This is my entire dump for Swamiji corpus analysis using data science and nlp tools.

The idea came to me when I was trying to use the same methods on Gospel of Sri Ramakrishna and the bilingual nature of the corpus made it very difficult to do any analysis since no formal mapping existed between Kathamrita and Gospel. I tried to use the “word-by-word” translation of kathamrit version (as opposed to the Nikhilanand ji translation) but even that had some major issues in trying to map the entire gospel even sentence by sentence one-to-one. So I decided to let go of the Gospel due to bilinguality and discovered that Swami Vivekananda complete works was just waiting there to be used, most of which was in English natively! And that is how this project began. The works of swamiji have been looked at with various angles, but using NLP and corpus analysis and data science is not at all one of them. I could not find a single work online doing anything remotely similar to what I wanted so I did the entire thing. It took about 8 hours of work over about two weeks to make all this possible.

Now basically, I had most of the work done for me by the grace of the God. On the wonderful site https://www.ramakrishnavivekananda.info/vivekananda/complete_works.htm, there was well annotated and rather updated dataset of complete works of swamiji. This was my main source of all the data I have analyzed in the following sections. From the url above, I was able to go to https://www.ramakrishnavivekananda.info/vivekananda/, which cleanly indexed (with hyperlinks) all the important pages as follows:

Name	Last modified	Size	Description
Parent Directory		-	
CompleteWorksDynamic..>	2018-09-05 21:48	941	
appendices/	2023-03-07 10:57	-	
complete_works.htm	2023-03-07 10:53	1.6K	
complete_works_conte..>	2008-09-05 22:20	1.9K	
completeworksindex.xlsx	2023-09-18 21:57	152K	
intro_frame.htm	2008-09-05 22:20	507	
introduction.htm	2023-08-04 14:39	17K	

master_index.htm	2023-10-02 18:53	135K
picosearch.htm	2008-09-05 22:20	440
swami_vivekananda_ja..>	2023-03-07 10:54	108K
unpublished/	2023-03-07 11:00	-
vivekananda.htm	2013-03-10 20:13	485
vivekananda.jpg	2013-03-10 20:13	24K
volume_1/	2014-01-24 17:42	-
volume_2/	2014-01-24 17:42	-
volume_3/	2013-08-25 18:52	-
volume_4/	2014-01-24 17:49	-
volume_5/	2013-08-25 18:52	-
volume_6/	2022-03-29 19:27	-
volume_7/	2013-08-25 18:55	-
volume_8/	2016-08-18 20:44	-
volume_9/	2013-08-25 18:56	-

As we can see, the data I will analyze is available on master_index and has last been updated on 2023, which would be 2.5 years ago from now. Quite recent. The master_index page has an array of 1100+ hyperlinks arranged in neat HTML. This is what I tried to parse initially using web scraping with beautiful soup, but then I discovered a hyperlink slyly waiting in the page heading:

COMPLETE WORKS OF SWAMI VIVEKANANDA

Master Index of All Volumes

A detailed index of the entire contents of the *Complete Works* is available as an .xlsx file.
This file can be downloaded and re-sorted as desired.
(Please note that the file has dates in the yyyy/mm/dd format with city names spelt and country assignments as prevalent in Swamiji's times)

The detailed index is available as an .xlsx file already! No need for data scraping (at least for now)! I immediately clicked on the link to download the xlsx file and nothing happened. I tried again, same result. Was the link broken? This would not mean failure, just that I would have to construct this dataset myself and it would be gruelling work. Then before giving up, I tried to download it using some code from my colab console and it worked! I could see the entire dataset of index in front of my eyes.

This was the point when I was convinced that this project has been waiting for so long! An entire full dataset is available in such an orderly format and no one thought of trying to get insights using data science? It is not like the page is dormant, it gets thousands of active users every single day (there is a stat at the bottom of the homepage that shows visitors to the site on a worldmap)! Even if someone did, they did not post it publicly since I tried to search extensively for something and did not find it. It is quite odd that the ENTIRE data is available in such an orderly fashion for so long and no one thought to do this.. Anyways, the ordered index was way better than I imagined, here is a glimpse.

```
Searching for XLSX link in master index...

Found XLSX file: http://ramakrishnavivekananda.info/vivekananda/completeworksindex.xlsx
Full URL: http://ramakrishnavivekananda.info/vivekananda/completeworksindex.xlsx

Downloading XLSX file...
✓ XLSX file saved to: complete_works_index.xlsx
✓ File size: 155,172 bytes

================================================================================
EXPLORING XLSX STRUCTURE
================================================================================

Shape: 1450 rows × 10 columns
```

Column names:

1. Index
2. Volume
3. Name
4. Category
5. Type
6. Date
7. Place
8. Source
9. Audience
10. Language

Index	Volume	Name	Category	Type	Date	Place	Source	Audience	Language
1	1.0	Introduction	Editorial Foreword	—	1907/07/04	—	—	—	English
2	1.0	Response to Welcome	Parliament of Religions	Lecture	1893/09/11	USA, IL, Chicago	Newspaper Reports	Lecture Attendees	English
3	1.0	Why We Disagree	Parliament of Religions	Lecture	1893/09/15	USA, IL, Chicago	—	Lecture Attendees	English

4	1.0	Paper on Hinduism	Parliament of Religions	Lecture	1893/09/19	USA, IL, Chicago	—	Lecture Attendees	English
5	1.0	Religion not the Crying Need of India	Parliament of Religions	Lecture	1893/09/20	USA, IL, Chicago	—	Lecture Attendees	English
6	1.0	Buddhism, the Fulfilment of Hinduism	Parliament of Religions	Lecture	1893/09/26	USA, IL, Chicago	—	Lecture Attendees	English
7	1.0	Address at the Final Session	Parliament of Religions	Lecture	1893/09/27	USA, IL, Chicago	—	Lecture Attendees	English
8	1.0	Karma in its Effect on Character	Karma-Yoga	Book	—	—	—	Book Readers	English
9	1.0	Each is great in his own place	Karma-Yoga	Book	—	—	—	Book Readers	English

10	1.0	The Secret of Work	Karma-Yoga	Book	—	—	—	Book Readers	English
11	1.0	What is Duty?	Karma-Yoga	Book	—	—	—	Book Readers	English
12	1.0	We help ourselves, not the world	Karma-Yoga	Book	—	—	—	Book Readers	English
13	1.0	Non-attachment is complete self-abnegation	Karma-Yoga	Book	—	—	—	Book Readers	English
14	1.0	Freedom	Karma-Yoga	Book	—	—	—	Book Readers	English
15	1.0	The Ideal of Karma-Yoga	Karma-Yoga	Book	—	—	—	Book Readers	English
16	1.0	Preface	Raja-Yoga	Book	—	—	—	Book Readers	English
17	1.0	Introductory	Raja-Yoga	Book	—	—	—	Book Readers	English

18	1.0	The First Steps	Raja-Yoga	Book	—	—	—	Book Readers	English
19	1.0	Prana	Raja-Yoga	Book	—	—	—	Book Readers	English
20	1.0	The Psychic Prana	Raja-Yoga	Book	—	—	—	Book Readers	English

```
================================================================================
DATA TYPES & MISSING VALUES
================================================================================
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1450 entries, 0 to 1449
Data columns (total 10 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   Index      1450 non-null   object
 1   Volume     1450 non-null   float64
 2   Name       1450 non-null   object
 3   Category   1447 non-null   object
 4   Type       1277 non-null   object
 5   Date       1201 non-null   object
 6   Place      1193 non-null   object
 7   Source     358 non-null    object
 8   Audience   985 non-null    object
 9   Language   1410 non-null   object
dtypes: float64(1), object(9)
memory usage: 113.4+ KB
None
```

Now immediately as soon as I saw this, I was very excited. The boring work of data collection and annotation had already been done and in such a good way, all I had to do was analyze! Looking at the dataset, some things are immediately visible, Index, Volume and Name columns are populated for each entry (I will be using columns or col for features and entry or row for datapoint) Others were not quite full, for example col Category has just 3 mising values for 3 rows. Source is very scarce, less than 10% of data has sources for it. Language, another one that should be easily inferrable from data itself, is missing for 40 entries. I decided to work on this later. But more importantly, than all these…where was the actual content?

Crucially, this is just the index, and not the complete works themselves. That means, the complete works themselves, the words of Swami Vivekananda, are not included in this table. If we did not have links to the actual works themselves, that entry was essentially unusable for us since the actual content would be missing. How would we get the actual content for where it exists?

Well, turns out I still had some work to do. We needed to make a bridge to actual content on the site, thankfully each single work (be it a single letter, a single lecture, a single book etc.) was all neatly placed onto one single html page, all of which were linked to in the master_index.htm. So I scraped all the links and tried to match them with the dataset of nice cleaned up index I already had using various exact matching, fuzzy matching and html-parsing techniques. This has all been shown in SwamijiAnalysis.ipynb notebook, which was the first notebook I made for this project. At the end, I had something quite neat.

```
================================================================================
FINAL MASTER DATASET SUMMARY
================================================================================

Total entries: 1466
With HTML links: 1418 (96.7%)
Without HTML links: 48

By genre (with HTML):
  Letters        :  774 documents
  Lectures       :  266 documents
  Reports        :  140 documents
  Writings       :   96 documents
  Other          :   53 documents
  Treatises      :   50 documents
  Conversations  :   39 documents

By language (with HTML):
  English        : 1139 documents
  Bengali        :  213 documents
```

```
  Sanskrit       :   11 documents
  Hindi          :    3 documents
  French         :    3 documents
   English       :    1 documents

Metadata coverage (with HTML):
  Has Date:     1174 / 1418 (82.8%)
  Has Place:    1177 / 1418 (83.0%)
  Has Audience:  973 / 1418 (68.6%)

================================================================================
MASTER DATASET IS READY FOR SCRAPING!
================================================================================

Next Steps:
1. Scrape HTML content for 1418 documents
2. Add word counts and text content to master dataset
3. Decide on prototype scope for initial analysis
4. Begin exploratory corpus analysis

The master dataset (master_dataset_final.csv) contains:
- Rich metadata (volume, category, type, date, place, audience, language)
- Genre classification (letters, lectures, treatises, etc.)
- HTML links for scraping
- Ready for content extraction
```

So as we can see, metadata was working correctly and I had a very rich dataset with high completion rates across all columns/features. The links were there, all I had to now do was scrape the links themselves and get the complete works.

Now we come into SwamijiAnalysis_1.ipynb, the second notebook in the explorations.
I noticed that there were many newspaper reports on Swamiji, now these reports were not written by Swamiji themselves but ABOUT swamiji, written by different newspapers. The aim of my project had clear scopes, we only take words written or spoken (later transcribed) by Swamiji himself ALONE, nothing else would qualify. So I removed the reports and I was left with this.

```
Documents after removing Reports: 1278

Genre distribution:
genre
Letters          774
Lectures         266
Writings          96
Other             53
Treatises         50
Conversations     39
Name: count, dtype: int64

============================================================
Volume distribution:
volume
1.0     51
2.0     34
3.0     56
4.0     65
5.0    179
6.0    216
7.0    155
8.0    246
9.0    276
Name: count, dtype: int64
```

Further in the SwamijiAnlaysis_1.ipynb, if one goes through the cells, one will see that I began to now scrape a sample of the html from the links I merged with the dataset previously. When one scrapes, one gets the html page, it is not in plaintext, it contains a lot of html code that renders a page nicely instead of just plaintext which is what we want for our purposes, so I begin to figure out how I can parse the html to only get the plaintext from each of the html pages I get. I realize furthermore that conversations are very difficult to parse only for Swamiji's words, since they contained Sharat Chandra Chatterjee's own observations and narratives along with Swamiji's words, so instead of focusing on extracting less than 1% of the total works, I removed conversations completely as well from the corpus to analyze. I then make a genre-wise strategy to parse all the html pages and successfully get the format of the actual textual data that I need

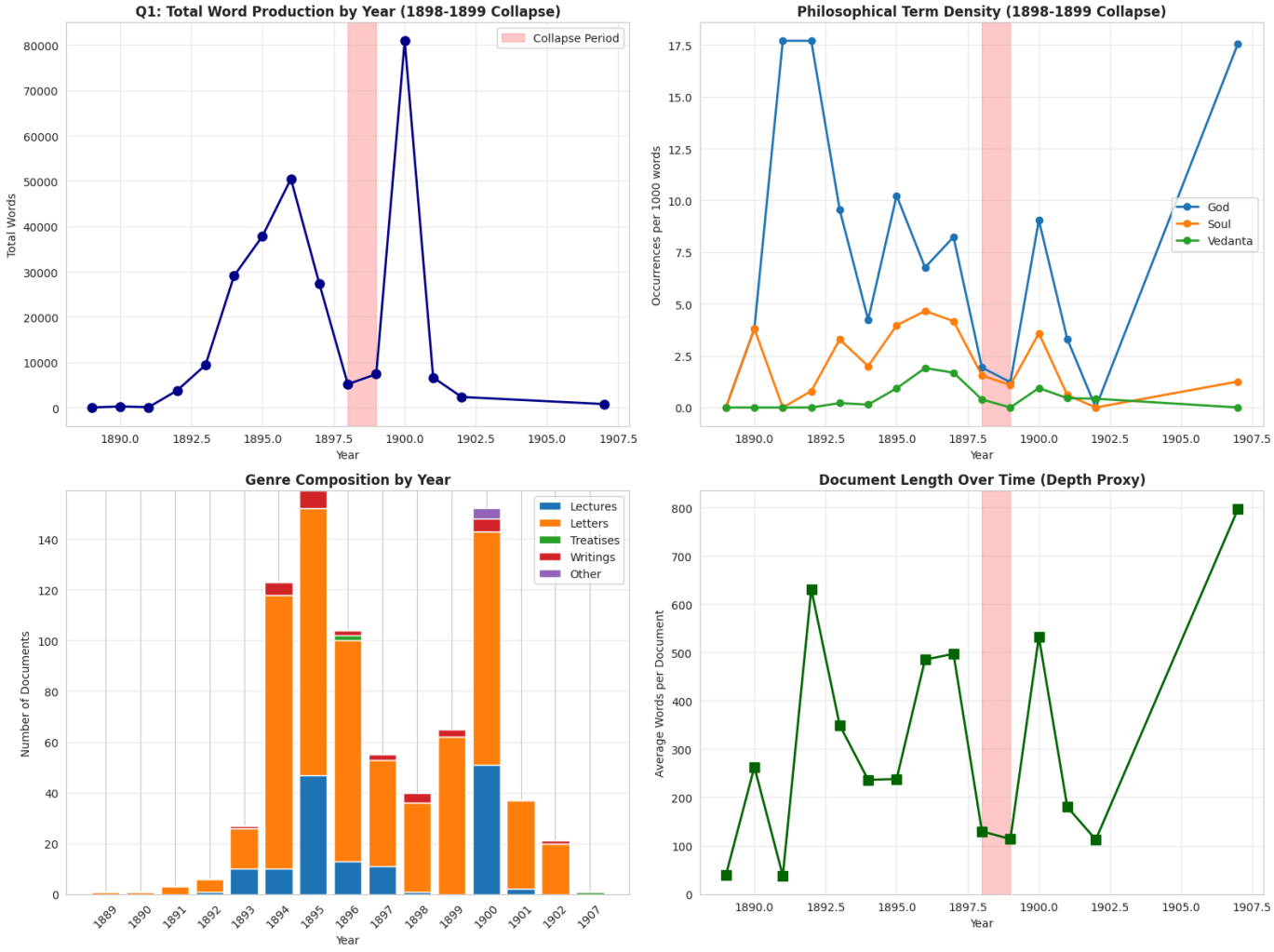
(plaintext), I merge it with the existing dataset, resulting in a master dataset that contained all the currently documented words of Swami Vivekananda along with rich metadata wherever available.

Furthermore, I removed poems as well (only 7-8 instances) since they contributed very little to the corpus I was interested in. In retrospect, it would have been interesting to analyze the poems separately and see where and when they were composed etc. in comparison to the rest of the corpus. Eitherway, there will always be something more to be done!

Now that we had the full dataset, I was ready to do some visualizations. Here is how the dataset looks:

```
============================================================
FINAL STATISTICS
------------------------------------------------------------
Documents with 0 words: 0
Total documents: 975
Total words: 1,033,863

✓ Saved to: swamiji_plaintext.csv
```

More than a million words, straight from an enlightened nitya-siddha sage Swami Vivekananda! What a treat indeed.
I made the following plots first:

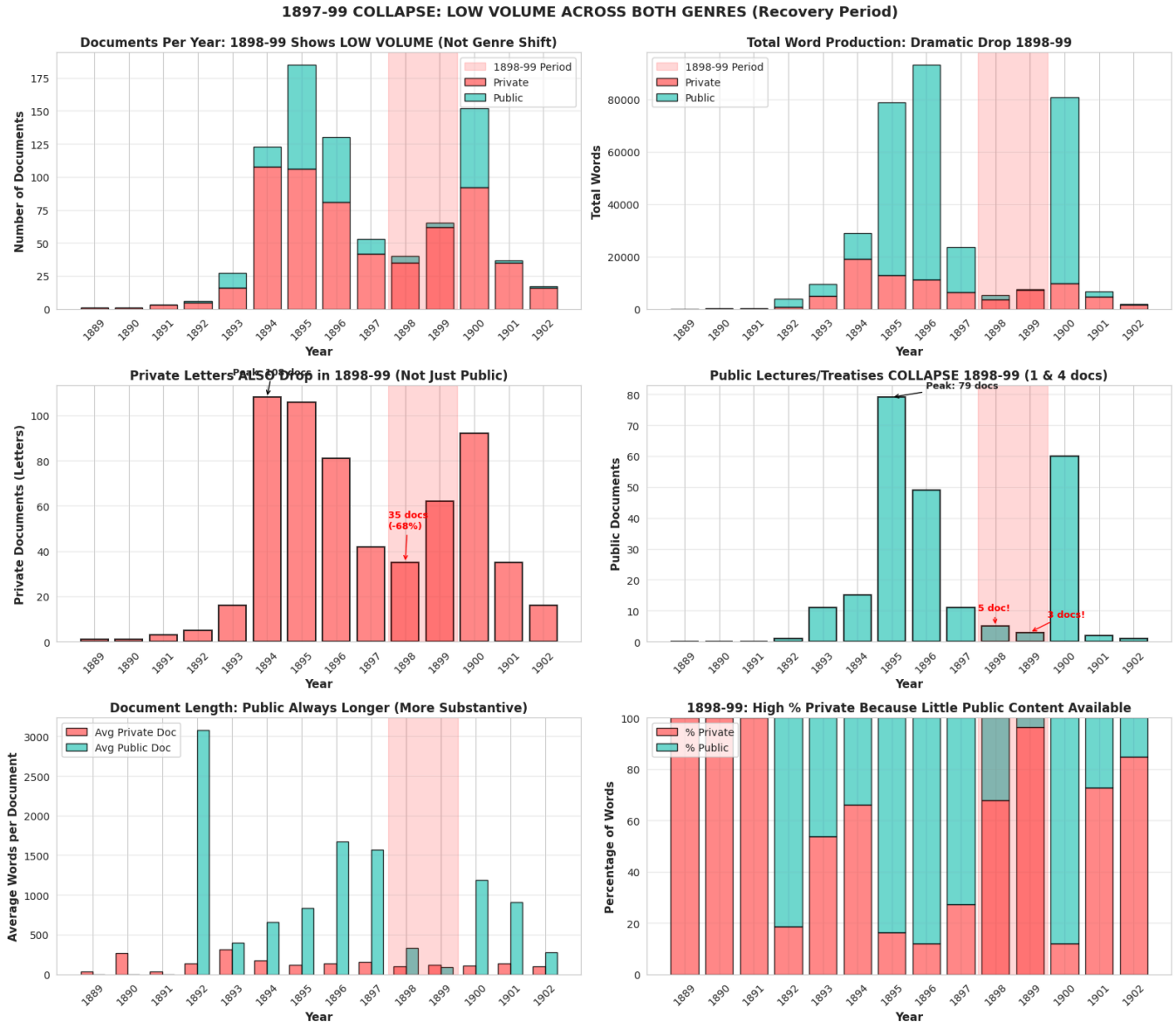


These plots are quite basic, just a volumetric analysis of the number of words by year, genre and depth proxy. Another basic thing is the “philosophical term density” using only three terms, almost childish. These are very basic, and not representative of the kind of analysis that will follow. I was just excited.
But one very nice thing that one can see immediately is that 1897 was the year that produced the least amount of Swamiji’s works documented works. The genrewise plot shows the number of works generated (one lecture, one letter, one essay counts as one work or document), and not the volume of words in document which could be a better indicator. So I made that in the next set of plots.

Private_vs_public_comprehensive.png

Now we can see, that the word volume of public works (treatises, lectures, writings etc.) is a lot more in general than epistles, this is natural since letters on average are quite small as compared to a lecture. But curiously, in the 1897 period of low works, the public works are lesser (word volume wise) than letter word volume, hinting that swamiji stopped public works in that period for some reason (we can talk about interpretations later) while the letters did not see that much of a volumetric decrease!
Also, let us ignore the “God” usage and “Brahman” usage etc. plots here since there will be a much more indepth analysis of these coming up. Now we can move

onto the third notebook in the series, Swamiji_Analysis2.ipynb. I immediately do some tokenization and jump into some nice analysis.



This has both document wise analysis (remember, document is any one complete body of work, like one letter, one lecture, one treatise, one booklet without duplication).1897-99 period again visually stands out, it cleanly corresponds to the period when Swamiji was back from a long stay in the west and was busy doing organisational work in India like founding the Ramakrishna Order, and recuperating from the busy life in the west.

The next analysis of the entire corpus as a whole will mark a decisive point in the analysis. I decided to do some syntactic and semantic analysis, much deeper than just word and document number analysis.

Here I use methods like TF-IDF, LDA and Outlier detection for some good analysis. Here is a very very basic rudimentary analysis of all the methods used here.

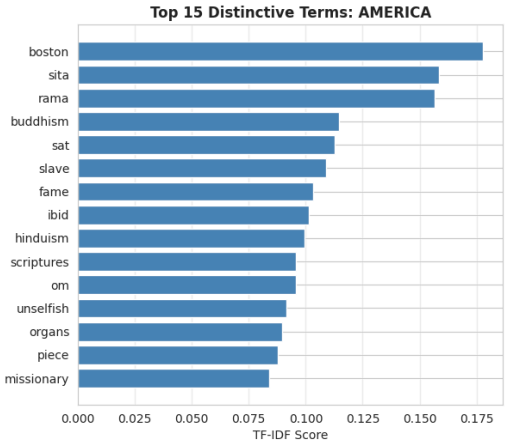
TF-IDF finds words that are important by rewarding words that appear often in one group of texts but not across all texts, helping identify what makes each location or genre linguistically unique.

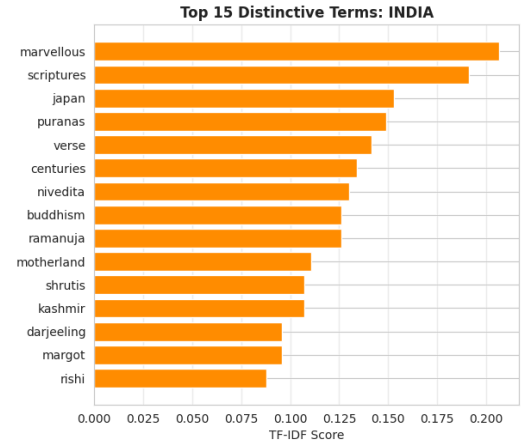
Topic modeling (LDA) automatically discovers hidden themes by grouping words that frequently appear together and estimating how much each theme appears in every document. LDA is the name of the algorithm that does this, it is quite famous.

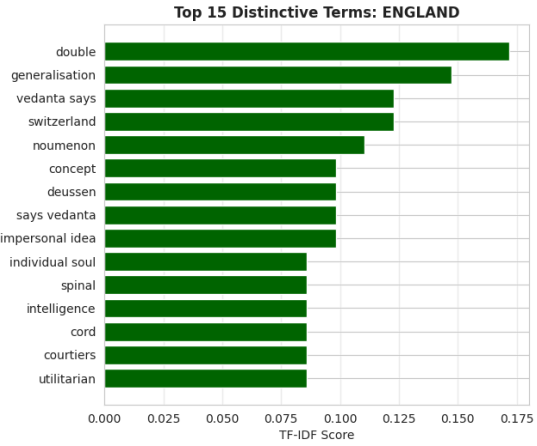
Outlier detection then finds documents whose topic patterns differ strongly from others in the same location, highlighting unusual or exceptional texts.

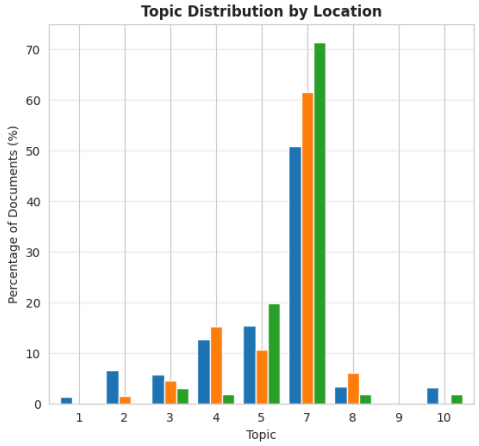
Together, these methods work like a layered understanding of a text collection: **TF-IDF** shows *what language is distinctive*, **topic modeling (LDA)** reveals *what ideas and themes are present*, and **outlier detection** identifies *which documents don't fit the usual patterns*. Combined, they move from surface vocabulary → deeper meaning → exceptions, giving a coherent, multi-level picture of how a corpus is structured, what it talks about, and where it diverges.

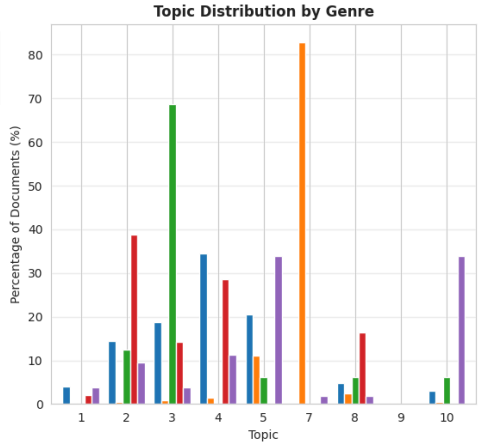
Here are the important plots that this section produced, which changed the course of the next analysis which followed.

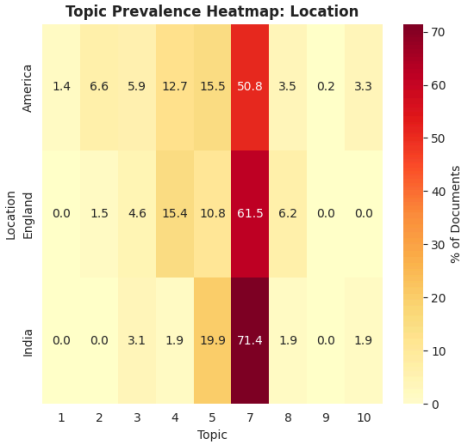

Top 15 Distinctive Terms: AMERICA
boston
sita
rama
buddhism
sat
slave
fame
ibid
hinduism
scriptures
om
unselfish
organs
piece
missionary
0.000 0.025 0.050 0.075 0.100 0.125 0.150 0.175
TF-IDF Score

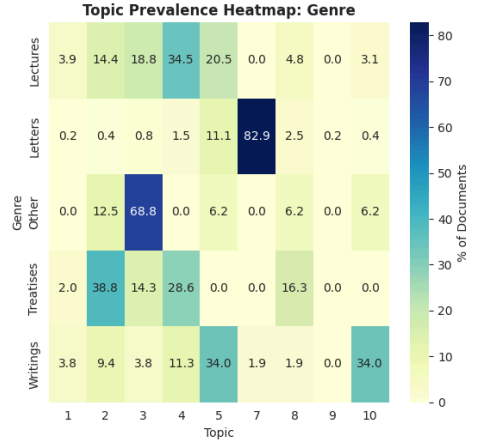

Top 15 Distinctive Terms: INDIA
marvellous
scriptures
japan
puranas
verse
centuries
nivedita
buddhism
ramanuja
motherland
shrutis
kashmir
darjeeling
margot
rishi
0.000 0.025 0.050 0.075 0.100 0.125 0.150 0.175 0.200
TF-IDF Score

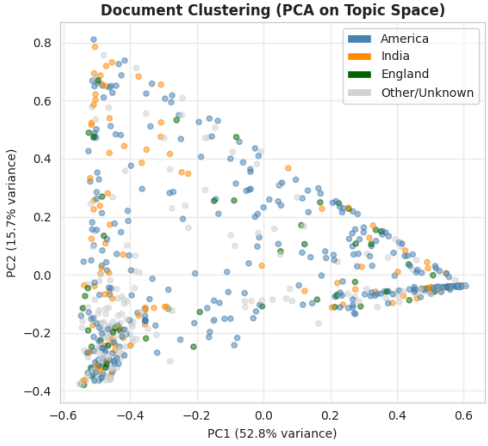

Top 15 Distinctive Terms: ENGLAND
double
generalisation
vedanta says
switzerland
noumenon
concept
deussen
says vedanta
impersonal idea
individual soul
spinal
intelligence
cord
courtiers
utilitarian
0.000 0.025 0.050 0.075 0.100 0.125 0.150 0.175
TF-IDF Score

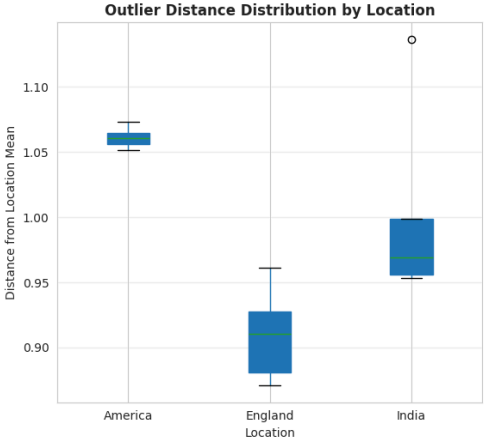

Topic Distribution by Location
Percentage of Documents (%)
Topic
Location
America
England
India


Topic Distribution by Genre
Percentage of Documents (%)
Topic
Genre
Lectures
Letters
Other
Treatises
Writings


Topic Prevalence Heatmap: Location
America
England
India
Location
Topic
% of Documents


Topic Prevalence Heatmap: Genre
Lectures
Letters
Other
Treatises
Writings
Genre
Topic
% of Documents


Document Clustering (PCA on Topic Space)
America
India
England
Other/Unknown
PC1 (52.8% variance)
PC2 (15.7% variance)


Outlier Distance Distribution by Location
Distance from Location Mean
America
England
India
Location

General_analysis_topicmodelling_10_topics.png

This has some real insights now. The top three graphs show the most distinctive terms in Swamiji's works by location. Immediately we can see that In America, swamiji has talked a lot about Lord Rama, Mother Sita, Buddhism and Hinduism. Since this is only one word without context, it is difficult to assume a great many things but something we can infer is that swamiji was talking about Indian epics and general terms like hinduism and buddhism as a whole in the west.
In India, in contrast, we see clearly more technical terms like puranas and shrutis and ramanuja! Seems like Swamiji reserved intricacies of his worldview for India.
In England, we can further see the differentiation with words like noumenon, deussen, utilitarian, he seems to be using western philosophy to make a case for Vedanta!

Now, as explained earlier, topic distribution tries to find topics across a corpus of texts based on things like word frequencies. This method is not semantic at all. LDA depends **entirely on word-frequency statistics**: how often words appear, how often they co-occur, and how those patterns repeat across documents. LDA has **no understanding of meaning, context, or syntax**; it does not know that "God" and "Divine" are related unless they frequently appear in the same documents.

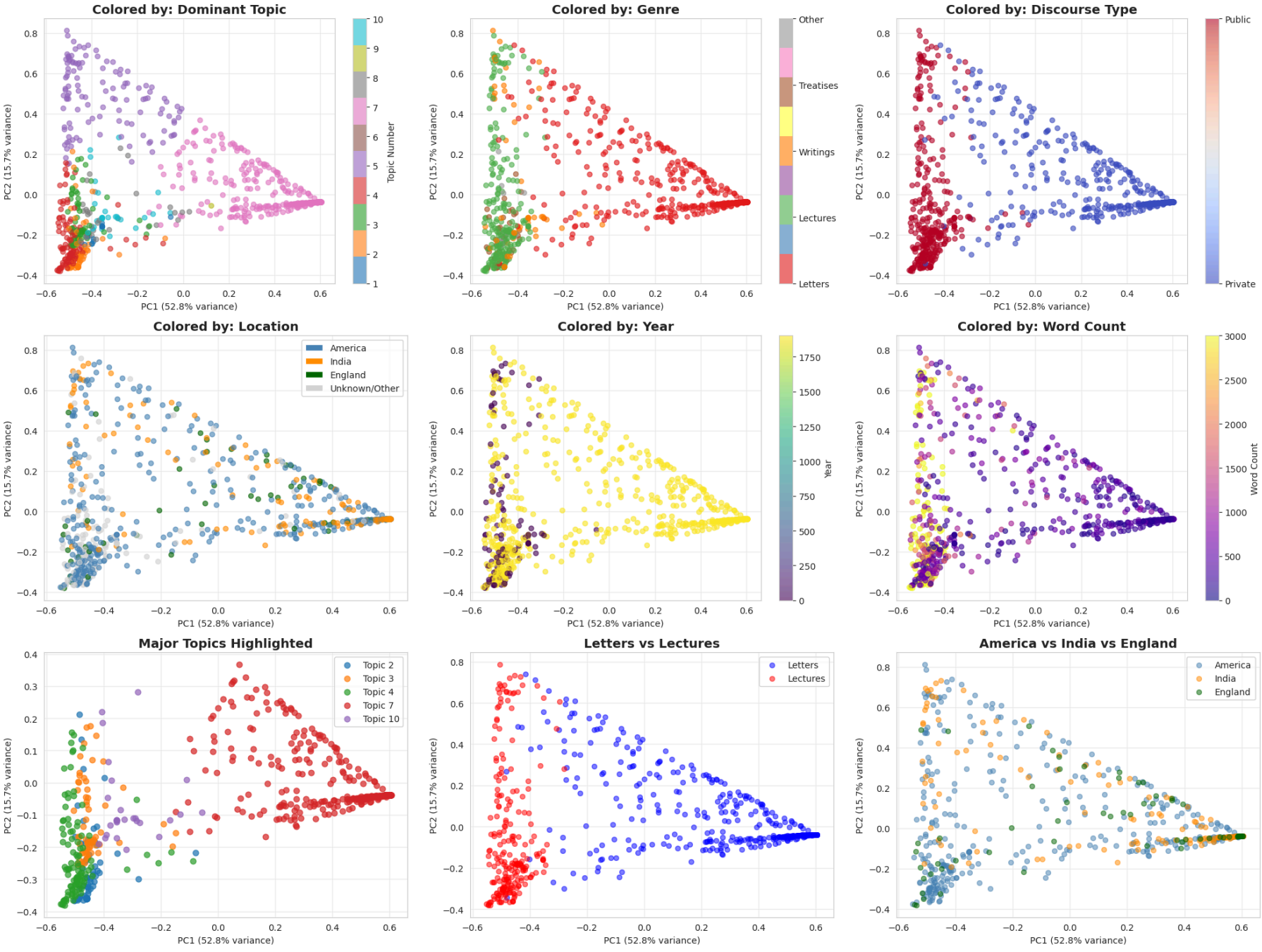
So what I did was I asked the code to find 10 topics based on these patterns across swamiji's corpus into which documents can be classified. It gave me 10 topics like I asked. It did not work very well since most of the volume of documents were classified into one topic, topic 7. This is reason I forego this analysis and later on do LDA with 5 topics instead of 10, which yields better results.

Now, the most striking thing was PCA or Principal Component Analysis. **PCA (Principal Component Analysis)** is just a way to **compress many numbers into two directions** so we can draw them on a 2-D plot.
Here, each document is originally a point in **topic space** (for example, 10 topic percentages), and PCA finds the **two strongest patterns of variation** (PC1 and PC2) that explain how documents differ from each other.

You notice how it forms a triangle, that is ont expected at all! Usually that same plot would be a blob of points with no real identifiable shape! This means something very significant. It means that something **strong and unusual** is happening:
The corpus is governed by **a small number of dominant topic archetypes**
Documents are **convex combinations** of those archetypes (mixtures, not random)
Topic modeling has captured **real, global structure**, not just noise

Any data scientist would look at this and be overjoyed. Now the next question was clearly this, **when we plot the documents of swamiji's works present in a 10-D topic space into a 2D plot on two most different (orthogonal) axes (PCA does exactly this), why does a triangle emerge out of nowhere?!**

So I try to do some detective work. I take that triangle and colour the documents according to the different features (columns) that we have, trying to find any pattern that is nontrivial.


Colored by: Dominant Topic
Topic Number
Colored by: Genre
Other
Treatises
Writings
Lectures
Letters
Colored by: Discourse Type
Public
Private
Colored by: Location
America
India
England
Unknown/Other
Colored by: Year
Year
Colored by: Word Count
Word Count
Major Topics Highlighted
Topic 2
Topic 3
Topic 4
Topic 7
Topic 10
Letters vs Lectures
Letters
Lectures
America vs India vs England
America
India
England
PC1 (52.8% variance)
PC2 (15.7% variance)

Corpus_pca_tentopics.png

This graph reveals several nontrivial details. There is some underlying structural uniformity. The most striking and obvious ones are the ones titles “lectures vs lectures” bottom middle and “discourse type” top right (which are essentially the same graphs just titled differently). We can draw a straight line through the triangle almost completely separating letters and lectures. This was a major discovery that led to the bifurcation of our analyses into two halves, one solely for public works and the next solely focusing on epistles, private letters which were not meant for the public.

Another nice thing is that, from the top left graph, we can see that topic 5 becomes the top corner of the triangle, topic 7 becomes the right corner and all the other topics are concentrated into the left corner. This would suggest that there are three major sections of topics, and in retrospect I should have only done LDA with 3 topics keeping this in mind but either way, I decided to have some nice variety so I went with 5 topics instead.

From this point on, I separated the dataset I was analyzing (complete corpus) into public and private and I focused on public part of the dataset first, since it was volumetrically more significant (despite being less interesting in comparison to swamiji’s letters, in my opinion at least).

==========
Part 1 : Public Corpus Analysis
==========

Now I do an n-gram analysis on public corpus. **N-gram analysis** looks for patterns of words that appear **together and in order** in a corpus, instead of analyzing single words one by one.
When the pattern has **two consecutive words**, it’s called a **bigram**; **three words** form a **trigram**; and in general, a sequence of **n words** is called an **n-gram**.
This helps capture **common phrases and expressions** (like “divine mother” or “pure consciousness”) that single-word analysis would miss. This was done manually before the era of computers with great effort and it was called collocation.

Here are the results pasted completely

```
================================================================================
LECTURES ONLY: CROSS-LOCATION LINGUISTIC ANALYSIS
================================================================================

Total lectures: 229
Lectures with known major locations: 169
  America: 131
```

```
  India: 25
  England: 13



================================================================================
4. SENTENCE COMPLEXITY: LOCATION AND TEMPORAL PATTERNS
================================================================================

--------------------------------------------------------------------------------
COMPLEXITY BY LOCATION
--------------------------------------------------------------------------------

location  avg_sentence_len  avg_word_len  type_token_ratio  num_lectures
 America         14.996899      4.270561          0.363518           131
   India         23.794655      4.353176          0.287026            25
 England         17.163469      4.283550          0.243449            13

--------------------------------------------------------------------------------
COMPLEXITY BY YEAR
--------------------------------------------------------------------------------

 year  avg_sentence_len  avg_word_len  type_token_ratio  num_lectures
 1893         21.881257      4.331425          0.446212            10
 1894         15.749513      4.333793          0.367424            10
 1895         13.675291      4.284366          0.426711            47
 1896         17.579385      4.226071          0.203986            13
 1897         24.274426      4.263933          0.281314             6
 1898         20.388350      4.277143          0.291429             1
 1900         13.521042      4.233838          0.308125            50
 1901         22.543921      4.430185          0.342655             2
```

Many insights can be gained from this. In India, these are the most significant bigrams found

```
INDIA:
   1. 'the puranas                  ' (n= 27, 27.0x more)
   2. 'national life                ' (n= 26, 8.7x more)
   3. 'our religion                 ' (n= 53, 7.6x more)
   4. 'the national                 ' (n= 25, 3.6x more)
   5. 'the brahmin                  ' (n= 37, 2.6x more)
```

It seems to suggest that swamiji was delving into the puranas more than the vedas in India, and his tone was very unifying and nationalistic “our religion” and “national life”. Also seems to be talking about caste issues by citing ‘the brahmin’ three times more than anywhere else.

The trigrams that appear, similarly, are

```
INDIA:
   1. 'the karma kanda                      ' (n=12, 12.0x more)
   2. 'this the land                        ' (n=11, 11.0x more)
   3. 'love the gopis                       ' (n= 8, 8.0x more)
   4. 'this motherland ours                 ' (n= 7, 7.0x more)
   5. 'from the west                        ' (n= 7, 7.0x more)
   6. 'india this the                       ' (n= 7, 7.0x more)
   7. 'our national life                    ' (n= 7, 7.0x more)
   8. 'every one who                        ' (n= 7, 7.0x more)
   9. 'nations the world                    ' (n=13, 6.5x more)
  10. 'the nations the                      ' (n=12, 6.0x more)
  11. 'religion and religion                ' (n= 6, 6.0x more)
  12. 'teach the world                      ' (n= 6, 6.0x more)
  13. 'and every one                        ' (n=10, 5.0x more)
  14. 'the different sects                  ' (n=10, 5.0x more)
  15. 'the western world                    ' (n= 9, 4.5x more)
```

As we can see, swamiji seems to be critiquing the karmakanda part of religion. Nationalistic tone is clear from “this the land” and “our national life” and “this motherland ours”. Gopis are a very interesting example here, he seems to have invoked the tenth canto of the Srmad Bhagavatam a lot in India.

In America, again some patterns emerge using ngrams.

```
AMERICA:
   1. 'the mother                   ' (n= 86, 10.8x more)
```

```
   2. 'the priests                ' (n= 64, 7.1x more)
   3. 'the spirit                 ' (n= 82, 4.3x more)
   4. 'the father                 ' (n= 52, 4.0x more)
   5. 'the mind                   ' (n=304, 3.8x more)
   6. 'let the                    ' (n= 54, 3.6x more)
   7. 'all this                   ' (n=108, 3.5x more)
   8. 'the hindu                  ' (n=123, 3.3x more)
   9. 'can never                  ' (n= 79, 3.2x more)
  10. 'not want                   ' (n= 59, 3.1x more)
  11. 'that god                   ' (n= 60, 3.0x more)
  12. 'they will                  ' (n= 86, 3.0x more)
  13. 'him and                    ' (n= 71, 3.0x more)
  14. 'but not                    ' (n= 73, 2.9x more)
  15. 'the king                   ' (n= 75, 2.9x more)
  16. 'the question               ' (n= 69, 2.9x more)
  17. 'the body                   ' (n=221, 2.8x more)
  18. 'nature and                 ' (n= 65, 2.8x more)
  19. 'would not                  ' (n= 62, 2.8x more)
  20. 'men and                    ' (n= 59, 2.8x more)
AMERICA:
   1. 'the power meditation                ' (n=13, 13.0x more)
   2. 'control the mind                    ' (n=11, 11.0x more)
   3. 'the time the                        ' (n=14, 7.0x more)
   4. 'the priests and                     ' (n=13, 6.5x more)
   5. 'the body that                       ' (n=13, 6.5x more)
   6. 'the name the                        ' (n=12, 6.0x more)
   7. 'this world and                      ' (n=12, 6.0x more)
   8. 'know the truth                      ' (n=17, 5.7x more)
   9. 'you are all                         ' (n=28, 5.6x more)
  10. 'man can not                         ' (n=11, 5.5x more)
  11. 'there such thing                    ' (n=16, 5.3x more)
  12. 'what the use                        ' (n=16, 5.3x more)
  13. 'but can not                         ' (n=20, 5.0x more)
  14. 'not know what                       ' (n=19, 4.8x more)
  15. 'the lord and                        ' (n=14, 4.7x more)
```

Swamiji can be seen shifting his vocabulary for his western audience, using words like “the father” and “the spirit”. However, he not afraid to appeal to the feminine using “the mother”. This theme of empowering women appears multiple times in our analyses, as we will see. One can assume that swamiji was very impressed to see the female suffragette movement taking initial form in the 1890s, and decided to vocally support the upliftment of women. We will see in future analyses that we can track swamiji’s revolutionary feminist stance during his first visit to the west.

Trigrams reveal some more themes that swamiji wanted to teach to the west, like controlling the mind and the power of meditation. This is also visible in his famously known works like Raja Yoga and Jnana Yoga among others.

```
ENGLAND:
   1. 'the vedanta says                        ' (n= 7, 7.0x more)
   2. 'the general the                         ' (n= 5, 5.0x more)
   3. 'that the personal                       ' (n= 5, 5.0x more)
   4. 'worship the impersonal                  ' (n= 5, 5.0x more)
   5. 'says the vedanta                        ' (n= 8, 4.0x more)
   6. 'the modern man                          ' (n= 4, 4.0x more)
   7. 'that the explanation                    ' (n= 4, 4.0x more)
   8. 'the idea personal                       ' (n= 4, 4.0x more)
   9. 'idea the impersonal                     ' (n= 4, 4.0x more)
  10. 'the same impersonal                     ' (n= 4, 4.0x more)
  11. 'that the impersonal                     ' (n= 4, 4.0x more)
  12. 'manifestation that one                  ' (n= 4, 4.0x more)
  13. 'one making for                          ' (n= 4, 4.0x more)
  14. 'this sameness for                       ' (n= 4, 4.0x more)
  15. 'the spinal cord                         ' (n= 7, 3.5x more)
```

England seems nice in this regard too, looking at the trigrams alone. One can see that number 15 is “the spinal cord”. I would believe that swamiji was praising the spine of the English folk, something he also expressed in his private letters. Number 4,9,10,11 show that Swamiji was very keen to stress the “impersonal” aspect of Vedantic truth, perhaps because in Britain a personal God (something that appears in number 3, perhaps used to contrast to the impersonal) was already out of fashion due to the falling of the church after the french revolution. I am not a historian, but anyways this is just me assuming. I am not a historian.

```
================================================================================
3. VOCABULARY RICHNESS OVER TIME (LECTURES)
================================================================================
```

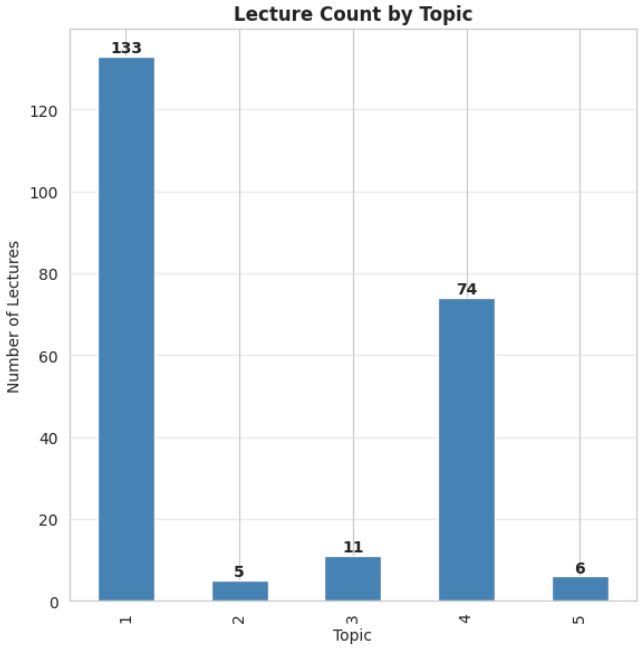
year	total_words	unique_words	type_token_ratio	num_lectures
1893	7082	1746	0.246541	10
1894	13156	2718	0.206598	10
1895	34012	4276	0.125720	47
1896	52796	4657	0.088207	13
1897	20456	3122	0.152620	6
1898	1600	584	0.365000	1
1900	106211	7181	0.067611	50
1901	3090	1078	0.348867	2

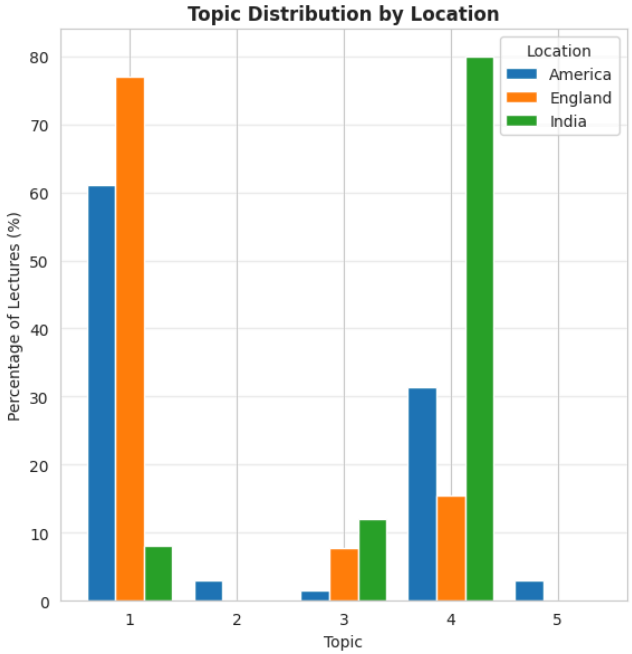
Now here is a table showing yearwise vocabulary richness. The TTR (which is unique words divided by total words) alone can be considered since it combines the two columns before.

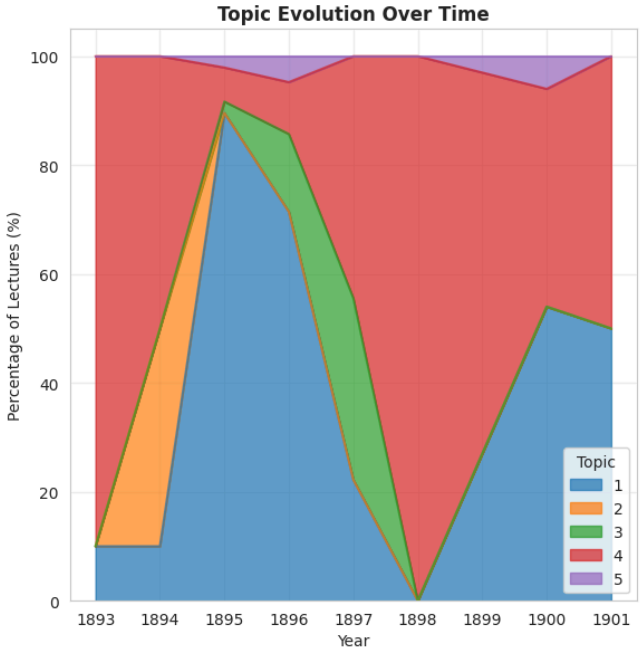
There is a strong inverse relationship between text size and TTR, Years with many words (1896, 1900) have low TTR, Years with few words (1898, 1901) have high TTR. This is exactly what TTR is known to do.

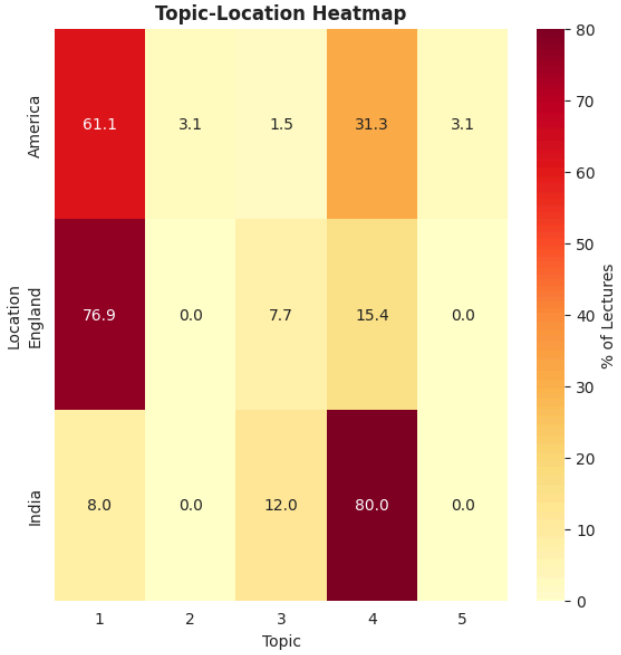
For serious comparison across years, one would typically use length-normalized measures (e.g., moving-average TTR, MTLD), but I was not interested in that since we had many more analyses that were more interesting.

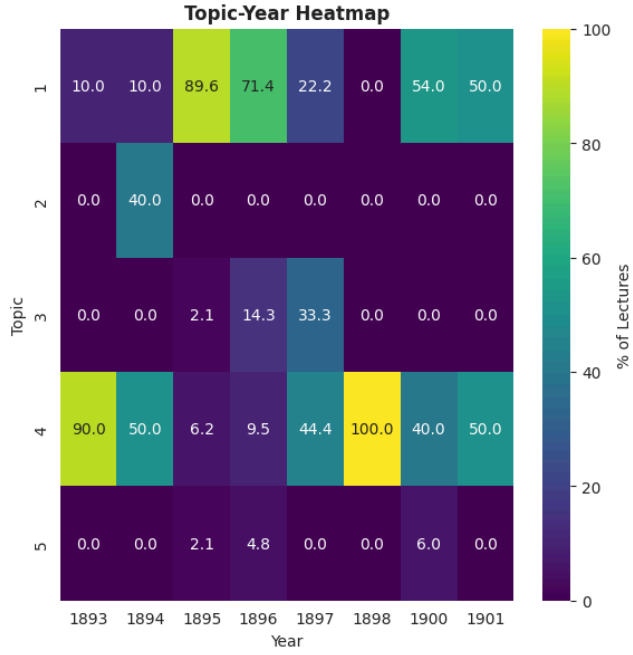
After this, I ran an LDA with 5 topics instead of 10. This is because I saw that 10 was too much, and 3 was too less for variety, although we had some statistical support that there were three strong directions of variation. Here is the graph

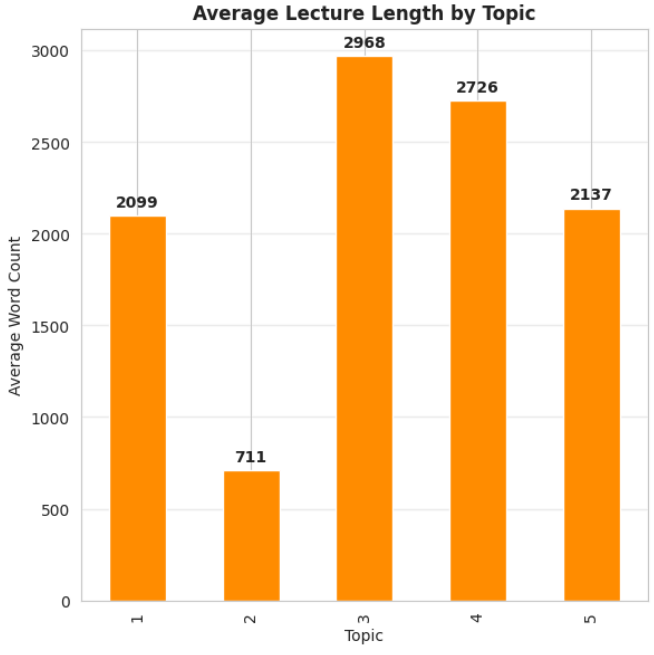

Lecture Count by Topic
Number of Lectures
Topic
133
5
11
74
6


Topic Distribution by Location
Percentage of Lectures (%)
Topic
Location
America
England
India


Topic Evolution Over Time
Percentage of Lectures (%)
Year
Topic
1
2
3
4
5


Topic-Location Heatmap
Location
America
England
India
61.1
3.1
1.5
31.3
3.1
76.9
0.0
7.7
15.4
0.0
8.0
0.0
12.0
80.0
0.0
Topic
% of Lectures


Topic-Year Heatmap
Topic
Year
10.0 10.0 89.6 71.4 22.2 0.0 54.0 50.0
0.0 40.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 2.1 14.3 33.3 0.0 0.0 0.0
90.0 50.0 6.2 9.5 44.4 100.0 40.0 50.0
0.0 0.0 2.1 4.8 0.0 0.0 6.0 0.0
% of Lectures


Average Lecture Length by Topic
Average Word Count
Topic
2099
711
2968
2726
2137

Now we get 5 topics, and the lecture content seems to be distributed across topics 1 and 4 in high volume. From the alluvial graph in top right, we can also see that red and blue (topic 4 and 1 resp.) are much more dominant than others. The heatmap on lower center is interesting.

We can see that through topic modeling alone we were able to separate most of the work done in 1893, 1894 and 1898 from works done in 1895 and 1896. This is interesting because remember, the dates were not fed into the LDA algorithm, it only worked based on patterns of word frequencies! This means there is something similar in the choice of words that Swamiji used (along with how he used it, in a basic syntactic sense) in 1893, 1894 and 1898 that can be differentiated from 1896 and 1895. Location heatmap on bottom left reveals similar things, 80% of lectures in India fall in topic 4 cleanly, this again reveals that there is something in the syntax of these 80% of lectures that swamiji gave in India which has been captured by the LDA algorithm under topic 4. Curiously, topic 2 and 5 contain only American lectures and no Indian or England ones.

After one does topic extraction, it is common practice to understand what these topics mean semantically by looking at the word patterns for each topic. This allows renaming these topics to something meaningful rather than arbitrary numeric names like “topic 1” etc. Here are the common word patterns for the five topics so extracted.

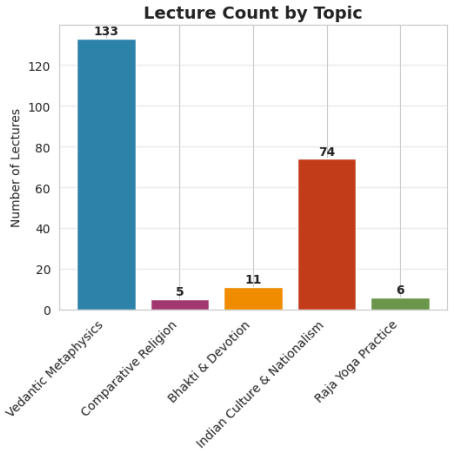
```
================================================================================
DISCOVERED TOPICS (Top 30 words per topic)
================================================================================

--------------------------------------------------------------------------------
TOPIC 1:
--------------------------------------------------------------------------------
love | knowledge | infinite | self | truth | free | little | matter | spirit | thought | human | says | thing | evil | want
| real | freedom | work | existence | form | vedanta | senses | highest | let | just | true | worship | lord | different |
reason

--------------------------------------------------------------------------------
TOPIC 2:
--------------------------------------------------------------------------------
said | sheep | lion | kananda | vive | vive kananda | believe | religions | existence | theory | came | speaker | space |
truth | away | motion | believed | creation | punishment | elephant | buddhism | better | christian | christians | night |
period | let | buddha | went | cause

--------------------------------------------------------------------------------
TOPIC 3:
--------------------------------------------------------------------------------
```

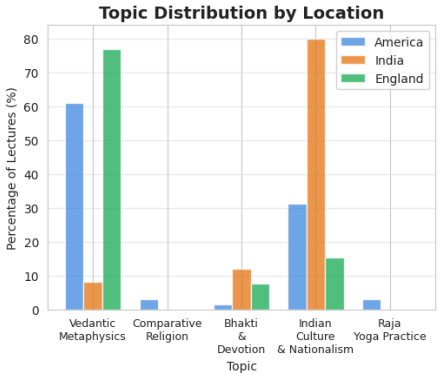
```
worship | form | bhakti | forms | love | gross | called | fine | finer | material | thought | image | eternal | just | cause
| certain | mother | different | ishta | worship god | external | gods | said | shiva | goes | human | temple | infinite |
effect | people

--------------------------------------------------------------------------------
TOPIC 4:
--------------------------------------------------------------------------------
people | love | work | hindu | country | came | race | little | said | mother | years | want | vedas | nation | let | ideal
| did | different | ideas | hindus | spiritual | woman | old | women | day | king | religions | went | worship | just

--------------------------------------------------------------------------------
TOPIC 5:
--------------------------------------------------------------------------------
ibid | senses | food | knowledge | work | misery | human | people | says | plane | sense | yogi | doing | control | want |
attachment | goal | animal | thought | known | animals | little | help | present | breath | arjuna | question | study |
standard | believe
```

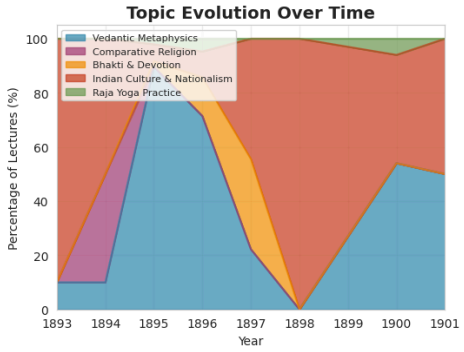
Based on these words, I gave the following appropriate names to the topics.

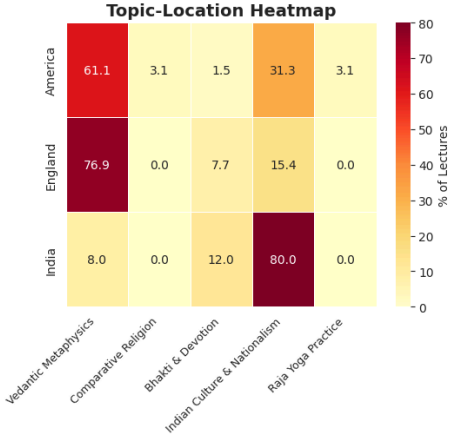
```
 Topic 1: Vedantic Metaphysics           (133 lectures, 58.1%)
  Topic 2: Comparative Religion           (  5 lectures, 2.2%)
  Topic 3: Bhakti & Devotion              ( 11 lectures, 4.8%)
  Topic 4: Indian Culture & Nationalism   ( 74 lectures, 32.3%)
  Topic 5: Raja Yoga Practice             (  6 lectures, 2.6%)
```

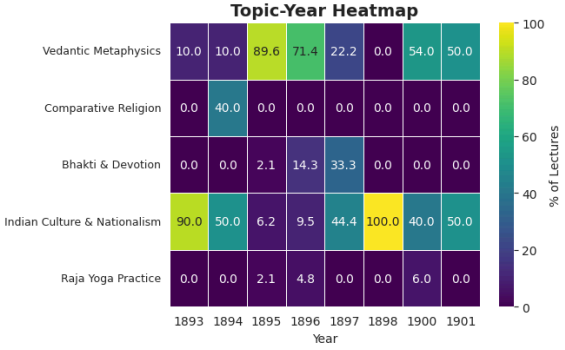
Now, remaking the same plots as above with these names instead of the arbitrary numbers makes more sense.

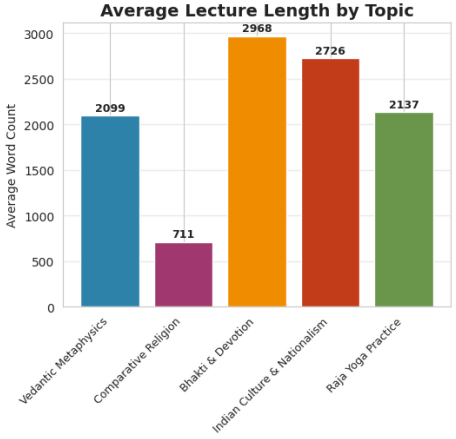

Lecture Count by Topic
Number of Lectures
133
5
11
74
6
Vedantic Metaphysics
Comparative Religion
Bhakti & Devotion
Indian Culture & Nationalism
Raja Yoga Practice

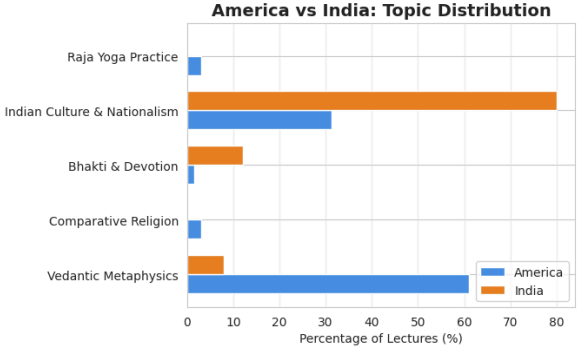

Topic Distribution by Location
Percentage of Lectures (%)
America
India
England
Vedantic Metaphysics
Comparative Religion
Bhakti & Devotion
Indian Culture & Nationalism
Raja Yoga Practice
Topic

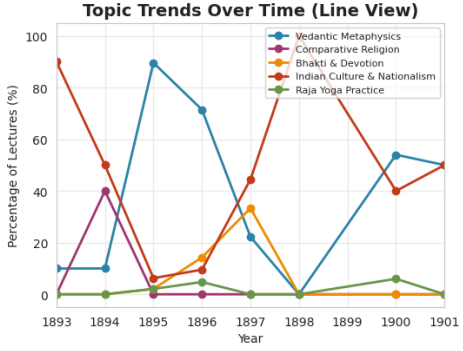

Topic Evolution Over Time
Percentage of Lectures (%)
Vedantic Metaphysics
Comparative Religion
Bhakti & Devotion
Indian Culture & Nationalism
Raja Yoga Practice
Year

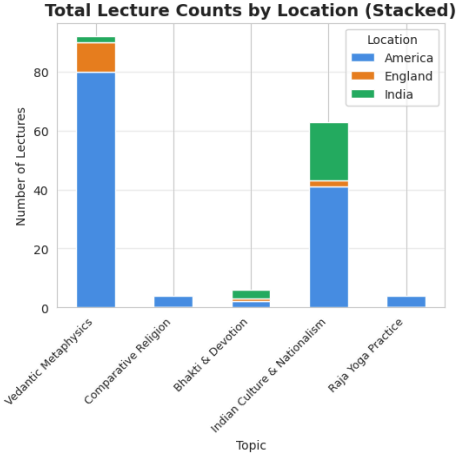

Topic-Location Heatmap
America
England
India
61.1
3.1
1.5
31.3
3.1
76.9
0.0
7.7
15.4
0.0
8.0
0.0
12.0
80.0
0.0
% of Lectures
Vedantic Metaphysics
Comparative Religion
Bhakti & Devotion
Indian Culture & Nationalism
Raja Yoga Practice


Topic-Year Heatmap
Vedantic Metaphysics
Comparative Religion
Bhakti & Devotion
Indian Culture & Nationalism
Raja Yoga Practice
10.0 10.0 89.6 71.4 22.2 0.0 54.0 50.0
0.0 40.0 0.0 0.0 0.0 0.0 0.0 0.0
0.0 0.0 2.1 14.3 33.3 0.0 0.0 0.0
90.0 50.0 6.2 9.5 44.4 100.0 40.0 50.0
0.0 0.0 2.1 4.8 0.0 0.0 6.0 0.0
1893 1894 1895 1896 1897 1898 1900 1901
Year
% of Lectures


Average Lecture Length by Topic
Average Word Count
2099
711
2968
2726
2137
Vedantic Metaphysics
Comparative Religion
Bhakti & Devotion
Indian Culture & Nationalism
Raja Yoga Practice


America vs India: Topic Distribution
Raja Yoga Practice
Indian Culture & Nationalism
Bhakti & Devotion
Comparative Religion
Vedantic Metaphysics
America
India
Percentage of Lectures (%)


Topic Trends Over Time (Line View)
Percentage of Lectures (%)
Vedantic Metaphysics
Comparative Religion
Bhakti & Devotion
Indian Culture & Nationalism
Raja Yoga Practice
Year


Total Lecture Counts by Location (Stacked)
Number of Lectures
Location
America
England
India
Vedantic Metaphysics
Comparative Religion
Bhakti & Devotion
Indian Culture & Nationalism
Raja Yoga Practice
Topic

The plots should be a lot more revealing and readable now. In short, the west gets vedantic metaphysics for liberation while India gets nationalist unification under vedanta. It is also clear that swamiji spoke the least on the topic of comparative religion, based on both volume and number of lectures. This is again a numeric verification to something we already know about swamiji, his main goal was unification and not differentiation.

Looking at plots in center left and bottom right, one can see that Swamiji apparently gave a lot of “Indian Nationalist” talks in America.

Looking at topic trends overtime in bottom center, we can derive some beautiful patterns across time. We can see that Indian nationalism was high when swamiji first visited the west in 1893, he had to appropriate and explain the land and culture of India to foreigners, who only had a colonial understanding of India. It drops down low in 1894-97 as the westerners understand Indian culture and land, and that is precisely when swamiji strategically just bombards them with Vedanta, the crest jewel of Indian philosophy. Then it surges again to an all-time high in 1888-89, when swamiji was in India. Understandably, he was trying to rouse the Indian nationalist sentiment. It remains rather high as he carries this energy to the west in his second visit in 1900, possibly wanting serious westerners to come and spend some time in India.

To understand this better, here are some example lectures under the Indian Nationalist topic that swamiji gave in America, sorted by topic strength.

```
--------------------------------------------------------------------------------
1. The Mahabharata
--------------------------------------------------------------------------------
Year:             1900
Place:            USA, CA, Pasadena, Shakespeare Club
Audience:         Unknown
Source:           Unknown
Word Count:       8016
Topic Strength:   1.000

Opening text:
  The other epic about which I am going to speak to you this
evening, is called the Mahâbhârata…

--------------------------------------------------------------------------------
2. Women of India
--------------------------------------------------------------------------------
Year:             1900
Place:            USA, CA, Pasadena, Shakespeare Club
Audience:         Unknown
```

Source: Unknown
Word Count: 6743
Topic Strength: 1.000

Opening text:
SWAMI VIVEKANANDA: "Some persons desire to ask questions about Hindu Philosophy before the lecture …

---

3. My Life and Mission

Year: 1900
Place: USA, CA, Pasadena, Shakespeare Club
Audience: Unknown
Source: Unknown
Word Count: 6706
Topic Strength: 1.000

Opening text:
Now, ladies and gentlemen, …

---

4. The Women of India

Year: 1894
Place: USA, MA, Cambridge
Audience: Lecture Attendees
Source: Ms. Frances Willard's Stenographer
Word Count: 6481
Topic Strength: 1.000

Opening text:
In speaking about the women of India…

---

5. The Ramayana

Year: 1900
Place: USA, CA, Pasadena, Shakespeare Club
Audience: Unknown
Source: Unknown
Word Count: 5045
Topic Strength: 0.999

Opening text:
  There are two great epics in the Sanskrit language, which are very ancient. Of course, there are hundreds of other epic poems. The Sanskrit language and literature have been continued down to the present day, although, for more than two thousand years, it has ceased to be a spoken language. I am now...

---

6. India

Year: 1894
Place: USA, MI, Detroit
Audience: Unknown
Source: Detroit Free Press
Word Count: 1475
Topic Strength: 0.998

Opening text:
  An audience that filled the Unitarian Church heard the renowned monk, Swami Vivekananda, deliver a lecture last night on the manners and customs of his country. His eloquent and graceful manner pleased his listeners, who followed him from beginning to end with the closest attention, showing approval…

As we can see, at this point we can confidently conclude that the topic modeling using LDA has done good work. Here are some more numbers to understand the topic distribution of Swamiji's public content. One can find an interesting trend in Swamiji's Indian Nationalist lectures in America

```
================================================================================
YEAR DISTRIBUTION OF NATIONALIST LECTURES IN AMERICA
================================================================================

Lectures by year:
  1893: 9 lectures
  1894: 5 lectures
  1895: 2 lectures
  1896: 1 lectures
  1900: 20 lectures
```

As we can see, the number of nationalist lectures gradually declines and resurges mightily in 1900. In his second visit, he gives more Indian Nationalist lectures in America than all previous years combined! One observation can be that since he was there for a shorter time, he wanted to give as many lectures as possible before he left. But the story is not as simple as that. We will explore his second visit shortly in some depth.

After this one might be tempted to ask, what did Swamiji lecture about in India? Well, as we discussed previously, he was quite Indian Nationalistic in India. India already had the means to liberation for several millennia, but this spirit had become drowsy during Mughal reign and Colonisation one after the other. Swamiji wanted to rouse this spirit, and wake it from slumber.

```
================================================================================
COMPARISON: WHAT TOPICS DID HE LECTURE ON IN INDIA?
================================================================================

Total lectures in India: 25

Topic distribution in India:
  Indian Culture & Nationalism       : 20 lectures (80.0%)
  Bhakti & Devotion                  :  3 lectures (12.0%)
  Vedantic Metaphysics               :  2 lectures (8.0%)

--------------------------------------------------------------------------------
```

```
OBSERVATION:
--------------------------------------------------------------------------------
In America: 41 out of 131 lectures (31.3%) on Indian Nationalism
In India:   20 out of 25 lectures (80.0%) on Indian Nationalism
```

As we can see, this further corroborates that Swamiji was majorly Nationalistic in India. Interestingly, the data shows that he was more Bhakti oriented in India than Vedantic.

Looking at the low percentage of Vedantic Metaphysics lectures in India, I wanted to see which were these minority in India.

```
================================================================================
REVERSE OUTLIER: VEDANTIC METAPHYSICS LECTURES IN INDIA
================================================================================

Total 'Vedantic Metaphysics' lectures in India: 2
(Only 8.0% of India lectures)

These rare Vedantic lectures in India:

1. The Vedanta
   Year: 1897
   Place: India, Lahore
   Audience: Unknown
   Words: 14851

2. What have I learnt?
   Year: 1901
   Place: India, Dacca
   Audience: Unknown
```

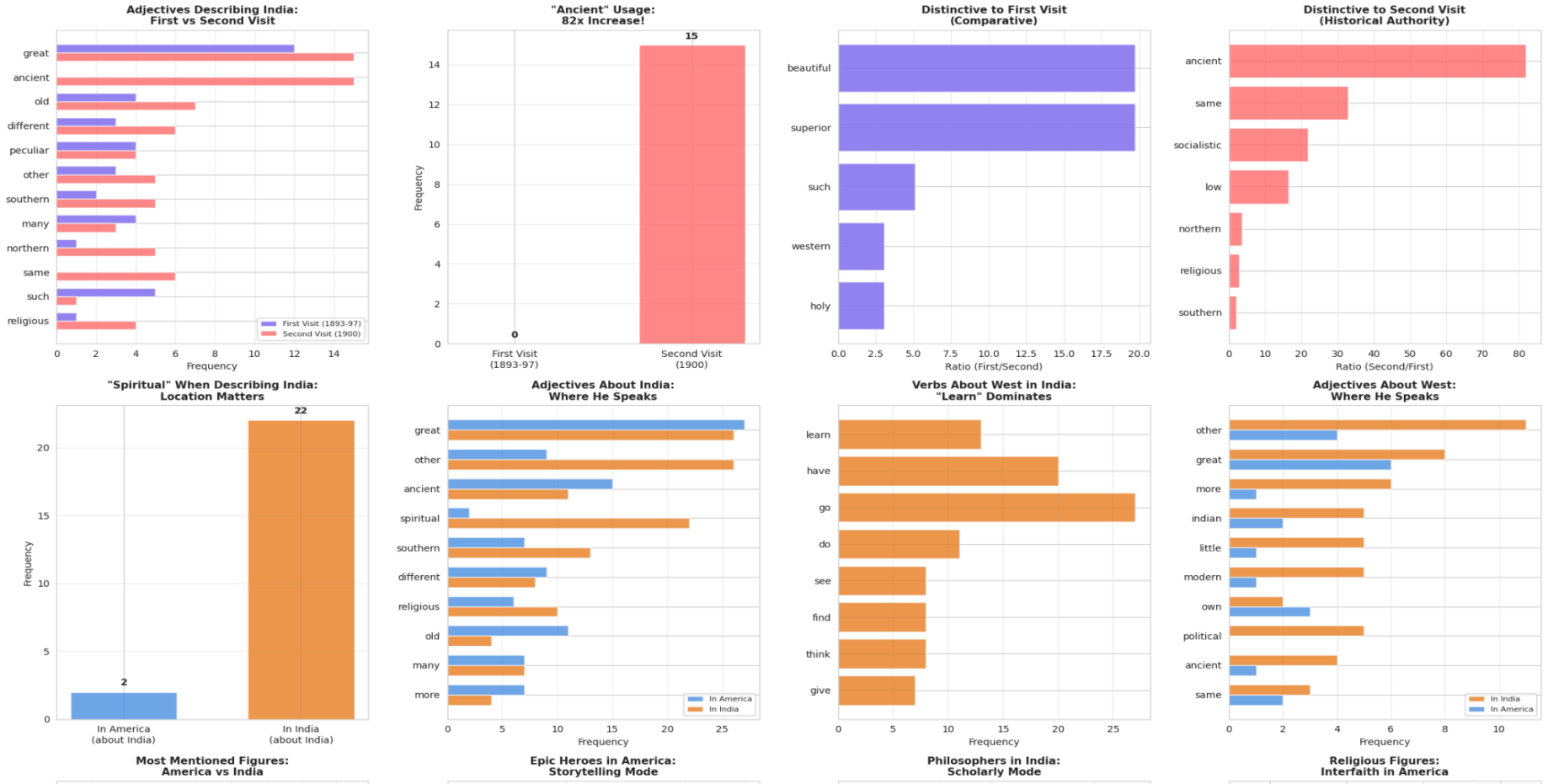
Words: 1514

These are the two outlier-ish Vedantic lectures given in India. Funnily enough, both of these places where he gave the lectures on Vedantic Metaphysics are no longer within the borders of India.

Now I basically wanted to answer some natural follow-up questions that arose during the analysis.

1) What was the difference in how Swamiji framed India to the West during first and second visit?
2) I had an intuition that Swamiji used to criticise India to Indians and praise west, while he would criticise the West to Westerners and praise India. I wanted to confirm this using data.
3) Which saints/figures does he invoke in America vs in India?

Here is one plot that shows all the answers.



Now we can get a clearer picture. For question 1, the answer is clear from top row, right two figures. He seems to evoke the ancient nature of India a lot more in his second visit, something he never did in his first visit. In the first visit, he can be seen challenging

colonial views by using adjectives like beautiful, superior, holy etc in relation to India. In the second visit, he uses ancient a LOT more, along with a more nuanced view of India by using directional adjectives like south and north.

For question 2, when Swamiji talked about India to Indians, he used to emphasize the spirituality of India. This aligns with our earlier comment that he wanted to rouse the spiritual aspect of India, and boost the confidence that had been trampled by foreign forces. In India when talking about the west, he seems to be using verbs like “learn” and “go” and “see”. Swamiji is building bridges between India and the West, something that would be walked by every spiritual revolutionary of future like Osho, Prabhupada, Paramhamsa Yogananda etc. This also answers the second question, he does not critique the country and culture he is in as much as he praises them. He is not scrutinizing, he is sanctifying.

Regarding question 3, the answer is quite clear again from graphs in bottom row. In America he uses mainstream ancient Indian epics to make a point. Curiously, he uses many instances of female figures like Mother Sita, Savitri, and Draupadi. This again aligns with the inception of the female suffragette movement that was bubbling up during that time, which would culminate in 1920 with the right to vote being granted to women. He uses the figures of Mother Sita, Savitri and Draupadi to show the dynamism of femininity that Indian epics celebrate. In India, everyone was quite aware of these figures. Sri Rama and Ramayana had already taken the northern belt by storm due to Goswami Tulsidasji, and Mahabharat was anyways commonplace within Indian collective memory. So instead, he moves to talk about philosophical figures, he wants Indians to dive deeper! Swamiji invokes Ramanuja and Sankara. It is curious to see that Swamiji talked in India mostly about Indian Nationalism, but when he evokes figures he talks about philosophers of India. What an integral vision! It is also well noted that Swamiji had said, “Vedanta has enough for the liberation of man, and that is what I teach to the west” while saying something like “the worship of divine mother is my special fad”. Our data supports it, since invoking Ramanuja and Bhakti topics exclusively in India are quite telling in that regards.

Another noticeable thing is mention of Manu in top 6 mentions of names people. In those days the Manusmrti was quite in vogue, since it had been used a few decades ago to codify the Hindu law by the colonial government. This roused debates about Manusmrti. Swamiji was addressing a cultural topic by invoking Manu.

After these analysis, I wanted to see what outlier analysis would reveal. Outlier analysis in data science means finding data points that don’t behave like the rest, things that are unusually different and therefore interesting.
For this text corpus, a document is treated as an outlier if its word usage is very different from other documents from the same place and/or if it belongs to a rare topic. The code turns each document into numbers, measures how far it is from the “average” document for its location, combines that with topic rarity, and assigns a score, higher score = more unusual document. For those well versed in data science and linguistics: Outlier detection is performed by embedding documents in TF-IDF space and computing their Euclidean distance from location-specific centroids, capturing lexical deviation from local norms. This signal is combined with topic-level rarity (inverse topic prevalence from LDA assignments), and the normalized sum yields a composite outlier score that highlights documents that are lexically atypical, topically rare, or both.

In my analysis in this way, something striking came out. I looked at the outliers grouped by locations where they were given (India, America and England). For America, where the volume of lectures is the highest total, what would you guess came out as the top outlier?

```
--------------------------------------------------------------------------------
TOP 5 OUTLIERS PER LOCATION
--------------------------------------------------------------------------------

AMERICA:
  1. Inspired Talks
     Year: 1895, Genre: Lectures, Topic: Bhakti & Devotion
     Outlier Score: 4.113, TF-IDF: 0.973, Words: 424
  2. Reincarnation
     Year: 1894, Genre: Lectures, Topic: Comparative Religion
     Outlier Score: 3.780, TF-IDF: 0.938, Words: 794
  3. Buddhism, The Religion of the Light of Asia
     Year: 1894, Genre: Lectures, Topic: Comparative Religion
     Outlier Score: 3.701, TF-IDF: 0.931, Words: 450
  4. The Practice of Religion
     Year: 1900, Genre: Lectures, Topic: Raja Yoga Practice
     Outlier Score: 3.654, TF-IDF: 0.927, Words: 354
  5. Comparative Theology
     Year: 1894, Genre: Lectures, Topic: Comparative Religion
     Outlier Score: 3.021, TF-IDF: 0.874, Words: 1030
```

It is a section from the Inspired talks. For those who are aware of the Swami Vivekananda literature, Inspired talks stand out like nothing else. Here is a brief introduction to Inspired Talks given on the Advaita Ashrama website:
“ …Notes of a series of inspiring class talks of Swami Vivekananda delivered during his seven weeks stay at Thousand Island park, U.S.A. on various religious and philosophical topics and recorded by Miss S.E. Waldo, an American disciple. This book records the most inspired thoughts of Swami Vivekananda during those days in class talks (and informal discussions) on wide ranging topics – spiritual, philosophical, cultural & social, on Saints and Prophets, History, people, etc. The language is very powerful…”

These were given to a very select group of close, inner-circle disciples by Swamiji in Thousand Island park. These are among the most beautiful and most oft-quoted works of Swamiji. To have found one of these lectures as an outlier purely from statistical methods conveys that the inspiration of Swamiji is not only aesthetically pleasing, but numerically, scientifically different from the rest of his works. We will explore this in some more depth shortly.

Here I also had an idea to do some network based outlier analysis of the public corpus. This sort of analysis tries to find documents that are typical, atypical using some different techniques than lexical methods above. Here is what it does:

Simple explanation:
The documents are treated like points in a network, where two documents are connected if they use similar language. Documents with many connections are very typical, while documents with high betweenness centrality act like bridges — they link otherwise separate groups of texts and often combine themes or styles from different clusters.

Technical explanation:
A document similarity graph is constructed using cosine similarity in TF-IDF space, with edges above a fixed threshold. Degree centrality identifies documents that are lexically similar to many others, while betweenness centrality highlights documents that lie on many shortest paths, indicating transitional, hybrid, or structurally important texts that connect distinct topical or stylistic communities.

Here are some bridge documents (high betweenness-centrality), these are the works that connect many different themes and form “bridges” between concepts.

```
--------------------------------------------------------------------------------
TOP 10 BRIDGE DOCUMENTS (High Betweenness Centrality)
--------------------------------------------------------------------------------

1. The Vedanta
   Location: India, Topic: Vedantic Metaphysics
   Betweenness: 0.0404, Degree: 0.6300

2. My Master
   Location: America, Topic: Indian Culture & Nationalism
   Betweenness: 0.0361, Degree: 0.6100
```

```
3. Practical Vedanta: Part II
   Location: England, Topic: Bhakti & Devotion
   Betweenness: 0.0319, Degree: 0.5200

4. The Soul and God
   Location: America, Topic: Vedantic Metaphysics
   Betweenness: 0.0305, Degree: 0.5800

5. The Sages of India
   Location: India, Topic: Indian Culture & Nationalism
   Betweenness: 0.0286, Degree: 0.4850

6. Gita III
   Location: America, Topic: Vedantic Metaphysics
   Betweenness: 0.0270, Degree: 0.5350
```

Hmmm… The Vedanta from India again, from Lahore. It showed up previously as one of the only two lectures given in India and classified under the topic of “Vedantic Metaphysics”. It shows up here as having the highest betweenness centrality among all documents. However, even the highest betweenness centrality is quite low (4% only).

Low betweenness values indicate a dense, highly connected network with many alternate paths, meaning no document acts as a critical bottleneck. This implies strong thematic overlap in Swamiji’s works overall, fuzzy clusters rather than sharp divisions, and an ideationally integrated corpus where texts blend smoothly across topics instead of forming isolated communities. That said, since there a range of documents, some of them exist which are least connected to the rest of the corpus. In a vague sense, they can also be called outliers. Here are the some of the most weakly connected documents that the algorithm found:

```
--------------------------------------------------------------------------------
ISOLATED/WEAKLY CONNECTED: 6 documents
--------------------------------------------------------------------------------
  Why We Disagree (America, Indian Culture & Nationalism)
  Vedic Teaching in Theory and Practice (India, Indian Culture & Nationalism)
  The Ramayana (America, Indian Culture & Nationalism)
  Congress of Religious Unity (America, Indian Culture & Nationalism)
```

Christ's Message to the World (America, Vedantic Metaphysics)

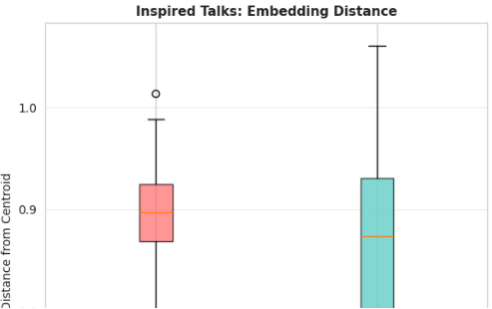
The Ramayana occurs again! It appeared earlier when we were looking at the Indian Nationalism lectures given during the first visit of Swamiji to the west. Most of these are explained away by the topic, Swamiji mostly reserved a Vedantic Metaphysics for the West, hence these appear as somewhat outlier-ish. It is interesting to see that the lecture titled “Christ’s Message to the World” is classified under Vedantic Metaphysics for some reason. We have a graph somewhere above which shows how infrequently Swamiji talks about Christ in America, this is in that minority. That said, when one reads this small talk, it becomes clear that this are not Swamiji’s words actually, but short notes of one of his disciples of a lecture.

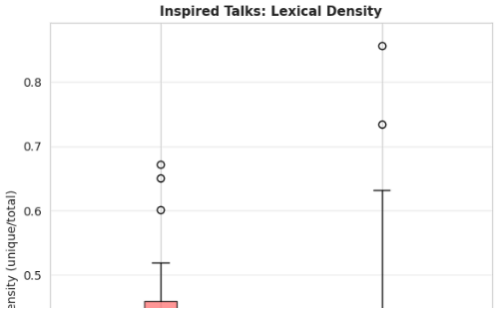
[From Mr. Frank Rhodehamel’s notes of a lecture delivered in San Francisco, California, on March 11, 1900]

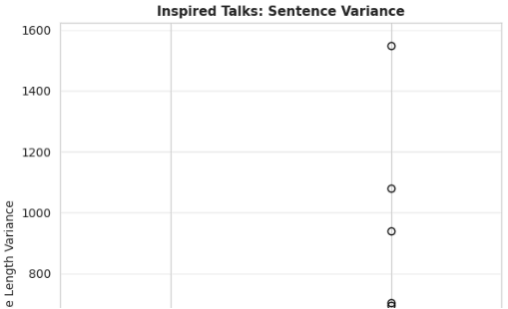
Hence this is an outlier.
Now, I was very interested to see how much of an Outlier the Bhakti talk from Inspired Talks was compared to the rest of the corpus. At the same time I decided to analyse if the mentions of Sri Ramakrishna had any effect on the outlier nature.

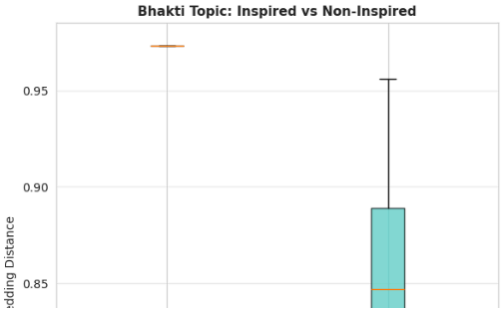
The techniques used here involved similar outlier analysis as before. Formally, the algorithm performs subgroup-conditioned anomaly analysis by combining lexical statistics (lexical density, sentence-length variance), TF-IDF–based topic entropy, centroid-distance embedding deviation, and network centrality metrics to compare “Inspired Talks”against controls, validating structural outliers via distributional statistics and multivariate visual diagnostics.

In simple terms, the algorithm quietly separates out Swami Vivekananda’s “Inspired Talks” and checks whether they naturally stand apart from the rest. It does this by looking at simple signs like how varied the words are, how uneven or flowing the sentences feel, how focused or wide-ranging the ideas are, and how different each talk is from the average talk given in the same place. It then compares these special talks with ordinary ones—especially within bhakti themes—to see if they are more distinctive or central.









We can see from the graphs on the above that show some interesting things. The first three graphs (embedding distance, lexical density, sentence variance) show how Inspired talks as a whole behave in comparison to the rest of the corpus, the “non inspired” parts. The differences are visually significant, but statistically they fall within the normal range. However, the rightmost graph on the top row shows the Bhakti Inspired talk, which was the outlier identified, and that lies significantly (both visually and statistically) away from the normal non-inspired bhakti talks. This means that the combination of this being an Inspired Talk along with being classified as a Bhakti talk is what gave it the status of a major outlier.

```
================================================================================
DIRECTION 1: INSPIRED TALKS ANOMALY ANALYSIS
================================================================================

Identifying Inspired Talks subset...
  Found 43 Inspired Talks documents

  Inspired Talks topic distribution:
    Vedantic Metaphysics: 40
    Indian Culture & Nationalism: 2
    Bhakti & Devotion: 1

  Computing linguistic metrics...
  Computing embedding distances to location centroids...
```

As we can see, even within Inspired talks, there is only one talk which was classified under the topic of Bhakti and Devotion, and that is what appeared as the outlier.

The talk was the one given on Naradiya Bhakti Sutras on June 24th, 1895. *

>>What talk was it actually? Need to find out. Gosh. Sloppy.

Now we move on to analyzing mentions of Sri Ramakrishna in public discourse. Normally, as readers within the Sri Ramakrishna-Vivekananda community, it is assumed that Swamiji did not speak too much about Sri Ramakrishna publicly. I wanted to see if the numbers agreed with this notion. Hence I decided to analyze collocation of Sri Ramakrishna mentions. The methods deployed for the same is quite simple and well known in natural language processing.

The algorithm performs rule-based named-entity pattern matching to detect and count Ramakrishna mentions, analyzes their temporal, geographic, and topical dispersion, extracts concordance windows for local context, applies POS tagging to those windows to derive verb/adjective distributions, and computes comparative frequency ratios against a corpus baseline to identify statistically distinctive framing patterns, supported by exploratory visualizations.

In simple terms, it finds every place where Sri Ramakrishna is mentioned in Swami Vivekananda's talks, counts how often and where those mentions occur (by year, place, and topic), shows the exact surrounding words for each mention, and then studies the kinds of actions and descriptions used near His name to understand *how* Sri Ramakrishna is spoken about, not just *how often*.

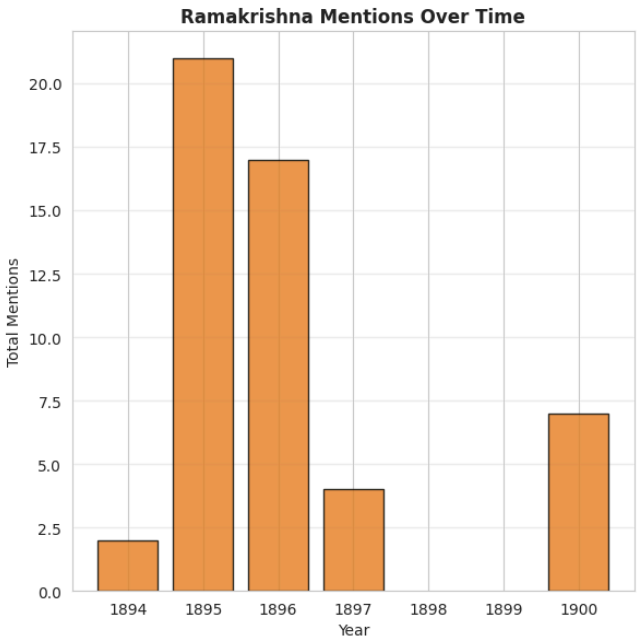
```
================================================================================
STEP 2: BASIC QUANTIFICATION
================================================================================

Overall Statistics:
  Total token mentions: 69
  Documents mentioning: 27 / 179 (15.1%)
  Average mentions per document (when present): 2.56
```

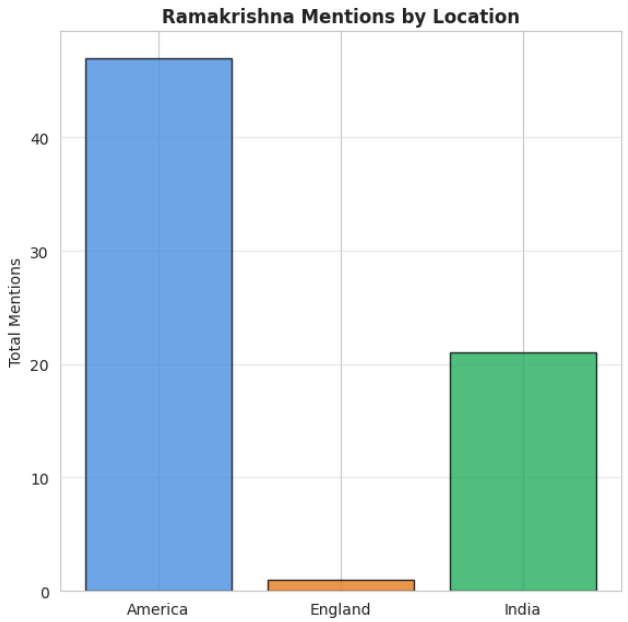
Seems like Swamiji mentions Sri Ramakrishna in 15% of the documents overall. Crucially, whenever Sri Ramakrishna is mentioned, he is mentioned 2.5 times on average. Here are the top talks that mention Sri Ramakrishna.

```
--------------------------------------------------------------------------------
TOP 10 TALKS BY RAMAKRISHNA MENTION FREQUENCY
--------------------------------------------------------------------------------
 1. My Master
    Year: 1896, Location: America, Mentions: 15
 2. Address of Welcome Presented at Calcutta and Reply
    Year: Unknown, Location: India, Mentions: 9
 3. Inspired Talks
    Year: 1895, Location: America, Mentions: 8
```

```
 4. Discipleship
    Year: 1900, Location: America, Mentions: 4
 5. My Plan of Campaign
    Year: Unknown, Location: India, Mentions: 3
 6. The Vedanta in all its phases
    Year: Unknown, Location: India, Mentions: 3
 7. Inspired Talks
    Year: 1895, Location: America, Mentions: 3
 8. Reply to the Address of Welcome at Ramnad
    Year: 1897, Location: India, Mentions: 2
 9. The Sages of India
    Year: Unknown, Location: India, Mentions: 2
10. Inspired Talks
    Year: 1895, Location: America, Mentions: 2
```

We can see that most of the mentions are in America. Also crucially, three Inspired talks make the list again! Seems like when swamiji was inspired, he was more likely to talk about his guru. Here are some basic visualisations:



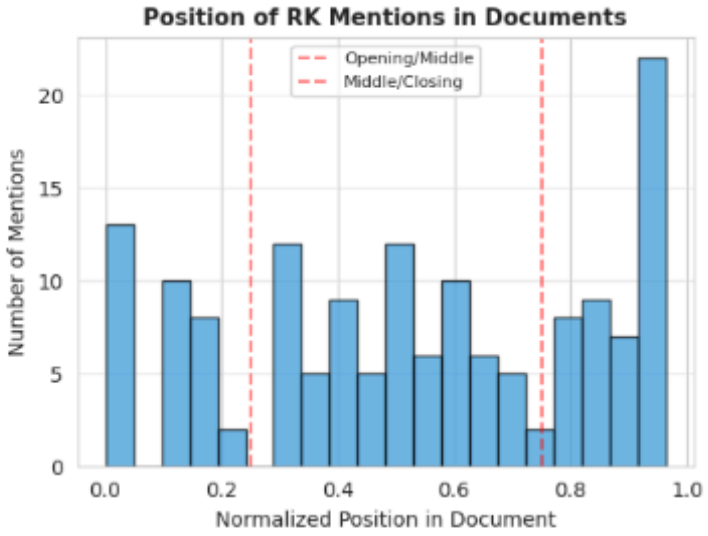


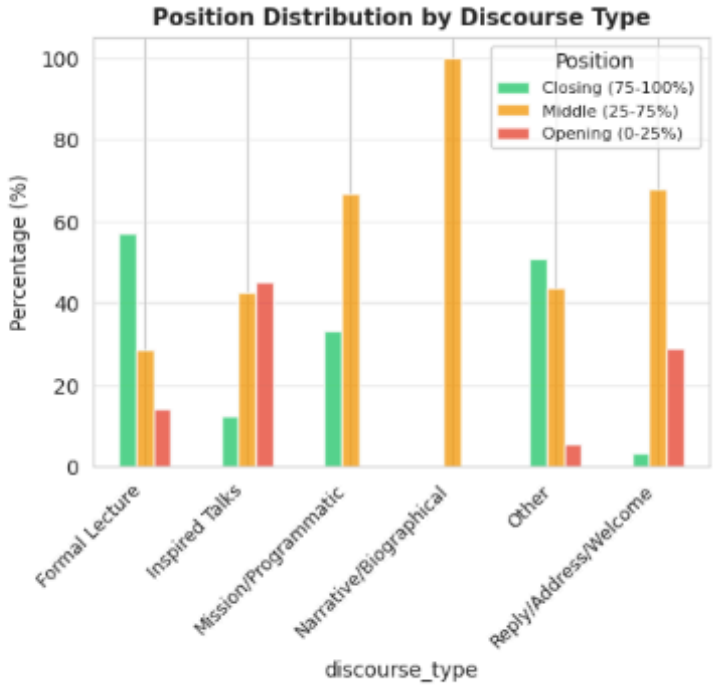
We can see that most of the occurrences are from 1895, the same year Inspired talks was collated. During his stay in India after first visit to the west, he seemed to have not used the example of Sri Ramakrishna much in public talks (except of course the address to various public gatherings). The text titled “Sages of India” mentions Sri Ramakrishna twice, and it was most likely given during 1898-99 though the exact date is not known.

We can see this pattern of invoking Sri Ramakrishna during Inspired talks and Addresses to welcomes more clearly. The mention rate in Replies/Address/Welcome genre of public corpus is quite high at 30%!

```
--------------------------------------------------------------------------------
Ramakrishna mention rates by discourse type:
--------------------------------------------------------------------------------
 discourse_type          sum  count_genre  mention_rate_percentage

Inspired Talks           14     44            31.818182
Reply/Address/Welcome     3     10            30.000000
```

Narrative/Biographical	1	5	20.000000
Formal Lecture	7	39	17.948718
Mission/Programmatic	1	6	16.666667





Another analysis was done to see where in the document Swamiji invokes Sri Ramakrishna (opening, middle, closing). Generally, he invoked his master in middle or end portions. In most of the Mission/Programmatic talks, Sri Ramakrishna was mentioned in the beginning as we can see. Inspired talks also deviate from the norm by having a high mention rate in the opening.

I also briefly looked at some devanagari text that was there in the works. Seems like he used it to quote the scripture mostly in sanskrit. Here are some examples.

DEVANAGARI SEGMENTS (with context):

---

[1]
attempt to get the solution of the deep problems of life from the material world. यस्यैते हिमवन्तो महित्वा — "Whose glory these Himalayas declare". This is a grand idea, but

[2]
matter to mind. There arose the cry, "When a man dies, what becomes of him?" अस्तीत्येके नायमस्तीति चैके — "Some say that he exists, others that he is gone; say,

[3]
doctrine of Adhikârabheda. It is true that the Upanishads have this one theme before them: कस्मिन्नु भगवो विज्ञाते सर्वमिदं विज्ञातं भवति। — "What is that knowing which we know everything

[4]
last both ideas are discarded, and whatever is real is He; there is no difference. तत्त्वमसि श्वेतकेतो — "Shvetaketu, That thou art." That Immanent One is at last declared to

[5]
into all these wave forms which we call suns, moons, and systems. We read again: यदिदं किंच जगत् सर्व प्राण एजति निःसृतम् — "Everything in this universe has been projected,

[6]
higher, until it attains perfection. We had that idea also. Declares our Yogi Patanjali — जात्यन्तरपरिणामः प्रकृत्यापूरात्। One species — the Jâti is species — changes into another species —

[7]
is all the difference. But that infinite power is there all the same. Says Patanjali: ततः क्षेत्रिकवत्। — "Like the peasant irrigating his field." Through a little corner of his

[8]
God fell to pieces. Truth, and nothing but truth, is the watchword of the Advaitist. सत्यमेव जयते नानृतं। सत्येन पन्था विततो देवयानः — "Truth alone triumphs, and not, untruth. Through

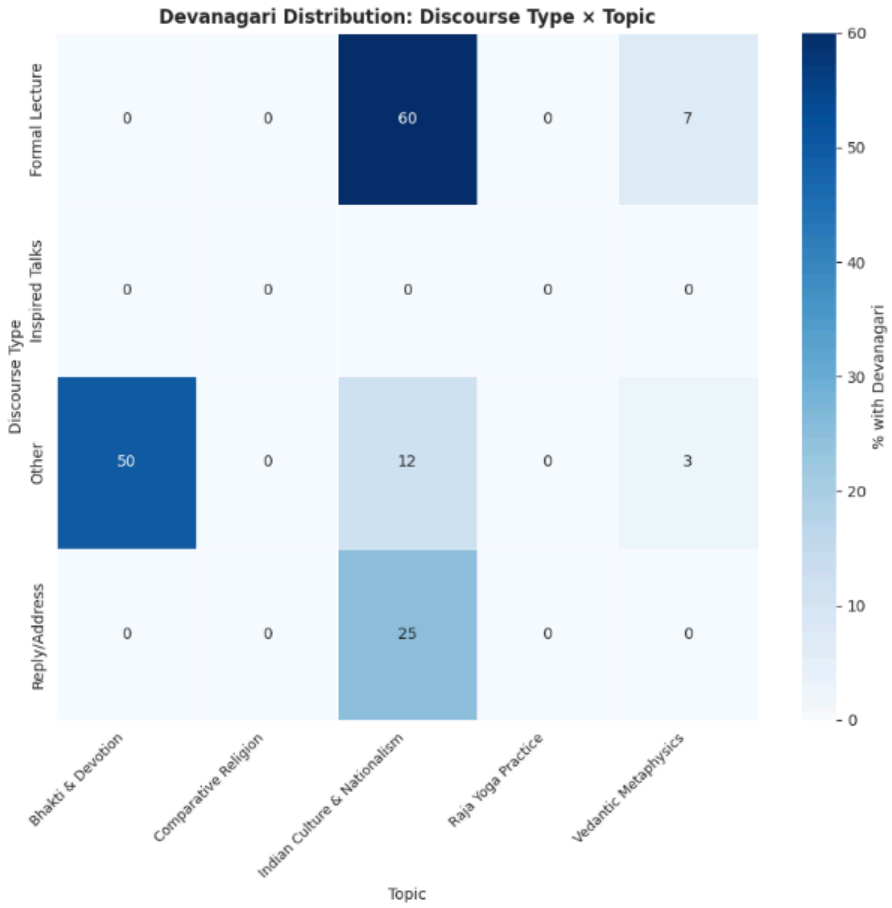
[9]
obey us or we will destroy you. That was the long and short of it. महद्भयं वज्रमुद्यतम्। — It is the idea of the thunderer who kills every one who

[10]
this doubt will come to you: If this is Brahman, how can we know it? विज्ञातारमरे केन विजानीयात् — "By what can the knower be known?" How can the knower

... and 3 more segments with Devanagari

As we can see clearly, these are all being used for quotation from scriptures. One would expect Swamiji to have used most of his Sanskrit references in India, since the Western audience did not know Sanskrit. This would only be the half truth, since Swamiji used sanskrit ONLY in India, I could not find a single place in the west where he quoted from Sanskrit. In comparison to the world today, we see Swamis in the west freely quoting directly from Sanskrit scriptures (Swami Medhananda, Swami Sarvapriyananda etc.), and many westerners actually follow too!

Almost all the occurrences of Sanskrit quotations were concentrated in two years, 1897 and 1901. Both are the years right after Swamiji's visit to the west.



Topicwise, Nationalist lectures get most quotations, followed by Devotional talks. Interesting to see that Swamiji does not quote from scriptures in comparative religion (again in stark contrast to comparative religion experts today who love to quote directly from scriptures), this is because first of all the volume of comparative religious talks were very low and secondly because Swamiji wanted to integrate, not differentiate. It is also to note that this does not mean Swamiji never quoted from scriptures in the west, he definitely quoted scriptures but by using english translations only, not the original sanskrit.

This is the summary in total:

```
SRI RAMAKRISHNA:
  Total mentions: 41
  Documents with mentions: 18
  Dominant discourse type: Inspired Talks
  Dominant rhetorical role: Exemplar/Personal
  Dominant position: Middle

DEVANAGARI:
  Documents with Devanagari: 13
  Percentage: 7.3%
  Dominant location: India
  Dominant topic: Indian Culture & Nationalism
```

This concludes our first section dealing with Public parts of the corpus. Now we move on swiftly to the private corpus containing letters/epistles.

==========
Part 2: Private Corpus Analysis
==========

Here we start with the notebook called Swamiji_Analysis_Private.ipynb. If one looks at the notebook, one will see that the entire dataset has been reconstructed from scratch. This is because before scraping for the corpus analysis, we did many eliminations. We removed poems from corpus, along with those entries that had no url, and reports since they were not Swamiji's direct words. One thing was that we only kept english content. There was very less Bengali content (mostly Diary of a Disciple of Sharatchandra Chattopadhyay, now known as "Conversations with Swami Vivekananda") to analyze in the corpus. However, among the private corpus of epistles, a significant number is in Bengali. Though doing any lexicographical (syntactic) analysis of translations would be of little insight (since Swamiji originally wrote these in Bengali and not in English), I realized we can at least incorporate the Bengali letters in semantic analyses of the epistles. It goes without saying here, these are only the letters of Swamiji that have been found and preserved until now. Swamiji probably wrote 5x-6x more letters than these in total. When one reads the analysis, for example where I say "the longest letter swamiji wrote", it should be kept in mind that these only take into account the surviving letters only and not the "absolute" longest letter written by swamiji.

Either way, the dataset is scraped and deduplicated. I scrape the HTML, and then I try to parse the letter. At this point I do something interesting, I try to parse it in a way that I can get only the important content (I do not need date, location, address, recipient name etc. since they are mentioned cleanly in different columns). What I try to do is to parse the HTML so I can separate out the salutation, the body, the signature and the post script of the letter into different columns. This would enable us to analyze Swamiji's signatures separately, along with trying to find patterns in where he writes post scripts etc. The trial is a semi success, since there are varying types of epistles with many different formats, I let the salutation and the body be clubbed together, and I am able to separate out the letter into body (including salutation), signature and post script with a high level of accuracy.

I print some random signatures to see the job done:

```
================================================================================
UNIQUE SIGNATURES (showing 15 random ones)
================================================================================
Total unique signatures: 247

Random sample of 50:
1. Ever yours gratefully,VIVEKANANDA.
2. Ever yours in love and affection, | VIVEKANANDA.
3. Yours in love, | VIVEKANANDA.
4. I remain yours faithfully, | VIVEKANANDA
5. Yours ever affectionately, | VIVEKANANDA.
6. Ever yours, | VIVEKANANDA.
7. Ever with love, | VIVEKANANDA.
8. With love and blessings to all, | VIVEKANANDA.
9. Your affectionate Son, | VIVEKANANDA.
10. Ever yours with love, | VIVEKANANDA.
11. Yours obediently, | VIVEKANANDA.
12. With all love as usual, | VIVEKANANDA.
13. Ever your devoted son, | VIVEKANANDA
14. I remain, yours faithfully, | VIVEKANANDA.
15. Your ever affectionate bro., | VIVEKANANDA.
```

Something I did not know at this point was that in this version of complete works available online, signatures of Swamiji were mostly regularised editorially before publication. A much better source is the new “Letters of Swami Vivekananda” which is in 4 volumes, and also has about 50-60 new letters that were discovered more recently. Either way, we were able to get some nice variation with 200+ unique signatures. Something that made me chuckle was the usage of the word “bro” by Swamiji on number 15, though he meant “brother” as signified by the contraction mark, it is still nice to see that he chose to use the lingo that Generation Z (my generation) would use in normal day-to-day as opposed a variation like “Br.” or something else.

Now we move onto the notebook called Swamiji_Private_2.ipynb. Some basic stats to look at a glance are as follows:

```
BASIC STATISTICS:
================================================================================
Total Letters: 762

Date column type issues - let's check unique values:
Non-null dates: 759
Null dates: 3

Missing Values by Column:
Category            2
Type                1
Date                3
Place               7
Source            762
Audience            4
Language            3
signature_text     29
ps_text           706
dtype: int64

Word Count Statistics:
       body_word_count  signature_word_count  ps_word_count
mean        332.551181              3.274278       2.585302
std         363.602143              1.973436      14.851298
min          16.000000              0.000000       0.000000
```

```
25%         115.000000              2.000000       0.000000
50%         222.000000              3.000000       0.000000
75%         410.750000              5.000000       0.000000
max        3147.000000             14.000000     263.000000

Extraction Success Summary:
Letters with body text: 762 (100.0%)
Letters with signatures: 733 (96.2%)
Letters with postscripts: 56 (7.3%)
```

Somethings that stand out immedeately are that “source” column is redundant since it is empty (null) for every entry/row. We can furthermore see the variation within the three columns we made, body, signature and ps (post script) on the basis of word count alone. The smallest letter is one that has 16 words only, and of course there are letters that do not have signatures (min=0) and ps (min=0). Interestingly, the longest post-script swamiji wrote is 263 words! It is longer than the average length of half the letters completely (the 50% stat). The longest signature, similarly, is 14 words long! Some natural questions follow. What are these longest post-scripts and signatures? Which is the longest letter and shortest letter? These are what we explore next.

```
LONGEST LETTER (Body: 3147 words):
To: Ramakrishnananda, Swami
Date: 1895/07/08E, Place: USA
Preview: MY DEAR SHASHI,

Yesterday I received a letter from you in which there was a smattering of news, but nothing in detail. I am much better now.
Through the grace of the Lord I am proof against the severe cold for this year. Oh, the terrible cold! But these people keep
all down through scientific knowl...


SHORTEST LETTER (Body: 16 words):
To: Bull, Sarah
Date: 1899/09/04, Place: USA, NY, Stone Ridge, Ridgely Manor
Full text: DEAR MRS. BULL,
```

```
. . .Mother knows best, that is all about me. . . .
```

Here are the longest and shortest letters. Not much to be said. It is interesting to see that Swami Ramakrishnananda has been written a letter that is longer than many lectures Swamiji gave. How fortunate must he be to receive such abundant blessings and affection of Swamiji! This letter will make a return towards the end of our analysis, hence it is best to be noted down.

The shortest letter has been written to Sarah Bull, also known as “dhira mata” or Swamiji’s spiritual mother in the west.

Now we can look at some long signatures. Which is the longest one?

```
LONGEST SIGNATURE (14 words):
To: Nivedita, Sister
Signature: Ever yours with all love and blessings,

[Stamp with Swamiji's portrait]

THE CALCUTTA BOY.
```

Ah, a picture says more than a thousand words. Here, the picture has been transcribed into 4 words only, but I believe even qualitatively this would be the longest signature, since Swamiji used a stamp with his portrait instead of his usual name. A note is that even without the 4 words describing the stamp, this would still be the longest signature with 10 words.

What are some other long signatures?

```
9 words to Sturdy, E.T.:
And I am ever yours in the Truth,

VIVEKANANDA.

9 words to Nivedita, Sister:
With all love, yours ever in the Truth,
```

VIVEKANANDA.

9 words to Sturdy, E.T.:
I remain as ever with love and blessings,

VIVEKANANDA.

These come second in line with 9 words each.

Now we can have a look similarly at the longest postscripts Swamiji wrote. Here they are, listed out in descending order of length.

TOP 5 POSTSCRIPTS >50 WORDS:

263 words to Bull, Sarah (1896/08/23):
PS. I have entire sympathy with the work of Dr. Janes and have written him so. If Goodwin and Saradananda can speed the work in U.S., Godspeed to them...

162 words to Bose, Balaram (1890/01/30):
PS. In my opinion, it will do you much good if you come and stay for some time at Ghazipur. Here Satish will be able to secure a bungalow for you,...

147 words to Brother Disciples, Swami (1894/07/31A):
PS. Remember my previous letter — we want both men and women. There is no distinction of sex in the soul. It won't do merely to call Shri Ramakrishna an Incarnation...

92 words to Sturdy, E.T. (1896/08/05):
PS. I hope you will consider well the plan for the big magazine. Some money can be raised in America, and we can keep the magazine all to ourselves at the same time...

85 words to Brahmananda, Swami (1895/07/08D):
PS. Sarada is talking of bringing out a Bengali magazine. Help it with all your might. It is not a bad idea. You must not throw cold water on anybody's project...

The longest postscript, one that is the length of an entire letter in itself, is written to Mrs. Bull (she also received the shortest letter Swamiji ever wrote as we saw earlier, quite the range!).

Now, Swamiji wrote in four languages. There was of course English, his mothertongue Bengali, but also there are three letters in French and three in Sanskrit! Here are some details of the recipients:

```
SANSKRIT (3 letters):
  Avg body length: 502.7 words
  Avg signature length: 4.3 words
  Letters with PS: 0 (0.0%)
  Top 3 recipients: {'Chakravarty, Sharat Chandra': 2, 'Shuddhananda, Swami': 1}

FRENCH (3 letters):
  Avg body length: 248.0 words
  Avg signature length: 1.0 words
  Letters with PS: 0 (0.0%)
  Top 3 recipients: {'Calve, Emma': 1, 'Christine, Sister': 1, 'Leggett, Francis Mrs.': 1}
```

Swami Shuddhananda was a direct disciple of Sri Ramakrishna as we know, and a brother disciple of Swami Vivekananda. Sharatchandra Chakravarty (whose writings were previously excluded from the public corpus analysis by us) was a disciple of Swami Vivekananda. He collated the conversations with Swami Vivekananda which are quite famous now, he also wrote the famous hymn in praise of Swamiji “Murta Maheswaram”, which was also in Sanskrit. The letters are 500 words in length, average.

For french, we have Sister Christine, Emma Calve and Mrs. Leggett. The average is a bit misleading here since one of them is only a few words, another is quite short and the last is quite long, which inflates the average. Also, before this I had no idea Swamiji experimented with French.

The first French communication from Swamiji that we have in written is the one to Mrs. Leggett, dated August 3rd 1900. In fact, this is not a letter at all, it is a telegram! Hence it is quite short.

```
ARRIVE A HUIT HRES STLAZARE — VIVEKANANDA

[Translation: "I arrive at eight o'clock (p.m.) St. Lazare — VIVEKANANDA".]
```

The rest of the two letters cannot be compared temporally since the one written to Emma Calve has no exact date, only that it was written in October of 1990. It is also not available in the original french in the source dataset, only the english translation is available as given:

MY DEAR MADEMOISELLE,

I have been very happy and content here. I am having the best of times after many years. I find life here with Mr. Bois very satisfactory — the books, the calm, and the absence of everything that usually troubles me.

But I don't know what kind of destiny is waiting for me now.

My letter is funny, isn't it? But it is my first attempt.

Here Swamiji indicates that he is trying to dabble in the French language.

The letter written to Sister Christine is the longest and it is dated 14th October 1900. This letter is quite heartfelt, I must say. Again, the original french is not available in the source dataset, but the translation seems to suggest that Swamiji had a good grasp of the language already, enough to form complex sentences.
Two very heartfelt excerpts from the letter, on the points of the French language are given:

… I have no time any more, nor the power to study a new language at my age. I am an old man, isn't it? …

Swamiji was in his late thirties. And here is the second chosen excerpt, the very end of the letter body.

…My knowledge of this language has not the power to express my emotion. But which language can really do so? So I drop it, leaving it to your heart to clothe my thought with a soft, loving, and shining language. Good night, gute Nacht!...

Forever shall we be in awe of the man Swamiji Maharaj is.

The link to the full letter is given below:
https://www.ramakrishnavivekananda.info/vivekananda/volume_8/epistles_fourth_series/195_sister_christine.htm

This was a bit of an emotional detour, but I wanted to share what touched my heart.I do not think I would have been able to find these letters (among more than 800 with the newly discovered ones) if I did not set out to use data analysis on the corpus.

I also thought to share these because being translations (from Sanskrit and French), and hence not the direct words of Swamiji, I decided not to include them in the syntactic analysis going forward, adhering to our strict rule of only taking the words of Swamiji directly. They were also only 6 in number, a very tiny minority of the 700+ corpus that we have. You might remember this is also the reason Bengali was also ultimately removed from syntactic analysis, giving us 600+ letters only. Bengali letters will be reintroduced during semantic analysis, when the words themselves don't matter as much as the meaning.

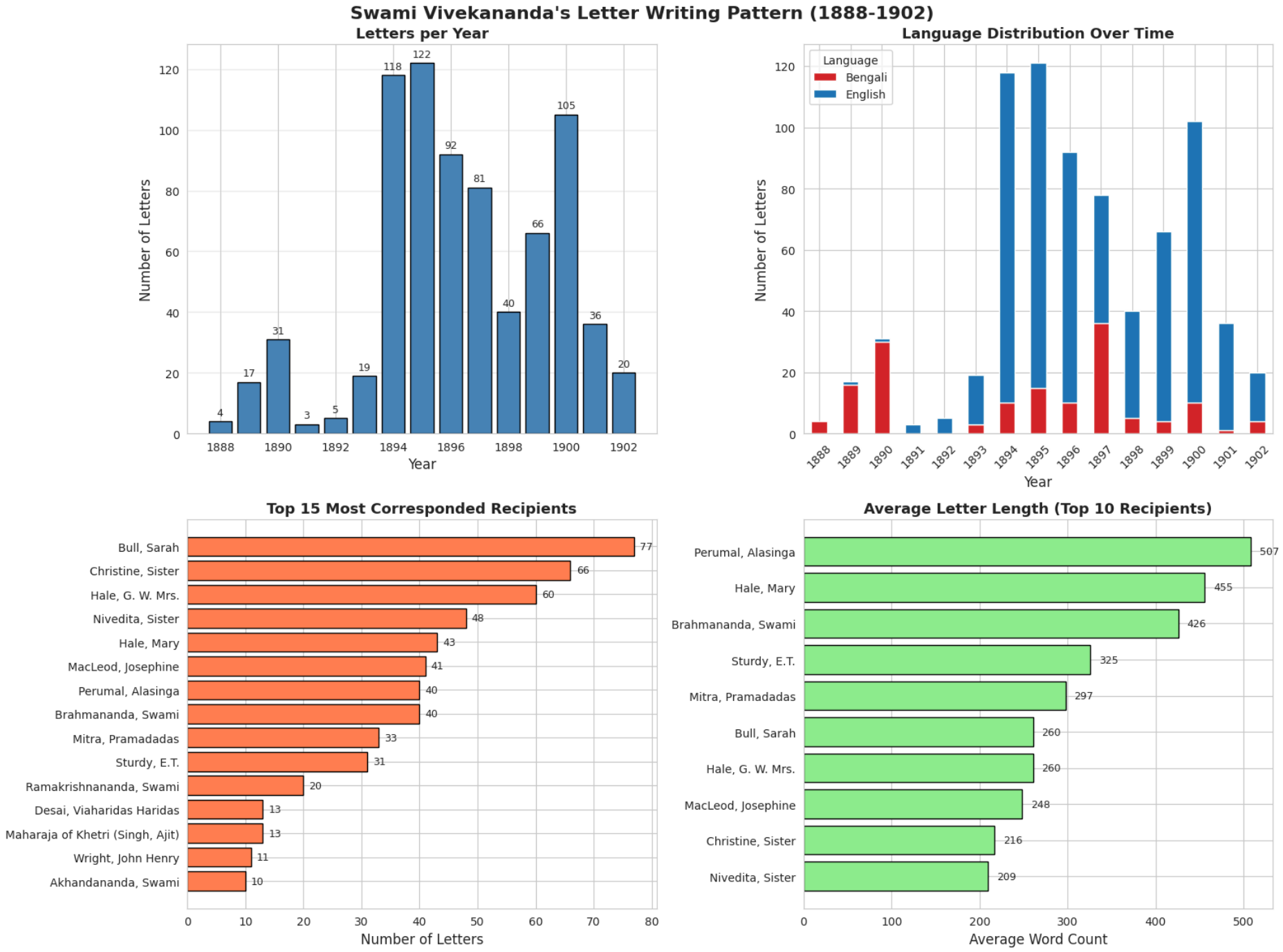
Either way, we can now move on and see which people Swamiji wrote to the most in English and Bengali and overall.

```
OVERALL (All Languages):
  Bull, Sarah: 77 letters
  Christine, Sister: 66 letters
  Hale, G. W. Mrs.: 60 letters
  Nivedita, Sister: 48 letters
  Hale, Mary: 43 letters


ENGLISH:
  Bull, Sarah: 77 letters
  Christine, Sister: 66 letters
  Hale, G. W. Mrs.: 60 letters
  Nivedita, Sister: 48 letters
  Hale, Mary: 43 letters

BENGALI:
  Mitra, Pramadadas: 33 letters
  Brahmananda, Swami: 33 letters
  Ramakrishnananda, Swami: 17 letters
  Bose, Balaram: 10 letters
```

```
Akhandananda, Swami: 9 letters
```

As we can see, the top 5 most corresponded recipients overall are the same as the top 5 in English. More than actual correspondence patterns of Swamiji, this shows that perhaps Mrs. Bull and other western disciples preserved the letters of Swamiji more fervently than their Indian counterparts.

Now here is a nice visualisation to sum up all the basic analyses we need.


Swami Vivekananda's Letter Writing Pattern (1888-1902)
Letters per Year
Number of Letters
Year
4
17
31
3
5
19
118
122
92
81
40
66
105
36
20
1888
1890
1892
1894
1896
1898
1900
1902
Language Distribution Over Time
Language
Bengali
English
Top 15 Most Corresponded Recipients
Bull, Sarah 77
Christine, Sister 66
Hale, G. W. Mrs. 60
Nivedita, Sister 48
Hale, Mary 43
MacLeod, Josephine 41
Perumal, Alasinga 40
Brahmananda, Swami 40
Mitra, Pramadadas 33
Sturdy, E.T. 31
Ramakrishnananda, Swami 20
Desai, Viaharidas Haridas 13
Maharaja of Khetri (Singh, Ajit) 13
Wright, John Henry 11
Akhandananda, Swami 10
Number of Letters
Average Letter Length (Top 10 Recipients)
Perumal, Alasinga 507
Hale, Mary 455
Brahmananda, Swami 426
Sturdy, E.T. 325
Mitra, Pramadadas 297
Bull, Sarah 260
Hale, G. W. Mrs. 260
MacLeod, Josephine 248
Christine, Sister 216
Nivedita, Sister 209
Average Word Count

Something interesting to see is the yearwise pattern of number of letters Swamiji wrote. Of course the top right graph shows the period of Swamiji's travels in India as a monk can be made out by the Bengali letters he wrote.

Another thing to note is that the number of letters is not an indicator of "how much" swamiji communicated, a better indicator of that is average letter length as shown on the bottom right. We can say that on average, Swamiji wrote the longest letters to Alasinga Perumal, who leads the average letter length race.

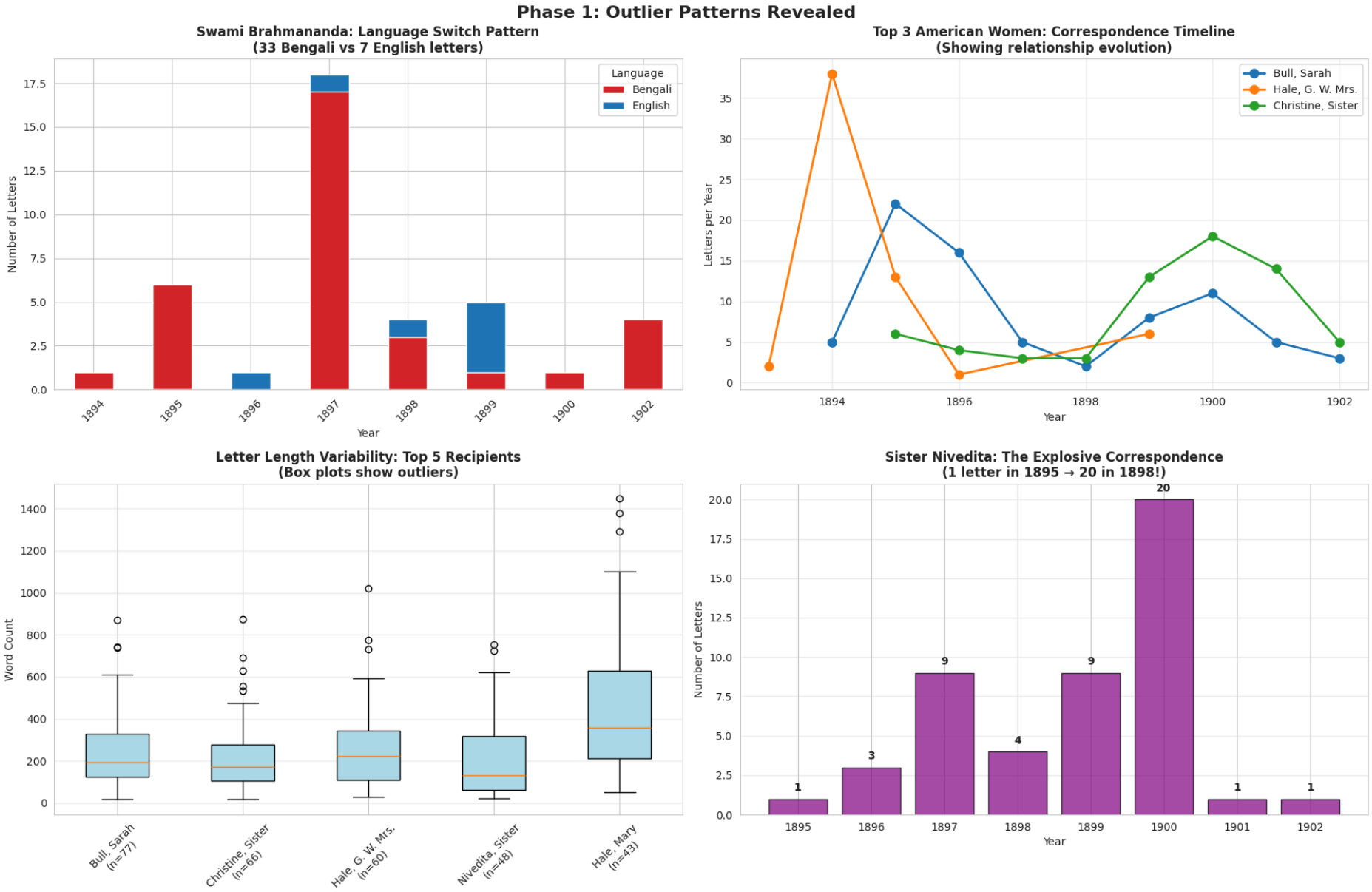
Now I wanted to see some length-based and language-based anomalies. Suppose A writes to B, 10 letters that are 100 words on average. Now, if suppose 9 of these letters were one word long and only one was 991 words long (still making the average 100 words per letter), then this 991 word letter is quite the outlier! This is what I wanted to find for Swamiji based on average letter length per recipient. A fancy word to measure this using statistical methods is z-score.

```
1. MOST UNUSUALLY LONG LETTERS:

   To: Brahmananda, Swami (1895/07/08A)
   Length: 2147.0 words (avg: 426.0) - 403% longer than average
   Z-score: 4.40

   To: Christine, Sister (1901/09/02)
   Length: 873.0 words (avg: 216.0) - 302% longer than average
   Z-score: 3.90

   To: Hale, G. W. Mrs. (1894/08/23)
   Length: 1022.0 words (avg: 260.0) - 291% longer than average
   Z-score: 3.87

   To: Perumal, Alasinga (1893/11/02)
   Length: 1810.0 words (avg: 507.0) - 256% longer than average
   Z-score: 3.42
```

As we can see, the topmost outlier is the 2000+ word letter written to Swami Brahmananda. If you remember from earlier, this is not the longest letter overall, that is the one written to Swami Ramakrishnananda at 3000+ words! The fact that it does not show up here means Swami

Ramakrishnananda was usually written long letters, whereas 2000+ was quite unusual for Swami Brahmananda, who was written shorter letters on average.

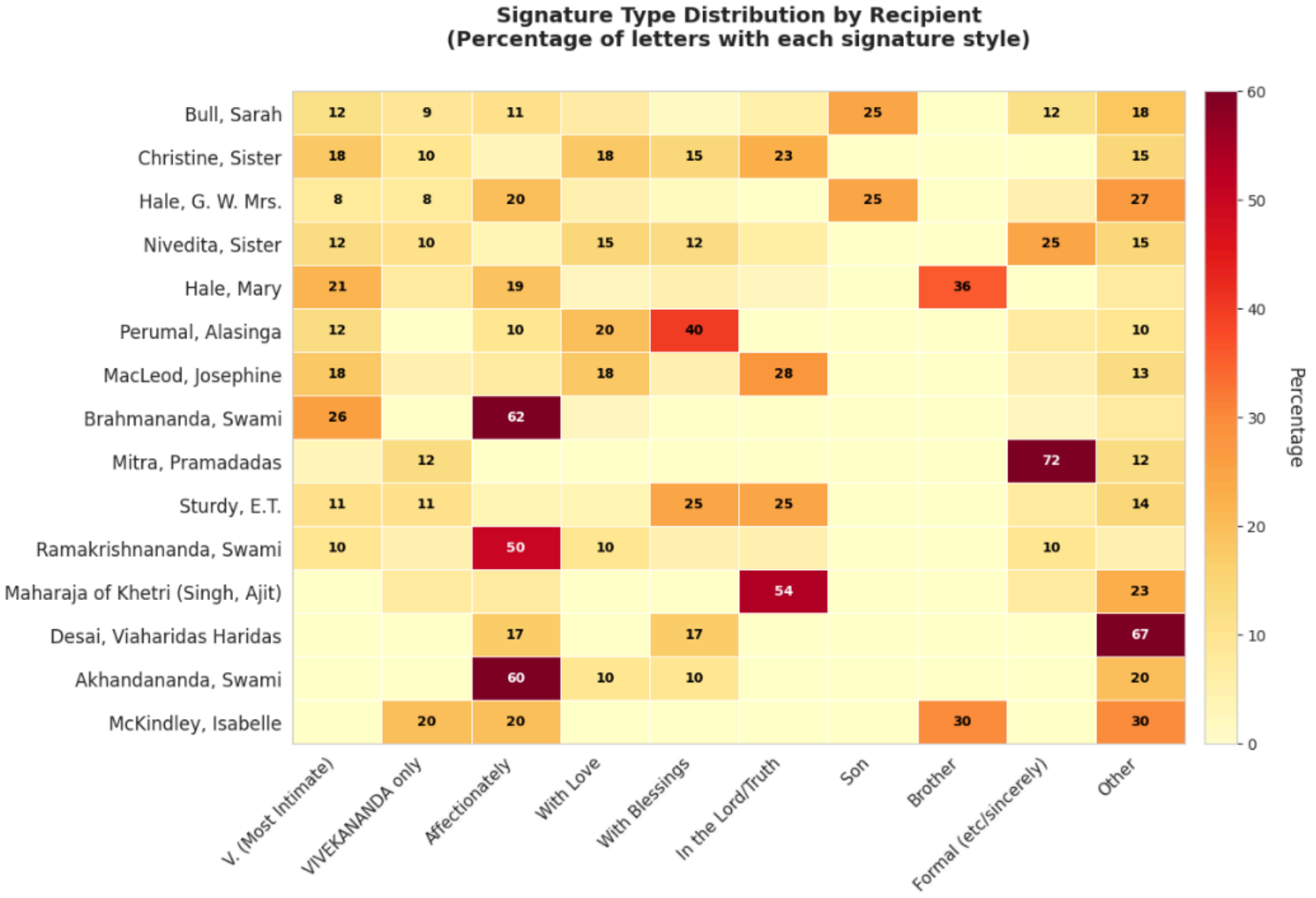
Here we can give a visualisation to summarise some more patterns.



This shows several interesting things.

Top left: Swami Brahmananda is written to quite less in English, but the English rate exploded for some reason in 1899 when Swamiji was in India. I do not have a clue as to why this is the case for this particular year.
Top right: Seeing letters per year for Sister Christine and Mrs. Bull (green and blue lines), we can see the bump right when Swamiji went on his second visit to the West. One would think Swamiji would correspond with western disciples more when he was in India, but the data seems to say otherwise.
It is similar in the bottom right graph with Sister Nivedita, where we can see two peaks. One in 1897 (when Swamiji returned to India from the west) and another in 1900, when he was on his second Western tour.

Now I got somewhat interested in seeing the signatures and any patterns that I could see in them. This was the reason why I spent extra effort in parsing the HTML pages to separate out the signatures from the bodies of the letter. Here are some visualisations of the same:

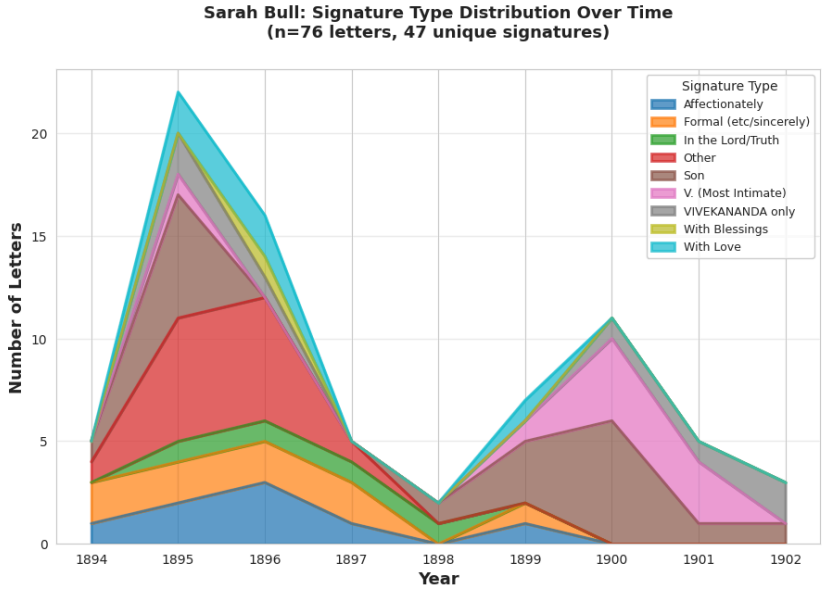


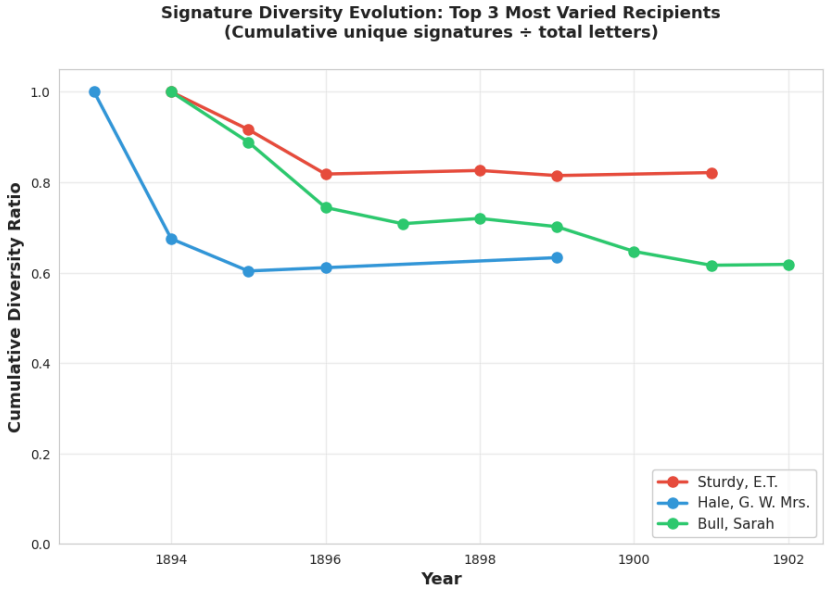
This is a heatmap of what signatures Swamiji used per recipient. At a glance we can make out some nice observations.

1) “Affectionately” is reserved mostly for brother disciples, Swamis Brahmananda, Akhandananda and Ramakrishnananda.

2) “With Blessings” is used mostly for Alasinga Perumal and E.T. Sturdy
3) “In the Lord/Truth” is used mostly for close western acquaintances, but most of all for Maharaja of Khetri!
4) The “etc.” lighting up Pramadadas Swami is only formal since it is an english translation, all the letters to Pramadadas Mitra were in Bengali originally.

Looking at Sarah Bull on top, being the one who preserved most letters of Swamiji, I wanted to look a bit more at the salutations used for her specifically.





On the left we can see an alluvial graph showing how his signatures changed with time for Mrs. Bull. Initially, being dominated by the red coloured “other category”, seems to suggest that Swamiji would experiment a lot with his signatures. Hence we can say many layers on the left half of the graph. On the right half (after 1898), Swamiji seems to have settled with a few options only like “V.”, “Son” and plain old “VIVEKANANDA” for Mrs. Bull.
A similar analysis of signature variation over time has been shown on the right for the three recipients with most signature variation. For all of them, as time went on, it seems like the signature variation dropped. This can be interpreted as a sign of stable relationship. In our own interactions, we

are not completely the same to everyone. We have different personas (formally called profiles) with which we speak with different people in our circle. This sort of analysis is one of many types of profile analysis that aims to capture difference in profiles as inferred written word.

After this I moved on to seeing if any patterns existed in postscripts of the letters. Here are the basic stats before we see it at length:

```
Total letters with PS: 56 out of 760 (7.4%)
Total PS word count: 1970 words
Average PS length: 35.2 words
Median PS length: 22 words
Longest PS: 263 words
Shortest PS: 6 words
```

The sample size is quite small, only 56 letters. This would not be enough to find any major pattern, but at least we can see which recipients got the most post-scripts.

```
Recipients who received letters with PS (Top 15):
  Bull, Sarah: 6 PS out of 77 letters (7.8%)
  Mitra, Pramadadas: 4 PS out of 33 letters (12.1%)
  Brahmananda, Swami: 4 PS out of 40 letters (10.0%)
  Christine, Sister: 4 PS out of 66 letters (6.1%)
  Maharaja of Khetri (Singh, Ajit): 3 PS out of 13 letters (23.1%)
  Bose, Balaram: 3 PS out of 10 letters (30.0%)
  Perumal, Alasinga: 3 PS out of 40 letters (7.5%)
  Sturdy, E.T.: 3 PS out of 31 letters (9.7%)
  Brother Disciples, Swami: 2 PS out of 5 letters (40.0%)
  Hale, Mary: 2 PS out of 43 letters (4.7%)
  Gupta, Mahendra Nath: 2 PS out of 3 letters (66.7%)
  Hale, Sisters: 2 PS out of 10 letters (20.0%)
  Hale, G. W. Mrs.: 2 PS out of 60 letters (3.3%)
  Desai, Viaharidas Haridas: 2 PS out of 13 letters (15.4%)
  MacLeod, Josephine: 2 PS out of 41 letters (4.9%)
```

As we can see, M. Gupta gets postscripts in 66% of his letters, but that is perhaps because only three letters of Swamiji to him survive. The highest absolute number of postscripts is 6 letters, as written to Mrs. Bull, but again that is perhaps only because she has the most surviving letters. I

decided that anything less than 10 letters in total (like M Gupta) is not a good representation. So I remade the list with recipients that had at least 10 surviving letters.

```
================================================================================
RECIPIENTS WITH HIGHEST PS RATE (min 10 letters)
================================================================================

Top 10 by PS rate:
  Bose, Balaram: 3/10 = 30.0%
  Maharaja of Khetri (Singh, Ajit): 3/13 = 23.1%
  Hale, Sisters: 2/10 = 20.0%
  Desai, Viaharidas Haridas: 2/13 = 15.4%
  Mitra, Pramadadas: 4/33 = 12.1%
  Brahmananda, Swami: 4/40 = 10.0%
  Sturdy, E.T.: 3/31 = 9.7%
  Wright, John Henry: 1/11 = 9.1%
  Bull, Sarah: 6/77 = 7.8%
  Perumal, Alasinga: 3/40 = 7.5%
```

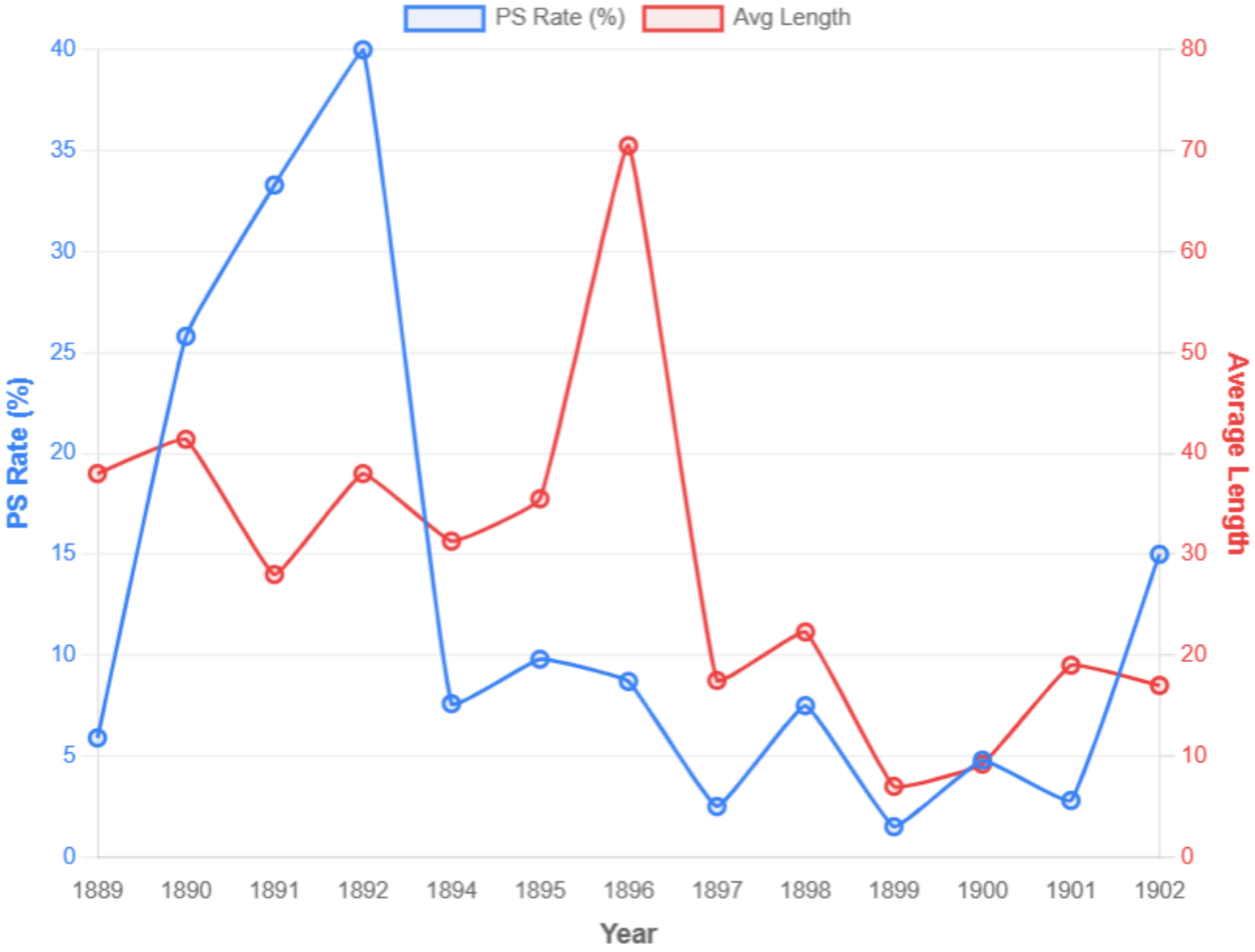
This is better, we can see that Balaram Bose has the highest postscript rate among recipients with more than 10 surviving letters.

Another interesting thing we can find from data is seeing post-scripts by language.

```
================================================================================
PS DISTRIBUTION BY LANGUAGE
================================================================================

English:
  Total letters: 604
  Letters with PS: 39 (6.5%)
  Average PS length: 32.3 words

Bengali:
  Total letters: 148
  Letters with PS: 17 (11.5%)
  Average PS length: 41.9 words
```

We can say by looking at this that Swamiji's rate of writing postscripts was higher for Bengali letters. One could note that the Bengali postscripts are longer in length too, but that might only be because we counted words for the English translation. Ideas can often be expressed in Bengali with fewer words than in English, so it is reasonable to assume that the Average PS Length for Bengali was similar to English ones, perhaps even lower.

Is there any temporal pattern?

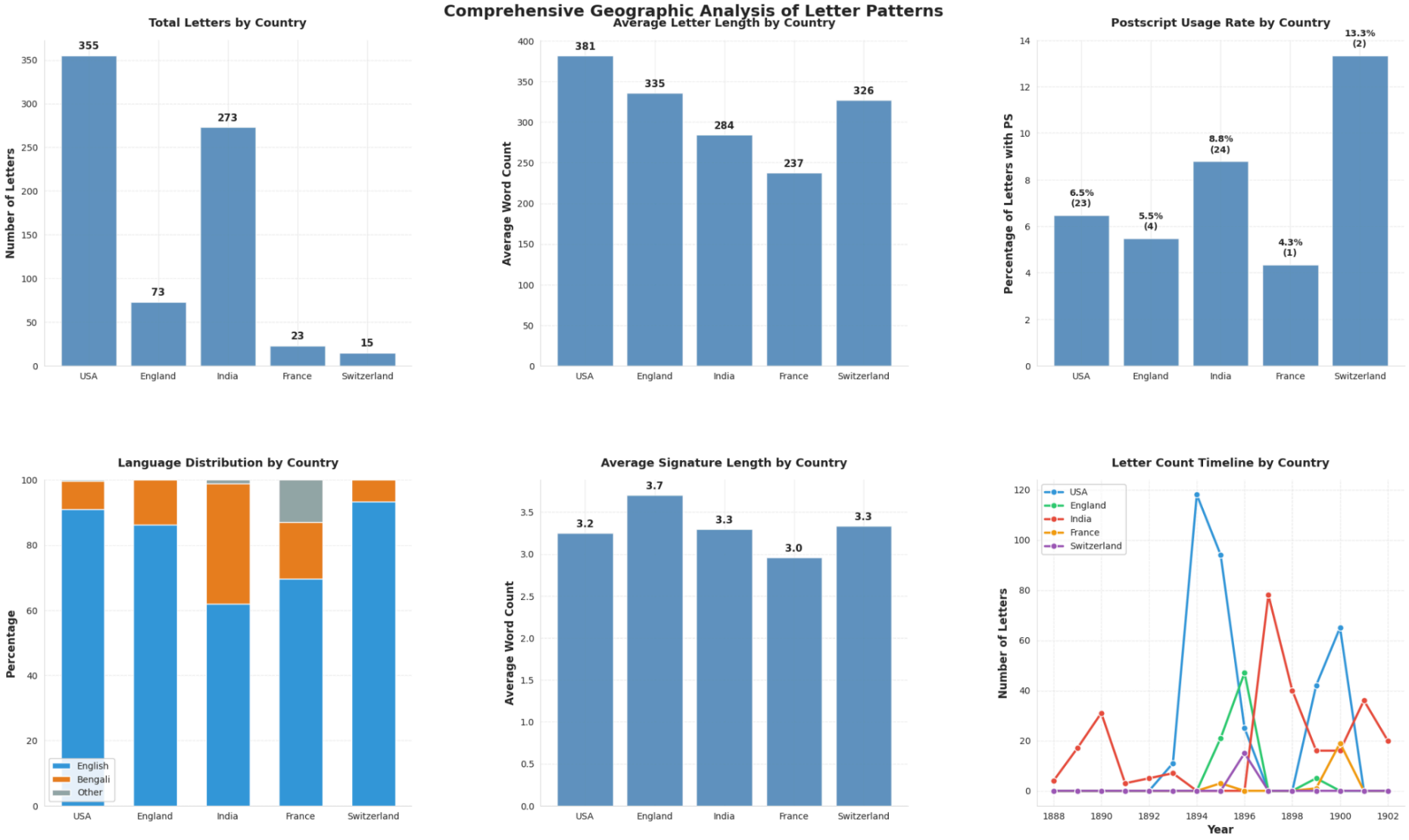

PS Rate (%)
Avg Length
PS Rate (%)
Average Length
Year
0
5
10
15
20
25
30
35
40
0
10
20
30
40
50
60
70
80
1889
1890
1891
1892
1894
1895
1896
1897
1898
1899
1900
1901
1902

Seems like the rate of Swamiji using post-script is highest (at 40%) in 1892, right before he left for America in 1893. However, the longest post-scripts on average occur in the year 1896, again just a year before he came back to India in 1897. Seems like Swamiiji had post-script based variation when any major period of his life was coming to an end.

A thematic analysis reveals that Swamiji mostly talked about health, giving instructions and other people (by taking their names) in the postscripts.

```
================================================================================
PS CONTENT ANALYSIS - COMMON THEMES
================================================================================

Thematic content in postscripts:
  magazine/publication: 6 mentions
  money/funds: 4 mentions
  people/names: 15 mentions
  instructions: 17 mentions
  health: 20 mentions
  work/mission: 11 mentions
```

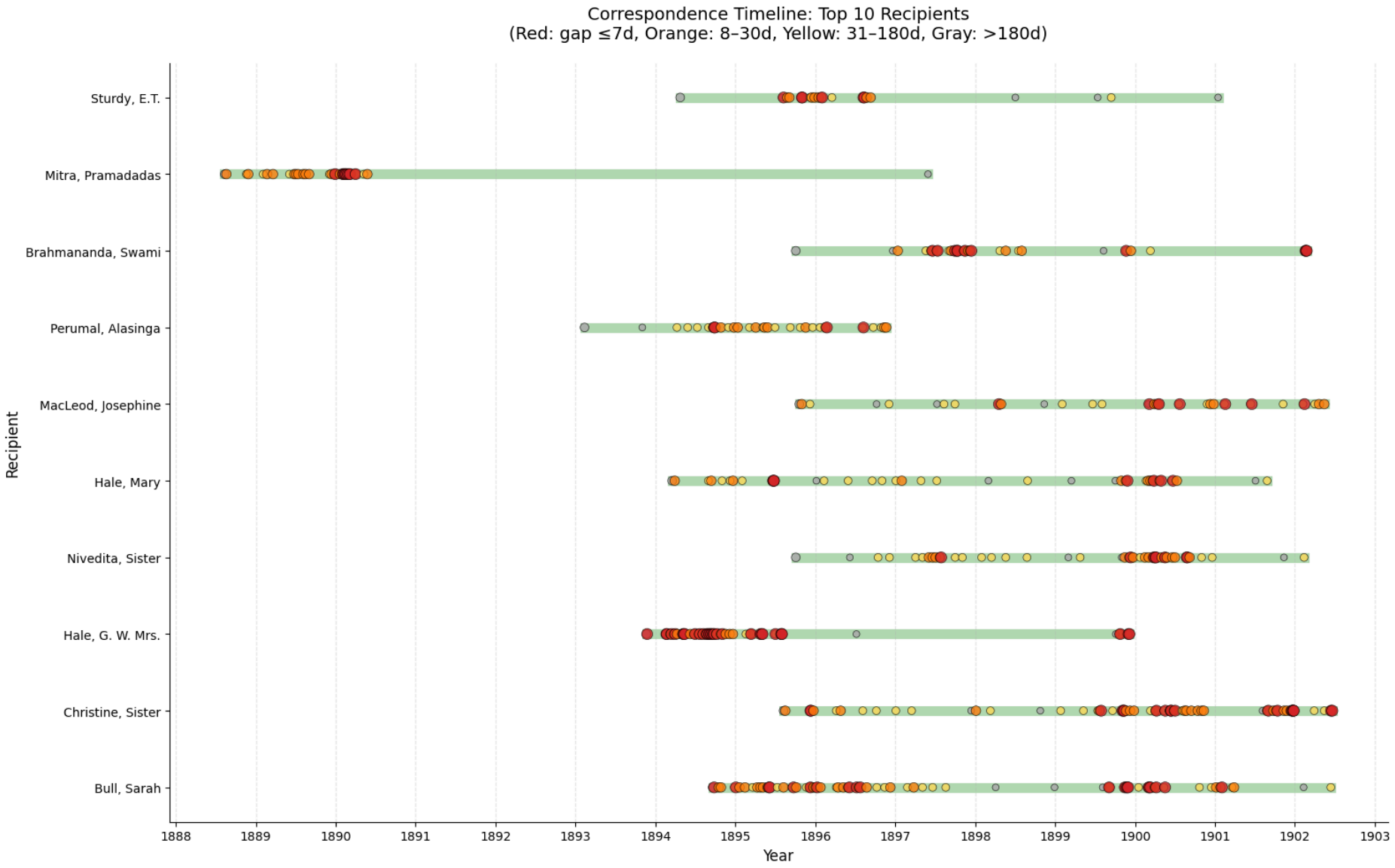
Some visuals for geographic patterns in letters would look like so:



One can see that England on average has letters of similar length to those written from USA by Swamiji, even though the total number of letters written in England is about a fifth of those written in the USA! Swamiji wrote longer letters on average in England. Swamiji's signature patterns are

quite stable across the countries. And as we saw earlier, grey part in France for language distribution (marked as “Other”) shows French while the sliver of grey in India is Sanskrit.

At this point, I got interested in seeing the correspondence pattern. From the surviving letters, what can we infer about the rate with which Swamiji wrote the letters to people? This would bring us to the next notebook, Swamiji_Private_3.ipynb. A lot of analysis was done to see stuff, but I was wondering how to put most of the information in one visual, and that is when this plot was made


Correspondence Timeline: Top 10 Recipients
(Red: gap ≤7d, Orange: 8–30d, Yellow: 31–180d, Gray: >180d)
Recipient
Sturdy, E.T.
Mitra, Pramadadas
Brahmananda, Swami
Perumal, Alasinga
MacLeod, Josephine
Hale, Mary
Nivedita, Sister
Hale, G. W. Mrs.
Christine, Sister
Bull, Sarah
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
Year

Basically this shows the correspondence timeline per recipient. Each dot is a letter, and it is coloured based on how recently the last letter was sent to the same recipient. The green area for each recipient highlights the time period between the first and the last letter written by Swamiji to them for ease of seeing the entire period of communication.

Immediately we can see that with Pramadadas Mitra, most of the correspondence took place during the wandering days. After mid 1890, correspondence came to a strong halt. However, the very last letter written to Pramadadas Mitra is in 1897! Among the surviving letters, this one stands out as being written after the longest period of no contact among all the recipients, more than 7 years!

```
================================================================================
RECIPIENT: Mitra, Pramadadas
Gap: 2561 days (7.0 years)
Previous letter: 1890-05-26
This letter: 1897-05-30
Place: India, Almora, Language: Bengali, Words: 914
URL: https://www.ramakrishnavivekananda.info/vivekananda/volume_6/epistles_second_series/124_sir.htm
```

It is a very bold letter, and I encourage the reader to go through it by clicking the url above. It was written to convey condolences for a “unavoidable domestic grief”. In the letter itself, Swamiji admits that he has been out of touch but also that he has been updated with the life of the recipient from others. Some interesting remarks also follow.

```
…Though for a long time I had no direct correspondence with you, yet I have often been receiving from others
almost all the news about you. Some time ago you kindly sent me to England a copy of a translation of the
Gita. The cover only bore a line of your handwriting. The few words in acknowledgment of this gift, I am
told, raised doubts in your mind about my old affection towards you.

Please know these doubts to be groundless…
```

The entire letter is quite interesting, and Swamiji uses this opportunity to bring a refreshing perspective to an orthodox Brahmin Hindu. The closing of the letter is particularly poignant, and has been left as an opportunity to explore oneself.

Here are some more examples of the longest no-contact periods with recipients:

```
================================================================================
```

RECIPIENT: McKindley, Isabelle
Gap: 1407 days (3.9 years)
Previous letter: 1895-10-24
This letter: 1899-08-31
Place: USA, NY, Stone Ridge, Ridgely Manor, Language: English, Words: 110
URL: https://www.ramakrishnavivekananda.info/vivekananda/volume_9/letters_fifth_series/144_isabel.htm

==========

RECIPIENT: Leggett, Francis H
Gap: 1220 days (3.3 years)
Previous letter: 1896-12-13
This letter: 1900-04-17
Place: USA, Language: English, Words: 196
URL: https://www.ramakrishnavivekananda.info/vivekananda/volume_7/epistles_third_series/51_mr_leggett.htm

==========

RECIPIENT: Hale, G. W. Mrs.
Gap: 1185 days (3.2 years)
Previous letter: 1896-07-07
This letter: 1899-10-05
Place: USA, NY, Stone Ridge, Ridgely Manor, Language: English, Words: 312
URL: https://www.ramakrishnavivekananda.info/vivekananda/volume_9/letters_fifth_series/146_mother_church.htm

On reading all of these letters, it does not seem like these were written after a long period of no contact, unlike the letter to Pramadadas Mitra. In the letter to Mrs Hale, Swamiji opens by thanking her for her “kind words”, suggesting frequent correspondence. One can infer then that the letters written by Swamiji in this period of a few years have been lost. Hence, these are gaps made by the limitations of the current dataset, which only includes surviving letters.

One particular letter coming up here, written to Francis Leggett, caught my attention and led me down a rabbithole. Here is the letter given in full.

17th April, 1900.

MY DEAR MR. LEGGETT,

Herewith I send the executed Will to you. It has been executed as desired by her, and of course, as usual, I am requesting you for the trouble of taking charge of it.

You and yours have been so uniformly kind to me. But you know, dear friend, it is human nature to ask for more favours (now that they have come) where it gets from.

I am only a man, your child.

I am so sorry A__ has made disturbances. He does that now and then, at least used to. I do not venture to meddle, for fear of creating more trouble. You know how to manage him best. By the time you receive this letter, I will be off from San Francisco. Will you kindly send my Indian mail C/o Mrs. Hale, 10 Aster Street, Chicago, and to Margot in the same place? Margot writes very thankfully of your gift of a thousand dollars for her school.

May all blessings ever follow you and yours for your uniform kindness to me and mine, is the constant prayer of

Yours affectionately, VIVEKANANDA.

As we can see, there is a redaction of the name by the editors. This piqued my interest a lot. I had no idea who this person was I was very curious to see if I could find the redacted person's identity just by looking at the data. I immediately checked for other redactions. From this letter alone we know that the person who was redacted is male (since Swamiji uses "he") and that he was in America in 1900. These clues will help us later on.

```
================================================================================
SYNTACTIC ANALYSIS OF EDITORIAL REDACTIONS
================================================================================

1️⃣ REDACTION TOKEN EXTRACTION
--------------------------------------------------------------------------------
Total redaction tokens found: 13
Unique redaction forms: 4

All redaction forms by frequency:
full_token
```

```
A__    10
S__     1
X__     1
M__     1
Name: count, dtype: int64

Redaction forms by initial:
initial
A    10
S     1
X     1
M     1
Name: count, dtype: int64
```

Immediately I saw that A__ is the most redacted person in the entire corpus. Here is a closer look:

```
All redaction forms with reuse statistics:
  full_token  frequency  num_letters  year_span  num_recipients
0        A__         10            5  1899-1900               4
1        M__          1            1  1895-1895               1
2        S__          1            1  1900-1900               1
3        X__          1            1        N/A               1

Redaction forms appearing in multiple letters: 1
Redaction forms appearing with multiple recipients: 1
```

One can immediately see that A__ has been redacted from 5 letters to 4 different recipients, all in the time period of 1899-1900. I checked the number of underscores after the name, and they were uniform across (only 2, no hints about the length of the name). At this point I thought that maybe the editors kept the first letter of the actual name in the redacted version, inferring that the actual name might start with A. This was difficult to assume since a redaction with X__ was also present, denoting that there was some person whose name started with ‘X’ but no such person was known to be in Swamiji’s circle.

I wanted to see all the 13 occurrences of the redaction with context so I could get some insights.

```
Found 13 redacted-name occurrences

================================================================================
Redaction : A__
Letter    : XLVI Mother
Date      : 1899-11-30 00:00:00 (1899.0)
Recipient : Leggett, Francis Mrs.
Sentence  : These lines in haste as A__ is waiting.
================================================================================
Redaction : A__
Letter    : CLXXIII Dhira Mata
Date      : 1900-04-08 00:00:00 (1900.0)
Recipient : Bull, Sarah
Sentence  : Ole Bull

DEAR DHIRA MATA,

Here is a long letter from A__.
================================================================================
Redaction : A__
Letter    : CLXXIII Dhira Mata
Date      : 1900-04-08 00:00:00 (1900.0)
Recipient : Bull, Sarah
Sentence  : A__ has done good work so far; and, of course, you know I do not meddle with my workers at all.
================================================================================
Redaction : A__
Letter    : CLXXIII Dhira Mata
Date      : 1900-04-08 00:00:00 (1900.0)
Recipient : Bull, Sarah
Sentence  : Did you reveal to A__ that I have given over to you the charge of the entire work?
================================================================================
Redaction : A__
Letter    : CLXXIV Joe
```

Date   : 1900-04-10 00:00:00 (1900.0)
Recipient : MacLeod, Josephine
Sentence : I got a letter from A__ stating that he was going to leave New York.
==========
Redaction : A__
Letter   : LI Mr. Leggett
Date   : 1900-04-17 00:00:00 (1900.0)
Recipient : Leggett, Francis H
Sentence : I am so sorry A__ has made disturbances.
==========
Redaction : A__
Letter   : CLXXVI Joe
Date   : 1900-04-20 00:00:00 (1900.0)
Recipient : MacLeod, Josephine
Sentence : I am so sorry this little quarrel came with A__.
==========
Redaction : A__
Letter   : CLXXVI Joe
Date   : 1900-04-20 00:00:00 (1900.0)
Recipient : MacLeod, Josephine
Sentence : You simply wrote me a general idea that A__ wanted to keep things in his hands.
==========
Redaction : A__
Letter   : CLXXVI Joe
Date   : 1900-04-20 00:00:00 (1900.0)
Recipient : MacLeod, Josephine
Sentence : All that is, sure, the outcome of this A__ business!
==========
Redaction : A__
Letter   : CLXXVI Joe
Date   : 1900-04-20 00:00:00 (1900.0)
Recipient : MacLeod, Josephine
Sentence : Now A__ disturbs Mrs.

As we can see, all of these are redacted names. We also get to know more context about A__ reading the letters themselves. Eitherway, having no other idea, I just assumed that the actual name of A__ began with ‘A’ itself and began to check for suspects. The idea was simple, here were the things I knew for sure:

1) A is in America in 1900, specifically New York (`CLXXIV Joe`)
2) Swamiji calls A ‘my worker’, terms he would only use for a disciple or a monastic brother. (`CLXXIII Dhira Mata`)
3) Swamiji is on friendly terms with A as he is traveling with A in America in 1899 (`XLVI Mother`)

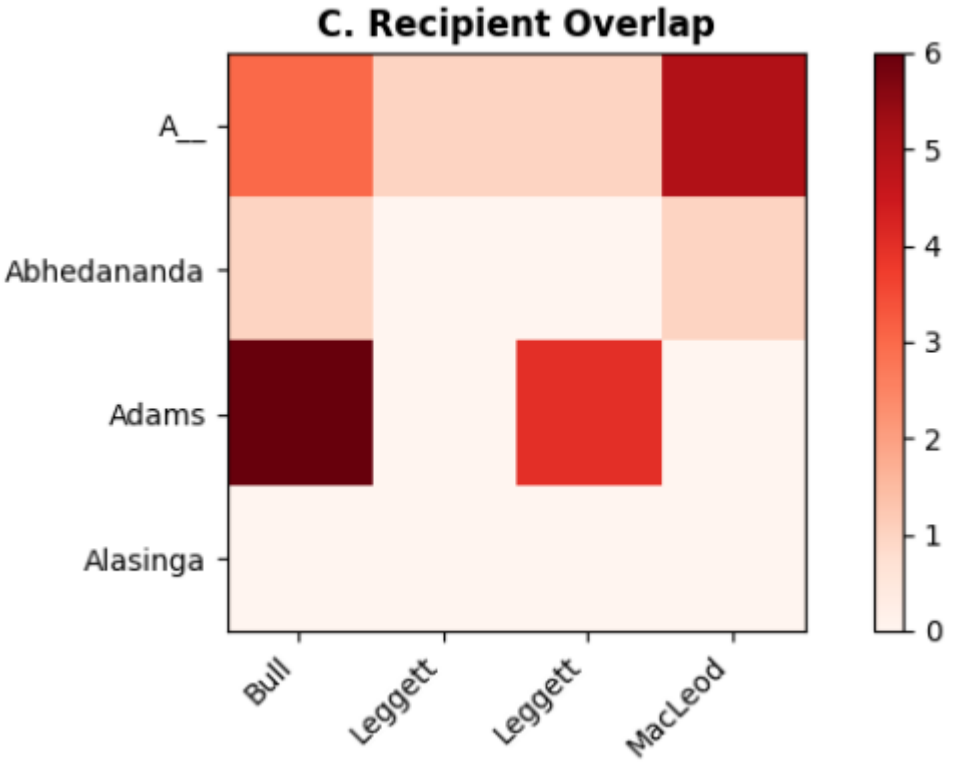
Honestly, anyone well versed with the Sri Ramakrishna-Vivekananda literature can deduce the person’s name from these clues alone. However, as a data scientist I cannot make deductions without statistical basis.

Apart from this, we have assumed that the actual name of the person starts from A. Another assumption we can make is that the same person, when not talked in a negative light, has his name uncensored. And we can also assume that Swamiji mostly uses the uncensored name of the person with the same recipients with which he uses the redacted names, namely Mrs. Bull, Mrs. MacLeod and Mr and Mrs. Leggett.

Now I begin to look for mentions of all people whose name starts from A, and matching these patterns from what I know. Here are the list of people whose names start with A:

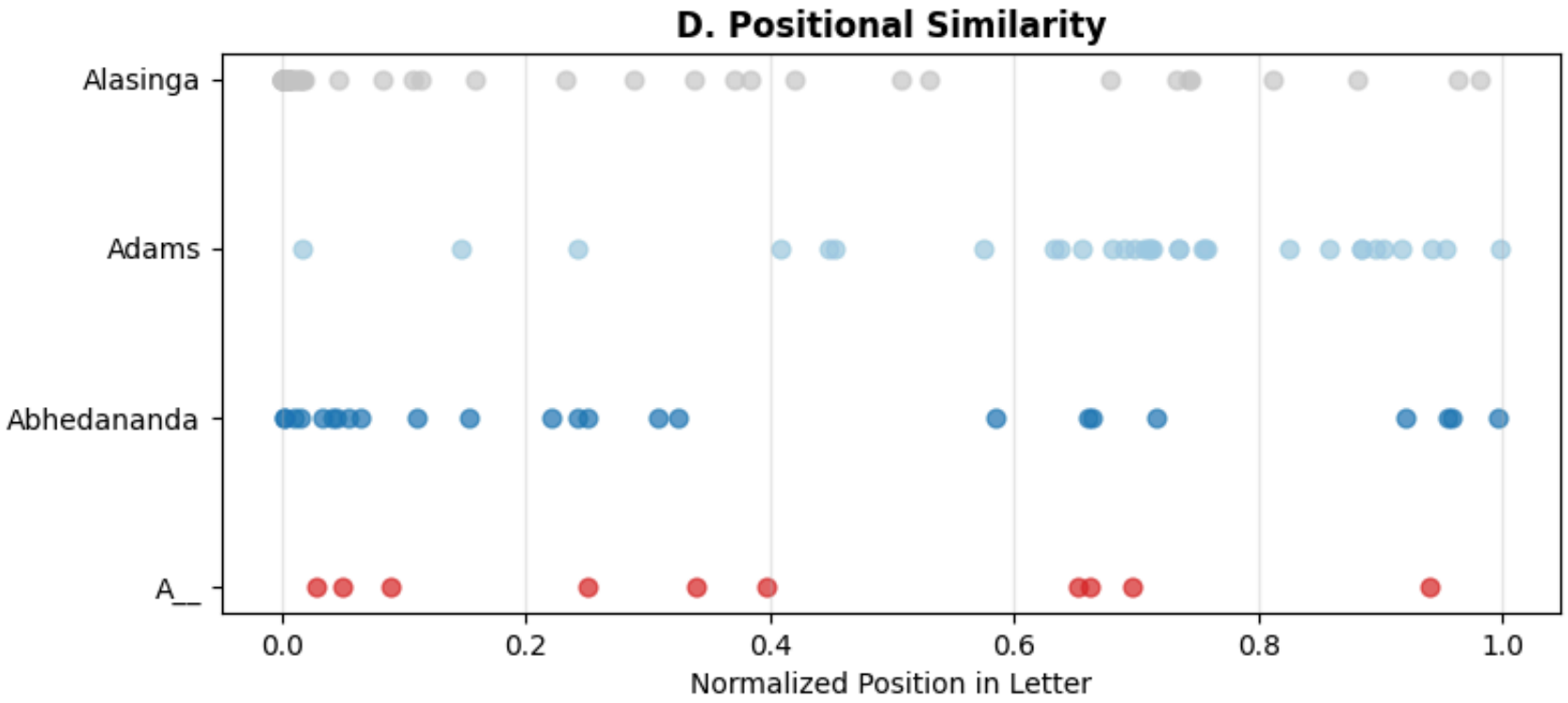
```
Audience names where ANY name starts with 'A' (total: 13):

Abhedananda, Swami
Akhandananda, Swami
Ghosh, Atul Chandra
Hansbrough, Alice (Shanti)
Maharaja of Khetri (Singh, Ajit)
Perumal, Alasinga
Florence Adams
Sturges, Alberta
```

Very quickly we can eliminate many of the options. None of Swami Akhandananda, Atulchandra Ghosh (youngest brother of Girishchandra Ghosh), Maharaja of Khetri and Alasinga Perumal have been known to have left India and been in America in 1899-1900. Alice, Adams and Alberta are females. This leaves us with only one option, Swami Abhedananda.



This is the Heatmap showing recipient overlap. As we can see, A__ has most concentrated occurrence with Bull and MacLeod, appearing to the Leggetts only once each. In exactly the two recipients with most A__ occurrences, Abhedananda also occurs. Adams does occur with Bull, but not so with MacLeod. Alasinga and Adams were controls.

I decided to also see positional similarity keeping Alasinga and Adams as controls. This is basically checking if their mentions come in the same points in the letter.



Looking at this plot, the closest match is that with Swami Abhedananda. I concluded at this point and I was happy with it and inferred that only Swami Abhedananda could have been the redacted person.

I was wondering whether or not to write this down, since the publishers redacted his name and I did not want to use data analysis to overwrite editorial restraint. But then I checked the latest official “Letters of Swami Vivekananda” Vols 1-4, published by the Advaita Ashrama in 2025, and I was happily surprised to see that in this version, the redacted names are given in full. A__ is indeed Swami Abhedananda. Since the redacted names were officially published by Ramakrishna Mission, Advaita Ashrama, I do not feel any guilt in writing this down now!

You might remember we had three more redactions, these were impossible to find since they only appeared once.
==========

```
Redaction : M__
Letter    : LXV Sharat
Date      : 1895-12-23 00:00:00 (1895.0)
Recipient : Saradananda, Swami
Sentence  : Think of the case of M__ Babu!
================================================================================
Redaction : S__
Letter    : LII Aunt Roxy
Date      : 1900-05-02 00:00:00 (1900.0)
Recipient : Blodgett, Mrs.
Sentence  : S__, and the other friends.
================================================================================
Redaction : X__
Letter    : LIII Kali
Date      : NaT (nan)
Recipient : Abhedananda, Swami
Sentence  : "Work is apt to cloud spiritual vision." X__ is very eager to come, but unless the foundation is strongly laid,
there is every likelihood of everything toppling down.
```

So I just checked the latest editions and found the uncensored names for the readers, as published officially by Advaita Ashrama now.

M__ is quite simply Mahendranath, the famous compiler of the Gospel, so here it means Mahendra babu.
S__ is Severance, referring to Mrs. Caroline M. Severance (1820-1914) who was an American abolitionist, suffragist etc. It seems that she has been mentioned once more in the letter written to Sister Nivedita on 23rd December, 1899.

Now the last redaction, the one with X__ redaction is quite interesting. First of all, the date was indeterminate. In the latest version, the date has been recovered from postal marks on the envelope, and given as 4th October 1895. Also more importantly, the latest version says that the recipient is Swami Ramakrishnananda, and not Swami Abhedananda! And this redaction has been preserved, for whatever reason, in the latest version as well. Instead of the “X__” redaction, the redaction simply says “T”.

This makes me wonder, apart from the number of total letters jumping from 763 to 815, what else has new evidence uncovered? Unfortunately, there is no digital version of the new 4-volume Letters to Swami Vivekananda that I can access.

Now we come to a new notebook called Swamiji_Private_4_semantics.ipynb.

As the name suggests, I now try to do some high order syntactic analysis of the letters hoping to find something more. I use similar techniques as I used on the public corpus like PCA.

Technically, this analysis extracted structural linguistic features from each letter: such as clause type proportions (imperative, interrogative, declarative), grammatical agency (first- vs second-person subjects), modality and passive voice usage, sentence length statistics, syntactic depth, coordination and subordination density, and part-of-speech ratios: using a dependency parser without semantic lexicons or embeddings.
These features are standardized and projected into a low-dimensional space via PCA to model the corpus' structural variation, enabling comparison across time and recipients, identification of stylistic clusters, and detection of outliers based on distance from the structural centroid and local density, all while remaining agnostic to lexical content and topic.

Simply put, his analysis looks at how the letters are written rather than what they say. It measures patterns in sentence structure, tone, and grammatical posture—such as whether the writer tends to give commands, ask questions, speak as "I" or address "you," write simply or with heavy structure, and how forceful or restrained the style is. Each letter was turned into a stylistic profile, and then compared these profiles to see how writing style shifts over time, differs by recipient, and which letters stand out as unusually written, all without using word meanings, topics, or interpretation of ideas.

I am putting some numbers and tables for completeness, the reader not interested in the technical details can skip to the interpretations after the tables of numbers.

```
================================================================================
DESCRIPTIVE STATISTICS
================================================================================

--- OVERALL CORPUS STATISTICS ---
                      count       mean         std        min        25%        50%         75%          max
pct_imperative        658.0   0.234447    0.151857   0.000000   0.140977   0.222222    0.312500     1.000000
pct_interrogative     658.0   0.033741    0.068867   0.000000   0.000000   0.000000    0.050000     0.562500
pct_declarative       658.0   0.702478    0.167131   0.000000   0.600000   0.707108    0.812500     1.000000
pct_exclamative       658.0   0.029334    0.054961   0.000000   0.000000   0.000000    0.043478     0.391304
```

first_person_agency	658.0	0.347985	0.182487	0.000000	0.219062	0.333333	0.454004	1.000000
second_person_agency	658.0	0.100032	0.098391	0.000000	0.018266	0.080000	0.142857	0.571429
other_agency	658.0	0.550464	0.200450	0.000000	0.424451	0.560488	0.683991	1.000000
modal_density	658.0	0.270282	0.168554	0.000000	0.153846	0.250000	0.368421	1.000000
passive_density	658.0	0.089231	0.092035	0.000000	0.000000	0.071429	0.136364	0.500000
sent_len_mean	658.0	16.669678	4.513117	6.500000	13.785000	16.033670	18.633523	48.000000
sent_len_var	658.0	96.720655	100.127335	0.000000	46.615748	74.135716	113.194420	1176.000000
sent_len_max	658.0	39.161094	16.880587	12.000000	28.000000	36.000000	46.000000	138.000000
max_tree_depth	658.0	8.044073	2.525885	3.000000	6.000000	8.000000	9.000000	23.000000
subordination_density	658.0	0.615092	0.368579	0.000000	0.363636	0.571429	0.789474	2.375000
coordination_density	658.0	0.544143	0.290889	0.000000	0.372321	0.528857	0.686134	2.125000
noun_density	658.0	0.120379	0.029396	0.017857	0.101828	0.119665	0.139385	0.218182
verb_density	658.0	0.098492	0.019108	0.000000	0.087436	0.098885	0.111036	0.169014
adj_density	658.0	0.051577	0.018883	0.000000	0.040328	0.052386	0.063816	0.110390
adv_density	658.0	0.053961	0.020906	0.000000	0.041330	0.053610	0.066408	0.128205

```
================================================================================
PCA DIMENSIONALITY REDUCTION
================================================================================

Explained variance by first 10 components:
  PC1: 0.2401 (24.01%)
  PC2: 0.1362 (13.62%)
  PC3: 0.1014 (10.14%)
  PC4: 0.0757 (7.57%)
  PC5: 0.0640 (6.40%)
  PC6: 0.0553 (5.53%)
  PC7: 0.0520 (5.20%)
  PC8: 0.0483 (4.83%)
  PC9: 0.0419 (4.19%)
  PC10: 0.0380 (3.80%)

Cumulative variance (first 5 PCs): 0.6174
```

```
--- TOP FEATURE LOADINGS ON FIRST 3 PCs ---

PC1 loadings:
sent_len_max             0.835572
sent_len_mean            0.774468
sent_len_var             0.738079
max_tree_depth           0.733004
coordination_density     0.679291
subordination_density    0.659718
noun_density             0.541091
other_agency             0.522364
passive_density          0.499274
first_person_agency      0.473943

PC2 loadings:
pct_imperative           0.776172
pct_declarative          0.675968
modal_density            0.662678
verb_density             0.473366
other_agency             0.472223
second_person_agency     0.410384
subordination_density    0.393001
first_person_agency      0.312924
adj_density              0.248241
noun_density             0.243545
```

For any data, interpretation varies from data scientist to data scientist. Here are my subjective interpretations of this data.

1. Swamiji's letters differ most in how complex the sentences are, not in emotional tone

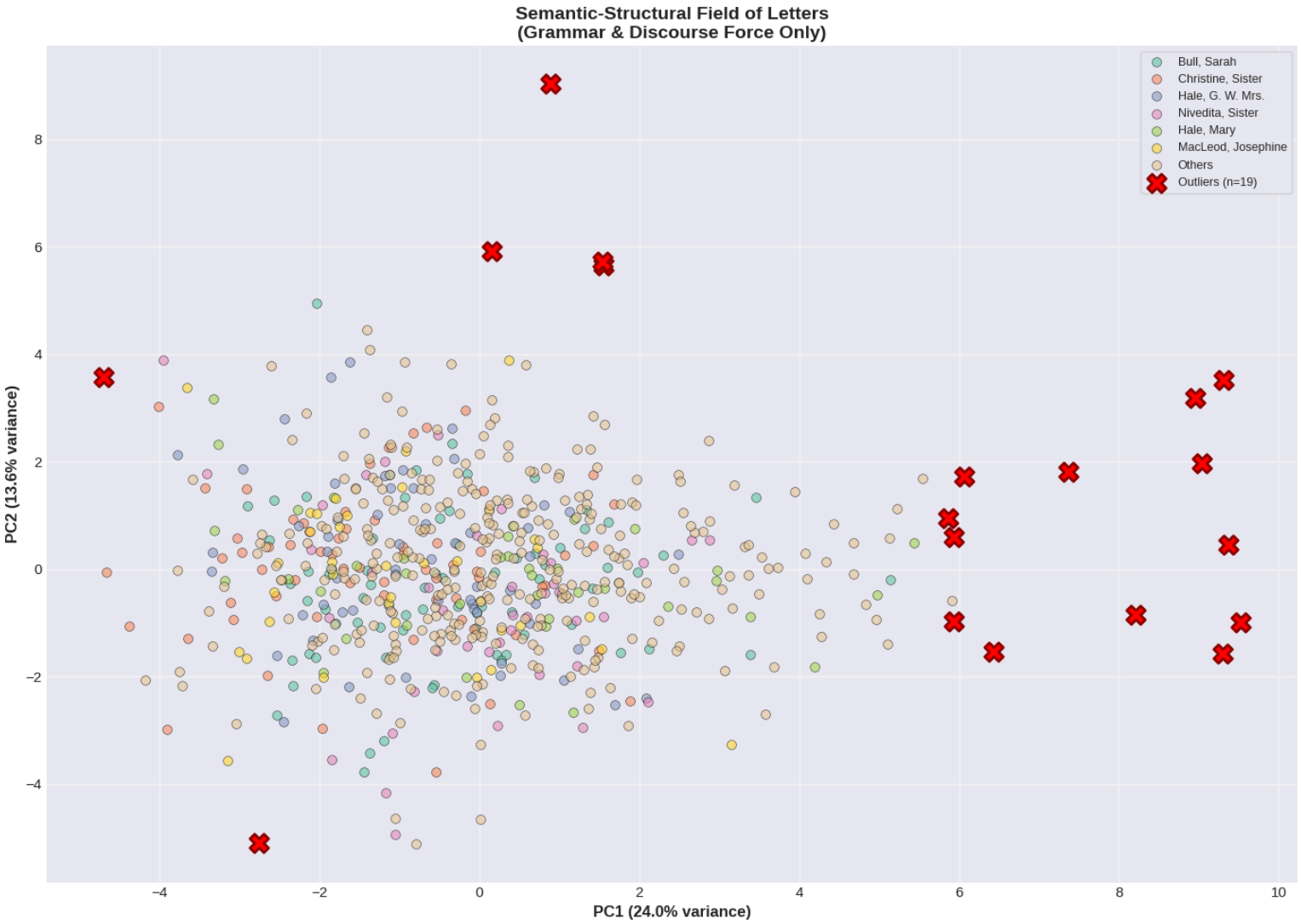
This comes directly from PCA results. The first principal component (PC1) explains 24.01% of all variation, and its strongest contributors are sentence length, sentence length variance, maximum sentence length, syntactic tree depth, coordination, and subordination. Features related to

tone contribute much less here. This shows that the biggest difference between Swamiji's letters is how structurally dense and elaborate the writing is, rather than whether it is forceful or emotional.

2. Swamiji systematically changes his style depending on the recipient

Recipient-level statistics show consistent differences across multiple features. For example, letters to Nivedita have a high imperative proportion (0.27) and high noun density (0.137), while letters to Sarah Bull show a much higher declarative proportion (0.78) and stronger first-person agency (0.43). Because these differences appear across several independent measures, not just one, they indicate deliberate audience adaptation, not random fluctuation.

Here are some outliers plotted along PC1 and PC2



This is the same technical process that produced a triangle out of nowhere in the public corpus. This is how most PCA plots look, like a random blob of points, so you can understand my awe when I saw a perfect triangle previously.
Eitherway, we can see that the outliers are visibly separate from the mainstream blob of letters in the center. Here are some outliers:

==========

```
OUTLIER DETECTION
================================================================================

Outliers identified (distance-based): 14
Outliers identified (density-based): 14
Total unique outliers: 19

--- TOP 10 OUTLIERS BY DISTANCE FROM CENTROID ---
  year                  audience  dist_from_centroid  local_density
1896.0   Light, Letter to Editor            9.950960       0.247287
1898.0          Nivedita, Sister            9.573569       0.291796
1894.0          Hume, R. A. Rev.            9.507951       0.273792
1896.0                Hale, Mary            9.427493       0.286272
1896.0             Indian Mirror            9.383173       0.324390
1893.0 Desai, Viaharidas Haridas            9.243977       0.320594
1896.0          Nivedita, Sister            9.080970       0.195723
1900.0     Leggett, Francis Mrs.            8.249734       0.386170
1893.0 Desai, Viaharidas Haridas            7.584951       0.424671
1898.0 Husain, Mohammed Sarfaraz            6.620359       0.515424
```

Some of these outliers are very surprising. Firstly, the structural differences in Letter to the Editor of the Light along with Indian Mirror surface up immediately. These are not regular letters, but those written to editors of newspapers like Indian Mirror and the Light.

Another striking letter is the one written to Reverend R. A. Hume. Rev. Hume wrote to Swamiji challenging his statements about Christian missionaries and said some indecent things about India. This fiery letter from Swamiji is a reply to that letter. I would recommend giving it a read. Here is the link:

https://www.ramakrishnavivekananda.info/vivekananda/volume_7/epistles_third_series/15_brother.htm

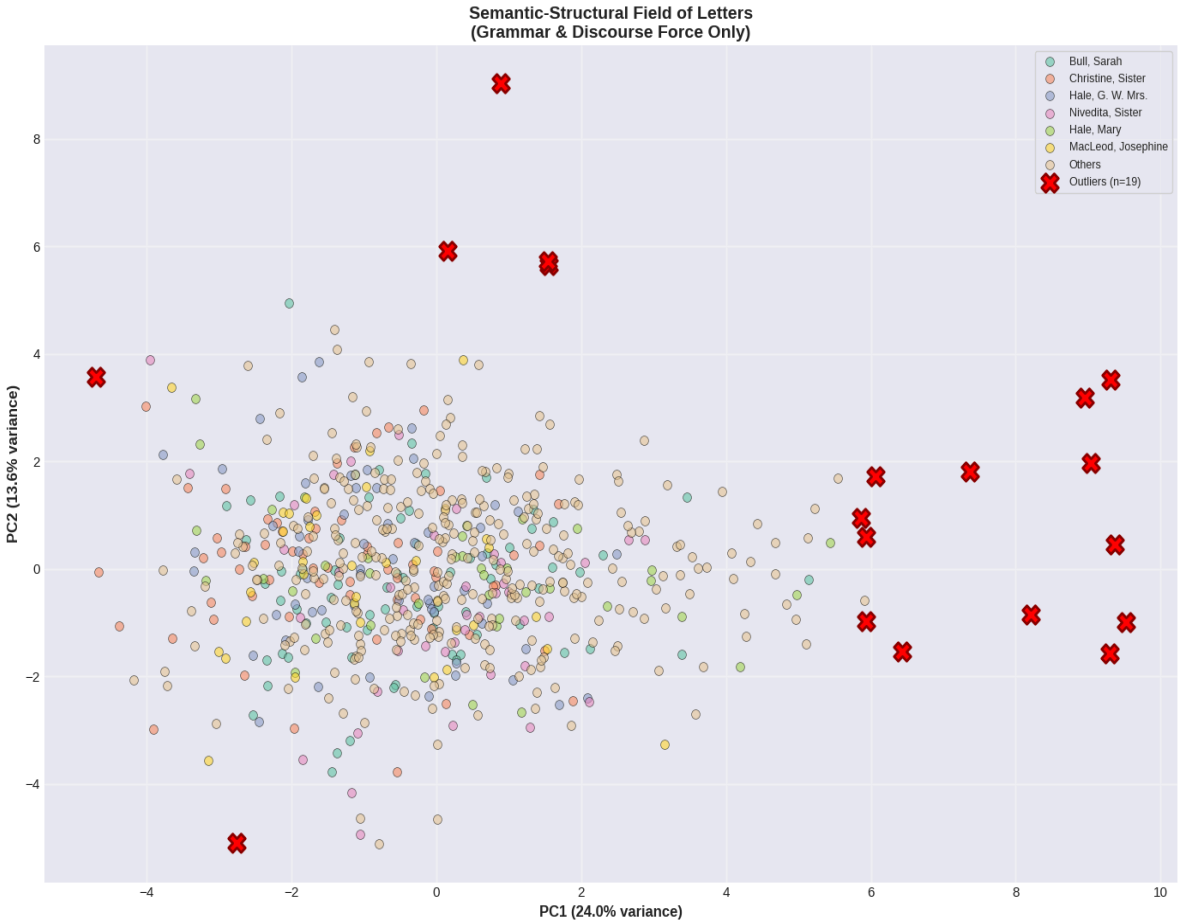
A famous letter to Mohammed Sarfaraz Husain also surfaces, this is one of the only known letters written by Swamiji talking about the unity of Advaita and the Islamic faith. The link is provided.

https://www.ramakrishnavivekananda.info/vivekananda/volume_6/epistles_second_series/142_friend.htm

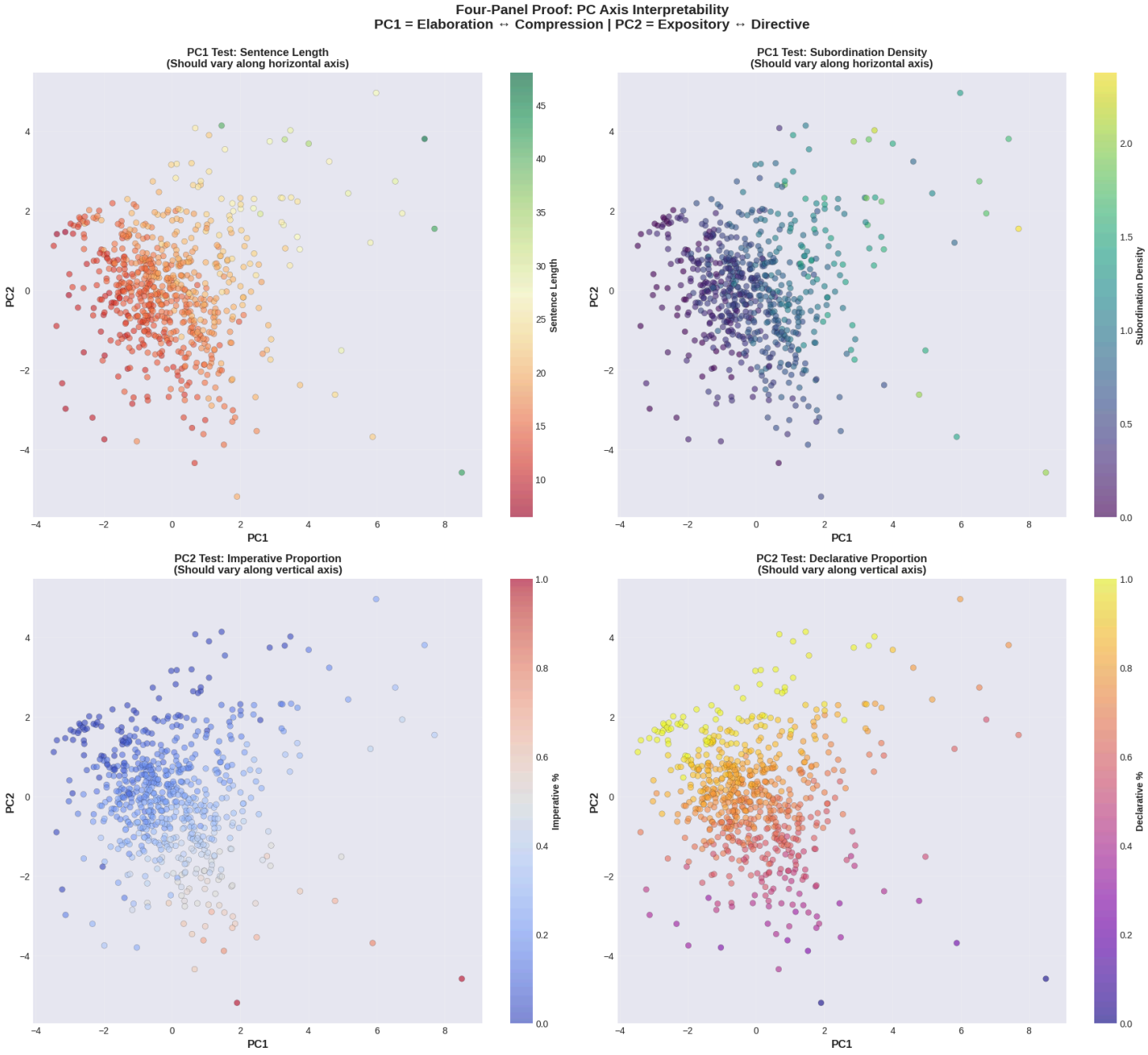
The other outlier letters are outliers from the perspective of metrics like sentence complexity and such as shown previously, not too much in the content of the letters themselves.

Now as before, PC1 and PC2 are not helpful names. PC1 and PC2 are basically the two axes along which the variation of the letters when plotted in a high dimensional state with respect to various metrics shown earlier (sentence complexity, max tree depth etc.) is the highest. Basically, these two axes capture the diversity of the high dimensional plots of letters the best, and since they are two in number, they can be plotted in 2d charts. Now I try to find a good name for PC1 and PC2 so things make more sense, similar to how we tried to find names for the topics derived in the public corpus analysis.

In trying to find good explanatory names for the two PCs, I ran some analysis and found that PC1 which was the x-axis basically demonstrated how compressed the writing of Swamiji was in that letter vs how elaborate. Basically, as you go from left to right in the plots, you can assume that letters are getting more and more elaborate. Similarly, PC2 (y-axis) basically meant how expository Swamiji's writing was ("This means X, that means Y") vs how directive ("Do X, do Y"). This means as you go from down to up in a graph, you are going from letters that are more expository to letters that are more directive.
It would be easy to prove this, if one plots the letters along PC1 and 2 and colours them based on metrics respectively, instead of a random plot of colours one should see smooth gradients along x and y axes. I am showing an example of the same plot with colours representing recipients, also shown previously, (which is not the pattern the PCs match, and hence they will be random) vs the same plots but coloured representing the metrics of respective PCs.



As we can see, this is quite random and noisy. Now let us see the same graph but coloured by the metrics that affect PC1 and 2 the most:


Four-Panel Proof: PC Axis Interpretability
PC1 = Elaboration ↔ Compression | PC2 = Expository ↔ Directive
PC1 Test: Sentence Length
(Should vary along horizontal axis)
PC1 Test: Subordination Density
(Should vary along horizontal axis)
PC2 Test: Imperative Proportion
(Should vary along vertical axis)
PC2 Test: Declarative Proportion
(Should vary along vertical axis)
PC1
PC2
Sentence Length
Subordination Density
Imperative %
Declarative %

Tada! Smooth gradients all over just as expected. Based on this, we give the names as follows:

PC1 → Syntactic Elaboration Axis

Subordination density: r = 0.81
Sentence length: r = 0.77
Modal density: r = 0.71
Interpretation: PC1 measures the degree of syntactic complexity and elaboration in discourse. High PC1 = complex, subordinated, lengthy sentences. Low PC1 = compressed, simple syntax.

PC2 → Illocutionary Force Axis

Declarative proportion: r = 0.88 (positive)
Imperative proportion: r = -0.81 (negative)
Interpretation: PC2 measures the speech act orientation of discourse. High PC2 = declarative, expository, explanatory. Low PC2 = imperative, directive, commanding.

Renamed Axes

Original	Renamed	Variance	Meaning
PC1	Syntactic Elaboration	32%	Complexity of sentence structure
PC2	Illocutionary Force	26%	Declarative ↔ Imperative orientation

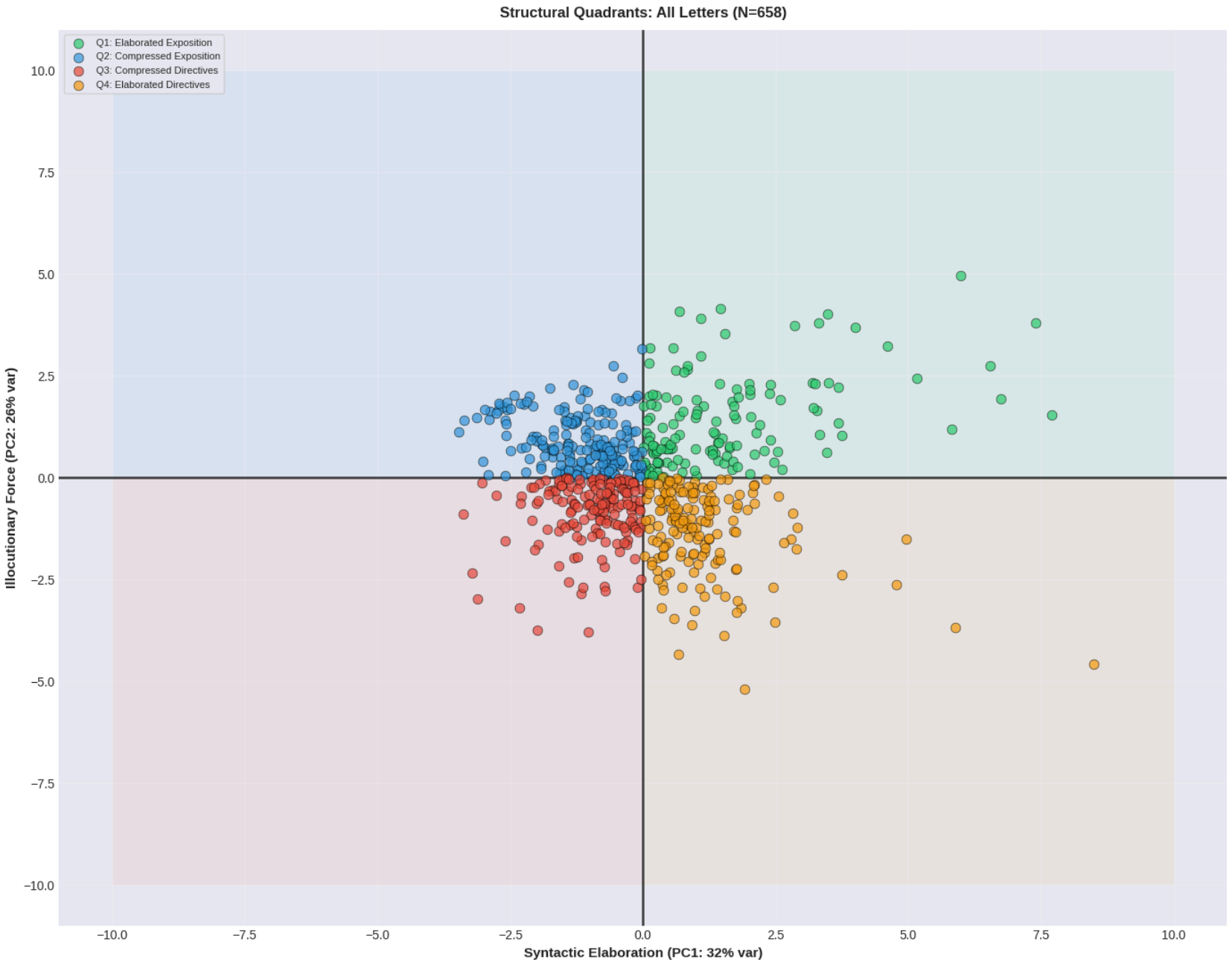
Conclusion:
The structural space of Swami Vivekananda's letters is defined by two orthogonal dimensions:

How complex the syntax is (elaboration)
What speech act is being performed (declaration vs. command)
These are not about content or topic—they are grammatical modes of discourse.
Now based on this, to ease our further interpretations, we can label the four quadrants of the 2D cartesian space produced by PC1 and PC2 accordingly.

```
================================================================================
QUADRANT ANALYSIS: STRUCTURAL MODES ACROSS CONTEXTS
================================================================================

Quadrant distribution:
quadrant
Q2: Compressed Exposition    194
Q3: Compressed Directives    165
Q4: Elaborated Directives    164
Q1: Elaborated Exposition    135
```

This basically means we have now divided the graph into four quadrants, four equal areas that is. If a letter falls in any of these areas, it means the letter is of a certain type based on Illocutionary Force and Syntactic Elaboration.
Letters that are explanatory or expository in nature, while being elaborate and lengthy fall in the first quadrant
Letters that are explanatory or expository in nature, while being quite restrained and concise fall in second quadrant.
Letters that are directive, action-demanding in nature while being restrained and compressed fall in third quadrant
Letters that are directive, action-demanding in nature while being elaborate and lengthy fall in the fourth quadrant

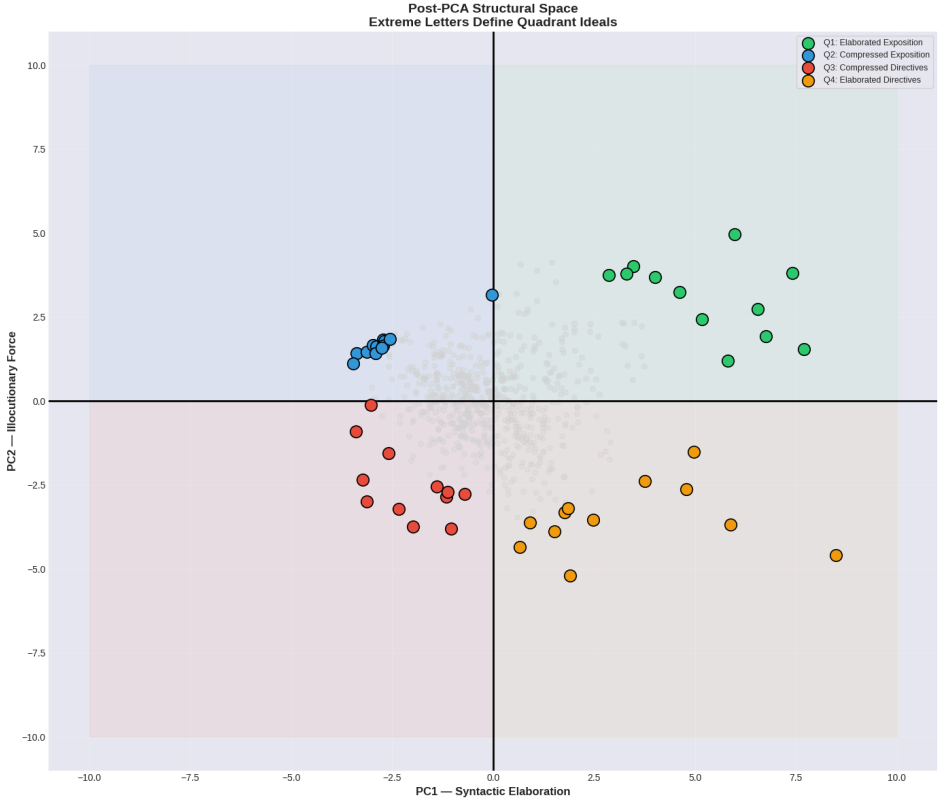
Hence the names. We can see that the letters are quite evenly distributed in the four quadrants. Now we can see how it looks if we plot all the letters with respect to these quadrants.


Structural Quadrants: All Letters (N=658)
Q1: Elaborated Exposition
Q2: Compressed Exposition
Q3: Compressed Directives
Q4: Elaborated Directives
Illocutionary Force (PC2: 26% var)
Syntactic Elaboration (PC1: 32% var)
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0

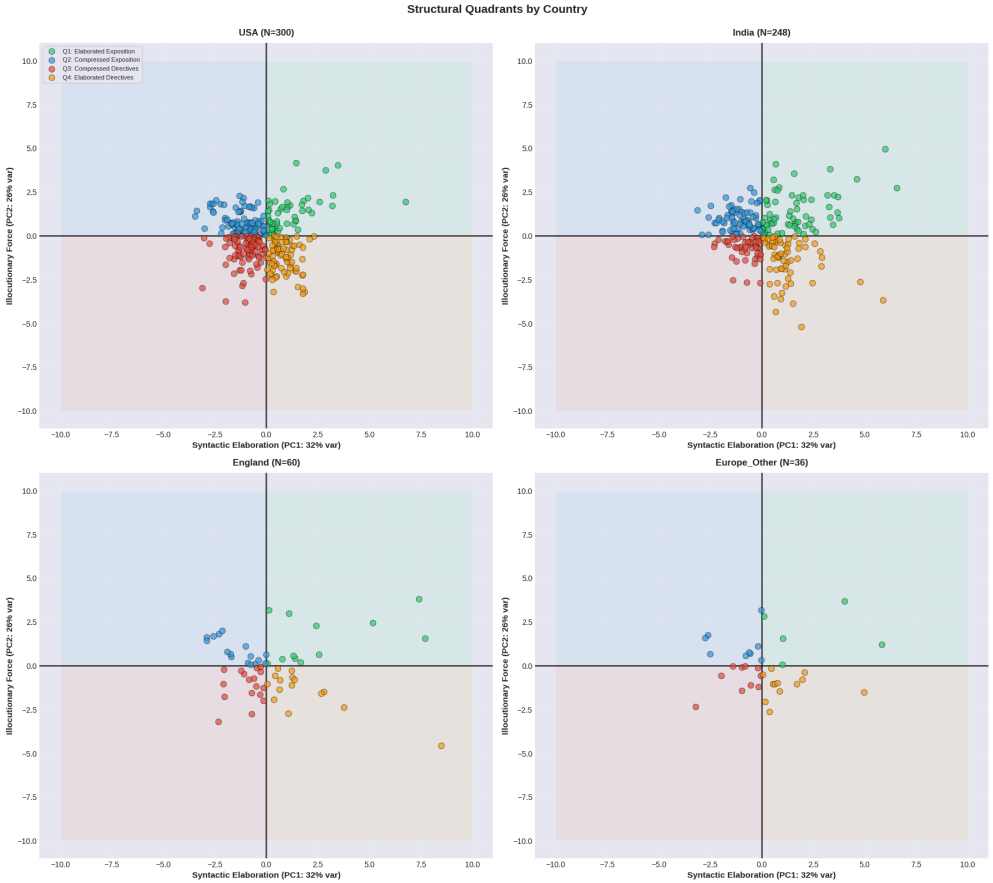
We can immediately see the “spread” of the letters is not even in the four quadrants. In Q4 and Q3, spread is quite less while it is a lot more in Q1 and Q2. What does this mean in simple terms?

“The most concise letters exhibit only modest compression, whereas the most elaborate letters display extreme expansion.”

If we only plot the most extreme points, we see that it forms a shape resembling a ring that is lopsided towards the right (positive x-axis), this supports the above statement visually.



Now we can play around within the constraints we have defined to our heart’s content! What is the nature of letters written from different countries?



Letters written in USA are all quite similar, they are clinging to the center (origin) of the space where the difference between the four categories is the least. While those written in England have the most variation along the Elaboration axis!

I have many more plots but I will just write the results here in natural language for clarity.

1. What changes by country

- In the USA and India, Swamiji most often writes brief explanations — he explains ideas, but keeps them compact.
- In England, he more often writes short, direct instructions.

- In other parts of Europe, his letters are most often long and instructive, combining detail with guidance.

In simple terms: where he is writing from or to shapes how much he explains and how forceful he is.

2. How his writing changes over time

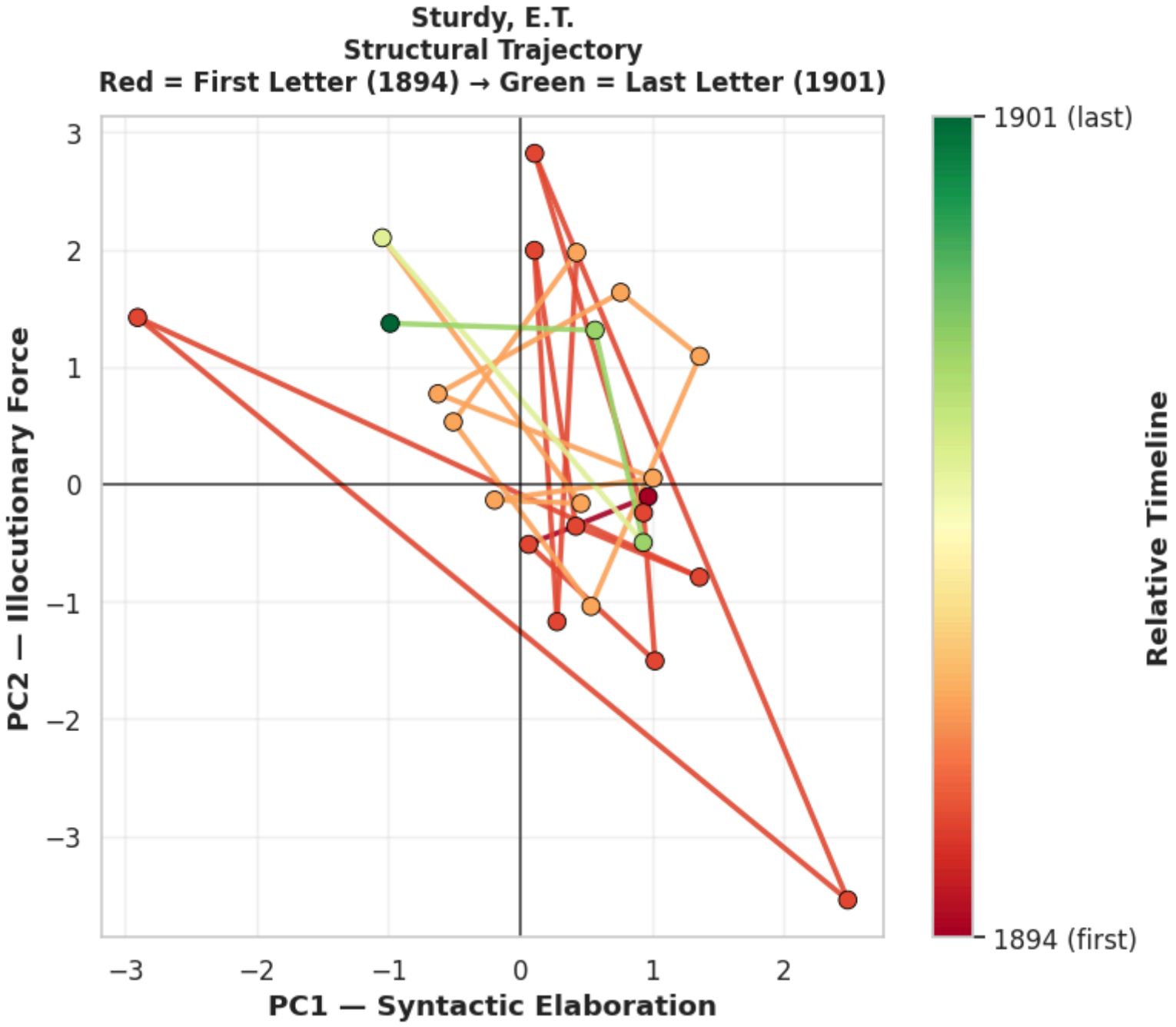
- In his early wandering years and first trip West, Swamiji mostly writes long, thoughtful explanations.
- In the middle period (1894–1897), his letters shift toward long, directive guidance, suggesting more responsibility and leadership.
- During the second Western trip, the tone becomes shorter and more directive.
- In his final years, the writing settles into brief, clear explanation rather than long instruction.

In simple terms: he moves from exploring and explaining, to directing and organizing, and finally to distilled clarity.

3. How Swamiji adapts to individuals

- To Sarah Bull and Mrs. Hale, he mainly gives clear, concise explanations.
- To Sister Christine, he writes short but strongly directive letters.
- To Sister Nivedita, he mostly explains, but not at great length.
- To Mary Hale, he writes long, reflective explanations.

Something interesting that we can now do is plot all the letters to a specific recipient in this fourfold categorized space and see how the letters jump around over time. That is, how the letters (to one recipient) go from compact to elaborate, or explanatory to directive, over the correspondence period of Swamiji and the recipient.
For example, here is how the plot for E.T. Sturdy looks like.

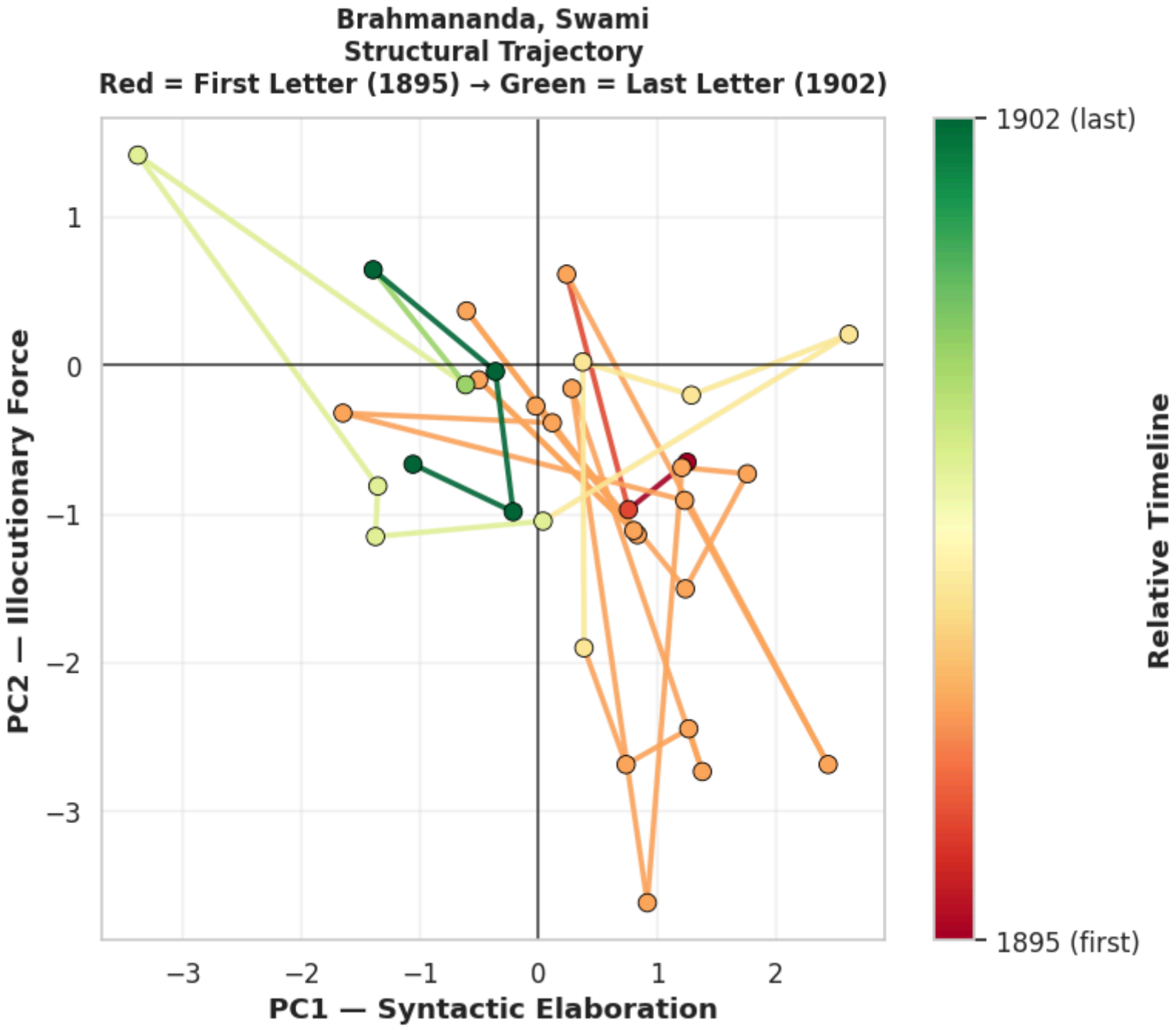

Sturdy, E.T.
Structural Trajectory
Red = First Letter (1894) → Green = Last Letter (1901)
PC2 — Illocutionary Force
PC1 — Syntactic Elaboration
Relative Timeline
1901 (last)
1894 (first)
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2

The colour shows the chronology of the letter, greenness means the letter is more recent while redness indicates oldness. We can see that the earliest letters (red points) include large jumps across quadrants, including excursions deep into Q3 (Compressed Directives) and Q4 (Elaborated Directives). This indicates that early on, Swamiji had no settled register with Sturdy—his tone shifts sharply between directive modes, sometimes brief and sharp, sometimes extended and forceful.
As time progresses (orange to yellow), the trajectory clusters closer to the **origin**, repeatedly crossing quadrant boundaries. These letters move between **Q1 and Q4**, suggesting alternating phases of explanation and instruction, but with **less extremity** than the earliest letters.
The latest letters (green points) settle mostly in the **upper half** of the space. This indicates a shift away from directive force and toward **explanation and reflection**, with syntax remaining moderately to highly elaborated.

Interestingly, when the graph is placed next to what we know about the real-life events, an unexpected contrast appears.

In the early letters to Sturdy, Swamiji's writing moves around a lot on the graph, showing that the tone and style were still finding their footing. As time goes on, the points stop jumping wildly and stay closer together, mostly in the parts of the graph linked with explanation rather than instruction. Even during the period when the relationship was breaking down in real life (with Sturdy accusing Swamiji of many ungrounded actions, and Swamiji ending the correspondence with an explosive letter that one should definitely read here https://www.ramakrishnavivekananda.info/vivekananda/volume_7/epistles_third_series/44_sturdy.htm), the graph does not show Swamiji's writing becoming tighter, sharper, or more one-sided. Instead, it keeps returning to the same calm, explanatory zones. In simple terms, while the relationship outside the letters became strained, the way Swamiji wrote stayed steady and balanced, as if the conflict did not fully take over how he expressed himself.

Here is another example with the same trajectory plot for Swami Brahmananda.


Brahmananda, Swami
Structural Trajectory
Red = First Letter (1895) → Green = Last Letter (1902)
PC2 — Illocutionary Force
PC1 — Syntactic Elaboration
Relative Timeline
1902 (last)
1895 (first)

In the early letters to Swami Brahmananda (red points), Swamiji's writing sits mostly in Q2: Compressed Exposition. This means the letters are brief, explanatory, and economical—ideas are shared without much instruction or elaboration. The tone is calm and matter-of-fact, suggesting shared understanding rather than the need to guide or persuade.

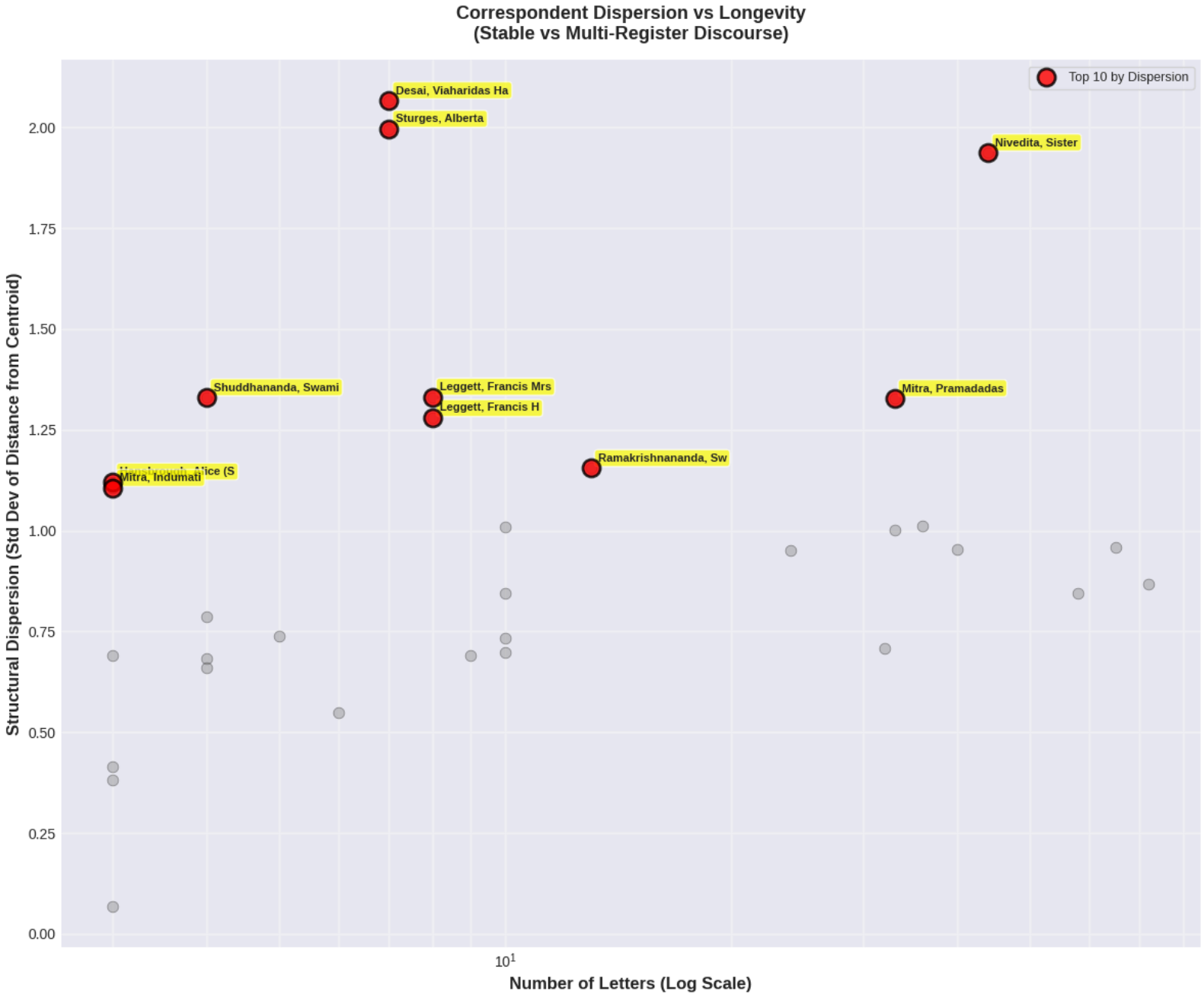
As time goes on, the trajectory shifts noticeably into Q4: Elaborated Directives. The later letters (orange to green points) move to the lower-right quadrant, showing longer, more carefully structured instruction. This is not abrupt or chaotic, but a sustained movement, indicating a change in what the letters are doing, not instability in how they are written.

So the key pattern is not "stability," but a clear directional shift:
from brief explanation → extended guidance.

This makes one thing that as Swamiji was approaching his final days (the last letter is in 1902, the year Swamiji left his mortal coil), he became more of a guide to Swami Brahmananda, knowing that the responsibility of the Ramakrishna Order was to be shouldered by Rakhal Maharaj.

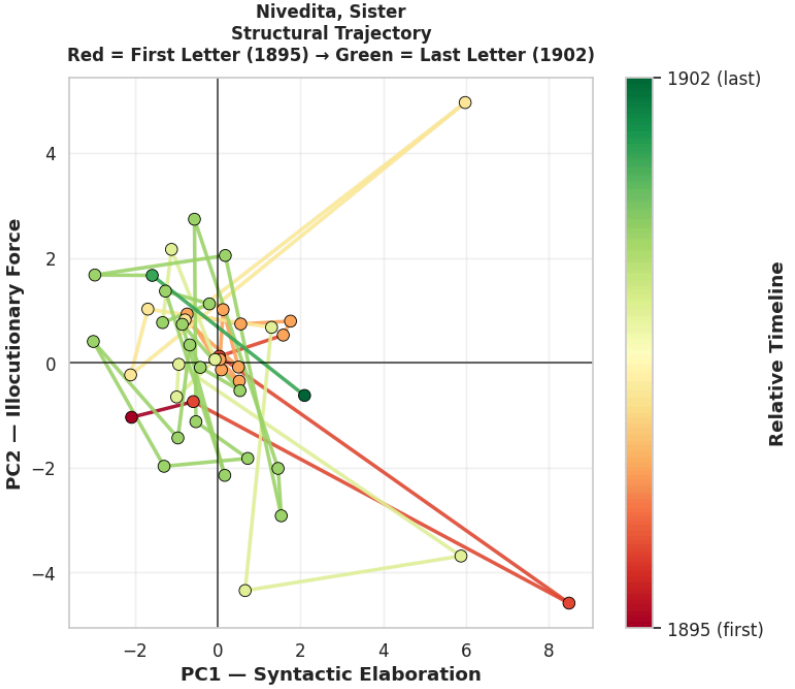
Many such plots have been made, they have not been shown in detail for clarity.

I also looked at recipients, comparing them across longevity and dispersion. Longevity tells us how long the correspondence lasted; dispersion tells us how many different "voices" or "writing styles" Swamiji used within it. Here is the graph:


Correspondent Dispersion vs Longevity
(Stable vs Multi-Register Discourse)
Top 10 by Dispersion
Desai, Viaharidas Ha
Sturges, Alberta
Nivedita, Sister
Shuddhananda, Swami
Leggett, Francis Mrs
Leggett, Francis H
Mitra, Pramadadas
Ramakrishnananda, Sw
Mitra, Indumati
Structural Dispersion (Std Dev of Distance from Centroid)
2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
10^1
Number of Letters (Log Scale)

Several correspondents with many letters sit low on the dispersion axis. This shows that even over long periods, Swamiji sometimes maintains a steady, settled mode of address. These relationships appear structurally stable rather than evolving across many registers.

In contrast, we see Sister Nivedita on the far top right corner of this graph. What does this mean in simple language?
This indicates a **long correspondence that also spans many distinct styles**. Structurally, this suggests a relationship that evolves across roles—teacher, guide, collaborator—rather than settling into a single mode.
The relationship of Swamiji and Sister Nivedita was among the most dynamic, based on analysis of Swamiji's letters to her!
This is also supported by her quadrant-letter trajectory, which moves all over the place fluidly throughout the correspondence without settling.



The graph itself seems to visually show how dynamic the relationship was between Swamiji and Sister Nivedita.

We come now to our final notebook, Swamiji_Private_FinalAnalysis.ipynb.

I wanted to do some stylometric analysis of Swamiji's works. Stylometry uses linguistic techniques to find stylistic invariants in an author's works, usually to determine authorship of contested documents. This is the field (among others) that is used, for example, to determine that most of the literature attributed to Adi Shankaracharya is most probably not written by him. The only 4 absolutely uncontested words of Adi Shankaracharya are his commentaries on the Brahma Sutras, the principal Upanishads, the Bhagavad Gita and a collection of a thousand teachings called Upadesa Sahasri.

Why would I do stylometry here, when authorship of all the letters are absolutely uncontested? Because it can help to find some stylometric outliers which I thought would be interesting to see. Also for the sake of completeness.

I used a pipeline of various algorithms, which are well known stylometric techniques. Technically, the pipeline constructs a normalized function-word frequency matrix, applies multiple Burrows' Delta distance variants, and projects the z-scored space via PCA to model stylistic variance. It performs Ward hierarchical clustering and detects stylistic outliers using Mahalanobis distance, Isolation Forest, and Local Outlier Factor, retaining consensus anomalies flagged by multiple methods. The outputs jointly characterize global stylistic structure, cluster topology, and robustly identified deviations within the corpus.

In simple words: This code compares Swamiji's letters based on **small, unconscious word choices** (like "and," "but," "to," "of") to see how similar or different they are in writing style. It maps the letters into a style space, groups similar ones together, and then uses several independent methods to find letters whose style stands out strongly from the rest. The result shows overall stylistic structure, clusters of similar letters, and a small set of letters that are unusually written compared to Swamiji's normal style.

The outliers I got were these:

```
Top 5 stylometric outliers:

Date        : 1895/05/02
Audience    : S_
Word count : 247
Place       : USA

Date        : 1895/12/08
Audience    : Christine, Sister
Word count : 108
```

```
Place       : USA, NY, New York City

Date        : 1900/01/24
Audience    : Nivedita, Sister
Word count : 107
Place       : USA, CA, Los Angeles

Date        : 1899/05/10
Audience    : Christine, Sister
Word count : 100
Place       : India, Belur

Date        : 1895/12/29
Audience    : Sturdy, E.T.
Word count : 109
Place       : USA, NY, Stone Ridge, Ridgely Manor
```

Again a letter written to a redacted correspondent comes up on top! I tried to find the recipient but I could not. Even in the latest version (Complete Words of Swami Vievekananda Vols 1-4, Advaita Ashrama 2025) this recipient is redacted. It is known that this recipient was a resident of Calcutta, he was close to Nag Mahashay, he was considering renouncing the world in 1895, and his spiritual master was not Swamiji but one of Swamiji's gurubhais. The link is provided to the letter below.

https://vivekavani.com/xxxviii-s-letters-swami-vivekananda/

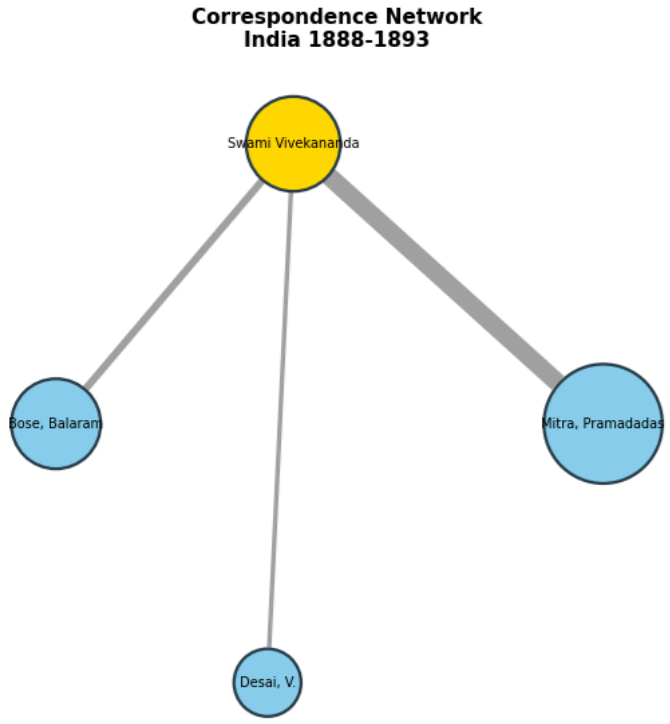
I now move onto co-mention and social network analysis, that is trying to see how often two people are mentioned together in a letter. This is something done in epistolary analysis quite often.
Basically, we extract all the proper nouns that we can find (using a technique called Named Entity Recognition or NER), isolate the names of people out of them and then map them onto recipients who corresponded with Swamiji. This is not as simple as it sounds since Swamiji often used nicknames of many people (like Kali, Gangadhar, Shashi for Swamis Abhedananda, Akhandananda, Ramakrishnananda respectively).

After manually intervening to make nickname - recipient mappings to most of the graphs, I plotted the social network of Swamiji based on preserved letters in many different ways.

Network analysis is a way of studying relationships between people by treating them as a network or web of connections. Think of it like a social map. Instead of reading letters one by one, network analysis lets us see the "big picture" of who Swamiji communicated with and how these people related to each other.

First I decided to temporally graph the social network of Swamiji based on major periods of his life. This would constitute to visually seeing how Swamiji's social network evolved over time.

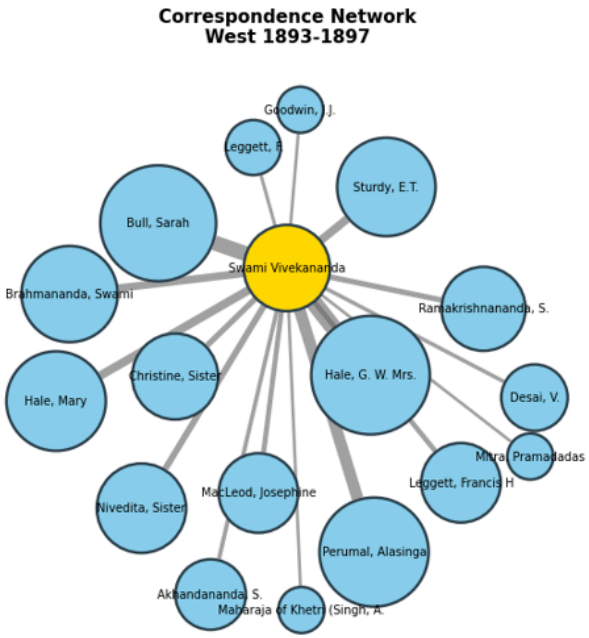
Starting from Swamiji's days as a wandering monk in India, we can see the social network is quite sparse.



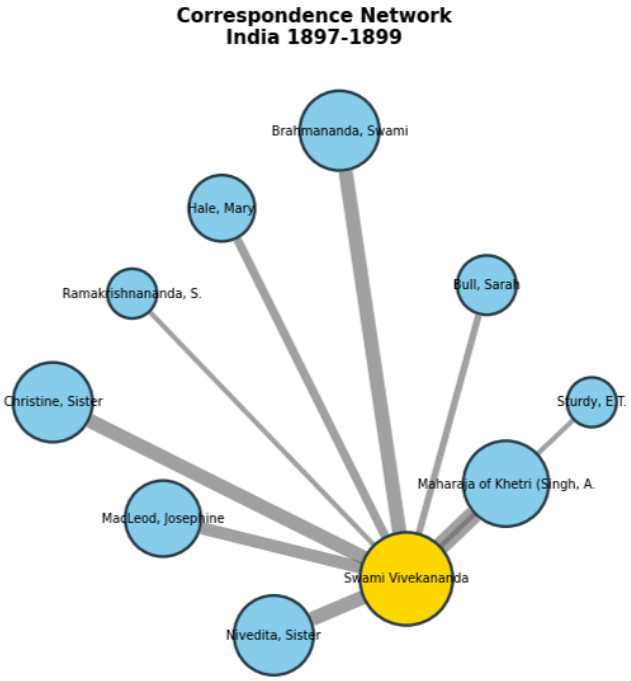
Here,
- Nodes (dots) represent people
- Edges (lines) represent relationships or interactions between them
- The thickness of lines shows how strong the connection is (more letters = thicker line)

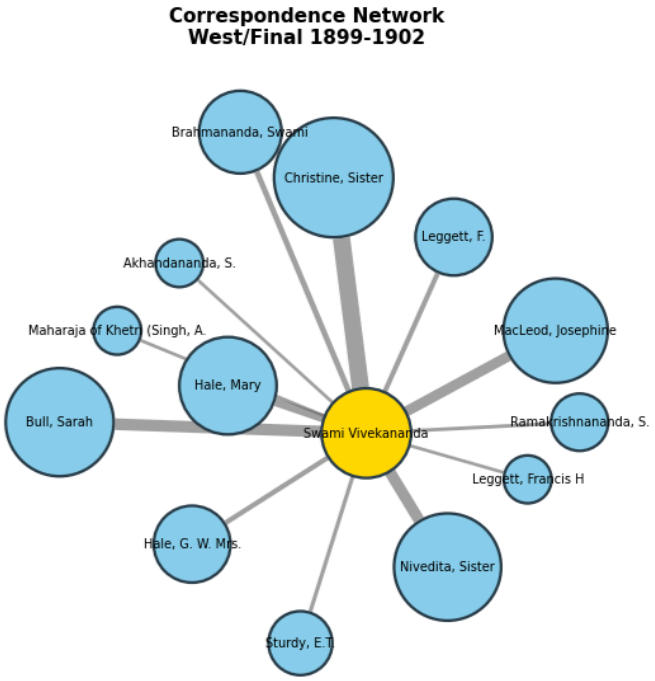
- The size of dots can show how important someone is in the network
We can immediately see that Swamiji corresponded mostly with these three individuals. Mostly with Pramadadas Mitra.

How did his social network look during the first visit to the west?



We can again immediately see the important figures by the size of the nodes and the thickness of the edges. Alasinga Preumal, Mrs. Hale and Mrs. Bull stand out the most. Here are the rest of the two graphs together.

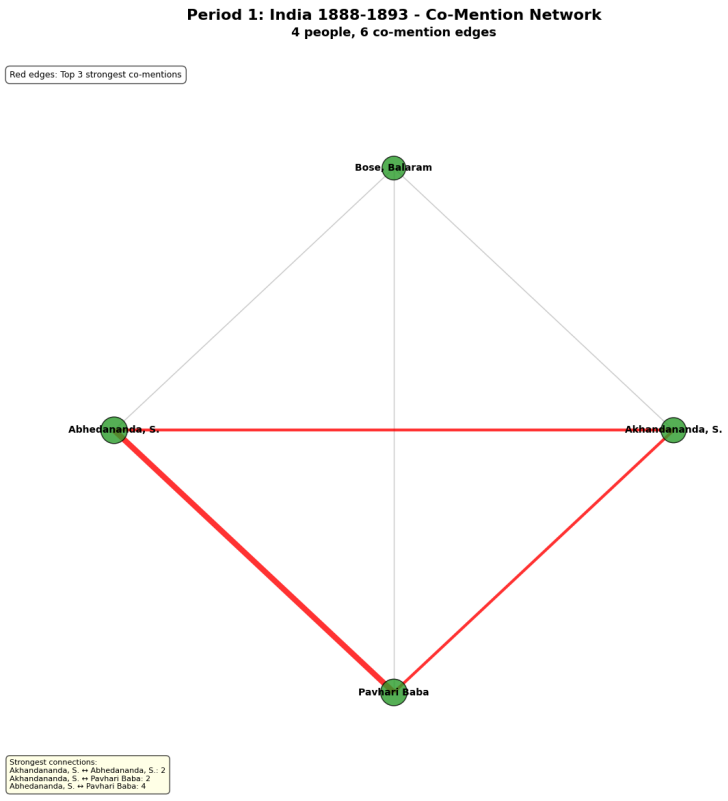




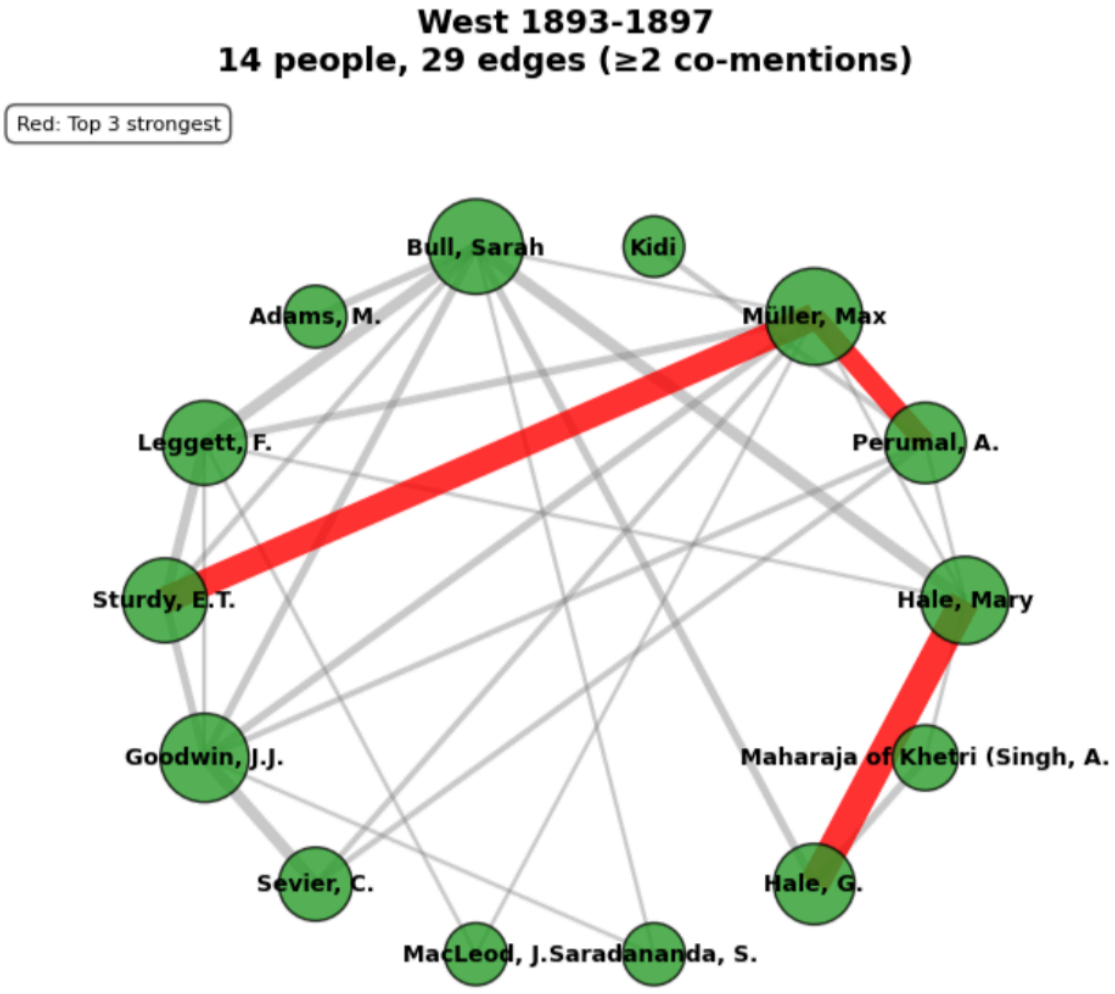
As we can see, During the final years, Swamiji increased correspondence with Sister Christine. We can also see how the relationship with Sister Nivedita intensified over time.

Now we are in a position to look at co-mention networks. The co-mention network shows how Swami Vivekananda mentally connected people in his correspondence by looking at how often two people are mentioned together in a letter.

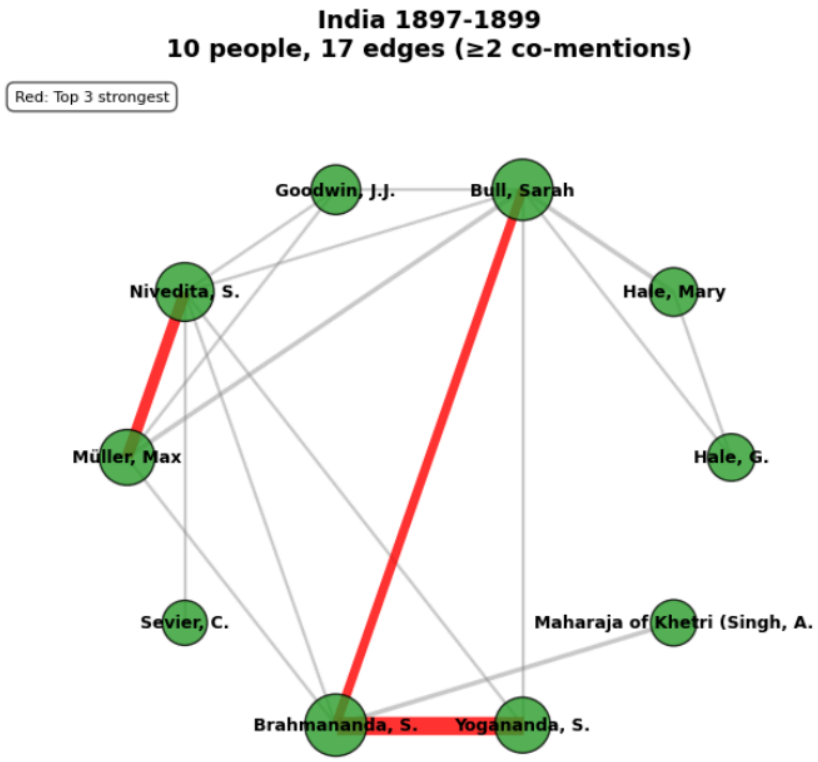
It revealed that he frequently discussed his correspondents in relation to one another—13 individuals forming 35 co-mention links with a network density of 44.9%—indicating a moderately dense, highly interconnected relational map where shared contexts, projects, and concerns naturally overlapped in his letters. Here is the one for the first Visit

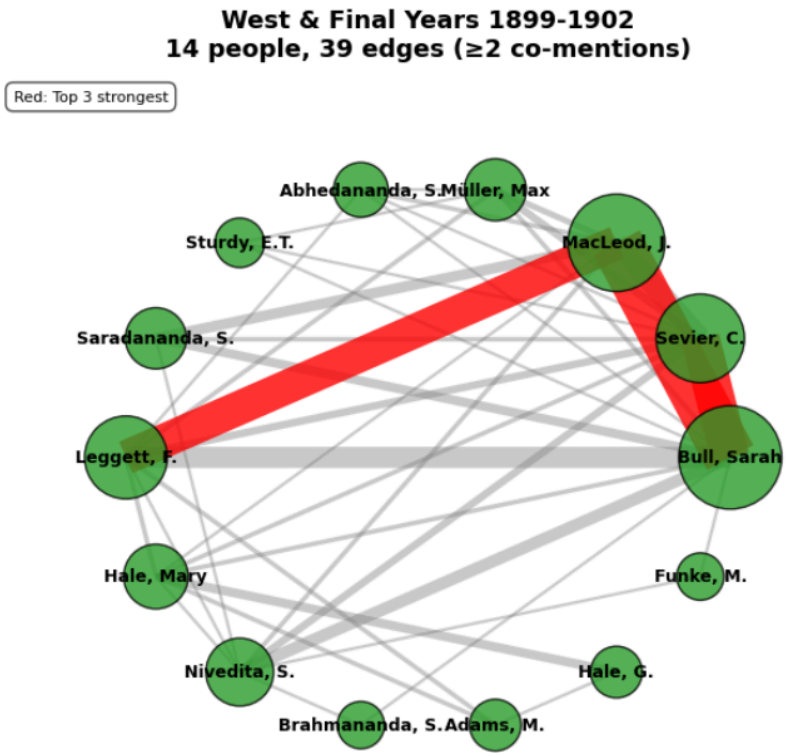


Again, the thickness of the lines indicate how often the two people were mentioned together in a letter. The size of a node shows how important that person is (the formula is node_size = 150 + degree * 15, also called degree centrality. Most letters in this period (wandering days) are to Pramadadas Mitra and we can see that the graph does not reveal too much. Swamiji mentioned many householder disciples to Mitra, like Gagan Babu, Chandra Basu etc. but since they are not very important, they have not been shown. Anyone reading Swamiji's letters will know of the events that took place surrounding Pavhari Baba from the letters to Pramadadas Mitra, hence he has been shown.



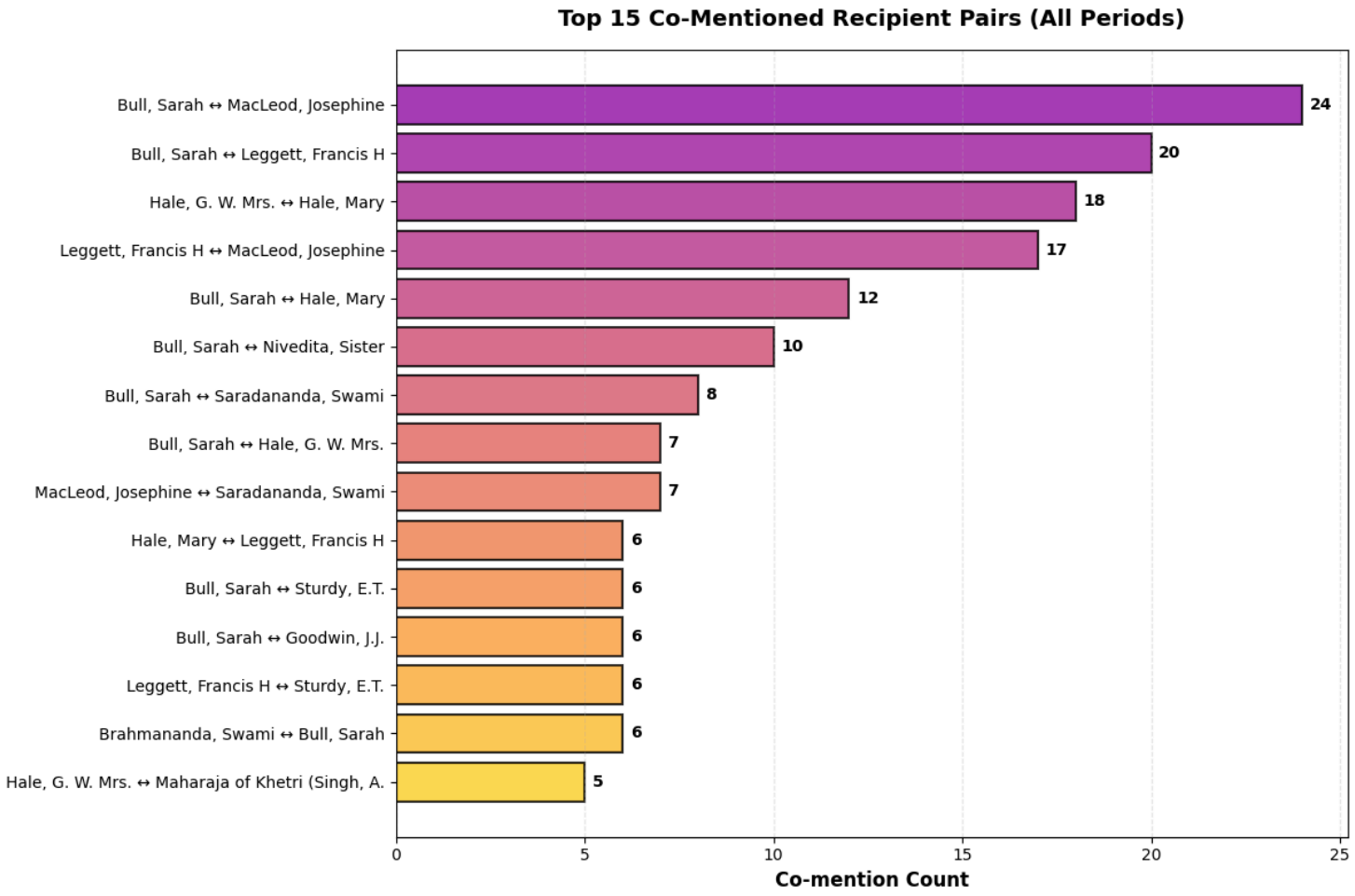
In the next one, we can see that The Hales are mentioned together often, this is unsurprising since they are family. Moreover, we can see that Max Muller is often mentioned with Alasinga and Sturdy! Here we can see no mentions at all for Sister Nivedita, who would become a central figure in Swamiji's mission after meeting him in 1895.



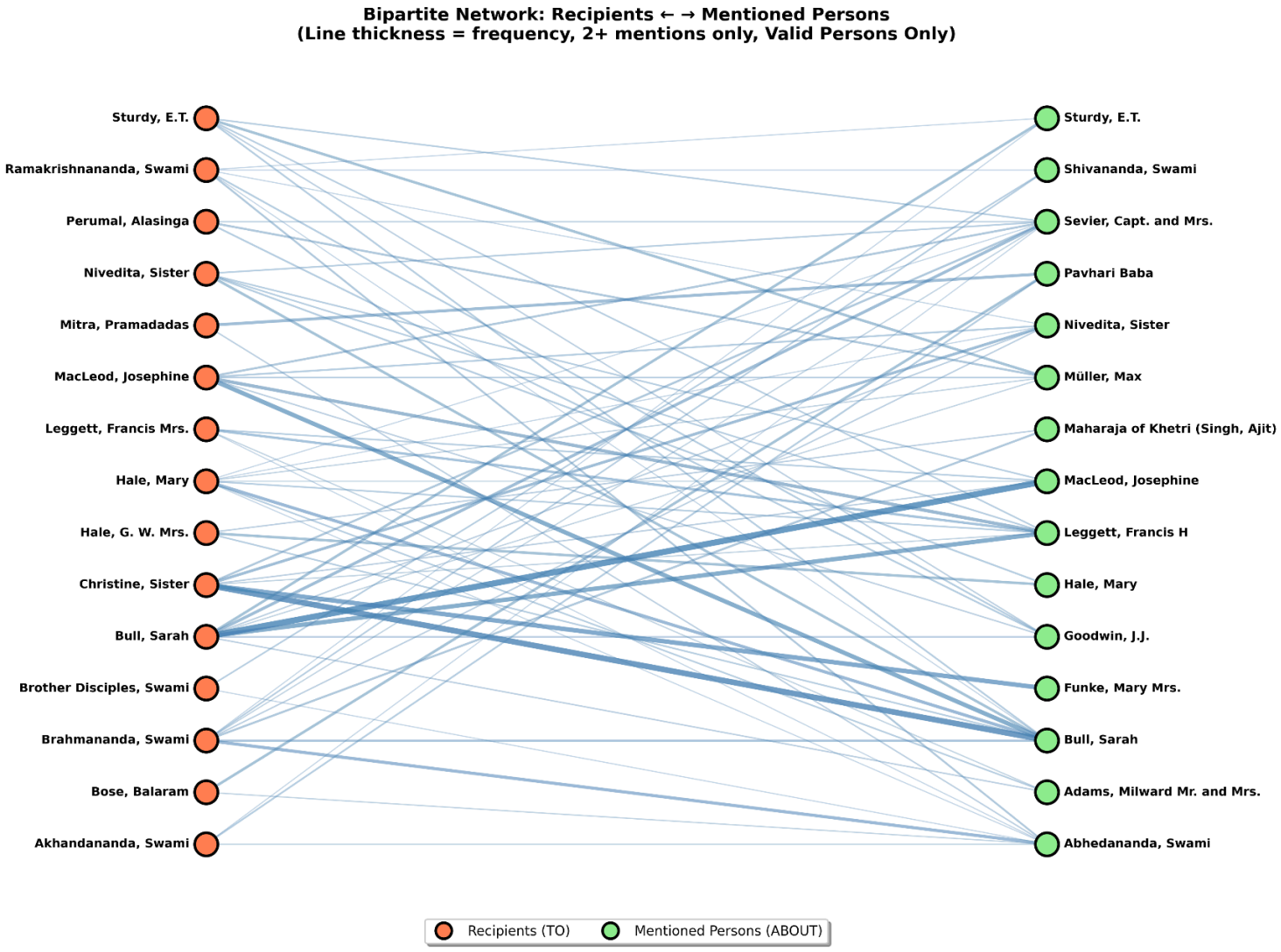


These are the rest of the two co-mention graphs for the titled periods. We can now see Sister Nivedita emerging as a prominent figure. Something interesting to note is the density of the edges (the lines) is very scarce during the India stay in 1897-99, but it is most dense (even denser than 1893-97 period) during his final years. Swamiji seems to be mentioning multiple people quite often in his letters during 1899-1902.

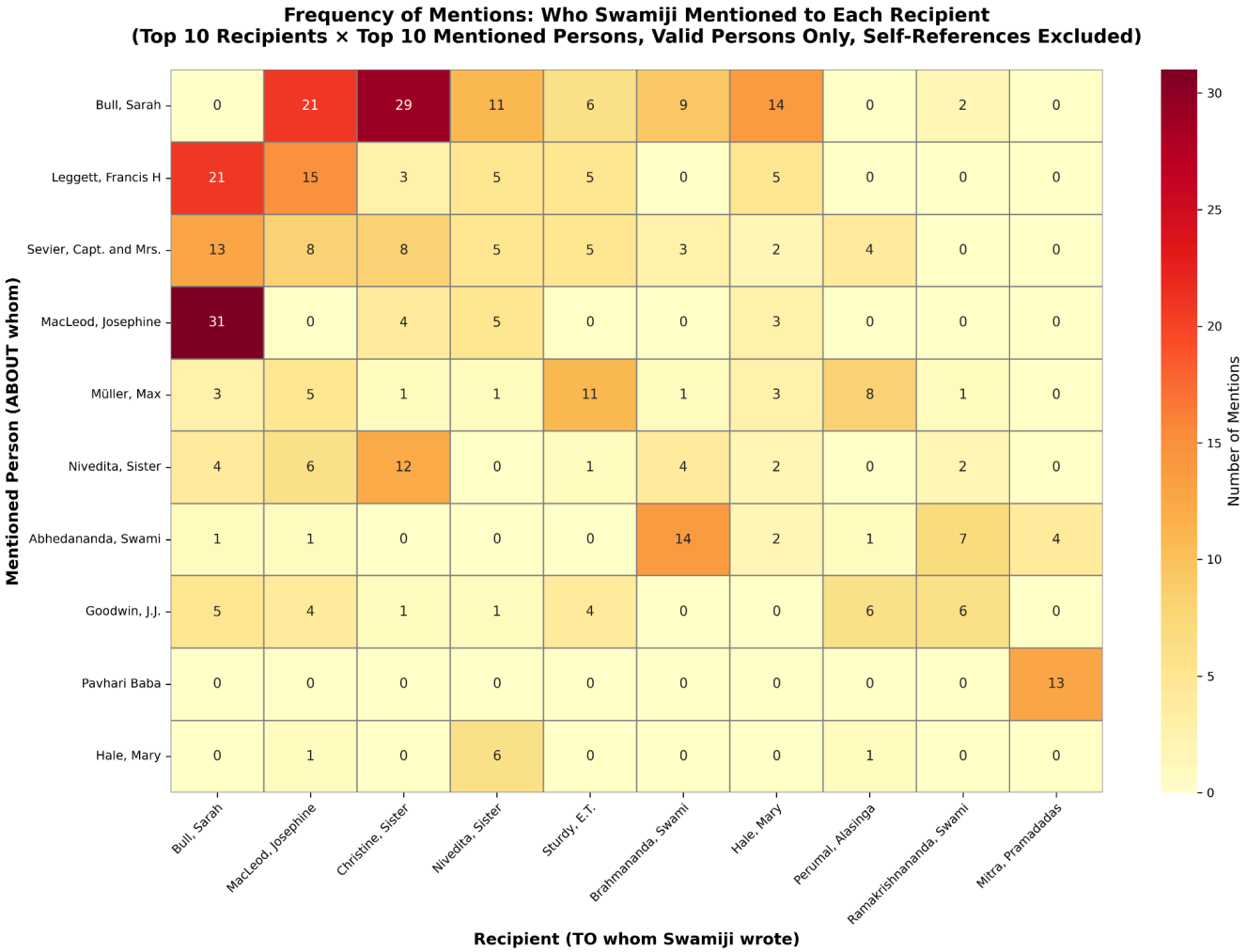
Here is a summary of most co-mentioned people across all periods.



Sarah Bull is most often comentioned with others as we can see.
We now refer to the notebook called Letter_Extra_Analysis.ipynb.
The above included comentions, that is how often two people are mentioned together in a letter. This does not mean that swamiji is talking about them together, he may just be talking about them in different contexts. But another question can be “Who did swamiji mention to whom?” This would answer which people Swamiji wanted to make bridges between, or introduce to each other. Here is a graph showing the same.

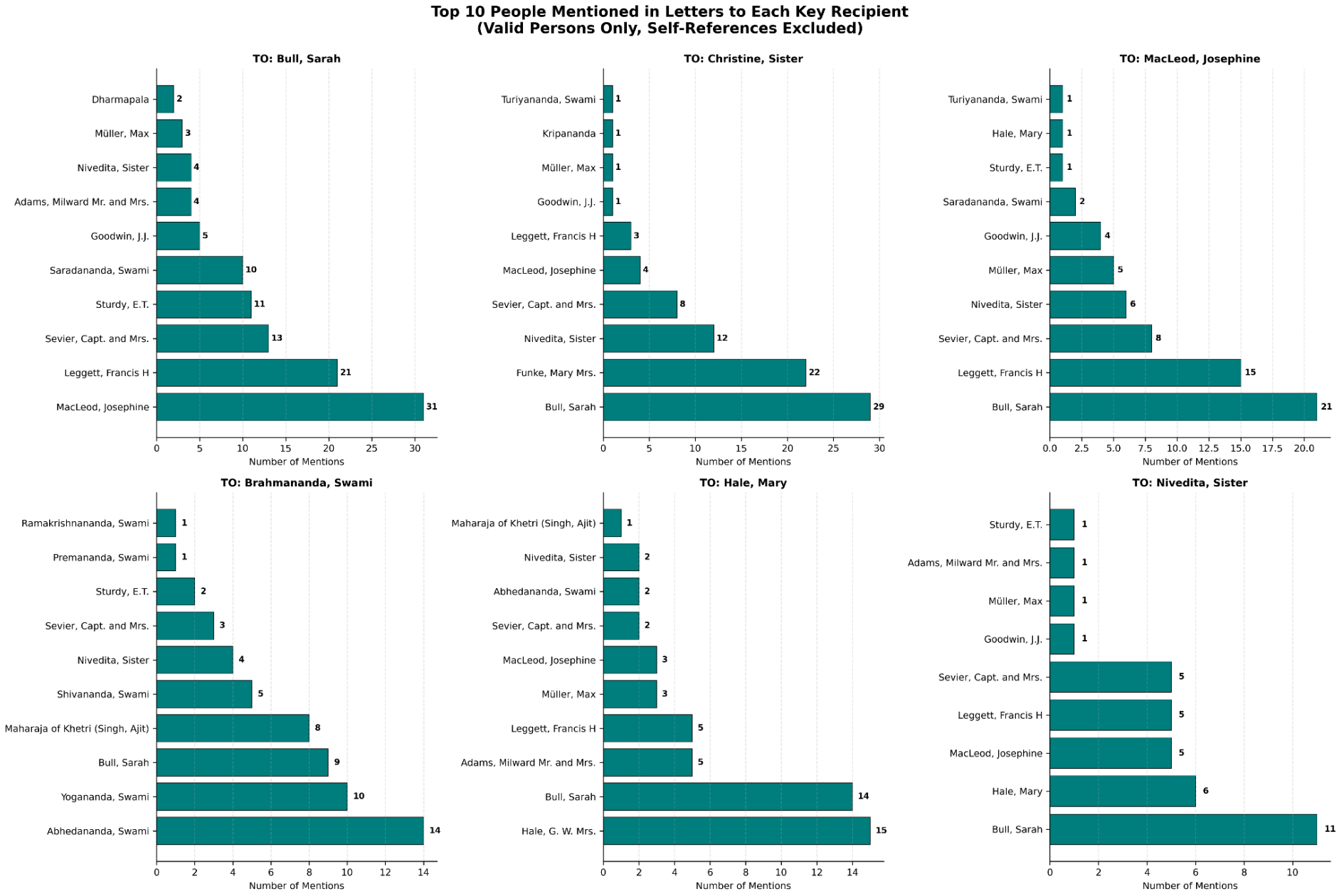


The red dots on the right show recipients (who swamiji wrote the letter to) and the left shows who was mentioned (who swamiji talked about). Sarah Bull on the right has quite a thick line connecting to Josephine MacLeod. The thickness represents that when Swamiji wrote to Sarah Bull, MacLeod was mentioned very often. Another representation of the same connection is this heatmap.



The columns are for recipients and mentioned person are in rows. We can clearly see 31 mentions of MacLeod when writing to Mrs. Bull.

If we want to see who was mentioned to whom recipient wise, here are some simple bargraphs.



We can see that while writing to Swami Brahmananda, Swami Abhedananda has been mentioned the most along with Swami Yogananda.

I would say the most basic observation these graphs go to show how central Sarah Bull was to Swamiji's bridge building process. More so than anyone else.

Using the trend from public corpus analysis, we can now look at how Swamiji uses devanagari in his letters. It would mostly be for either salutation (Swamiji often wrote sanskrit salutations to his brother-monks and those who knew sanskrit to some extent like Sharatchandra Chakravarty) or to quote scripture.

```
================================================================================
EXAMPLE 1
================================================================================
Letter Name: XXIV Kali
Year: 1894
Recipient: Abhedananda, Swami

Sentence with Devanagari:
  "शिवा वः सन्तु पन्थानः — May blessings attend your path

URL: https://www.ramakrishnavivekananda.info/vivekananda/volume_7/epistles_third_series/24_kali.htm

================================================================================
EXAMPLE 2
================================================================================
Letter Name: LV Akhandananda
Year: 1894
Recipient: Akhandananda, Swami

Sentence with Devanagari:
  You have read —"मातृदेवो भव, पितृदेवो भव— Look upon your mother as God, look upon your father as God" — but I say"दरिद्रदेवो भव, मूर्खदेवो भव—
The poor, the illiterate, the ignorant..

URL: https://www.ramakrishnavivekananda.info/vivekananda/volume_6/epistles_second_series/055_akhandananda.htm

================================================================================
EXAMPLE 3
================================================================================
Letter Name: XV Sir
```

Year: 1890
Recipient: Bose, Balaram

Sentence with Devanagari:
  "आत्मानं सततं रक्षेत् — One must save oneself under any circumstances

URL: https://www.ramakrishnavivekananda.info/vivekananda/volume_6/epistles_second_series/015_sir.htm

==
EXAMPLE 4
==
Letter Name: LXXXIX Mother
Year: 1899
Recipient: Bose, Mrinalini

Sentence with Devanagari:
  उद्धरेदात्मनात्मानं — One should raise the self by the self

URL: https://www.ramakrishnavivekananda.info/vivekananda/volume_5/epistles_first_series/089_mother.htm

==
EXAMPLE 5
==
Letter Name: LXXI Rakhal
Year: 1895
Recipient: Brahmananda, Swami

Sentence with Devanagari:
  —

"मनसि वचसि काये पुण्यपीयूषपूर्णाःत्रिभुवनमुपकारश्रेणिभिः प्रीणयन्तः।परगुणपरमाणुं पर्वतीकृत्य नित्यंनिजहृदि विकसन्तः सन्ति सन्तः कियन्तः

— There are some saints who full of holiness in thought, wor...

URL: https://www.ramakrishnavivekananda.info/vivekananda/volume_6/epistles_second_series/071_rakhal.htm

==
EXAMPLE 6

==========
Letter Name: XLV Brothers
Year: 1894
Recipient: Brother Disciples, Swami

Sentence with Devanagari:
  "सन्निमित्ते वरं त्यागो विनाशे नियते सति — When death is so certain, it is better to die for a good cause

URL: https://www.ramakrishnavivekananda.info/vivekananda/volume_6/epistles_second_series/045_brothers.htm

==========
EXAMPLE 7
==========
Letter Name: XXXIII Diwanji Saheb
Year: 1894
Recipient: Desai, Viaharidas Haridas

Sentence with Devanagari:
  मूकं करोति वाचालं etc

URL: https://www.ramakrishnavivekananda.info/vivekananda/volume_8/epistles_fourth_series/033_diwanji_saheb.htm

==========
EXAMPLE 8
==========
Letter Name: VII Atul Babu
Year: 1890
Recipient: Ghosh, Atul Chandra

Sentence with Devanagari:
  यावज्जननं तावन्मरणंतावज्जननीजठरे शयनम्।इति संसारे स्फुटतरदोषःकथमिह मानव तव सन्तोषः॥

— "While there is birth there is death, and again entering the mother's womb

URL: https://www.ramakrishnavivekananda.info/vivekananda/volume_7/epistles_third_series/07_atul_babu.htm

==========
EXAMPLE 9

==
Letter Name: LXXV Doctor Shashi
Year: 1897
Recipient: Ghosh, Shashi Bhushan

Sentence with Devanagari:
  You ought to see me, Doctor, when I sit meditating in front of the beautiful snow-peaks and repeat from the Upanishads: "न तस्य रोगो न जरा न मृत्युः प्राप्तस्य योगाग्निमयं शरीरम् —He has neither disea...

URL: https://www.ramakrishnavivekananda.info/vivekananda/volume_5/epistles_first_series/075_doctor_shashi.htm

==
EXAMPLE 10
==
Letter Name: LXXIV Honoured Madam
Year: 1897
Recipient: Ghoshal, Sarala

Sentence with Devanagari:
  ॐ तत् सत्

HONOURED MADAM, (Shrimati Sarala Ghoshal — Editor,Bharati)

I feel much obliged for theBhâratisent by you, and consider myself fortunate that the cause, to which my humble life has been ded...

URL: https://www.ramakrishnavivekananda.info/vivekananda/volume_5/epistles_first_series/074_honoured_madam.htm

==
MORE EXAMPLES - TOP 5 RECIPIENTS WITH MOST DEVANAGARI LETTERS
==

==
TO: Brahmananda, Swami (6 letters with Devanagari)
==

  Example 1:
  Letter: LXXI Rakhal (1895)

Sentence: —

"मनसि वचसि काये पुण्यपीयूषपूर्णाःत्रिभुवनमुपकारश्रेणिभिः प्रीणयन्तः।परगुणपरमाणुं पर्वतीकृत्य नित्यंनिजहृदि विकसन्तः सन्ति सन्तः कियन्तः

— There ar...
URL: https://www.ramakrishnavivekananda.info/vivekananda/volume_6/epistles_second_series/071_rakhal.htm

Example 2:
Letter: LXXVIII Rakhal (1895)
Sentence: However, I have nothing to say against any particular course which any one may be led to adopt; on the contrary, God-speed —"शिवाः वः सन्तु पन्थानः— M...
URL: https://www.ramakrishnavivekananda.info/vivekananda/volume_6/epistles_second_series/078_rakhal.htm

Example 3:
Letter: LXXXIII Rakhal (1895)
Sentence: "मयैवैते निहताः पूर्वमेव निमित्तमात्रं भव सव्यसाचिन्— These have verily been killed by Myself long ago, be only the instrument, O Arjuna...
URL: https://www.ramakrishnavivekananda.info/vivekananda/volume_6/epistles_second_series/083_rakhal.htm

==========
TO: Ramakrishnananda, Swami (5 letters with Devanagari)
==========

Example 1:
Letter: XLI Shashi (1894)
Sentence: ये निध्नन्ति निरर्थकं परहितं ते के न जानीमहे— We do not know what sort of people they are who for nothing hinder the welfare of others" (Bhartrihari)...
URL: https://www.ramakrishnavivekananda.info/vivekananda/volume_6/epistles_second_series/041_shashi.htm

Example 2:
Letter: IL Swami Ramakrishnananda (1894)
Sentence: Wherever the seed of his power will find its way, there it will fructify —अद्य वा अब्दशतान्ते वा— be it today, or in a hundred years...
URL: https://www.ramakrishnavivekananda.info/vivekananda/volume_6/epistles_second_series/049_swami_ramakrishnananda.htm

Example 3:
Letter: LIV Swami Ramakrishnananda (1894)
Sentence: Never mind,उपेक्षितव्यं तद्वचनं भवत्सट्टशानां महात्मनाम्। अपि कीटदंशनभीरुका वयं रामकृष्ण तनयास्तदूहृदयरुधिरपोषिताः। "अलोकसामान्यमचिन्त्यहेतुकंनिन्दन्त...

URL: https://www.ramakrishnavivekananda.info/vivekananda/volume_6/epistles_second_series/054_swami_ramakrishnananda.htm

==========
TO: Mitra, Pramadadas (4 letters with Devanagari)
==========

Example 1:
Letter: VII Sir (1889)
Sentence: Another perhaps is:

तच्चेतसा स्मरति नूनमबोधपूर्वं भावस्थिरानि जननान्तरसौह्रदानि ।

(Kalidasa's Shakuntalam, Act V: "It must be the memories, unwittingl...
URL: https://www.ramakrishnavivekananda.info/vivekananda/volume_6/epistles_second_series/007_sir.htm

Example 2:
Letter: VIII Sir (1889)
Sentence: He only quotes "यज्ञेऽनवक्लृप्तः" ("The Shudra is not conceived of as a performer of Yajna or Vedic sacrifices...
URL: https://www.ramakrishnavivekananda.info/vivekananda/volume_6/epistles_second_series/008_sir.htm

Example 3:
Letter: XIII Sir (1889)
Sentence: And my resolve is something like "either to lay down my life or realise my ideal" "शरीरं वा पातयामि मन्त्रं वा साधयामि ।" — so help me the Lord of Kas...
URL: https://www.ramakrishnavivekananda.info/vivekananda/volume_6/epistles_second_series/013_sir.htm

==========
TO: Akhandananda, Swami (4 letters with Devanagari)
==========

Example 1:
Letter: LV Akhandananda (1894)
Sentence: You have read —"मातृदेवो भव, पितृदेवो भव— Look upon your mother as God, look upon your father as God" — but I say"दरिद्रदेवो भव, मूर्खदेवो भव— The poo...
URL: https://www.ramakrishnavivekananda.info/vivekananda/volume_6/epistles_second_series/055_akhandananda.htm

Example 2:
Letter: CXXVIII Akhandananda (1897)

Sentence: If in the attempt to carry morsels of food to starving mouths, name and possession and all be doomed even —अहो भाग्यमहो भाग्यम्— thrice blessed art th...
URL: https://www.ramakrishnavivekananda.info/vivekananda/volume_6/epistles_second_series/128_akhandananda.htm

Example 3:
Letter: CXXXVI Akhandananda (1897)
Sentence: "स ईशः अनिर्वचनीयप्रेमस्वरूपः— The Lord is the Essence of unutterable love...
URL: https://www.ramakrishnavivekananda.info/vivekananda/volume_6/epistles_second_series/136_akhandananda.htm

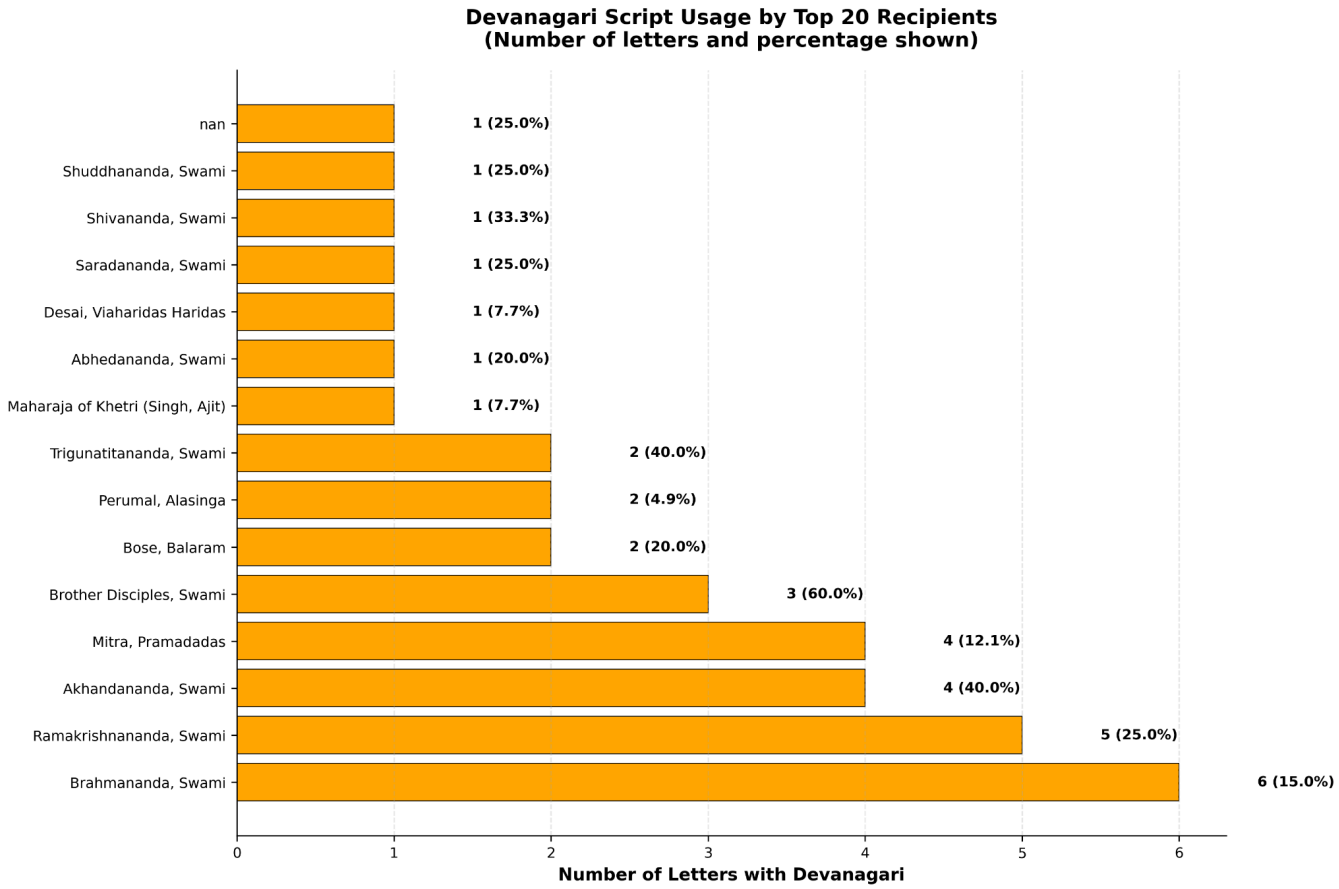
==========
TO: Brother Disciples, Swami (3 letters with Devanagari)
==========

Example 1:
Letter: XLV Brothers (1894)
Sentence: "सन्निमित्ते वरं त्यागो विनाशे नियते सति — When death is so certain, it is better to die for a good cause...
URL: https://www.ramakrishnavivekananda.info/vivekananda/volume_6/epistles_second_series/045_brothers.htm

Example 2:
Letter: XLVII Brother Disciples (1894)
Sentence: You know,श्रेयांसि बहुविघ्नानि— Great undertakings are always fraught with many obstacles...
URL: https://www.ramakrishnavivekananda.info/vivekananda/volume_6/epistles_second_series/047_brother_disciples.htm

Example 3:
Letter: LVI Dear and Beloved (1894)
Sentence: "उत्तिष्ठत जाग्रत प्राप्य वरान्निबोधत— Arise...
URL: https://www.ramakrishnavivekananda.info/vivekananda/volume_6/epistles_second_series/056_dear_and_beloved.htm

Just as expected, let us plot this recipient wise.



Just as expected. Note that this does not mean that Swamiji does not quote scriptures to Westerners, anyone who has read the letters knows Swamiji's repertoire was immense, to the westerners he would often quote from the Bible and English literature like Shakespeare. These can be found too, but it would not be as trivial as these analyses.

Now we can move onto the last two analyses that I have done for this project. From the notebook called Letter_Extra_Analysis.ipynb. Embeddings have risen into popularity as methods of encoding natural language data. This is basically turning a text into a long list of numbers (projection into higher dimensions) such that words that are semantically similar end up being closer together in the high-dimensional space. This is exactly how LLMs like ChatGPT etc. work internally. The text you write is first turned into a long list of numbers (embedding) and then it finds

semantically similar things in the embedding space (high dimensional space) and prints it out using complex algorithms perfected over many years now. Since these words have become a long list of numbers (embeddings) using a very specialized model, the semantic meaning of the word is intact. Interesting things can now be done with these words-as-lists-of-numbers (called vectors), for example vector math. If you ask this model “What is king - man + woman?” it would immediately say “queen”. This is an overly simplified explanation for (arguably) the most technologically advanced thing that we have been able to build. Take it with a grain of salt.

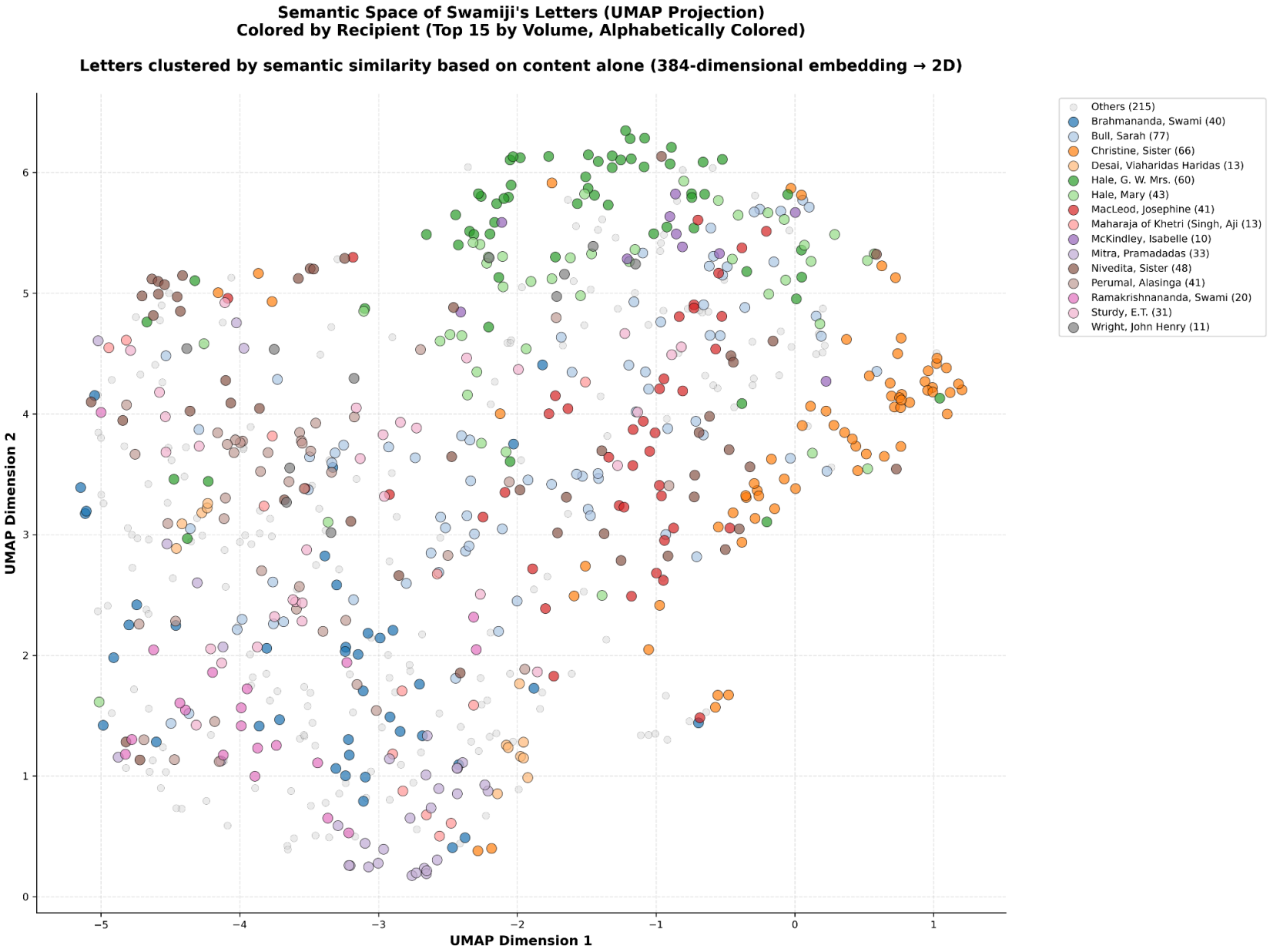
What we have now done is projected Swamiji’s letters into a 384 - dimensional hyperspace (ChatGPT operates in a 20,000 dimensional hyperspace), where semantically similar letters cluster together based on some internal semantic pattern that we are unable to notice ourselves. We then try to see what sort of letters cluster together (semantically similar as detected in this 384-D hyperspace) and which are outliers and, plot these back into familiar 2-dimensions so we can visualize them (techniques like exist to plot things in higher dimensional space back in 2d space).

Formally, for technical people, each letter is encoded using a pretrained sentence-transformer model (`all-MiniLM-L6-v2`), which maps variable-length text into fixed-length 384-dimensional semantic embeddings. The model is trained with contrastive objectives such that semantic similarity between texts corresponds to proximity in the embedding space, measured using cosine similarity.

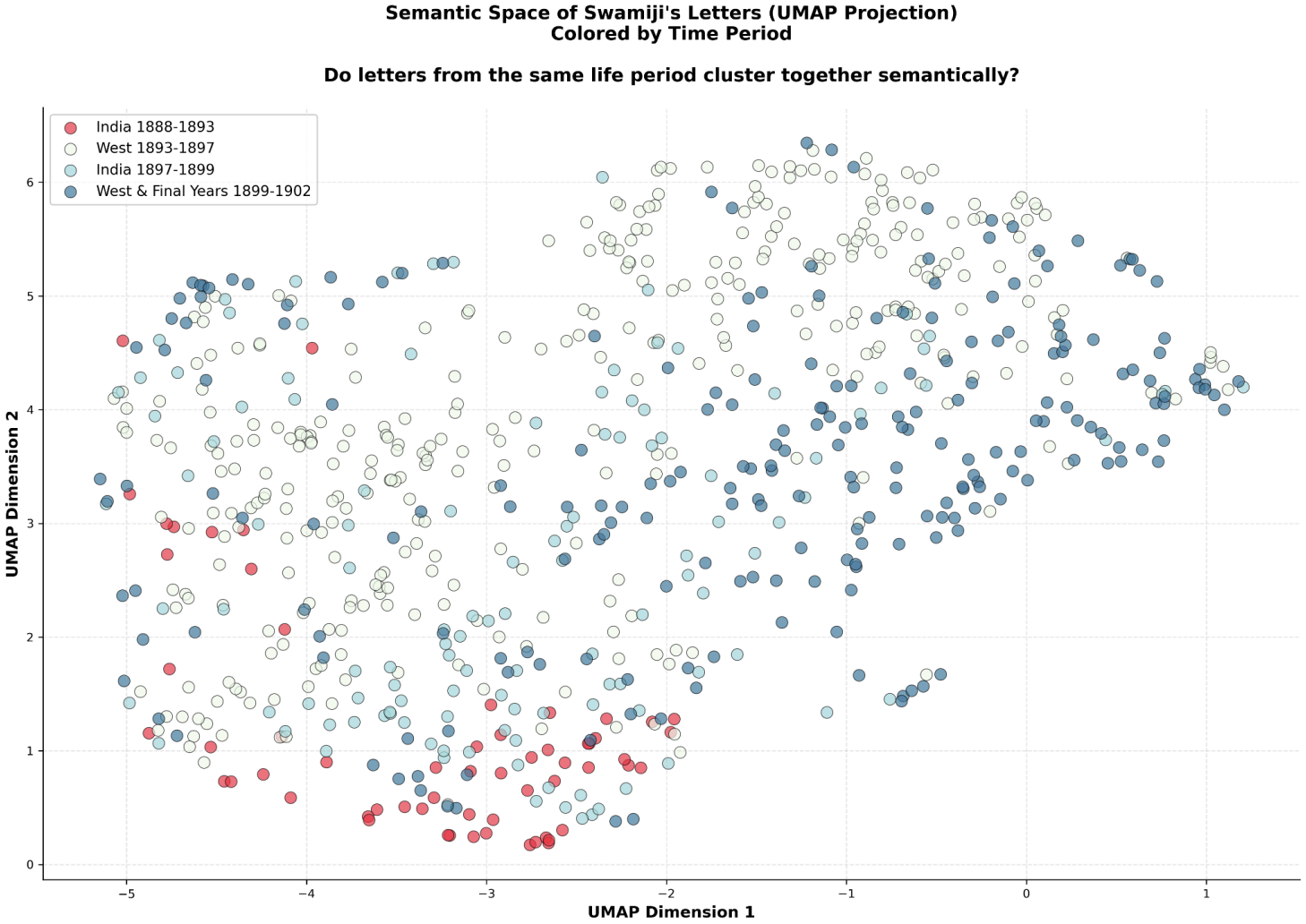
All analysis is performed directly in this high-dimensional space. Clustering methods (K-means, DBSCAN, hierarchical clustering) identify latent semantic groupings; mean embedding vectors define recipient-level semantic profiles; cosine similarity matrices quantify relationships between letters and between recipients; and isolation- and distance-based methods identify semantic outliers.

For visualization only, embeddings are projected from 384 dimensions into two dimensions using UMAP and t-SNE, which preserve local neighborhood structure to enable qualitative inspection of clusters, temporal patterns, and anomalous letters. Quantitative results and inferences are computed in the original embedding space.

We can now see what the letters look like when projected into this 384-dimensional space.

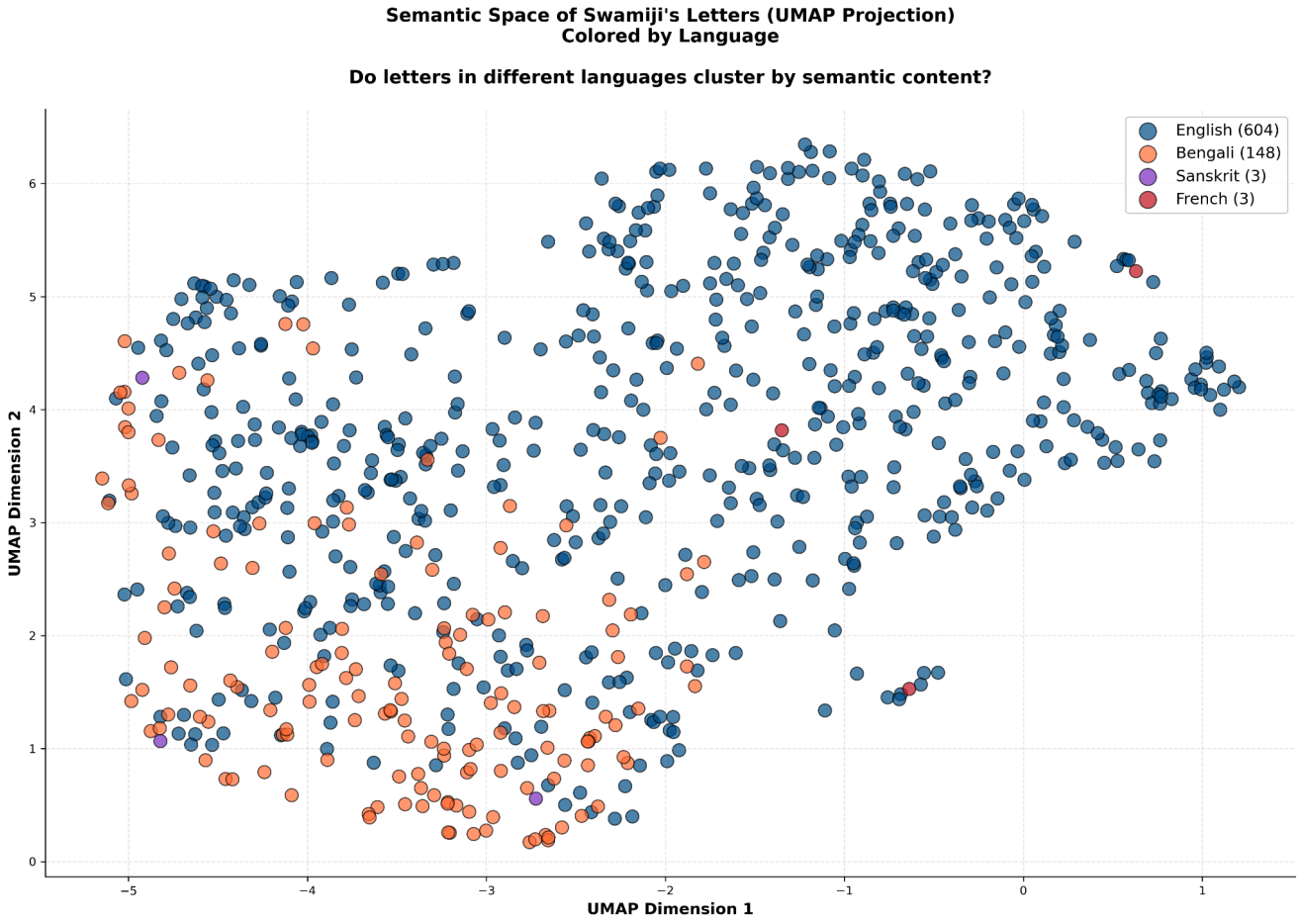

Semantic Space of Swamiji's Letters (UMAP Projection)
Colored by Recipient (Top 15 by Volume, Alphabetically Colored)
Letters clustered by semantic similarity based on content alone (384-dimensional embedding → 2D)
UMAP Dimension 1
UMAP Dimension 2
Others (215)
Brahmananda, Swami (40)
Bull, Sarah (77)
Christine, Sister (66)
Desai, Viaharidas Haridas (13)
Hale, G. W. Mrs. (60)
Hale, Mary (43)
MacLeod, Josephine (41)
Maharaja of Khetri (Singh, Aji (13)
McKindley, Isabelle (10)
Mitra, Pramadadas (33)
Nivedita, Sister (48)
Perumal, Alasinga (41)
Ramakrishnananda, Swami (20)
Sturdy, E.T. (31)
Wright, John Henry (11)

What a beauty. This is the analysis that makes one be in awe. We can see that there is some pattern that is being captured by the projection of Swamiji's letters in 384 dimensions. When these points are coloured recipient wise, we can see letters to the Hales (green dots) appear together on the top, and orange dots (to Sister Christine) appear near the right edge of the graph.
Let us colour this same graph a bit differently by the period in which the letters were written.



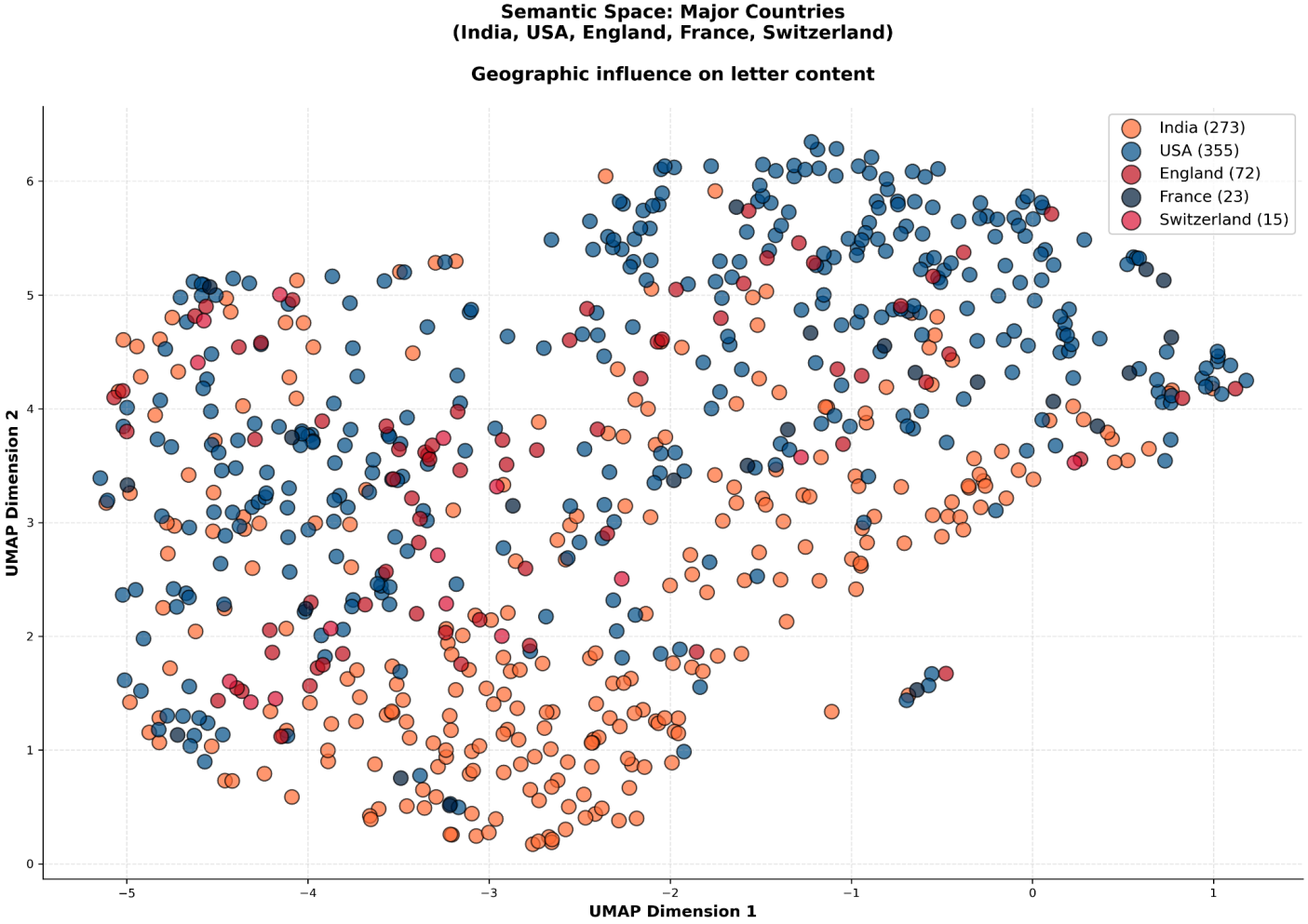
Look at this! We can clearly see how nicely period1, period2 and period 4 separate! Of course it is not perfect, but it is still quite separable! Period 3 seems to be scattered unevenly.

What if we colour the same by language?



Most of the letters are either in English or Bengali, as we saw earlier, and these are also highly separate from one another, the Bengali “zone” is in bottom left while English letters are in top right.

We can also colour the same embeddings plot by country.



Again, USA and India are quite separable from each other, blue and orange dots form specific areas. We must remember that this is an abstracted low dimensional plot, only limited by the fact that we cannot perceive anything more than 2d here. In the original 384-dimensional space, these are likely to be a lot more separable.

This basically means that there is a lot of semantic structure in Swamiji's letters being captured in a high-dimensional hyperspace, which makes it possible to infer much of the information about the letter! Let me rephrase that a bit differently so you know exactly what this means:

Given only a letter by Swamiji with no other information, things like the year of writing, recipient, country, whether the letter is originally english or translated from Bengali etc. can be inferred without any other support with reasonable accuracy.

This can be done by finding clusters of recipients, year of writing, country etc. and seeing which cluster is closest or most similar to the provided body of letter using various algorithms. It was keeping this in mind that I decided to try and find a letter that caught my eye in the newly released "Letters of Swami Vivekananda" vols 1-4 (Advaita Ashrama, 2025). There is a letter called "Letter to the Artist" whose addressee has not been found yet.

```
    Index: 73
    Name: LXXIII Madam
    Audience: American Lady
    Date: 1896/12/13
    Place: England, London
    Language: English
    Year: 1896.0

  Letter content preview:

DEAR MADAM, (An American lady.)

We have only to grasp the idea of gradation of morality and everything becomes clear.

Renunciation — non-resistance — non-destructiveness — are the ideals to be attained through less and less worldliness,
```

I wanted to see if I could find something using this technique, but this letter is quite the outlier. So I tried to find the most similar letters to this one in the high-dimensional space.

```
================================================================================
[STEP 3] Finding semantically similar letters...
```

```
================================================================================
  ✓ Calculated similarities for 761 letters

  Top 20 most semantically similar letters:
================================================================================================================
   1. Sim: 0.4932 | XC Sister                | To: Farmer, S Miss              | Year: 1895
   2. Sim: 0.4810 | CXLIII Margot             | To: Nivedita, Sister            | Year: 1898
   3. Sim: 0.4741 | XXXVI Miss Noble          | To: Nivedita, Sister            | Year: 1896
   4. Sim: 0.4703 | CXLVI Mother Church       | To: Hale, G. W. Mrs.            | Year: 1899
   5. Sim: 0.4629 | XLIII Madam               | To: Ghoshal, Sarala             | Year: 1899
   6. Sim: 0.4462 | XL Friend                 | To: Sturdy, E.T.                | Year: 1894
   7. Sim: 0.4455 | LIV Alasinga              | To: Perumal, Alasinga           | Year: 1895
```

As we can see, the MOST similar letter is the only one written to Miss S Farmer in 1895, but even that has a similarity score of about 0.5. Two of the next most similar letters are to Sister Nivedita, who is unlikely to be a recipient for many reasons (primarily, she is not American).

```
================================================================================
[ANALYSIS 4] Computing isolation score for mystery letter...
================================================================================

  Mystery Letter Isolation Metrics:
    Average similarity to all letters: 0.2017
    Median similarity: 0.1947
    Std deviation: 0.1008
    Max similarity (best match): 0.4932
    90th percentile: 0.3371
    75th percentile: 0.2699
    50th percentile: 0.1947

  Interpretation:
    ⚠️  HIGHLY ISOLATED: Max similarity is only 0.4932
    This suggests UNIQUE content → likely ONE-OFF correspondent
```

This shows some other numerical evidence. The letter's average similarity to all the letters is only 0.2. It is more than twice as similar to the one letter written to Miss S. Farmer. Here is the only known letter to Miss Sarah Farmer written by Swamiji, which is the closest semantic match to the mystery letter.

```
   Letter: XC Sister
   To: Farmer, S Miss
   Similarity: 0.4932
   Year: 1895
   Place: USA, NY, New York City
   Content preview:

DEAR SISTER, (Miss S. Farmer)

In this universe where nothing is lost, where we live in the midst of death in life, every thought that is thought, in …
```

In the second volume of "Letters of Swami Vivekananda" (Advaita Ashrama, 2025), the footnote below this letter says that Frank Dvorak (a noted painter, disciple of Swami Abhedananda) sent pictures of this letter (in Swamiji's handwriting) to Vedanta Center, NY many years after Swamiji left his mortal coil. This would suggest that whoever Swamiji wrote this letter to also somehow knew Frank Dvorak, or perhaps gave it to him for safekeeping. There is some circumstantial evidence for this, looking at timelines of Farmer and Dvorak.

Frank Dvorak (active in USA: 1888–1905): He worked as a painter in the U.S., and during this period he also encountered "New Spirituality" and attempted to integrate Eastern spirituality into his painting (including portraits of Ramakrishna and Sarada Devi). He became a disciple of Swami Abhedananda later on and extensively painted Sri Ramakrishna, Holy Mother and other direct disciples.

Sarah Farmer (active in USA: 1894-1900): She launched the Green Acre Conferences (Eliot, Maine). By the late 1890s, Green Acre faced internal conflicts and financial strain. In 1900 he traveled abroad (notably meeting 'Abdu'l-Bahá and becoming associated with the Bahá'í Faith during that trip, depending on the account).

There could have been some overlap. But this is purely hypothetical at this point. One strong evidence against Sarah Farmer being the same person as the mystery recipient is the use of "Madam" in the salutation in 1896 for the mystery letter, while he used "Sister" to address Miss Farmer in 1895. It would not make sense to use "Madam" for someone when one is already comfortable enough to address her as "Sister".

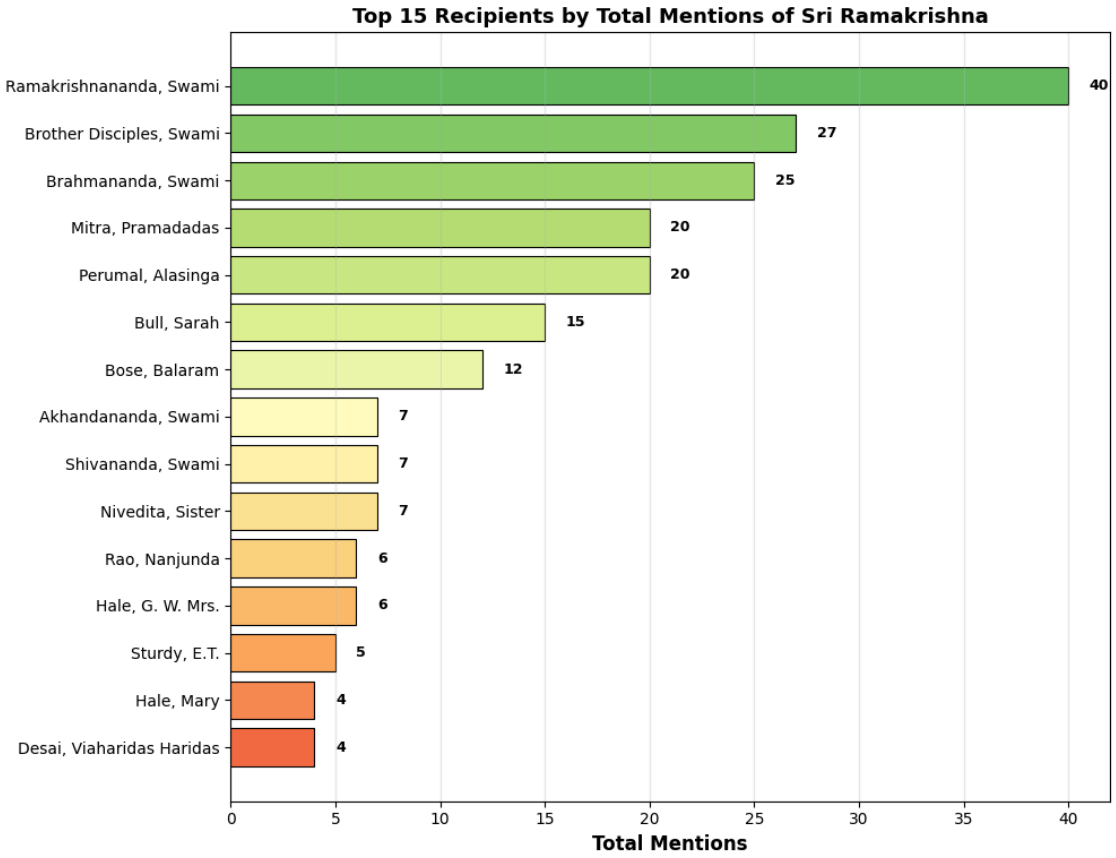
Notably, this letter is the ONLY one where Swamiji addresses the recipient as “Madam” (there is one to Sarala Ghosal but that is a translation from original Bengali). This further suggests that it would be someone else entirely.

Now moving on to our last analysis from Swamiji_Private_FinalAnalysis.ipynb. Now we will see in his private correspondence, how much does Swamiji mention his guru, Sri Ramakrishna.

Letter_Index	Name	Recipient	Year	Total_Mentions	Mentioned_Variations
197	LXXV Shashi	Ramakrishnananda, Swami	1895	23	ramakrishna paramahamsa;
167	XLV Brothers	Brother Disciples, Swami	1894	8	ramakrishna paramahamsa;
193	LXXI Rakhal	Brahmananda, Swami	1895	8	ramakrishna paramahamsa;
155	XXXIII Sir	Mitra, Pramadadas	1890	7	ramakrishna; shri ramakrishna
22	XXII Alasinga	Perumal, Alasinga	1894	6	ramakrishna; shri ramakrishna
173	LI Dear and Beloved	Rao, Nanjunda	1894	6	my guru; ramakrishna
178	LVI Dear and Beloved	Brother Disciples, Swami	1894	6	ramakrishna; sri ramakrishna;
169	XLVII Brother Disciples	Brother Disciples, Swami	1894	6	ramakrishna; our guru; shri
323	XXXII Dear	Brother Disciples, Swami	1896	6	ramakrishna; shri ramakrishna;
316	XXV Brother Shivananda	Shivananda, Swami	1894	6	ramakrishna paramahamsa;

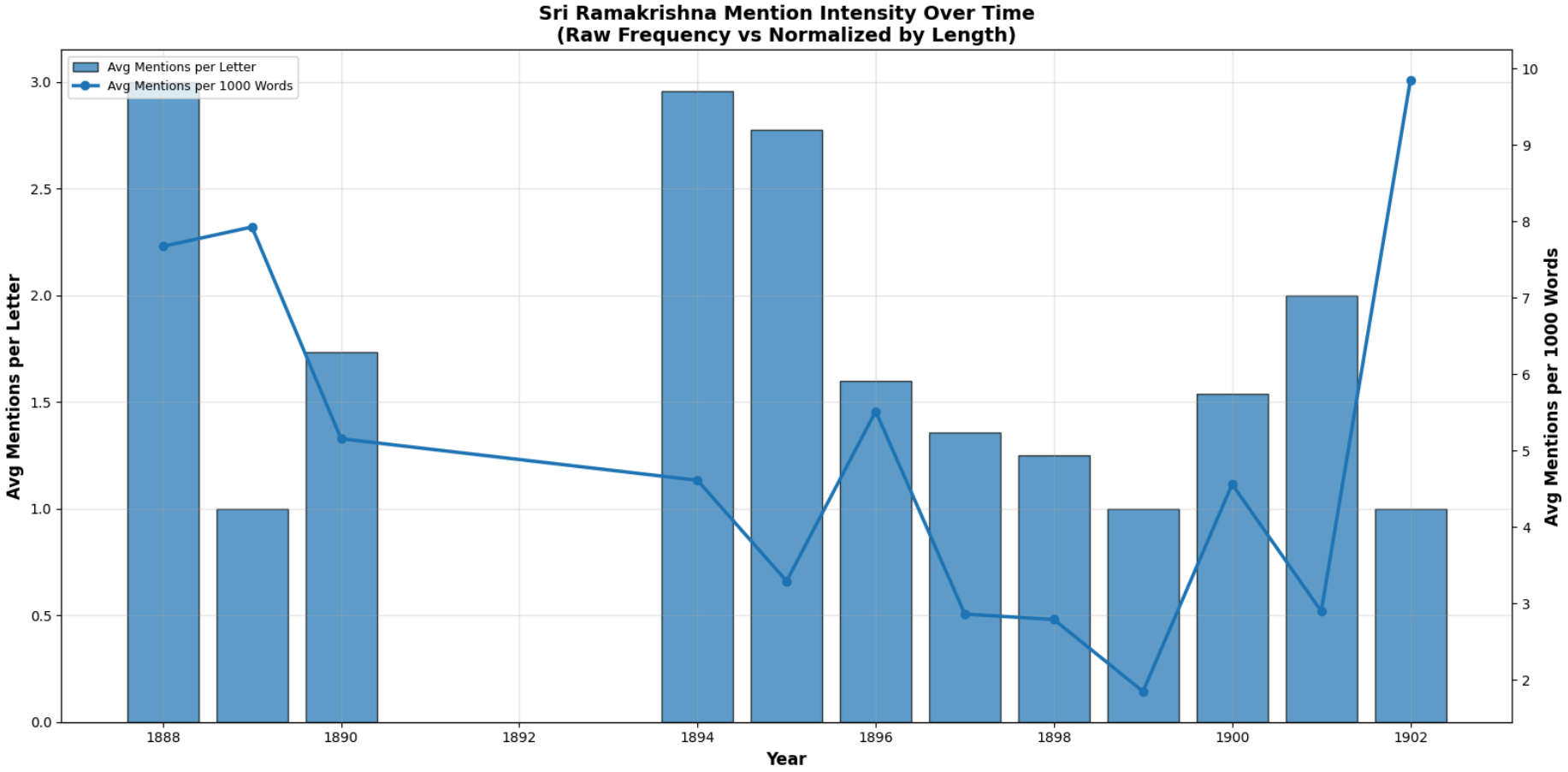
As we can see, one letter to Shashi Maharaj (Swami Ramakrishnananda) contains mentions of Sri Ramakrishna equal to the next three combined!

What about when we see total mentions of Sri Ramakrishna per recipient?

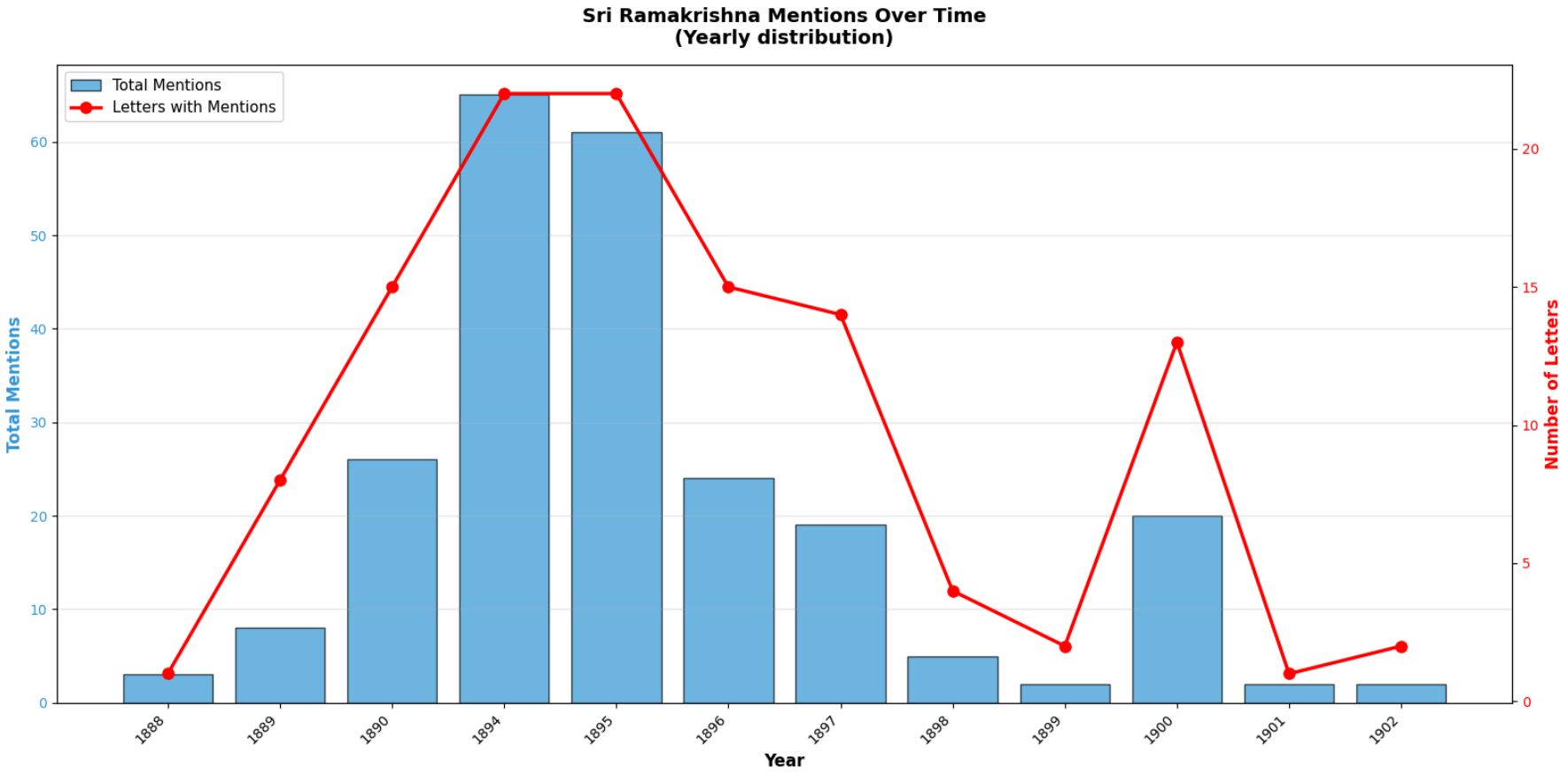


Shashi Maharaj tops the list again, followed closely by Brother Disciples. Seems like Swamiji did not talk much about Sri Ramakrishna to western disciples (Sister Christine, Josephine MacLeod don’t show up in top 15).

What if we graph the normalized mentions temporally by years?



Swamiji is mentioning Sri Ramakrishna a lot during his wandering days, then in America average mentions per letter increase but average mention per 1000 words decreases, this basically means Swamiji used to mention Sri Ramakrishna more often per letter than per word. It skyrockets in 1902, the final year. What if we just graph the mentions of Sri Ramakrishna using absolute mentions?



As we can see, the number of mentions closely follows the red line, which represents the number of letters. This just means that

Sri Ramakrishna was mentioned by Swami Vivekananda most highly to Shashi Maharaj (Swami Ramakrishnananda), while the yearly mention rate linearly follows the number of letters written that year.

With this, I conclude the long and drawn out data-science based discussion of Swami Vivekananda's complete works corpus. Thanks for following along, I really enjoyed discovering things in the corpus.

Namah srî yati râjâya vivekânanda sûraye

Saccit sukha swarûpâya swâmine tâpahârine

"Salutation to that king of renouncers and controller of passions, the sage Vivekananda,
who is Satcitananda (Existence, Knowledge and Bliss Absolute) Itself,
the spiritual preceptor and the remover of distress."

- Swami Ramakrishnananda

Sri Râmakrishnârpanamastû