4] T. Ruprechter, K. Burghardt, and D. Helic, "Poor attention: The wealth and regional gaps in event attention and coverage on Wikipedia," *PLOS ONE*, vol. 18, no. 11, p. e0289325, Nov. 2023.

[5] P. Gildersleve, R. Lambiotte, and T. Yasseri, "Between news and history: Identifying networked topics of collective attention on Wikipedia," *Journal of Computational Social Science*, vol. 6, no. 2, pp. 845–875, Oct. 2023.

[6] J. Lin, E. Keogh, L. Wei, and S. Lonardi, "Experiencing SAX: A novel symbolic representation of time series," *Data Mining and Knowledge Discovery*, vol. 15, no. 2, pp. 107–144, Aug. 2007.

[7] P. Geurts, "Pattern Extraction for Time Series Classification," in *Principles of Data Mining and Knowledge Discovery*, G. Goos, J. Hartmanis, J. Van Leeuwen, L. De Raedt, and A. Siebes, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2001, vol. 2168, pp. 115–127.

[8] L. Ye and E. Keogh, "Time series shapelets: A new primitive for data mining," in *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Paris France: ACM, Jun. 2009, pp. 947–956.

[9] L. Ulanova, N. Begum, and E. Keogh, "Scalable Clustering of Time Series with U-Shapelets," in *Proceedings of the 2015 SIAM International Conference on Data Mining*. Society for Industrial and Applied Mathematics, Jun. 2015, pp. 900–908.

[10] M. Laouenan, P. Bhargava, J.-B. Eyméoud, O. Gergaud, G. Plique, and E. Wasmer, "A cross-verified database of notable people, 3500BC-2018AD," *Scientific Data*, vol. 9, no. 1, p. 290, Jun. 2022. [Online]. Available: https://www.nature.com/articles/s41597-022-01369-4

[11] G. M. Ljung and G. E. P. Box, "On a measure of lack of fit in time series models," *Biometrika*, vol. 65, no. 2, pp. 297–303, Aug. 1978.

[12] Pola-Rs, "Pola-rs/polars: Dataframes powered by a multithreaded, vectorized query engine, written in rust." [Online]. Available: https://github.com/pola-rs/polars

[13] E. Segev and A. Baram-Tsabari, "Seeking science information online: Data mining Google to better understand the roles of the media and the education system," *Public Understanding of Science*, vol. 21, no. 7, pp. 813–829, Oct. 2012.