1. What I did so far, where we are at
   1. Describe the data
   2. Transformation and analysis
2. Possible better way to analyse the data

Note:

Meeting with prof. Hamid Pezeshk (LTA Assistant Professor, Mathematics and Statistics) Tomorrow afternoon

Dr Yiming Xiao

Assistant Professor, Computer Science and Software Engineering

**\_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% |
| % of >0.3 per col [0.4-1] | 0% | 6% | 4% | 22% | 1% | 3% | 1% | | 0% | 22% | 3% | 0% | 13% |
| **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 | **asynchronou**s | **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% |
| % of >0.2 per col [0.4-1] | 0% | 6% | 31% | 6% | 1% | | 0% | 22% | 4% | 16% |
| 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. How should we treat the data: continuous, ordinal, distinct

Q2. Should we only consider the senses shown in Table 1 and Table2 (‘senses of interest’)

Q3. 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

Q4. Is it custom to transform such data into binary data?

Q5. Is this an acceptable/custom way to transform data into binary data?

Q6. Do we have enough non-zero data to perform correlation analysis?

Q7. What is the best method to analyse the binary data (Phi Coefficient (φ), Cramér's V, Kendall's)

Q8. Should we consider P value?

**\_Processing Compositional Data**

Q9. 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~~ |

Q10. 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) |  |

Q11. Considering fig 3 below, what is the best approach to analyze correlation

Q12. Could we consider the data to be ordinal so that we use correlation coefficients such as 'spearman' and kendall directly without transformation?

|  |
| --- |
| **Fig 3. Subset of Pairwise Scatter Plots for Leaves Level – CLR Transformation** |
|  |

|  |
| --- |
| **Fig 4. Subset of Pairwise Scatter Plots for Level Two – CLR Transformation** |
|  |

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.)

|  |  |  |
| --- | --- | --- |
| 10 annotators in general | S1 | S2 |
| 5 annotators | X |  |
| 5 annotators |  | X |

* There is a confusion between S1 and S2
* >=50% 🡺 1 else 0
* Allow some ‘noise’ we set it at 40%

|  |  |  |
| --- | --- | --- |
| **S1** | **S2** | Other S’s |
| 50% | 50% | 0% |
| 40% | 60% | 0% |
| 60% | 40% | 0% |
| 50% | 40% | 10% |
| 40% | 50% | 10% |
| 40% | 40% | 20% |

Only looking at >40%

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Leaves level** | **precedence** | **result** | **conjunction** | **arg2-as-detail** |
| Non-zero value% per col - % of >0 i.e. [0.1-1] | 29% | 67% | 76% | 60% |
| % of >0.1 per col [0.2-1] | 15% | 46% | 53% | 37% |
| % of >0.2 per col [0.3-1] | 9% | 32% | 35% | 22% |
| % of >0.3 per col [0.4-1] | 6% | 22% | 22% | 13% |
| **Col total val/ data total val** | 7% | 22% | 23% | 16% |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **asynchronou**s | **cause** | **concession** | **conjunction** | **level-of-detail** |
| Non-zero value% per col - % of >0 i.e. [0.1-1] | 32% | 77% | 37% | 76% | 67% |
| % of >0.1 per col [0.2-1] | 16% | 59% | 17% | 53% | 43% |
| % of >0.2 per col [0.3-1] | 10% | 44% | 9% | 35% | 26% |
| % of >0.2 per col [0.4-1] | 6% | 31% | 6% | 22% | 16% |
| Col total val/ data total val | 8% | 28% | 8% | 23% | 18% |

+ makes more sense

+ focus on the data that does not have a lot of zeros

(-) We still have the zero problem

(-) we only examining few senses

|  |  |  |
| --- | --- | --- |
| S1 | S2 | Confusion |
| 0 | 0 | ??? |
| 0 | 0 | ?? |
| . | . | ?? |
| . | . | ?? |
| . | . | ?? |
| . | . | ?? |
| 1 | 0 | No |
| 1 | 1 | Yes |
| 1 | 1 | Yes |
| 0 | 0 | ?? |
| . | . | ?? |
| 0 | 0 | ?? |
| 0 | 1 | No |
| 0 | 0 | ?? |
| . | . | ?? |
| 0 | 0 | ?? |
| 1 | 1 | Yes |
| 1 | 1 | Yes |
|  |  |  |

Remove all double zeros

|  |  |
| --- | --- |
| S1 | S2 |
| 0.1 | 0.4 |
| 0.3 | 1 |
| 0.2 | 0.1 |
| 0.7 | 0 |
| 0.5 | .1 |

* Confusion (+ corr)

Practical use :

Nelson gives Nawar a sentence(s)

Nawar annotates as S1

Nelson consider that the answer could be S1 OR S2

--- ---- ----

|  |  |
| --- | --- |
| S1 | S2 |
| 0.1 | 0 |
| 0 | 0.5 |
| 1 | 0 |
| 0.3 | 0 |
| 0 | 0.7 |

No confusion (- corr)

Practical use :

Nelson gives Nawar a sentence(s)

Nawar annotates as S1

Nelson consider the answer to be S1 OR not S2

--- ---- ---- ----

|  |  |
| --- | --- |
| S1 | S2 |
| 0.3 | 0 |
| 0 | 0.8 |
| 0.2 | 0.3 |
| 0.5 | 0.1 |
| 0 | 1 |
| 0.2 | 0.3 |

No corr found (not no corr)

Try another method, or another dataset

(+) get rid of the zero problem

(+) we can consider all senses in theory

(-) removing the double zeros will lead to much fewer instances for most of the senses

(-) treating the confusion between 4-5 annotators