* Veröffentlichung als Open-Source?
* Binärer Vergleich für nominale Fragen?
* Gewichtung bei nominalen Fragen bzw. bei binärem Vergleich?
* Genauigkeit? Double genügend?
* Umgang mit fehlenden Werten?
* Umgang mit ignore Werten?

Inhalt

[Organisatorisches 3](#_Toc284918098)

[Planung 3](#_Toc284918099)

[Spezifikation 4](#_Toc284918100)

[Datenstrukturen 4](#_Toc284918101)

[Ordinal-reelle Fragen 4](#_Toc284918102)

[Ordinal-diskrete Fragen 4](#_Toc284918103)

[Nominale Fragen 4](#_Toc284918104)

[Eingabe-/Ausgabedateien 5](#_Toc284918105)

[Input-Datei 5](#_Toc284918106)

[Projektdatei 5](#_Toc284918107)

[Outputdatei 5](#_Toc284918108)

[Benutzerschnittstellen 6](#_Toc284918109)

[Grafische Oberfläche 7](#_Toc284918110)

[Navigation 7](#_Toc284918111)

[S1 Datenselektion 7](#_Toc284918112)

[S2 Fragenspezifikation 8](#_Toc284918113)

[S3 Matrixspezifikation 9](#_Toc284918114)

[S4 Matrix speichern und analysieren 10](#_Toc284918115)

[Algorithmus 10](#_Toc284918116)

[P1 Datenanalyse 10](#_Toc284918117)

[P2 Berechnung Vergleichsvektor 10](#_Toc284918118)

[P3 Berechnung Ähnlichkeitsmatrix 10](#_Toc284918119)

# Organisatorisches

## Planung

## Vorgehen

1. Entwicklung des grundlegenden Algorithmus mit ein paar Verfahren
2. Datenimport/Export
3. Entwicklung der Screens

# Spezifikation

## Begriffe

### Ordinal-reelle Fragen

Bei diesen Fragen. Fragen mit mehr als 20 verschiedenen gegebenen Antworten sollen vom Programm als ordinal-reele Fragen kategorisiert werden.

### Ordinal-diskrete Fragen

Bei diesen Fragen gibt es nur eine beschränkte Anzahl möglicher Antworten. Die Antworten lassen sich aber Vergleichen und auf einer Achse anordnen.

### Nominale Fragen

Bei diesen Fragen handelt es sich ebenfalls um Fragen mit einer beschränkten Anzahl Anwortmöglichkeit. Der Unterschied zu den ordinal-diskreten Fragen liegt darin, dass sich die nominalen Fragen nicht auf einer Achse anordnen lassen.

### Fragebogen

Bei einem Fragebogen handelt es sich um eine Ansammlung von Antworten auf Fragen. Im Kontext dieser Applikation entspricht ein Fragebogen einem Datensatz.

## Eingabe-/Ausgabedateien

### Input-Datei

Bei der Input-Datei handelt es sich um eine CSV-Datei. Es sollen jedoch verschiedene Trennzeichen unterstützt werden. Es wird angenommen, dass sämtliche Werte bereits als Zahlen codiert wurden. In der Datei befinden sich also nur Zahlen. Jede Zeile enthält einen Fragebogen, dementsprechend enthält jede Spalte eine Frage. In der ersten Zeile befinden sich eventuell die Spaltentitel der Fragen. Es kann davon ausgegangen werden, dass die Dateien noch genügend klein sind um komplett in den Arbeitsspeicher geladen zu werden.

### Projektdatei

Es soll möglich sein die im Programm getätigten Einstellungen in einer Projektdatei abzulegen und die Datei zu einem späteren Zeitpunkt neu zu laden.

### Outputdatei

Die Ausgabe der Distanzmatrix soll eine herkömmliche CSV-Datei sein. Das Trennzeichen soll vom Benutzer festgelegt werden können.

### Formelspezifikation

Eventuell nötige Datei um eine Berechnungsformel zu spezifizieren.

## Benutzerschnittstellen

Es sollen nach Möglichkeit 2 Benutzerschnittstellen entworfen werden:

* Eine grafische Benutzeroberfläche
* Eine Kommandozeile

Die grafische Benutzeroberfläche soll sich an Benutzer wenden welche über kein technisches Hintergrundwissen verfügen. Konkret soll sie sich an Benutzer aus dem Bereich der Sozial- und Politikwissenschaften richten.

Die Kommandozeile kann sich an ein eher technisch versiertes Publikum richten. Die Kommandozeile soll verwendet werden um das Testing zu erleichtern. Ausserdem könnten mit Hilfe der Kommandozeile eventuell Batch-Analysen ermöglicht werden d.h. Analysen die Beispielsweise sämtliche Distanzmasse durchrechnen.

## Grafische Oberfläche

### Ablauf

### Navigation

Bei der Berechnung der Ähnlichkeitsmatrix handelt es sich um einen Prozess. Dies soll auf der grafischen Oberfläche auch sichtbar sein. Aus diesem Grund wir auf jedem Screen ein Navigationselement eingeblendet auf dem der Benutzer erkennen kann wo er innerhalb des Prozesses steht.

In einem ersten Schritt soll dieses Element nur ein visuelles Hilfsmittel sein. Es wäre auch denkbar, dass der Benutzer durch Anklicken zum entsprechenden Prozessschritt springen kann. Dies würde dem Benutzer jedoch ermöglichen gewisse Prozessschritte auszulassen und ist aus diesem Grund problematisch.

### S1 Datenselektion

Im ersten Schritt muss der Benutzer definieren aufgrund welcher Daten er die Ähnlichkeitsmatrix berechnen möchte. Dazu muss er eine CSV-Datei aussuchen. Zusätzlich kann er aussuchen welche Fragen und welche Fragebögen er berücksichtigen möchte.

#### Offene Fragen/Ideen/Erweiterungen

* Man könnte für die Auswahl der Spalten ein besseres GUI implementieren.

### S2 Fragenspezifikation

Auf diesem Screen soll der Benutzer die Details zu den Fragen spezifizieren können. Für diesen Zweck werden die Fragen in tabellarischer Form aufgelistet. Einige Spalten sind für den Benutzer editierbar, andere nicht.

|  |  |  |  |
| --- | --- | --- | --- |
| **Spaltenname** | **Inhalt** | **Editierbar** | **Kommentar** |
| Bezeichung | Bezeichnung der Frage (Text oder Nummer) |  | Eventuell Editierbar? |
| Werte | Auflistung aller eindeutiger Werte |  | "Metrisch" sobald mehr als 20 Werte vorhanden sind |
| Min | Kleinster Wert |  |  |
| Max | Grösser Wert |  |  |
| Ausschluss | Werte die ausgeschlossen werden sollen | X | Hat einen Einfluss auf die Spalten Min & Max |
| Normierung |  | X |  |
| Gewicht Frage | Optionale Gewichtung der Frage | X | Default = 1 |
| Typ | Metrisch/Ordinal/Nominal | X |  |
| Gewicht Distanz | Optionale Gewichtung der berechneten Distanz | X | Default = 1 |

Weitere Eigenschaften der Tabelle:

* Die Tabelle soll nach Möglichkeit sortierbar sein.
* Weiterhin soll Copy/Paste von und nach Excel möglich sein.

#### Offene Fragen/Ideen/Erweiterungen

* Bezeichnung Frage editierbar?
* Anzeige von Mittelwert/Standardabweichung?
* Nicht nur Normierungsfaktor sondern auch ein Faktor um einen Offset zu entfernen?
* Aussschlusskriterium eventuell auch mit grösser als? z.B. alle Werte > 20 ausschliessen?
* Automatische Berechnung einer Gewichtungsfunktion z.B. exponentiell absteigende Gewichtung der Fragen.

### S3 Matrixspezifikation

Auf diesem Screen soll der Benutzer definieren mit welchem Algorithmus er die Ähnlichkeitsmatrix erstellen möchte. Je nach Algorithmus muss der Benutzer eventuell zusätzliche Parameterwerte angeben.

Da unter Umständen die Berechnung bereits längere Zeit benötigt, soll es möglich sein nur ein kleines Zufallssample auszuwerten. Dies soll es ermöglichen in kurzer Zeit herauszufinden ob sich ein Distanzmass für die Datenauswertung eignet. Diese Funktion soll jedoch erst implementiert werden, wenn es nötig wird.

#### Offene Fragen/Ideen/Erweiterungen

* Die Möglichkeit für verschiedene Fragetypen unterschiedliche
* Falls nominale Fragen vorhanden sind können gewisse Algorithmen nicht verwendet werden

### S4 Matrix speichern und analysieren

### S21 Fragenanalyse

### S31 Formeleditor

### S41 Datenmatrix analysieren

## Algorithmus

### P1 Datenanalyse

In diesem Schritt sollen die Daten auf Konsistenz geprüft und analysiert werden:

* Sind nur Zahlen in der Datei vorhanden?
* Welche Fragen sind vermutlich vom Typ "metrisch/ordinal"?
* Sind die vom User angegebenen Spalten/Zeilen vorhanden?

### P2 Berechnung Vergleichsvektor

In diesem Schritt werden die Vektoren berechnet, welche anschliessend für die Berechnung der Ähnlichkeitsmatrix verwendet werden.

Aus dem Antwortvektor x wird in diesem Schritt der Vergleichsvektor v berechnet. Es wird folgende Formel angewendet:

### P3 Berechnung Distanzmatrix

In diesem Schritt wird die eigentliche Distanzmatrix berechnet. Grundlage für die Berechnung sind die Vergleichsvektoren plus das gewählte Distanzmass inklusive allfälliger Parameter.

Es kann davon ausgegangen werden, dass die Matrix symmetrisch ist. Aus diesem Grund entfällt die Berechnung für die Hälfte der Fragebogen.