**\_Data Source**

|  |  |
| --- | --- |
| * Discogem | * ~~qadc~~ |

**\_Data Structure**

* The data contains 27 columns (at **leaves level**)
* Around 6500 rows
* They could be grouped into 16 columns (at **level 2**)
* ~~And also could be grouped into 4 columns (at level 1)~~

**\_Data Values**

**[ LEAVES LEVEL ]**

* The data has values between 0 and 1 with **a step of 0.1**
* However, the majority of the cells are of **value 0**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Table 1 | | | | | |  |
| = 0 | = 0.1 | = 0.2 | = 0.3 | =0.4 | = 0.5 | >= 0.6 |
| ≈85% | ≈7% | ≈3% | ≈2% | ≈1% | ≈1% | ≈1% |

* Column wise; only 12 with 10% or more of their values are non-zero

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 2 | | | | | | | | | | | | | |
|  | synchronous | precedence | reason | result | arg1-as-denier | arg2-as-denier | contrast | | similarity | conjunction | arg2-as-instance | arg1-as-detail | arg2-as-detail |
| Non-zero value% per col - % of >0 i.e. [0.1-1] | 10% | 29% | 29% | 67% | 17% | 28% | 18% | | 12% | 76% | 30% | 18% | 60% |
| % of >0.1 per col [0.2-1] | 3% | 15% | 13% | 46% | 5% | 12% | 6% | | 3% | 53% | 12% | 4% | 37% |
| % of >0.2 per col [0.3-1] | 1% | 9% | 7% | 32% | 2% | 5% | 3% | | 1% | 35% | 6% | 1% | 22% |
| **Col total val/ data total val** | 1% | 7% | 6% | 22% | 3% | 6% | 3% | | 2% | 23% | 6% | 2% | 16% |
| constitute 97% of total value | | | | | |

**[ LEVEL TWO ]**

* A little better, butthe majority of the cells are of **value 0**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Table 3 | | | | |  |  |
| = 0 | = 0.1 | = 0.2 | = 0.3 | = 0,4 | = 0.5 | = 0.6 |
| ≈78% | ≈9% | ≈5% | ≈3% | ≈5% | ≈2% | ≈ 1% |

* Column wise; only **6** columns contain **6%** to **28%** of the total value of the data (Rest are less than **3%**)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 4 | | | | | | | | | | |
|  | synchronous | asynchronous | cause | concession | contrast | | similarity | conjunction | instantiation | level-of-detail |
| Non-zero value% per col - % of >0 i.e. [0.1-1] | 10% | 32% | 77% | 37% | 18% | | 12% | 76% | 31% | 67% |
| % of >0.1 per col [0.2-1] | 3% | 16% | 59% | 17% | 6% | | 3% | 53% | 13% | 43% |
| % of >0.2 per col [0.3-1] | 1% | 10% | 44% | 9% | 3% | | 1% | 35% | 6% | 26% |
| Col total val/ data total val | 1% | 8% | 28% | 8% | 3% | | 2% | 23% | 6% | 18% |
| constitute 97% of total value | | | | |

**\_Data Visualization**

|  |
| --- |
| **Fig 1. Subset of Pairwise Scatter Plots for Leaves Level** |
|  |

|  |
| --- |
| **Fig 1. Subset of Pairwise Scatter Plots for Level 2** |
|  |

Q1. Should we only consider the senses shown in Table 1 and Table2 (‘senses of interest’)

Q2. Can we visually choose which senses to study based on the figures above (fig1 and fig2)

**\_Transform to Binary**

Different threshold values used to transform the data into binary data (1 or 0). The threshold vector is

exist\_thresholds = [0.1, 0.2, 0.3, 0.4]

Methods used to calculate the correlation

* Phi Coefficient (φ): [It's suitable for measuring the degree of association between two binary variables.]

**Note: Incorrect Implementation in Python**

* Others could be used:
  + Cramér's V: more commonly used for categorical data. Q can I use it since binary is a special case of categorical
  + Kendall's: can I use it [Suitable for both **continuous and ordinal** (ranked) data] Maybe I transfer the data into three ranks A for [0, 0.1], B for ]0.1, 02] and C for ]0.3, 1]

\_Questions

Q3. Is it custom to transform such data into binary data?

Q4. Is this an acceptable/custom way to transform data into binary data?

Q5. Do we have enough non-zero data to perform correlation analysis?

Q6. What is the best method to analyse the binary data (Phi Coefficient (φ), Cramér's V, Kendall's)

Q7. Should we consider P value?

**\_Processing Compositional Data**

Q8. Is the data compositional data?

* 1. Originally each row sums up 1
  2. We removed the last two columns for a reason and not not all the rows sum up to 1
     1. Discogem:

|  |  |
| --- | --- |
| **# of entities** (out of 6505) | **Sum of the row** |
| 3,871 (60%) | = 1 |
| 1,547 (24%) | = 0.9 |
| 1,087 (16%) | <= 0.8 |

* + 1. ~~Quadc~~

|  |  |
| --- | --- |
| **~~# of entities~~** ~~(out of 900)~~ | **~~Sum of the row~~** |
| ~~830 (92%)~~ | ~~= 1~~ |
| ~~26 (3%)~~ | ~~= 0.9~~ |
| ~~44 (5%)~~ | ~~<= 0.8~~ |

Q9. If transformation is needed, are any of the three below suitable for the transformation?

ALR, ILR, CLR

|  |  |
| --- | --- |
| Additive logratio transform (ALR) |  |
| Isometric logratio transform (ILR) |  |
| Center log ratio transform (CLR) |  |

|  |
| --- |
| **Fig 1. Subset of Pairwise Scatter Plots for Leaves Level** |
|  |

1. After transformation, what is the best correlation coefficient to use

['pearson', 'spearman', 'kendall']

|  |  |
| --- | --- |
| 'pearson' | * Assumption: Assumes a **linear relationship** and that data is normally distributed. * Use Case: Suitable for continuous data when you want to **measure linear associations.**   *[A* ***linear relationship*** *is a specific type of monotonic relationship where the* ***rate of increase*** *or decrease between two variables is* ***constant****]* |
| 'spearman' | * Assumption: **Non-parametric** and does not assume a linear relationship but assumes a **monotonic** relationship. * Use Case: Appropriate for both **continuous and ordinal data**. Particularly useful when the relationship is expected to be **monotonic** but not necessarily linear.   [ *A monotonic relationship, either* ***consistently increases or decreases*** *but* ***not necessarily*** *at* ***a constant rate***  ***[***Non-parametric Non-parametric, *do not make assumptions about the underlying distribution of the data.* |
| 'kendall' | * Assumption: **Non-parametric** and makes **no assumptions about the data distribution.** * Use Case: Suitable for both **continuous and ordinal** (ranked) data. Useful when the data may not follow a linear relationship. |

[source](Assumption:%20Non-parametric%20and%20makes%20no%20assumptions%20about%20the%20data%20distribution.%20Use%20Case:%20Suitable%20for%20both%20continuous%20and%20ordinal%20(ranked)%20data.%20Useful%20when%20the%20data%20may%20not%20follow%20a%20linear%20relationship.)

1. It is safe to say that our data are not liner? Examples

Check if they are normal distributes

Plot transformaiton