Überarbeitung wissenschaftlicher Arbeiten anhand einer Kombination von Text Mining und Visualistik

Sascha Lemke
Matr.Nr.: 11074933

Dennis Dubbert
Matr.Nr.: 11089478

ABSTRACT

Dieses Dokument beschreibt eine Projektarbeit, welche versucht den Prozess der Textverbesserung zu unterstützen, indem Verfahren aus dem Bereich Text Mining und Visualistik kombiniert werden. So wird im ersten Schritt definiert, wie und welche Daten gesammelt wurden und wie diese daraufhin veranschaulicht werden. Dabei wird durch eine Kombination aus visuellen und textuellen Darstellungen versucht, eine direkte Brcke zwischen diesen Repräsentationen zu schaffen

1 EINLEITUNG

In den letzten Jahren haben immer mehr digitale Medien Einzug in das alltägliche Leben gehalten. Insbesondere das Internet ist dabei für einen Groteil der Veränderung verantwortlich. Doch neben Bildern, Videos und Audio ist der Text, gerade auch durch das Internet, immer noch das wichtigste Kommunikationsmittel für Wissen.

Doch nicht nur im alltäglichen Leben, insbesondere im wissenschaftlichen Umfeld ist Schrift und Sprache wichtig. Das Verfassen von Hausarbeiten oder das spätere Schreiben von Artikeln für Journale erfordert dabei ein gewisses Ma an Qualität. Diese Qualität wird oft durch Erfahrung, einhalten von diversen Richtlinien und Zeit in Form von wiederholten Iterationen erreicht.

Dieses Projekt versucht diesen Prozess mit Hilfe von Visualisierungen von Text Mining Daten zu unterstützen. In vielen Leitfäden zu wissenschaftlichem Schreiben finden sich Richtlinien wie Präzision oder das vermeiden von Wortwiederholungen. Diese Werte lassen sich mit Text Mining und Mitteln der deskriptiven Statistik leicht ermitteln und dienen als Grundlage für die, in diesem Dokument beschriebenen Visualisierungen. Aus diesem Grund sollten die Texte auf folgende Eigenschaften untersucht werden:

- Worthäufigkeit (Redundanz)
- Satzlänge (Komplexität)
- Satzzeichen (Komplexität)
- Füllwörter (Präzision)

Im folgenden Kapitel wird beschrieben, welche Text Mining Verfahren eingesetzt wurden, um die Daten fr die Visualisierung der oben beschrieben Eigenschaften zu gewinnen.

2 TEXT MINING UND DATENSTRUKTUREN

Um die oben beschriebenen Eigenschaften visualisieren zu können muss eine entsprechende Datenbasis geschaffen werden, welches durch Text Mining ermöglicht wird. Der Prozess teilt sich dabei in folgende Schritte auf [1] :

- Aufgabendefinition
- Dokumentselektion
- Dokumentaufbereitung
- (Text) Mining Methoden
- Interpretation / Evaluation
- Anwendung

Die in der Einleitung beschrieben Motivation bildet die Aufgabendefinition dieses Projekts. Der nächste Schritt umfasst die Recherche und Sammlung von Dokumenten, welche dem Erreichen der vorher formulierten Ziele der Aufgabendefinition dienen. Oft werden dabei Dokumente aus verschiedenen Quellen mit verschiedenen Formaten entnommen. Die Aufbereitung der Dokumente nimmt in einem Text Mining Vorhaben den gröten Zeitaufwand ein, da hier Formate, Codierungen und Zeichen standardisiert werden müssen. Vor der Anwendung der Ergebnisse müssen die Daten noch interpretiert und evaluiert werden. Dabei kommen oft Experten mit entsprechenden Domänenwissen zum Einsatz, denn auch heute noch kommt Text Mining nicht komplett ohne menschliche Komponente aus.

2.1 Vorverarbeitung

Da sich dieses Projekt primär auf die Visualisierung der Daten konzentriert, wurde die Menge der Dokumente begrenzt und die Vorverarbeitung entsprechend angepasst um Zeit zu sparen.

Hierbei wurde aus Zeitgründen nur ein Dokument gewählt, welches die Möglichkeiten von Text Mining und Visualisierung verdeutlichen soll. Dieses wurde entsprechend mit Hilfe von XML standardisiert. So fällt ein Groteil der Zeit für die Sammlung von Dokumenten und dessen Standardisierung weg, welche stattdessen für die Visualisierung genutzt werden kann.
Durch diese Standardisierung lässt sich das Dokument nun auf mehreren Ebenen analysieren [2] , [1] .

2.1.1 Morphologisch

Die morphologische Untersuchung behandelt die Texte auf Zeichenebene. Dabei werden folgende Verfahren auf Wortformen (sogenannte Token) angewendet. Hier einige Beispiele:

- Tokenisierung
- Finden von Satzenden
- Stammformreduktion (Stemming)

Üblicherweise beginnt man damit den Text in einzelne Wortformen (Token) zu zerlegen. Dabei wird in westlichen Sprachen das ”white-space-tokenizing” verwendet. Dadurch werden Strings anhand von Leerzeichen, Tabs, Zeilenumbrüchen und ähnlichen Trennzeichen separiert. Hier treten erste Herausforderungen auf, da nicht jede Sprache über Leerzeichen verfügt (Chinesisch) oder Wörter wie ”New York” werden als einzelne Worte betrachtet, obwohl sie zusammen gehören.

Der nächste Schritt umfasst das Finden von Satzgrenzen. Auch hier treten einige Herausforderungen auf, da beispielsweise nicht jeder Punkt ein Satzende darstellt. So bilden Abkürzungen wie ”Dr.” oder ”et al.” keine Satzgrenzen und dürfen dementsprechend nicht markiert werden..

Ein weiterer Schritt ist das Stemming, welcher allerdings oft in Abhängigkeit der Text Mining Ziele durchgeführt wird. So gibt es hier einfache Normalisierungen wie das Auflösen von Ein- und Mehrzahl, welches man auch als ”Inflectional Stemming” bezeichnet [2]. Das Ziel dabei ist es die Genauigkeit der darauf folgenden Analyse zu erhöhen, etwa bei statistischen Analysen wie Worthäufigkeiten. Als Gegensatz dazu existiert auch das ”Root Stemming”, welches eine aggressivere Form des Stemmings darstellt. Dabei geht das Stemming so weit, das Token wie ”Application” auf ihren eigentlichen Wortstamm zurückgeführt werden (”apply”).

Da diese Untersuchungen oft die Grundlage von Text Mining Vorhaben bilden werden auch in diesem Projekt einige der oben genannten Methoden verwendet. So wurden die Texte tokenisiert und die Satzgrenzen entsprechend gekennzeichnet.

Durch die Tokenisierung des Dokuments lassen sich die notwendigen Worthäufigkeiten für die Visualisierung ermitteln. Zusätzlich kann dadurch die Satzlänge und das Auftreten von Satzzeichen ermittelt werden, was die Daten für die Ermittlung der oben beschriebenen Komplexität liefert.

2.1.2 Syntaktisch

Die syntaktische Ebene befasst sich mit der Untersuchung von Texten auf Satzebene. Dabei werden die Zeichen in Beziehung zueinander gestellt und analysiert. Beispiele dafür sind folgende:

- Part-of-Speech
- Phrase Recognition

Bei dem Part-of-Speech Tagging handelt es sich um das Erkennen von Wortklassen (Nomen, Verben, ...) und die Annotation der Token mit diesen. Aufgrund der Vielfältigkeit von Sprache (und dem Umstand das eine Sprache lebt) werden hier oft statistische Ansätze verwendet.

Bei der Phrase Recognition geht es darum Phrasen in Texten zu erkennen. So können POS-Tags dabei helfen bestimmte Wortgruppen zu finden, oder einfache Form wie das ”Noun phrase chunking” ermöglichen.

Wie bereits oben beschrieben liegt der Fokus auf diversen Messwerten, daher ist eine syntaktische Untersuchung der Texte in diesem Projekt nicht zielführend. Dennoch wurden mit Hilfe von Part-of-Speech Tagging versucht Werte fr eine mögliche Erkennung des Nominalstils zu sammeln, welche im Ausblick diskutiert werden.

2.1.3 Semantisch

Die semantische Analyse befasst sich mit der Bedeutung von Wörtern und ihrer digitalen Darstellung. Hier wird oft auf Hintergrundwissen zurück gegriffen, so werden beispielsweise Fachsprachen (Informatik, Medizin, etc) über Ontologien oder Taxonomien modelliert und in den Prozess eingebunden. Damit sollen Probleme wie beispielsweise Mehrdeutigkeit (Word Sense Disambiguation) umgangen werden.

Diese Art der Untersuchung erfordert allerdings einen gröeren Umfang, als es in diesem Projekt möglich war, daher wurden keine semantischen Untersuchungen durchgeführt.

2.2 Analyse

Der Bereich Text und Data Mining umfasst eine groe Mengen an Methoden um diverse Informationen aus Texten zu extrahieren bzw. zugänglich zu machen. So werden diese oft in verschiedene Kategorien eingeteilt, wie beispielsweise das Information Retrieval oder Information Extraction. Diese Methoden sind sehr weitreichend und aufwendig. Eine entsprechende Vorstellung und Durchführung sind nicht Ziel dieses Projekts, daher wird an dieser Stelle darauf verzichtet.

Stattdessen werden statistische Hilfsmittel genutzt um Daten für die Visualisierungen zu gewinnen. Dabei werden Mittel der deskriptiven Statistik verwendet um die in der Einleitung beschrieben Werte zu ermitteln.

So ist insbesondere die Hufigkeit diverser Token ein wichtiger Messwert, welcher als Grundlage für eine qualitative Aussage dienen soll. Dabei handelt es sich um folgende einfache Formel [2]:

$$tf(t,d) = f_{t,d}$$

Dabei beschreibt t den Term und d das Dokument. Da hier nur ein Dokument als Grundlage dient handelt es sich um eine einfache Summenfunktion:

$$\sum_{i=1}^{n} t_i$$

Neben diesen recht einfachen Werten sollen zusätzliche statistische Werte einen Überblick geben. So sollen Median und Durchschnitt zusätzliche Informationen anbieten, beispielsweise die durchschnittliche Satzlänge oder Verwendung von Satzzeichen.

2.3 Datenstruktur und Implementierung

Da die Visualisierung mit Hilfe von d3.js implementiert wird, werden die Daten in JSON bereitgestellt. Die Texte werden mit Hilfe von Open NLP der Apache Software Foundation aufbereitet, welches Java als Programmiersprache nutzt. Da das Dokument bereits in XML standardisiert ist, lässt es sich leicht mit Hilfe von JAXB und Data Binding in Java Objekte übersetzen.

Open NLP stellt auerdem einige Modelle bereit um Aufgaben wie Tokenisierung oder Part-of-Speech Tagging durchzuführen. Die Daten werden im Anschluss mit einem Node.js Server gekoppelt um diverse Filteroptionen zu ermöglichen.

In den folgenden Abschnitten werden die gesammelten Daten und deren Datenstrukturen im Detail beschrieben.

Worthäufigkeit

Die Worthäufigkeit wird durch durch ein Schlüssel-Wert-Paar dargestellt und mit Informationen über den Fundort angereichert. So wird nicht nur das Wort und die Anzahl gespeichert, sondern das Kapitel und die entsprechenden Unterkapitel mit gespeichert. Zusätzlich wurde der Wert in Abhngigkeit der gesamten Wortanzahl des Fundorts normalisiert. Daraus ergibt sich folgende Datenstruktur:

```
{
  "word" : "visualization",
  "count" : 1,
  "chapterID" : 1,
  "sectionID" : 2,
  "subsectionID" : 0,
  "subsubsectionID" : 0,
  "normalized" : 0.4484304932735426
}
```

Satzlänge

Die Sätzlänge wird ebenfalls durch die Häufigkeit dargestellt und mit ID's versehen. Hier erhält zusätzlich jeder Satz noch eine eigene ID und der Satz wird im Original mit gespeichert.

```
{
  "chapterID" : 1,
  "sectionID" : 2,
  "subsectionID" : 0,
  "subsubsectionID" : 0,
  "sentenceID" : 135,
  "length" : 17,
  "sentence" : "Visual analytics (VA) is typically
       applied in scenarios where complex data has
       to be analyzed."
}
```

Satzzeichen

Die Struktur Satzzeichen ist an der Struktur der Satzlänge angelehnt. Der wichtige Unterschied ist das hier eine Liste der Position der Token mit angegeben wird, um im Nachhinein aufzuzeigen, an welcher Stelle sich die Satzzeichen befinden.

```
{
  "sentenceID" : 135,
  "count" : 2,
  "sentence" : "Visual analytics (VA) is typically
       applied in scenarios where complex data has
       to be analyzed.",
  "token" : [ 4, 17 ],
  "chapterID" : 1,
  "sectionID" : 2,
  "subsectionID" : 0,
  "subsubsectionID" : 0
}
```

Füllwörter

Da die Füllwörter nicht durch einen einfachen Zähler dargestellt werden können, unterscheidet sich diese Datenstruktur stark von den bisher vorgestellten. Da die Füllwörter pro Paragraph berechnet wurden erscheint hier eine zusätzliche ID. Auch eine Liste der Token ID's wurde übernommen, um die Position der Füllwörter aufzuzeigen. Da Paragraphen weder durch Text noch durch Zahlen eindeutig zu beschreiben sind, werden hier auch die Titel der diversen Überschriften mit angegeben um so eine verständliche Zuordnung zu ermöglichen.

```
{
  "paragraphID" : 3,
  "count" : 94,
  "token" : [ 5, 8, 10, 13, 14, 15, 20, 21, 22,
      25, 26, 28, 29, 31, 32, 34, 35, 37, 39, 44,
      47, 48, 51, 52, 54, 59, 62, 65, 67, 70, 71,
      77, 79, 81, 82, 85, 86, 87, 89, 90, 92, 95,
      100, 101, 103, 108, 110, 111, 114, 118, 120,
       121, 123, 126, 128, 132, 133, 142, 143,
      145, 146, 148, 150, 152, 154, 156, 159, 164,
       168, 169, 171, 173, 176, 177, 181, 184,
      188, 191, 196, 197, 198, 199, 200, 201, 203,
       205, 206, 207, 208, 211, 214, 218, 219, 221
       ],
  "chapterID" : 1,
  "sectionID" : 2,
  "subsectionID" : 0,
  "subsubsectionID" : 0,
  "chaptername" : "abstract",
  "sectionname" : null,
  "subsectionname" : null,
  "subsubsectionname" : null,
  "idInChapter" : 1
}
```

3 Aufbau der Visualisierung

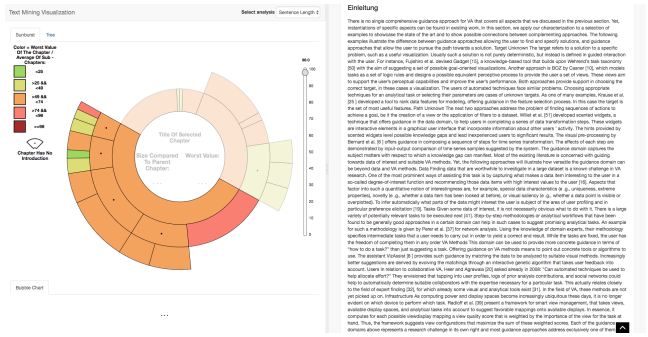

Figure 1: Startansicht der Visualisierung (links: Sunburst-Diagramm als Navigation, rechts: Betrachtetes Dokument im Text-Viewer-Fenster)

Die Visualisierung ist in zwei Abschnitte eingeteilt, welche vertikal voneinander Abgetrennt sind und in einer starken Synergie stehen (siehe Abbildung 2). Die Linke Seite des Bildschirms wird von der eigentlichen Visualisierung eingenommen und die rechte Seite stellt einen Text-View dar. In der Visualisierung können verschiedenen Auswahlen getroffen und hierdurch innerhalb des Text-Views markiert werden. Anhand der textuellen Darstellung und Nähe der markierten Begriffe kann der Nutzer nun Rückschlüsse bezüglich der Wissenschaftlichkeit seines Dokuments bilden und gegebenenfalls Verbesserungen an diesem vornehmen. Die Abschnitte sind durch einen Balken klar abgetrennt, welcher jedoch, jeh nach Anforderung des Nutzers, verschoben werden kann, um die Größenverhältnisse anzupassen. Wird gerade eine Auswahl getroffen, so kann es nützlich sein die Visualisierungsseite zu vergrößern. Wurde bereits eine Auswahl getroffen so kann die Textseite vergößert werden.

Um eine geeignete und feingranulare Auswahl dieser Begriffe gewährleisten zu können, ist auch der Visualisierungsbereich in weitere, aufeinander aufbauende, Einzelvisualisierungen gegliedert.

Somit beinhaltet sie:

- eine Navigation,
- eine Kapitelübersicht / -auswahl,
- und eine Detailansicht der Selektion.

Jede dieser Visualisierung hat ein festes Seitenverhältnis, sodass eine Verzerrung durch Größenänderungen vermieden wird.

Innerhalb der Navigation wird dem Nutzer ein Drop-Down-Menü geboten, über welches sich die Filterungsmethode definieren lässt (Redundanzen, Satzzeichen, Satzlänge oder Füllwörter). Wurde hier eine Auswahl getroffen, so wird die gesamte Visualisierung auf diese abgestimmt und dem Nutzer eine gezieltere Suche ermöglicht.

Der zweite Abschnitt veranschaulicht die Hierarchie des ausgewählten Dokuments und wird für die Auswahl zu betrachtender Kapitel genutzt. Hier werden zwei alternative Ansichten geboten, eine in Form eines Sunburst und eine als Baumstruktur, welche teilweise unterschiedliche Ausprägungen hervorheben und sich somit gegenseitig ergänzen. Über Tabs kann der Nutzer zu jedem Zeitpunkt die für ihn nützlichere Ansicht selektieren. Wichtig ist hierbei, dass die, in den Ansichten getätigten, Auswahlen synchronisiert werden, wodurch ein fließender Wechsel unterstützt wird.

Wurde in der Kapitelübersicht eine Auswahl getroffen, so wird diese in einer Detailansicht aufgegriffen. Hier sind die Begriffe der Ausgewählten Kapitel anhand der Filterungsmethode eingegrenzt und in Form eines Blasen-Diagramms visualisiert. Wurden beispielsweise Redundanzen selektiert, so steht jede Blase für ein Wort, welches innerhalb der Kapitelauswahl redundant auftritt. Zu jedem dieser Begriffe wird auch die entsprechende Ausprägung der Filterungsmethode bereitgestellt, sodass eine persöhnliche Problemeinschätzung des Nutzers erfolgen kann. Ist nun eine Überarbeitung bestimmter Begriffe gewünscht, so können diese selektiert und hierdurch jedes Vorkommnis, begrenzt durch die ausgewählten Kapitel, in dem Text-View markiert werden.

Das folgende Kapitel dient somit zunächst der Erläuterung dieser Einzelvisualisierungen. Hierbei werden besonders die Entscheidungsfindungen und Kompromisse beleuchtet, welche zu dem Aufbau und den Interaktionsmöglichkeiten der jeweiligen Visualisierung führten. Wichtig ist hier zudem zu beachten, dass sämtliche Visualisierung unter dem Aspekt der Anpassbarkeit entwickelt wurden. Auch wenn vorerst nur ein Dokument analysiert wurde, so wurden diese dennoch nicht von diesem Abhängig gemacht. Ziel war stets die Möglichkeit jeden Text analysierbar zu machen.

3.1 Kapitelübersicht: Sunburst-Diagramm

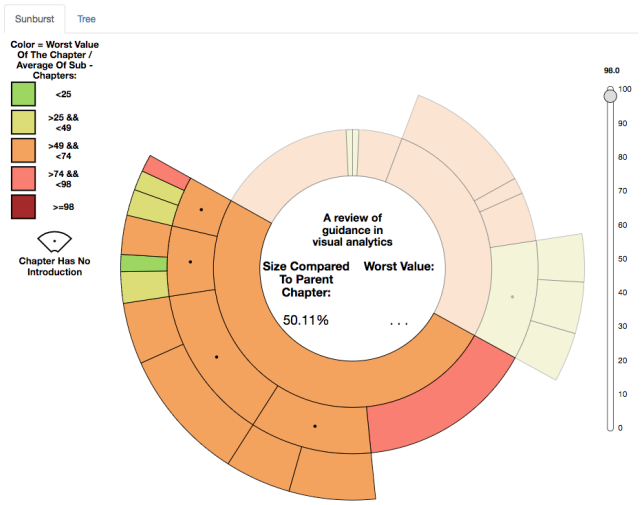

Figure 2: Sunburst-Diagramm als Kapitelauswahl mit Farblegende und Spezifizierungs-Slider (der Mauszeiger befindet sich über dem Kapitel "A review of guidance in visual analytics")

Dem Sunburst-Diagramm liegt der Gedanke zugrunde, dass der Nutzer den Aufbau des Dokuments bereits kennt, sich also im Vorhinein damit auseinander gesetzt hat und nun mithilfe der Visualisierung eine weitere Iteration vornehmen möchte. Im Fokus steht hierbei die Zentrierte und auf einen Blick Sichtbare Darzustellung der Dokumentstruktur um so eine schnelle Navigation und Auswahl zu ermöglichen. Da bei solch einem Diagramm nahezu keine Freiflächen zwischen den Elementen gegeben sind, kann den Elementen selber weitaus mehr Fläche zugeordnet werden, ohne die Übersichtlichkeit der Visualisierung zu gefährden. Dies führt zu einer Maximierung der Datendichte. Die verfügbare Visualisierungsfläche wird also bestmöglich ausgenutzt, da lediglich das Zentrum und die Ecken des Diagramms Freiflächen bilden, welche jedoch für Informationstexte nutzbar sind. Bei einer Verkleinerung des Diagramms sind einzelne Abschnitte somit klar erkenn- und trennbar, sodass die Visualisierung problemlos auf unterschiedlichen Bildschirmgrößen angezeigt und genutzt werden kann.

Ein weiterer Beweggrund für diese Art der Darstellung war, dass dem Nutzer ein Auswahlmedium bereitgestellt werden soll, welches die Dokumenthierarchie sowie Kapitel-Abhängigkeiten veranschaulicht. Durch die Aufeinanderschichtung der unterschiedlichen Ringlagen dieser Diagrammart sind die genannten Abhängigkeiten schnell erkennbar. Es werden keine weiteren visuellen Mittel benötigt um die Zugehörigkeit eines Kapitels zu verdeutlichen, da die Position der Elemente ausreichend Aufschluss hierüber bietet. Somit wird auch die Data-Ink-Ratio erhöht und Chartjunk vermieden. Der mittlere Kreis stellt hierbei das Dokument dar, welches zunächst von den Hauptkapiteln (Abstract, Einleitung, Hauptteil, Fazit, etc.) umringt wird. Die Größe des jeweiligen Ringabschnitts, in Abhängigkeit des Radius, ist hierbei der prozentuelle Anteil des Kapitels im Verhältnis zum gesamten Dokument. Als Maß gilt hier die Wortanzahl der Elemente, welche sich aus dem eigenen Inhalt, sowie dem Inhalt aller Unterkapitel erschließen lässt und anschließend auf den entsprechenden Teilwinkel des Kreises umgerechnet wird. Dieses Schemata erstreckt sich auch auf die Unterkapitel eines Abschnittes, welche jedoch nun ihr jeweiliges Elternkapitel als Maximum ansehen und nicht das gesamte Dokument (sowohl deren radialen

Abschnitt in der Visualisierung als auch deren Wortanzahl).

Ein Vorteil der hierdurch entsteht ist, dass neben der sichtbaren Hierarchie auch ein Größenverhältnis zum jeweiligen Oberkapitel ersichtlich wird. Für wissenschaftliche Arbeiten bestehen häufig formale Anforderungen, sodass auch die Größe einzelner Kapitel vorgegeben sein kann. Somit könnte es also nützlich sein solche Verhältnisse aus der Visualisierung entnehmen zu können. Hierbei wurde jedoch bedacht, dass Größenverhältnisse in Kreisdiagrammen trügen können, da der radiale Abschnitt die Größe eines Ringabschnitts definiert und somit äußere Abschnitte aufgrund ihrer Fläche automatisch größer erscheinen als Innere (auch bekannt als Lie Factor oder Lügenfaktor). Da bei der Überarbeitung eines wissenschaftlichen Textes jedoch vorwiegend das direkte Verhältniss eines Kapitels zu Nachbarkapiteln des selben Oberkapitels betrachtet wird, wurde dieser Aspekt im Kontext des Projektes als vernachlässigbar eingeschätzt. Im Sunburst-Diagramm besitzen zudem alle Lagen die selbe Höhe, sodass der Flächeninhalt hier eine geringere Bedeutung als der dargestellte Radius besitzt und somit trotzdem eine akkurate visuelle Einschätzung getätigt werden kann. Damit der eigene Textanteil eines Oberkapitels ebenfalls mit den Textanteilen der Unterkapitel verglichen werden kann, wurde entschieden auch diesen Anteil als Unterkapitel aufzulisten und ihn somit auf deren Ebene zu verschieben. Diese werden jedoch als Einleitung betitelt und somit durch ihre Namensgebung speziell gekennzeichnet, sodass die Zugehörigkeit bestehen bleibt. Dies ist jedoch nur benötigt, wenn ein Kapitel weitere Kapitel umschließt. Ansonsten wird der Textinhalt weiterhin im eigenen Element dargestellt.

Da in einem wissenschaftlichen Dokument jedes Kapitel eine Einleitung besitzen sollte und dieses nicht direkt aus der bisherigen Visualisierung hervorgeht, wurde zudem eine Möglichkeit gesucht, fehlenden Text aufzuzeigen. Sollte also ein Kapitel keinen eigenen Text beinhalten, so wird in der Mitte des zugehörigen Elements ein schwarzer Kreis gezeichnet. Dieser nimmt wenig Platz ein, wodurch die Data-Ink-Ratio nahezu unbeschadet bleibt. Durch die nun entstandene Unregelmäßigkeit innerhalb der Visualierung ist er jedoch auf den ersten Blick klar erkennbar, sodass dem Nutzer das Problem direkt übermittelt wird.

Ein weiterer wichtiger Aspekt dieser Visualisierung findet sich in der Farbgebung der Abschnitte. Da die Farbe ein gutes visuelles Mittel zur Übermittlung der Qualität ist, wird sie hier zu exakt jenem Zweck verwendet. Anhand der ausgewählten Filterungsmethode werden die jeweiligen Ausprägungen der Kapitel berechnet und diese anschließend in dem zugehörigen Farbton dargestellt. Bei der Auswahl der Farbe gilt der höchste ermittelte Wert innerhalb eines Kapitels als maßgebend für diese Einfärbung. Dies ist damit begründet, dass besonders Ausreißer erfasst und sichtbar gemacht werden sollten, welche bei der Bildung eines Durchschnitts verborgen bleiben könnten. Wurde beispielsweise die Filterungsmethode bezüglich der Satzlänge ausgewählt, so könnte bei der Durchschnittsbildung ein sehr langer Satz übersehen werden, welcher jedoch die Qualität des Dokumentes beeinträchtigt. Diese Einfärbung wird für jedes Kapitel eigenständig vorgenommen. Dies ist damit begründet, dass es Worthäufigkeiten gibt, welche einen hohen Einfluss auf die Textqualität eines Unterkapitels haben, jedoch verglichen mit dem Oberkapitel unauffällig erscheinen. Diese Aspekte sind nun direkt aus der Visualisierung entnehmbar, sodass die Trennung dem Nutzer eine gezieltere und schnellere Suche ermöglicht.

Bei der Farbwahl wurde von einer linearen Interpolation zwischen zwei Farbtönen abgesehen und sich stattdessen für eine klar unterteilte Skala mit den Farbtönen Grün, Gelb, Orange, Hellrot und Dunkelrot entschieden. Die Farbtöne wurden so gewählt, da die Endtöne im Komplimentärkontrast stehen, also eine einfache Unterscheidung der Qualität stattfinden kann. Zudem soll dem Nutzer über diese Farbgebung eine direkte Assoziation mit der gewünschten Aussage ermöglicht werden, da ein Grünton allgemein positives Feedback symbolisiert und ein Rotton vorwiegend als Gefahrensignal aufgefasst wird (Ampelprinzip). Auch wenn die letzten beiden Farbtöne den Grundton Rot besitzen, sind sie dennoch durch ihre Sättigung klar trennbar, sodass eine Qualitätsminderung erkenntlich wird. Weiterhin wurde die Skala auf fünf Untertöne reduziert. Auch hier ist wieder ein Argument, dass die Töne klar trennbar sein sollten und dies bei einer größeren Anzahl von Unterfarben nicht mehr gegeben wäre. Weiterhin ist es Ziel dieser Visualisierung dem Nutzer eine klare Einteilung der Qualität zu bieten, ebenso wie eine kurze Eingewöhnungsphase. Eine höhere Anzahl der Unterteilungen könnte die Eingewöhnungszeit verlängern.
Nun kann es vorkommen, dass unterschiedliche Nutzer auch unterschiedliche Anforderungen an ihr Dokument stellen. Ist die Zeit knapp, so steht beispielsweise die schnelle Identifikation der schwerwiegendsten Probleme im Vordergrund. Hat der Nutzer mehr Zeit oder sind strengere Regelungen an das Dokument gekettet, so sollten auch geringere Problemstellen aufgezeigt und beseitigt werden. Aus diesem Grund wurde den Farben kein fester numerischer Wert zugeordnet. Für die Zuordnung dieser Werte wurde ein Slider integriert, mit welchem der Nutzer definieren kann, ab welcher Ausprägung ein Kapitel dunkelrot eingefärbt wird. Diese Obergrenze wird nun zudem in vier gleichmäßig verteilte Unterbereiche aufgeteilt und den einzelnen Farbwerten zugeordnet. Durch die Anpassung der Einfärbung und die Größe der Kapitel kann der Nutzer nun eine geeignete Strategie bzw. Reihenfolge der Textbearbeitung festlegen. Wurde beispielsweise der Wert zehn als Obergrenze definiert, so zeigen sich folgende Zuordnungen:

Grün:	Werte von 0 bis 2,5
Gelb:	Werte von 2,5 bis 5
Orange:	Werte von 5 bis 7,5
Hellrot:	Werte von 7,5 bis 10
Dunkelrot:	Werte ab 10

Verändert sich jedoch der Kontext durch die Auswahl einer anderen Filterungsmethode, so werden auch andere Ausprägungen betrachtet. Tritt häufig Füllwörter auf, sollte die Anzahl im Verhältnis zum jeweiligen Paragraphen betrachtet werden. Wird hingegen nach der Satzlänge oder der Anzahl von Satzzeichen gesucht, so ist eine Angabe der maximalen Wort- oder Satzzeichenanzahl eines kritischen Satzes sinnvoller. Da die Werte des Sliders somit nicht auf jede Filterungsmethode gleichzeitig abgestimmt werden kann, wurden jeweils Obergrenzen und Inkrementierungsschritte in Form von unterschiedlichen Skalen definiert. Durch diese Vorauswahl wird der Slider nun auf den jeweiligen Kontext angepasst, jedoch weiterhin eine starke Anpassung des Nutzers ermöglicht. Innerhalb einer Recherche wurden leider keine Richtwerte für die einzelnen Kategorien gefunden, sodass vorerst Beispielwerte anhand der vorliegenden Datenstrukturen definiert wurden. In weiteren Schritten sollten hier durch Studien genauere Richtlinien ermittelt und die bisherigen Skalen angepasst werden.

Damit dem Nutzer stets bekannt ist, welcher Farbe welcher Wertebereich zugeordnet ist, sind diese Zusammenhänge auf der linken Seite in Form einer Legende aufgelistet. Wird der Slider bewegt und losgelassen, so werden die neuen Unterabschnitte direkt berechnet und innerhalb der Legende angepasst. Neben der Farbgebung werden in dieser Legende zudem ungewöhnliche

Aspekte dieser Visualisierung aufgegriffen. So findet sich hier beispielsweise eine kurze erklärung der Punkte, welche fehlenden Text darstellen.

Weitere wichtiger Aspekt jeder Visualisierung sind dessen Interaktionsmöglichkeiten. Eine der Interaktionen bildet in dieser Visualisierung das Hervorheben eines Kapitels sobald der Mauszeiger das zugehörige Element betritt. Standardmäßig besitzen die Kapitelelemente eine niedrige Sättigung, welche nun erhöht wird und somit ein klare Trennung von selektierten und nicht selektierten Kapiteln bewirken soll. Hierdurch wird dem Nutzer eine direkte Rückmeldung und dadurch ein Auffordrungscharakter geboten, welcher ihn zu weiteren Interaktionen und Erkundungen animieren soll. Um diesen Effekt zu verstärken wurde an den einzelnen Objekten ein schmaler schwarzer Rahmen angebracht. Auch wenn dieser wiederum Chartjunk darstellt, welcher in dieser Visualisierung weitestgehend umgangen wurde, so sorgt er durch die dunklen Abgrenzungen für eine stärkere Hevorhebung der Auswahl und wurde somit als sinnvoll erachtet.

Durch den Mouse-Over werden zusätzlich die Texte im Zentrum des Sunburst-Diagramms angepasst, welche nun den Titel des Ausgewählten Kapitels, sowie dessen höchste Ausprägung im Bezug auf die Filterungsmethode und den prozentualen Anteil zum Oberkapitel anzeigen. Ist der Titel zu lang für den inneren Bereich des Diagramms, so wird er abgekürzt, dieses jedoch durch die Beifügung dreier Punkte gekennzeichnet. Da dem Nutzer jedoch die Struktur bekannt ist, sollte dieser Ausschnitt für eine Zuordnung ausreichend sein. Somit erhält der Nutzer stets ausreichend Auskunft über dieses Kapitel, was die angesprochenen optischen Verhältnisse komplettiert. Besitzt das ausgewählte Kapitel Unterkapitel, so werden auch diese Hervorgehoben und die Gesamtgröße dieser als Grundlage für die Informationstexte genommen. Auf diese Weise kann eine möglichst flexible Erkundung des Dokuments gewährleistet werden. Verlässt der Mauszeiger das Kapitel, so wird auch die Auswahl rückgängig gemacht.

Hat sich der Nutzer jedoch entschieden ein Kapitel genauer zu betrachten, so kann er es, zusammen allen Unterkapiteln, über einen Linksklick an seine Auswahl binden und den restlichen Visualisierungen zur Verfügung stellen. Verlässt der Mauszeiger nun das Kapitel, so wird dieses nicht mehr Abgewählt. Ist ein Element bereits ausgewählt, so wird dieses durch einen weiteren Klick aus der Auswahl entfernt. Auf diese Weise können nahe gelegene Abschnitte parallel in den anderen Visualisierungen analysiert werden.

Eine parallele Auswahl mehrerer Hauptkapitel wird hierbei jedoch nicht unterstützt. Dies liegt zunächst an der Form der Datenstruktur. Solch eine Auswahl hätte den Nachteil, dass eine Menge Vorberechnungen auf Seiten des Front-Ends durchgeführt werden müssten, was zu ungewollten Ladezeiten führen könnte. Weiterhin ist eine Einzelselektion sinnvoll, da Unterkapitel meist das selbe Thema aufgreifen und beleuchten. Hierbei besteht eine hohe Wahrscheinlichkeit für ungewollte Wiederholungen, welche es nun in dem ganzen Abschnitt zu beseitigen gilt. Wird jedoch ein Unterkapitel eines anderen Oberkapitels ausgewählt, so ist dieser Zusammenhang oft nicht gegeben. Um Verfälschungen oder Verdeckungen von Problemstellen entgegenzuwirken, wird in diesem Falle die vorherige Auswahl verworfen und der angeklickte Abschnitt als neue Auswahl gespeichert. Findet der Klick außerhalb oder im Zentrum des Diagramms statt, so wird die gesamte Auswahl zurückgesetzt. Hierdurch werden dem Nutzer alle Möglichkeiten geboten, welche er für eine erste grobe Einschätzung und Auswahl der Kapitel benötigt.
Möchte der Nutzer nun ein weiters Unterkapitel des gleichen Hauptkapitels (dargestellt durch die innerste Ringschicht) über einen Klick an die Auswahl binden, so wird dieses der bereits getätigten Sammlung beigefügt

Eine letzte Interaktion betrifft die eigentliche Navigation durch das Textdokument. Wird ein Kapitel innerhalb des Diagramms angeklickt, so wird auch der Text-Viewer auf die Auswahl abgestimmt. Da sich das Sunburst-Diagramm auch bei geringer Größe nutzen lässt, kann es zudem als eine Art Inhaltsverzeichnis fungieren, über welches nun das Dokument erkundbar ist. In jedem Falle kann die Kapitelübersicht und somit auch dieses Sunburst-Diagramm als Brückenstück zwischen dem Textdokument und der Visualisierung gesehen werden, welches den Startpunkt jeglicher Interaktion darstellt.

3.2 Kapitelübersicht: Baum-Diagramm

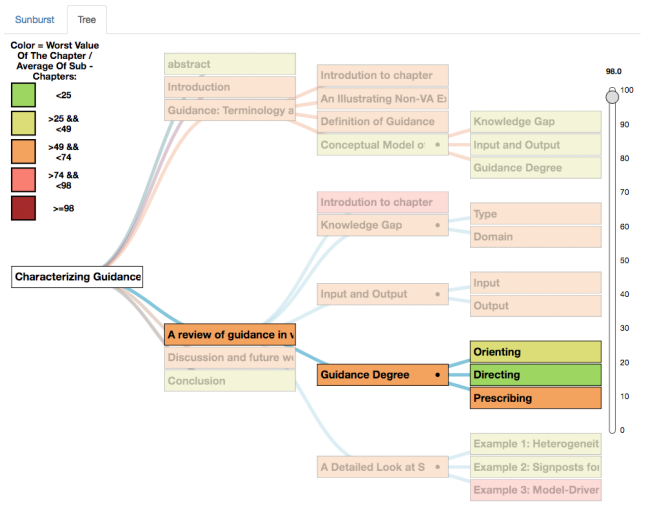

Figure 3: Baum-Diagramm als Kapitelauswahl mit Farblegende und Spezifizierungs-Slider

Ähnlich dem Sunburst-Diagramm steht auch bei dieser Visualisierung die Darstellung der Kapitelhierarchie im Vordergrund. Dem Nutzer soll ein Überblick über das Dokument geboten werden, sodass eine strukturierte Bearbeitung ermöglicht wird. Der größte und maßgebliche Unterschied zur vorherigen Visualisierung findet sich jedoch in der fokussierten Zielgruppe. Das Sunburst-Diagramm ist auf jene Personen ausgelegt, welche ihr eigenes Dokument überarbeiten wollen und sich somit exzellent in diesem Auskennen. Das ist der Grund, warum solch eine zentrierte Form und minimalistische Beschriftung als Ausreichend erachtet wurde. Wird sie jedoch einer Drittperson angeboten, so benötigt diese eine längere Einarbeitungsphase um sich zunächst mit den einzelnen Kapiteln vertraut zu machen. Da jedoch auch diese Personengruppe zu berücksichtigen ist, wurde an einer weiteren Visualisierung gearbeitet, welche nun Zweitprüfern oder anderen Korrekturlesern dienlich ist. Hierbei wurde versucht möglichst gleiche Visualisierungsmittel und Interaktionsmethoden zu verwenden, sodass ein fließender Übergang zwischen den zwei Ansichten stattfinden kann. Hat sich ein Nutzer nun beispielsweise durch die Baumstruktur ausreichend mit den Kapiteln auseinander gesetzt, so kann dieser ohne Probleme zu der Sunburst-Ansicht wechseln. Aufgrund der weitgehenden Übereinstimmungen befasst sich der folgende Abschnitt nun vorwiegend mit den Maßgebenden Änderungen, welche auf diese neue Zielgruppe abgestimmt sind.

Die Baum-Struktur wurde hier gewählt, da sie die Vernetzung der Kapitelstruktur übersichtlich aufgliedert. Ähnlich dem Sunburst spielt auch hier die Position der Knoten eine wichtige Rolle. Diese ist optisch am einfachsten wahrzunehmen und somit der Orientierung des Nutzers sehr dienlich. Dieser Aspekt wird durch die Anzeige der Äste unterstützt. Auch wenn solche Äste einen hohen Anteil des Chartjunks der Visualisierung ausmachen, bieten sie in dem Kontext mehr positive als negative Aspekte. Der unwissende Nutzer findet sich schnell zurecht, da er lediglich den einzelnen Verästelungen folgen muss um das gewünschte Kapitel zu erreichen. Hierdurch bieten sie bereits auf den ersten Blick ausreichend Orientierungspunkte, sodass Zusammenhänge zwischen den Kapiteln schnell erkannt werden können. Um diesen Aspekt weiterhin zu unterstützen, wurden dezente, jedoch klar voneinander trennbare Farben ausgewählt. Diese sind in Form einer ordinalen Skala eingebunden, sodass sie eindeutig einem Hauptkapitel und somit dessen Verästelungen zugeordnet werden können.

Weiterhin wurde ein horizontaler Aufbau der Baumstruktur gewählt, welcher die gewohnte Leserichtung aufgreift und dem Nutzer somit unterbewusst eine Hilfestellung bietet. Ein weiterer maßgeblicher Unterschied findet sich in der Kapitelbeschriftung. Im Sunburst wurde diese lediglich im Zentrum angebracht, da einzelne Ringe unter umständen zu klein für eine ausgiebige Beschriftung sein und die Krümmungen den Lesefluss beeinflussen könnten. Da diese Aspekte jedoch nicht auf die Knoten eines Baumes zutreffen, sind hier die Beschriftungen stets innerhalb der Knoten angebracht. Im Gegensatz zu dem stark zentrierten Aufbau des Sunburst-Diagramms ist eine Baumstruktur von Natur aus breitflächiger. Da die x-Achse bei horizontalen Bäumen deren Tiefe bestimmt und diese in einem wissenschaftlichen Dokument eine maximale Ausprägung von vier oder (in Ausnahmefällen) fünf besitzt, kann den einzelnen Knoten eine ausreichende Breite zur Verfügung gestellt werden, wodurch auch die Titelbeschriftung profitiert. Sollte jedoch der Fall eintreten, dass ein Titel über den Rand des Knotens hinaus ragen würde, so wird dieser mithilfe eines Clip-Path eingeschränkt. Für solche Fälle besteht die Möglichkeit, sich den vollständigen Titel über einen Hover-Effekt anzeigen zu lassen. Hierzu muss die Maus lediglich für kurze Zeit auf dem Element ruhen.

Im Gegensatz zum Sunburst wird die Kapitelgröße in dieser Visualisierung nicht beachtet. Einer Einfindung in die Datenstruktur wird hier der höchste Stellenwert zugesprochen. Die repräsentativste Veranschaulichung der Kapitelgößen wäre anhand der proportionalen Größe oder Position des jeweiligen Knotens. Dies würde jedoch die Kapitelhierarchie oder Anzeige der Titel beeinträchtigen, welche als Hauptorientierungspunkte gelten. Fehlt einem Kapitel jedoch die Einleitung oder allgemein Text, so wird dies auch hier mit einem Kreis markiert. Dieser befindet sich jedoch nicht im Zentrum des Elements, sondern an dessen Ende, sodass auch hierdurch kein Titel beeinträchtigt wird.

Abgesehen von diesen Aspekten finden sich nur geringe Änderungen. Die Farbgebung der Knoten wird anhand der selben Aspekte ermittelt und auch der Slider zur explorativen Spezifikation der Ausprägungen wird beibehalten. Auch die Legende findet sich an der selben Stelle. Lediglich die Auswirkungen der Interaktion mit dem Diagramm wurden geringfügig angepasst. Wird der Mauszeiger über ein Element bewegt, so folgt weiterhin die Anzeige des gefilterten Wertes dieses Kapitels, da er für eine Verfeinerung der Auswahl benötigt wird. Dieser ist hier jedoch an den Hover-Text des jeweiligen Elementes gebunden, welcher zudem den Titel des Kapitels Anzeigt. Das Grössenverhältnis zum jeweiligen Oberkapitel wird hierbei jedoch nicht angebracht, da der Größe hier im Allgemeinen zu Gunsten der Leserlichkeit weniger Beachtung geschenkt wird. Eine letzte Änderung findet sich in der Markierung von Kapiteln. Im Gegensatz zum Sunburst-Diagramm wird hier nicht nur das ausgewählte Element zusammen mit dessen Unterkapiteln hervorgehoben, sondern auch der Pfad vom Root-Element zu diesem Knoten, sowie alle zugehörigen Verästelungen der Auswahl. Dies ist wiederum damit begründet, dass diese Visualisierung vorwiegend von dokumentfremden Personen genutzt wird und ihnen die Auswahl auf diese Weise nachvollziehbarer dargestellt wird.

3.3 Detailansicht: Blasen-Diagramm

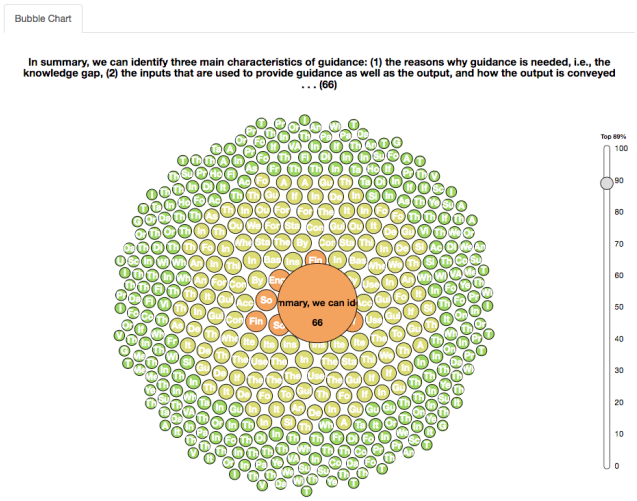

Figure 4: Blasen-Diagramm als Detailansicht (der Mauszeiger befindet sich über dem vergrößerten Element)

Wurde eine Auswahl in der Kapitelübersicht getroffen, werden dem Nutzer nun innerhalb der Detailansicht genauere Informationen geboten. Hier sind nun, in Abhängigkeit der Filterungsmethode, sämtliche Begriffe in Form eines Blasen-Diagramms aufgelistet (siehe Abbildung 4). Wird nach den Redundanzen gesucht, so symbolisiert jede Blase ein Wort, welches redundant in dem ausgewählten Bereich auftritt. Wurden Satzlänge oder Satzzeichen selektiert, so bildet jede Blase einen Satz ab. Die Suche nach Füllworten hebt sich hier jedoch durch dessen Präsentation etwas von den restlichen Filterungen ab. Hierbei steht jede Blase für einen Paragraphen innerhalb der Kapitelselektion. Diese Variante wurde gewählt, da Füllwörte eine eigene Kategorie bilden, welche durch die Masse aller beinhalteter Einzelworte schädlich ist. Sind mehrere Füllwörter häufig verwendet worden, so ist dies ebenso schlimm, wie das Auftreten eines einzelnen Füllwortes in der selben Häufigkeit.

Damit die einzelnen Blasen voneinander unterschieden werden können, wurden diese mit einem Schriftzug des jeweiligen Inhaltes versehen. Dieser wird in Abhängigkeit der Größe und Inhaltslänge jedoch verkürzt, damit er innerhalb der jeweilige Blase dargestellt werden kann. Auf diese Weise wird dem Nutzer eine erste Zugehörigkeit ersichtlich gemacht. Die Inhalte sind innerhalb der einzelnen Kategorien wie folgt aufgebaut:

Redundanz: Das jeweilige Wort gefolgt von der Worthäufigkeit und dem prozentuellen Anteil.

Satzlänge: Der vollständige Satz gefolgt von der Wortanzahl.

Satzzeichen: Der vollständige Satz gefolgt von der Anzahl an Satzzeichen.

Füllwörter: Der Pfad zu dem Paragraphen gefolgt von der Anzahl an Füllwörtern und dem prozentuellen Anteil. ($Kapitel > Unterkapitel > ... > Paragraphennummer + Anzahl$)

Da die Blasengröße jedoch von der Anzahl an Worten abhängt, können diese unter Umständen sehr klein ausfallen, wodurch diese Beschriftung unerkenntlich wäre. Aus diesem Grund wurde weiterhin eine Hover-Animation implementiert. Wird der Mauszeiger über eine Blase bewegt, so erhöht sich dessen Radius und Textgröße auf einen festgelegten Anteil der Visualisierung. Zudem wird diese in den Vordergrund gehoben, sodass Sie nicht von den umliegenden Blasen verdeckt wird und der Text problemlos wahrgenommen werden kann. Der zuvor weiße Text wird hierbei durch eine schwarze Einfärbung weiter hervorgehoben. Bewegt sich der Mauszeiger nun von der Blase weg, so wird auch diese Vergrößerung rückgängig gemacht. Als Ankerpunkt für diesen Effekt wurde jedoch nicht die Blase selbst, sondern der textliche Inhalt ausgewählt. Wäre hier die Blase selbst das Ziel, so wäre eine Betrachtung umliegender Elemente nahezu unmöglich, da diese von der momentanen Auswahl überdeckt sind.

Da bei langen Sätzen hier jedoch nur der Anfang präsentiert werden kann, wurde die Visualisierung um ein Textfeld ergänzt, welches sich direkt über dem Blasen-Diagramm aufstreckt. Innerhalb dieses Textfelds werden nun, parallel zu der Vergrößerung der Blasen, durch die Hover-Bewegung der jeweilige Inhalt weitaus größer und lesbarer präsentiert. Dieses Textfeld wird lediglich durch drei Punkte gekennzeichnet, welche den zunächst fehlenden Inhalt repräsentieren. Von Umrandungen, Hintergrundfarbe oder anderen Hervorhebungsmethoden wurde hier abgesehen, da diese vorwiegend Chartjunk darstellen, und somit der Data-Ink-Ratio schaden würden. Zudem ist das Textfeld auch ohne diese Hervorhebung klar erkennbar, spätestens bei der ersten Interaktion, innerhalb welcher es zudem erst an Bedeutung gewinnt. Dem Textfeld wurde zudem eine feste Gößе zugeordnet. Wäre dies nicht der Fall, so würde sich die Position des Blasen-Diagramm stets in Abhängigkeit der Textlänge verschieben. Wäre das Textfeld unterhalb des Diagramms angeordnet, so könnte der Text aus dem Sichtbereich hinaus ragen. Für den Fall, dass nun ein langer Satz zu groß für dieses Textfeld ist, wird stets die Textlänge überprüft und dieser Satz gegebenenfalls abgeschnitten. Diese Verkürzung wird wiederum mit drei Punkten gekennzeichnet, gefolgt von dem Ausprägungswert in Klammern, sodass dem Nutzer dennoch alle wichtigen Informationen gegeben werden.

Neben dem textuellen Inhalt sind die Blasen zudem durch eine Hintergrundfarbe gekennzeichnet, welche auch hier Aufschluss über die Qualität bzw. Problematik des Inhaltes gibt. Diese Farben richten sich, ebenso wie die der Kapitelauswahl, nach dem Spezifizierten Wert des Nutzers, sodass seine Eintscheidungen und Ansprüche auch hier aufgegriffen werden. Durch die einheitliche Hervorhebung von Eigenschaften wird auch für ungeschulte Nutzer ein möglichst schneller Gebrauch der Anwendung ermöglicht.

Weiterhin werden die Knoten vorsortiert, sodass sich die Begriffe mit dem höchsten Problempotential in der Mitte des Diagramms ansammeln und dieses Risiko bis hin zum äußersten Ring abnimmt. Anhand der Position eines Elements und dessen Farbe kann der Nutzer also gezielt nach Problemstellen suchen. Durch diese Kombination bilden sich zudem unterschiedliche Lagen, deren Breite Aufschluss über die allgemeine Qualität des ausgewählten Dokumentbereichs liefern, sodass der Nutzer bereits auf den ersten Blick Einschätzungen tätigen kann.

Um diesen Aspekt weiter zu fördern, wurde zudem die Göße der Blasen beachtet. Auch wenn mehreren Worten die selbe Farbe zugeordnet wird, so können sich diese dennoch in ihrer Auspr̈agung stark unterscheiden. Um auch diesen Aspekt einzubeziehen, verändert sich die Größe der Elemente nun Proportional zu ihrem Wert. Hierbei wurde bedacht, dass radiale Veränderungen die Wahrnehmung eines Menschen trügen können. Eine Erh¨ung ist von diesem nur schwer zu interpretieren, da der Flächeninhalt eines Kreises hierbei, subjektiv betrachtet, unproportional stark anwächst. Dieser Lügenfaktor wurde jedoch zu Gunsten des Projektes gesehen. Ziel ist die klare Übermittlung der Gefahrenstellen, welche nun weitaus deutlicher wahrgenommen werden. Anhand der Größe, Position und Signalfarbe ist dem Nutzer sofort bekannt, welche Textelemente überarbeitet werden sollten. Um einen Überblick über den gesamten Textabschnitt zu ermöglichen, werden unbedeutende Abschnitte immer noch dargestellt, jedoch nicht so stark hervorgehoben, sodass diese nicht von den wichtigen Abschnitten ablenken. Da diese, nun um einiges kleineren, Abschnitte für den Nutzer somit eine geringere Bedeutung haben, wurde sich zudem dafür entschieden, einen minimalen Radius zu definieren. Liegen Objekte unter einem bestimmten Wert, so wird die Interaktion mit diesen abgestellt. Sie befinden sich durch die Vorsortierung am äußeren Ende der Visualisierung und könnten durch eine Interaktion den Arbeitsfluss des Nutzers mit unnötigen Informationen stören. Diese enthalten nun ebenfalls keinen Text mehr, und sollen lediglich durch ihre Farbe eine allgemeine Einschätzung der Probleme dieses Abschnitts ermöglichen. Durch die so entstandene Hervorhebung wichtiger Aspekte könnte zudem, in Verbindung mit der Redundanzauswahl, ein roter Faden des Dokuments ersichtlich werden. Da keine semantische Analyse stattfindet und dieser Aspekt nicht den Fokus des Projektes darstellt, sollte er jedoch vorwiegend als kleine Hilfestellung gesehen werden.

Wie bereits angedeutet, kann solch eine Visualisierung schnell überfüllt sein. Wurde beispielsweise ein großes Hauptkapitel zusammen mit der Redundanzenfilterung ausgewählt, so befinden sich schnell tausend Kreise in dem Diagramm. Damit dem Nutzer trotzdem eine gezielte Suche ermöglicht ist, wurde auch hier ein Slider integriert, mithilfe dem die Auswahl nun weiterhin verfeinert werden kann. Dieser bildet einen Bereich von 0 bis 100 Prozent ab und ist abhängig von der Anzahl der visualisierten Objekte. Zunächst war hier eine Abhängigkeit von der maximalen Ausprägung angedacht, diese Idee wurde jedoch schnell verworfen da eine gleichmäßige Unterauswahl durch Ausreißer unmöglich gewesen wäre.

Bei dem Aufbau aller Slider dieser Visualisierung wurde die gewohnte Inkrementierungsrichtung bedacht, sodass diese am unteren Ende den Nullpunkt besitzen und sich bei einer Aufwärtsbewegung den Maximalwerten nähern. Als Inkrementierungsschritt wurde sich hier auf den Wert 1 geeinigt, da so eine weitgreifende Unterkategorisierung stattfinden kann. Oberhalb der Skala befindet sich auch bei diesem Slider eine Anzeige der momentanen Prozentzahl, sodass die Achsenbeschriftung auf elf Werte reduziert und somit unnötiger Chartjunk vermieden werden konnte. Nach dem Motto ”overview first” beginnt die Interaktion

des Nutzers mit einer 100%gen Ansicht der Daten, welche dieser jedoch nun mithilfe des Sliders eingrenzen kann.

3.4 Text-Viewer

Um den Zusammenhang zwischen Visualisierung und Text möglichst deutlich zu machen und den Prozess der Iteration zu unterstützen, wurde zusätzlich ein Text Viewer implementiert. So wurden die Daten nach dem Text Mining wieder zusammengesetzt und in HTML mit den gewonnen Informationen versehen. So wurden entsprechende Headline- und DIV-Elemente definiert um die Dokumentstruktur wiederherzustellen. Die vergebenen ID's wurden genutzt um diese über das Document Object Model anzusprechen.

Die Paragraphen und Sätze wurden hier bewusst durch die einzelnen Token rekonstruiert, anstatt eine einfach Ausgabe des ursprünglichen Textes anzuzeigen. Dadurch lassen sich alle Token, Sätze, usw. gezielt ansprechen und können mit Klassen versehen bzw. manipuliert werden. Dadurch lassen sich Inhalte aufgrund verschiedener Eigenschaften hervorheben wie in der folgenden Grafik zu sehen ist.

Figure 5: Text-Viewer (Markierungen verweisen auf Füllwörter in dem selektierten Unterkapitel "Orienting")

Durch diesen Effekt lässt sich eine direkte Brücke von den ermittelten Werten zum Text schlagen, was einen erheblichen Informationsgewinn ermöglicht, welcher mit den Visualisierungen allein nicht möglich ist. Durch diese Struktur lässt sich nicht nur der Text anhand von Token und ID's hervorheben, sondern auch zusätzliche Informationen wie Part-of-Speech Tags können genutzt werden, da diese in entsprechenden Data-Attributen vorhanden sind.

4 Diskussion

Dieser Abschnitt widmet sich der Diskussion von Vor- und Nachteilen, welche sich durch den Gebrauch dieser Visualisierung bzw. dieser Anwendung ergeben. Weiterhin werden hier Designentscheidungen kritisch betrachtet und im Kontext beurteilt.

Die Auswahl der einzelnen Visualisierungen ist ein Punkt der sowohl Vor- als auch Nachteile aufweist. Normalerweise sollte eine Visualisierung für einen bestimmten Anwendungszweck aussagekräftig genug sein und der Nutzer keine zweite Ansicht benötigen. In diesem Projekt werden jedoch zwei Ansichten als Kapitelauswahl angeboten. Ein Punkt der hier jedoch oft nicht beachtet wird ist, dass auch die Benutzergruppen eine Rolle bei der Erstellung solch einer Anwendung spielen. Eine Visualisierung kann nicht auf eine Benutzergruppe abgestimmt werden und dabei allen anderen Nutzern die selben Informationen bieten. Hierbei sind allgemeine Kenntnisse und ein Kontext bezogenes Vorwissen zu berücksichtigen. Dies ist der Grund, warum die Aufspaltung als notwendig und voranbringend angesehen wird.

Auch abseits der Benutzergruppe sind Argumente für eine Doppelansicht zu finden. Wie in Kapitel 3.2 erwähnt, ist in diesem Projekt für den Baum lediglich eine Tiefe von maximal vier bis fünf Ebenen zu erwarten (Dokument / Root, Kapitel, Unterkapitel, Unterunterkapitel und ggf. Unterunterunterkapitel). Die Höhe jedoch wird von der Anzahl an Nachbarkapiteln einer Ebene bestimmt, welche je nach Dokument sehr unterschiedlich ausfallen können. In der Vollansicht stellt dies kein Problem dar, wird diese jedoch verkleinert, so können einzelne Elemente einer breit gefächerten Dokumentstruktur gegebenenfalls nur noch schwer auseinander gehalten werden. Vor allem hier zeigt sich wiederum der Vorteil einer Aufteilung in zwei Visualisierungen. Aufgrund der hohen Data-Density des Sunbursts würde nun diese Ansicht ausgewählt werden um weitere Navigationen zu vollziehen. Der Baum hingegen besitzt eine geringe Datendichte und ist somit für solch eine Navigation weniger geeignet. Der Fokus des Baumes liegt jedoch auch mehr in der strukturierten, als in einer zentrierten Darstellung. Dies ist auch einer der Gründe, warum das Sunburst-Diagramm als Hauptansicht und die Baumstruktur vorwiegend als Alternative gewählt wurde.

Ein weiterer möglicher Kritikpunkt ist die Verwendung von Diagrammen, welche im Alltag eher selten auftreten, wie beispielsweise das Sunburst-Diagramm. Die gezielte Verwendung der Visualisierung kann für bestimmte Nutzergruppen durch eine längere Einarbeitungszeit verzögert werden, sodass diese im schlimmsten Falle eine Einführung benötigen könnten. Auf den ersten Blick ist nicht umbedingt für jeden direkt ersichtlich, wofür die einzelnen Elemente stehen oder wie sie verwendet werden können. Eine mögliche Hilfestellung könnte hier durch eine ausführlichere Legende geboten werden. Diese stellen jedoch für gewöhnlich unnötigen Chartjunk dar, sodass eine aussagekräftige Visualisierung zu bevorzugen ist. Aus diesem Grund wurden Legenden vorwiegend für die Auflistung von Farbwerten verwendet und weniger zur Erläuterung von Interaktionen. Diese Farbwerte visualisieren selber Daten, ohne welche der Nutzen einen weitaus geringeren Informationsgewinn haben würde. Zudem wurde bedacht, dass die Interaktion mit der Visualisierung einen Dominoeffekt aufweist. Auch wenn der Nutzer nicht auf den ersten Blick die Funktionsweise erkennt, so wird diese spätestens durch das Zusammenspiel der einzelnen Elemente deutlich. Auswahlen in einem Diagramm haben direkte Auswirkungen auf die anderen Diagramme, sodass ihre Zusammengehörigkeit schnell ersichtlich wird. Einzeln betrachtet bieten die Visualisierungen bereits einige Informationen, zusammen jedoch bieten sie dem Nutzer ein weites Repertoire an Möglichkeiten.

Unter dem Aspekt der Verständlichkeit sollte auch das Blasen-Diagramm genannt werden. Im Grunde stellt dieses eine Rangliste der gezeigten Begriffe dar. Aufgrund der, in Kapitel 3.3 beschriebenen, Verzerrungsfaktoren werden zu so einem Zweck meist andere Visualisierungen verwendet, über welche die Größenverhältnisse wahrheitsgemäßer interpretiert werden können. Somit war auch für dieses Projekt zunächst ein Balkendiagramm angedacht, da dieses

eine gute Grundlage für vergleichbare Elemente bietet. Im Bezug auf den Kontext dieser Visualisierung wurde sich dann schließlich doch gegen den Gebrauch eines Balkendiagramms und für die Nutzung des Blasen-Diagramms entschieden. Zwei Gründe wurden bereits in Kapitel 3.3 genannt, sodass die verzerrte Wahrnehmung unterschiedlicher Radien zu einer gewollten Hervorhebung der kritischen Bereiche führt und ein roter Faden ersichtlich werden könnte. Einen weiteren Grund stellt hier jedoch die Tatsache dar, das ungewohnte Visualisierung zur Exploration auffordern. Ähnlich zu hellen und verspielten Einfärbungen wird auch hierdurch der Nutzer indirekt aufgefordert die Möglichkeiten der Anwendung auszutesten, sodass der zuvor genannte Dominoeffekt ausgelöst wird. Da ein Fokus dieser Anwendung die explorative Erkundung von Dokumenten darstellt, wurde versucht, diesen Aspekt in der kompletten Visualisierung beizubehalten, was zusätzlich die Wahl des Sunburst-Diagramms unterstützt.

Letztlich bleibt zu erwähnen, dass durch die Nutzung der Anwendung großflächige Überarbeitungen eines Dokumentes in geringer Zeit statt finden können. Anstatt diese vollständig und geradlinig zu durchsuchen, können diese Problemstellen nun gezielt aufgefunden und beseitigt werden. Durch den reduzierten Bereinigungsaufwand von morphologischen Fehlern, kann nun mehr Zeit für die inhaltliche Überarbeitung des Dokumenteninhalts aufgebracht werden. Die unterschiedlichen Arten der Visualisierung übertragen diesen Vorteil auch auf Personen, welche als Korrekturleser eingesetzt werden und sich in dem Dokument nicht auskennen. Gerade solche Personen verbringen einen Großteil der Bearbeitungszeit mit dem Ausbessern von redundanten Fehlern, welche dem Schriftsteller bis dato nicht bekannt waren. Da gerade solche durch diese Visualisierung auffindbar sind, können auch diese Nutzer ihre Zeit auf inhaltliche Verbesserungen fokussieren, wodurch die Qualität des Textes potentiell ansteigt.

Unbedacht genutzt kann dies jedoch auch ein Nachteil sein. Es könnte vorkommen, das sich Korrekturleser ausschließlich auf die so gefundenen Problemstellen konzentrieren und keinen Gesamteindruck erlangen, welcher jedoch für eine endgültige Beurteilung des Dokuments unerlasslich ist. Hier werden keine semantischen Zusammenhänge erfasst, sodass ein Text nach der Korrektur zwar syntaktisch fehlerfrei sein könnte, diesem jedoch vollständig der Inhalt fehlt. Zudem besteht ein wissenschaftliches Dokument zu grßen Teilen aus dem Zusammenspiel von Bildern und Text. Diese Zusammenhänge gehen nur aus einer semantischen Analyse hervor und müssen somit weiterhin manuel durchgeführt werden.

5 Conclusion

Zum Abschluss lässt sich sagen, dass die in der Aufgabendefinition formulierten Ziele gut erreicht worden sind. Durch die Kombination aus Visualisierung und Textdarstellung lassen sich mögliche Probleme durch selbstdefinierte Filter erkennen und bieten einen Mehrwert, welcher beim Verfassen von wissenschaftlichen Arbeiten oder ähnlichen Texten Anwendung finden kann.

Im Verlauf der Projektdurchführung gab es einige Punkte die sich nicht komplett umsetzen lieen und hier nochmal aufgegriffen werden.

Nominalstil

In der ursprünglichen Aufgabendefinition war auch die Markierung von Sätzen, welche eine Tendenz zum Nominalstil aufweisen, geplant. Dazu wurden die Token mit entsprechenden Part-of-Speech Tags versehen, welche sich durch den Text Viewer hervorheben lassen. Entsprechende Kennwerte währen eine gute Erweiterung für das Projekt gewesen, wurden allerdings aus Zeitgründen und dem notwendigen sprachlichen Wissen nicht weiter verfolgt.

Einlesen von Dokumenten

Zum aktuellen Zeitpunkt ist es nur möglich, vorher standardisierte Texte in XML einzulesen, zu analysieren und am Ende zu visualisieren. Ein weiteres mögliches Feature wäre die Implementierung einer automatischen Analyse, welche hochgeladene Dokumente bearbeitet und im Anschluss anzeigt.

Visualisierung des Roten Fadens

Eine weitere berlegung war die Umsetzung einer gezielten Visualisierung des Roten Fadens. Dabei sollten mit Hilfe von Part-of-Speech Tagging die Token entsprechend annotiert und gefiltert werden. Ziel war es eine einfache Alternative zum Topic Modelling zu implementieren, bei dem es darum geht, aus dem Text das unbekannte Thema zu finden. Die POS-Tags sollten dabei primär nach Nomen gefiltert werden, da diese Orte, Personen oder andere Entitäten beschreiben.

Die Häufigkeit dieser annotierten Token sollte als alternative zum Topic Modelling nun die einzelnen Abschnitte des Textes beschreiben. Die Häufigkeiten könnten dann in einer Baumstruktur oder einer anderen hierarchischen Visualisierung dargestellt werden und einen möglichen Roten Faden aufzeigen.

Statistik

Die bereits im Text Mining Kapitel angesprochenen Mittel der deskriptiven Statistik, welche nicht umgesetzt wurden, können einen zusätzlichen Einblick in das Dokument bieten. Dabei sind einfache Visualisierungen wie Boxplots hilfreich, welche ein Verständnis für die Verteilung der Werte bieten könnten und so Ausreier leichter erkennen lieen. Solch eine Visualisierung währe somit eine geeignete Erweiterung der bestehenden Detailansicht.

Texteditor

Neben dem Einlesen von Dokumenten könnte auch eine Erweiterung des Text Viewers implementiert werden. So könnten mögliche Probleme live behoben werden bzw. direktes Feedback zu dem geschriebenen Text vermittelt werden. Auf diese Weise würde ein stetiger Arbeitszyklus entstehen, an dessen Ende das fertige Textdokument steht.

Mehrfachauswahl in Visualisierung

Ein weiterer, wichtiger Punkt wird durch die Möglichkeit abgebildet, die verschiedenen Inhalte zu vergleichen. Aktuell lassen sich zwar Kapitel und deren Unterkapitel auswählen, allerdings können keine zwei (oder mehr) Hauptkapitel gleichzeitig angewhlt werden. Durch eine Mehrfachauswahl und der entsprechenden Berechnung der Daten lassen sich möglicherweise Zusammenhänge zwischen einzelnen Kapiteln darstellen, welche wiederum auf einen roten Faden verweisen könnten.

Zustzliche Filteroptionen

Zusätzliche Filteroptionen können eine noch tiefere Exploration der Daten ermöglichen. Neben einfachen Filtern wie das aktivieren bzw. deaktivieren von Füllwörtern oder Worthäufigkeiten, kann auch ein Filtern nach POS-Tags einen anderen Einblick ermöglichen.

Wichtige Ërgänzungen zu dem Projekt

Da die Zeit begrenzt war und wir uns vorwiegend auf eine Funktionstüchtige Visualisierung konzentriert haben, konnte dem Text Mining leider zu wenig Zeit zugeordnet werden. Aus diesem Grund befinden sich einige kleine Fehler innerhalb der Datenquellen, welche nun anhand der fertigen Visualisierung ersichtlich wurden.

So befinden sich in Oberkapiteln bei der Suche nach Redundanzen alle Begriffe der Kindknoten, jedoch keine zusammengefasste

6 Personal Contribution

Im Allgemeinen fand nahezu täglich eine gemeinsame Bearbeitung des Projektes statt, sodass folgende Matrix die Arbeitsverteilung beschreibt:

	Dennis Dubbert	Sascha Lemke
Text Mining	50%	50%
Visualisierung	50%	50%
Dokumentation	50%	50%

References

[1] R. R. Hajo Hipper. Text mining. *Informatik Spektrum*, 2006.

[2] S. M.Weiss. *FundamentalsofPredicitiveTextMining*. Springer-Verlag London Limited, London, 2010.