Institut für Visualisierung und Interaktive Systeme

Universität Stuttgart
Universitätsstraße 38
D–70569 Stuttgart

Bachelorarbeit

Visuelle Kombination von Lupen zur Textexploration

Ba-Anh Vu

Studiengang:	Softwaretechnik
Prüfer/in:	Prof. Dr. Thomas Ertl
Betreuer/in:	Dipl.-Ling. Florian Heimerl, M. Sc. Markus John
Beginn am:	27. Oktober 2016
Beendet am:	28. April 2017
CR-Nummer:	H.5.2, J.5

Kurzfassung

In Anbetracht des heutigen Informationszeitalters kann es in vielen Bereichen vorkommen, dass zur Wissensgewinnung eine umfangreiche Menge von Dokumenten analysiert werden muss. Solche Dokumentenmengen weisen häufig keine oder eine nur schwache Strukturierung auf, so dass es schwierig ist, hieraus auf effiziente Weise die gewünschten Informationen zu gewinnen. Für diese Problemstellung gibt es bereits Lösungsansätze, die auf Lupen basieren. Mit Hilfe dieser ist es möglich, die Spatialisierung der Dokumente, d. h. deren räumliche Darstellung, auf einfache und effiziente Weise zu explorieren und relevante Dokumente zu identifizieren.

Die meisten lupenbasierten Ansätze, die hierzu existieren, weisen jedoch die Einschränkung auf, dass nur eine Lupe gleichzeitig verwendet werden kann. Hierdurch werden vergleichende Analysen zwischen verschiedenen fokussierten Mengen ausgeschlossen. Deshalb stellt diese Arbeit einen Ansatz vor, wie solche lupenbasierten Ansätze um die Möglichkeit erweitert werden können, mit mehreren Lupen gleichzeitig zu arbeiten. Gemeinsamkeiten und Unterschiede zwischen den fokussierten Mengen sollen dabei visuell hervorgehoben und anschaulich dargestellt werden. In dieser Arbeit stehen hierbei bei vergleichenden Analysen vor allem die enthaltenen Personen und Orte im Vordergrund.

Die Arbeit dokumentiert hierzu die Entwicklung und Erläuterung der Konzepte zur visuellen Kombination der Lupen für die Textexploration. Hierbei werden unter anderem neue Ansichten eingeführt, welche verschiedene Informationen und Details zu den fokussierten Mengen visuell darstellen. Weiterhin werden entsprechende Interaktionsmöglichkeiten vorgestellt, welche die Exploration der Texte unterstützen. Im Rahmen dieser Arbeit wird außerdem eine prototypische Implementierung der vorgestellten Konzepte umgesetzt. Zur besseren Veranschaulichung der enthaltenen Funktionen und dessen Bedienung wird zusätzlich in mehreren Anwendungsszenarien beispielhaft die Exploration von verschiedenen Dokumentensätzen durchgeführt, und es werden in einer anschließenden Diskussion mögliche Erweiterungen und Verbesserungen herausgearbeitet.

Inhaltsverzeichnis

Abbildungsverzeichnis

1 Einleitung

In unserer heutigen Informationsgesellschaft und in Zeiten von „Big Data“ wächst die Menge an anfallenden Daten immer mehr. So kommt es in verschiedenen Bereichen immer öfter vor, dass eine umfangreiche Dokumentenmenge vorliegt, die jedoch keine oder nur eine schwache Strukturierung aufweist. Somit kann es sich als schwierig gestalten, aus dieser Dokumentenmenge die gewünschten, relevanten Informationen zu extrahieren.

An dieser Stelle greifen die Methoden der Informationsvisualisierung: Speziell für die Untersuchung von großen Dokumentenmengen und entsprechenden Dokumentenspatialisierungen gibt es lupenbasierte Werkzeuge, um eine Exploration dieser Dokumentenmengen durchzuführen. Wenn man z. B. eine Recherche in einem bestimmten Themengebiet durchführen möchte, kann es von Vorteil sein, vorher die hierfür wichtigen und relevanten Dokumente aus einer sonst unsortierten Menge herauszufiltern. Mit Hilfe der lupenbasierten Werkzeuge lässt sich dies effizient und flexibel bewerkstelligen. Auch im Allgemeinen eignen sich diese hervorragend zur Textexploration.

Die meisten dieser lupenbasierten Werkzeuge, die zur Textexploration existieren, haben jedoch gemeinsam, dass sie die Nutzerinteraktion auf nur eine Lupe gleichzeitig beschränken. Somit werden vergleichende Analysen zwischen mehreren fokussierten Mengen ausgeschlossen. An dieser Stelle will diese Arbeit anknüpfen: Bestehende lupenbasierte Ansätze zur Textexploration sollen um die Funktion erweitert werden, mit mehreren Lupen gleichzeitig zu arbeiten. Dies ermöglicht vergleichende Analysen, die auf verschiedenen Aspekten der Dokumentenmengen basieren können, wie z. B. den enthaltenen Personen und Orten und deren Verbindungen untereinander. Hierbei sollen diese Relationen zwischen den verschiedenen fokussierten Mengen hervorgehoben und anschaulich dargestellt werden. Somit wäre dem Nutzer erlaubt, effizient und auf eine einfache Weise solche vergleichende Analysen auf gegebenen Dokumentenmengen durchzuführen, wobei der Fokus in dieser Arbeit vor allem auf den enthaltenen Personen und Orten liegen soll.

Hierzu dokumentiert diese Arbeit den Weiterentwicklungsprozess von *DocuCompass*. DocuCompass ist ein Werkzeug, das bereits Funktionen für die lupenbasierte Textexploration anbietet, die mit einer Lupe gleichzeitig arbeiten. Die in dieser Arbeit beinhaltete Weiterentwicklung dessen stellt eine der Möglichkeiten dar, wie man eine visuelle Kombination von Lupen zur Textexploration umsetzen kann (unter Einbeziehung der vorher genannten Aspekte).

Es wird hierbei wie folgt vorgegangen:

Kapitel 2 — Verwandte Arbeiten beleuchtet andere, bereits existierende Arbeiten, die relevant für diese Arbeit sind.

Kapitel 3 — Grundlagen: In diesem Kapitel werden die nötigen theoretischen Grundlagen erarbeitet, auf die diese Arbeit aufbaut.

Kapitel 4 — Konzept: Dieses Kapitel enthält die Beschreibung des Konzepts, auf welchem die Weiterentwicklung von DocuCompass in dieser Arbeit basiert.

Kapitel 5 — Implementierung: Hier werden die Details zur Implementierung des Prototypen erklärt, der im Rahmen dieser Arbeit umgesetzt wurde.

Kapitel 6 — Anwendungsbeispiele und Diskussion: In diesem Kapitel werden beispielhaft Anwendungsfälle beschrieben, um das Verständnis und die Verwendung der entwickelten Funktionen zu erleichtern. Zusätzlich findet hier anhand der Beispiele eine Diskussion des entwickelten Ansatzes statt und es werden eventuelle Verbesserungsmöglichkeiten betrachtet.

Kapitel 7 — Zusammenfassung und Ausblick: Hier erfolgt eine Rekapitulation dieser Arbeit und es erfolgt ein Ausblick auf weitere, potentielle Weiterentwicklungsmöglichkeiten.

2 Verwandte Arbeiten

In diesem Kapitel werden einige ausgewählte Arbeiten beleuchtet, die verwandte oder ähnliche Themengebiete wie das dieser Arbeit abdecken. Sie sind insofern hilfreich für diese Arbeit, weil sie in den jeweiligen Themengebieten ein besseres Verständnis für die relevante Materie geben, und können dadurch in verschiedenen Bereichen als Orientierung dienen und uns zeigen, inwiefern wir uns von diesen Arbeiten abheben können.

Wir gehen dazu in diesem Kapitel wie folgt vor: Zunächst wird für jede Arbeit beschrieben, wovon sie handelt. Anschließend wird der Zusammenhang zu unserer Arbeit dargelegt und inwiefern sie hilfreich für unser Ziel ist — der visuellen Kombination von Lupen zur Textexploration.

2.1 DocuCompass

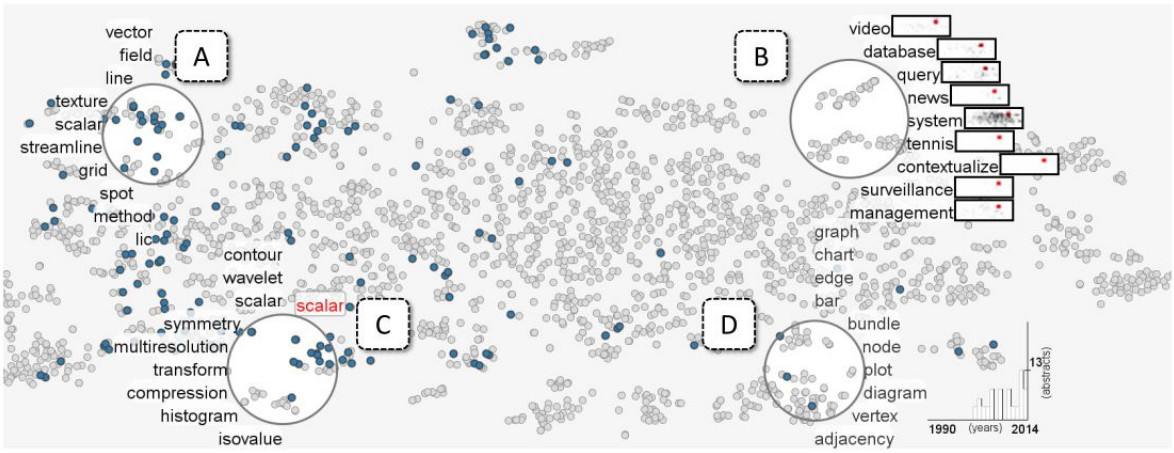


Abbildung 2.1: Ein Beispiel für ein Anwendungsszenario in DocuCompass: Durch die Lupen werden verschiedene Details zu ihrer fokussierten Dokumentenmenge aufgerufen [HJH+ar].

DocuCompass [HJH+ar] ist ein Ansatz, der Funktionen zur lupenbasierten Textexploration anbietet. Hierzu arbeitet dieser auf Spatialisierungen von gegebenen Dokumentenmengen, d. h. deren räumlichen Darstellungen. Diese Darstellungen beinhalten mehrere Glyphen im zweidimensionalen Raum, welche die verschiedenen Dokumente der Dokumentenmenge symbolisieren.

Um diese Spatialisierungen zu explorieren und zu untersuchen, können interaktive Lupen eingesetzt werden. Unter Verwendung des sogenannten „Fokus+Kontext"-Ansatzes können mit den Lupen verschiedene Untermengen durch den Nutzer fokussiert werden und zu diesen Untermengen bestimmte Details angezeigt werden. Hierbei können diese Details aus verschiedenen Informationen bestehen: Ein Überblick über den Inhalt der fokussierten Dokumentenmenge (z. B. häufigste Wörter), der Kontext auf lokaler Ebene (z. B. Verteilung der Wörter über die Dokumente) oder der Kontext auf globaler Ebene (z. B. Dokumente, die ebenfalls ein Wort enthalten), um einige Beispiele zu nennen. Ein Beispiel für ein Anwendungsszenario ist in Abbildung 2.1 zu sehen: Hier sieht man mehrere Lupen, die sich auf einer Dokumentenspatialisierung befinden und neben sich jeweils verschiedene Informationen zu ihrer fokussierten Dokumentenmenge anzeigen.

Mit Hilfe der angebotenen Interaktionsmöglichkeiten und der genannten, anpassbaren Anzeige von zusätzlichen Details kann der Nutzer auf einfache Weise einen Überblick und ein grobes Verständnis der Dokumentememenge oder von bestimmten Untermengen gewinnen. Ebenso wird dem Nutzer ermöglicht, initiale, explorative Analysen der Dokumentenspatialisierungen durchzuführen, um so sein Suchgebiet einzuschränken und die für ihn relevante Untermenge der Dokumente für weitere potentielle Analysen zu finden.

Um den wichtigen Bezug von DocuCompass zu dieser Arbeit darzulegen, fassen wir nochmal kurz dessen vorher beschriebene Merkmale zusammen: Bei DocuCompass handelt es sich um einen Ansatz, womit sich lupenbasierte Textexplorationen durchführen lassen. Dieses ist ein wichtiges Merkmal, was es gemeinsam mit dieser Arbeit hat, denn auch in unserem Fall haben wir das Ziel, lupenbasierte Funktionen bereitzustellen, die Textexplorationen unterstützen sollen. Die Art der verschiedenen Features von DocuCompass kann deshalb auf abstrakter Ebene als Orientierung für unser Ziel dienen: Das ist z. B. die Vielzahl an möglichen Arten von Details der Dokumente und deren Darstellung, die der Nutzer sich zusätzlich neben den Lupen anzeigen lassen kann. Auch die Interaktionsmöglichkeiten mit den Lupen auf lokaler und globaler Ebene können uns als Vorbild dienen.

Es gibt jedoch einen markanten Unterschied: In DocuCompass steht es zwar offen mit mehreren Lupen gleichzeitig zu arbeiten, jedoch haben diese keinerlei Verbindung untereinander, so dass vergleichende Analysen der fokussierten Untermengen nicht möglich sind. Diese Kombination der Lupen soll in unserem Fall erarbeitet und entwickelt werden und ist ein zentrales Merkmal dieser Arbeit. Obwohl diese Funktionalität in DocuCompass fehlt, kann DocuCompass trotzdem auch hier als Vorbild und Orientierung dienen: Denn auch bei der visuellen Kombination der Lupen können die Ansätze zur Interaktion mit den Lupen und den zusätzlichen angezeigten Informationen zu den Dokumenten hilfreich sein, wenn auch diese Ansätze nur bis zu einem gewissen Maße und angepasst auf unseren Fall anwendbar sind (mehr dazu in den Kapiteln 4 und 5).

Neben diesen eher theoretischen Gründen, weshalb DocuCompass relevant für diese Arbeit ist, gibt es auch einige praktische Gründe: DocuCompass dient in unserem Rahmen als Basis für die Entwicklung und Implementierung der Zielfunktionen. Hierfür sprechen weitere Funktionen, die DocuCompass besitzt:

1. Das Einlesen von Datensätzen und die anschließende Spatialisierung, wobei hier verschiedene Dokumententypen und auch verschiedene Spatialisierungsarten und dazugehörige Charakterisierungen der Dokumente möglich sind

2. Grundlegende Funktionen der Lupen, wie z. B. die Fokussierung von Dokumentenuntermengen oder die Interaktionsmöglichkeiten, wozu das Verschieben, Vergrößern, Verkleinern der Lupe usw. dazugehören

Diese Basisfunktionen hat DocuCompass mit dieser Arbeit gemeinsam, weshalb es sich anbietet DocuCompass als Fundament für den weiteren Verlauf der Arbeit zu verwenden (mehr dazu in den Kapiteln 4 und 5). In diesem Rahmen werden die Software-Komponenten von DocuCompass freundlicherweise durch die Betreuer dieser Arbeit zur Verfügung gestellt.

2.2 TrajectoryLenses

TrajectoryLenses [KTW+13] ist ein Ansatz, der Verwendung bei der Untersuchung von spatiotemporalen Bewegungsbahnen auf geopraphischen Karten findet. Hierbei sind die Daten der Bewegungsbahnen meist so umfangreich und unübersichtlich, dass herkömmliche Methoden diese effizient zu analysieren, nicht ausreichen. TrajectoryLenses löst dieses Problem, indem es interaktive Lupen verwendet, um komplexe Filterfunktionen umzusetzen, die für die Analyse der Daten eingesetzt werden können.

Hierzu kann der Nutzer mehrere dieser Lupen auf der geographischen Karte platzieren und diese mit mengenbasierten Operationen zu bestimmten Gruppen kombinieren, um so „maßgeschneiderte“ Filter für die Bewegungsbahnen zu erstellen. Hierbei handelt es sich um Filter für den Ursprung, das Ziel und den Zwischenwegpunkten der Bewegungsbahnen. Diese Lupen können nach Bedarf vergrößert, verkleinert und verschoben werden, wobei die Bedienbarkeit bei diesen Interaktionen durch eine performante Implementierung sichergestellt wird. Ein Beispiel für eine solche beschriebene Kombination von Lupen ist in Abbildung 2.2 [KTW+13] zu sehen. Dort erkennt man die Lupengruppierungen, welche farblich an den Lupenrändern gekennzeichnet sind, und wie sie die entsprechenden Bewegungsbahnen fokussieren.

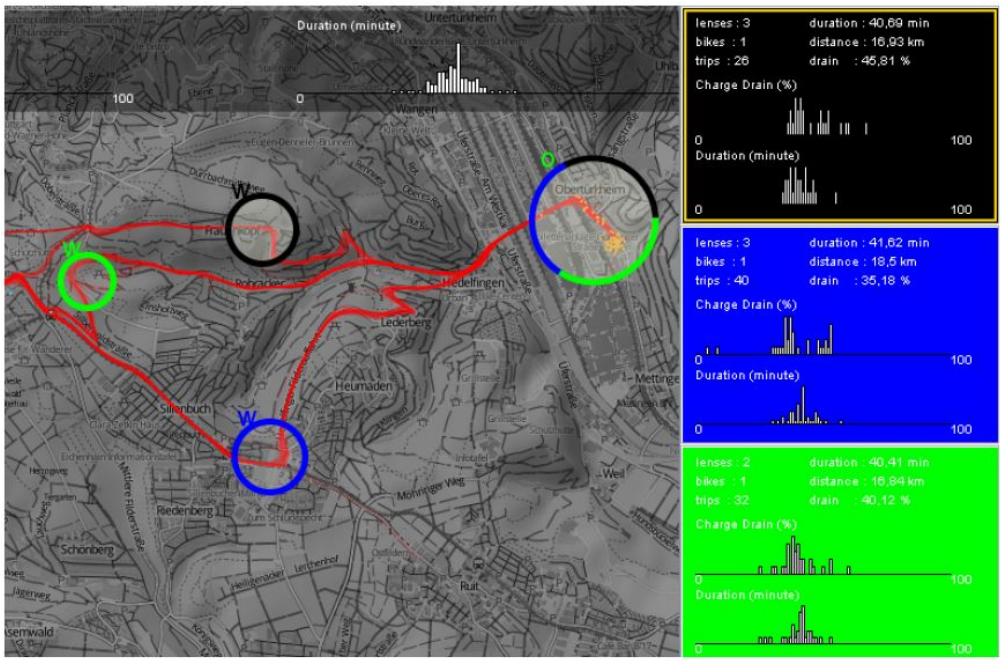


Abbildung 2.2: Ein Beispiel, wie Lupen kombiniert eingesetzt werden können, um Daten zu filtern: Die Lupengruppierungen sind durch eingefärbte Lupenränder gekennzeichnet und heben die entsprechenden Bewegungsbahnen hervor [KTW+13].

Mit Hilfe dieser Filterung kann der Nutzer zusätzliche Informationen zu den fokussierten Bewegungsbahnen aufrufen, welche aus deren verschiedenen Attributen bestehen können (z. B. Länge, Zeitraum, Anzahl) oder domänenspezifisch sein können. Um mit der Fülle der aggregierten Daten nicht die Übersichtlichkeit negativ zu beeinflussen, können diese komprimiert in kompakten Diagrammen neben den Lupen angezeigt werden. Ebenso muss der Nutzer bei der Exploration mit den Lupen seine Aufmerksamkeit nicht zweiteilen, da die Informationen in unmittelbarer Nähe zu den Lupen angezeigt werden. Um zu vermeiden, dass die fokussierte Menge ihrem (semantischen) Kontext entrissen wird, werden auch diejenigen Bewegungsbahnen angezeigt, die alle Filterbedingungen bis auf eine erfüllen, nämlich die der jeweiligen Lupe. Diese werden entsprechend hervorgehoben. So wird es dem Nutzer erleichtert, den lokalen und globalen Kontext der fokussierten Menge zu erfassen, um so leichter bestimmte Informationen zu finden, die noch nicht abgedeckt sind. In Abbildung 2.3 [KTW+13] ist beispielhaft dargestellt, wie die gefilterten Informationen neben den Lupen angezeigt werden: Neben dem kompakten Diagramm, das Informationen zu den fokussierten Bewegungsbahnen enthält, werden auch die Endpunkte der Bewegungsbahnen aus dem dazugehörigen Kontext wie vorher beschrieben hervorgehoben, gekennzeichnet durch cyan-blaue Markierungen.

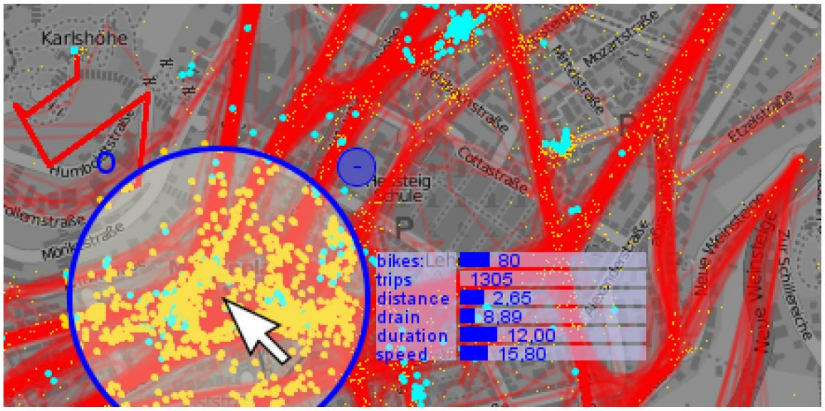


Abbildung 2.3: Ein Beispiel für die Darstellung von Informationen neben den Lupen: Ein kompaktes Diagramm neben der Lupe zeigt Informationen zu der fokussierten Menge an. Cyan-blaue Markierungen helfen außerdem den Kontext der Bewegungsbahnen zu erkennen [KTW+13].

Ebenso wie DocuCompass (Abschnitt 2.1) behandelt TrajectoryLenses eine ähnliche Problemstellung wie unsere Arbeit: Mittels lupenbasierter Hilfsmittel sollen räumliche Darstellungen von Daten exploriert und analysiert werden. In diesem Fall handelt es sich aber um eine andere Art von Daten, nämlich spatiotemporalen Verläufen von Bewegungen, weshalb die Funktionsweise der Lupen sich nicht pauschal auf unseren Fall anwenden lässt. Jedoch ist der verwendete Lösungsansatz trotzdem interessant für uns, denn verschiedene Designprinzipien von TrajectoryLenses können einen Anstoß geben, wie die Zielfunktionalitäten in unserer Arbeit sinnvoll gestaltet werden können.

Hierbei spielt vor allem die Verwendung von Lupengruppierungen eine wichtige Rolle, da auch in dieser Arbeit die Kombination von Lupen Hilfsmittel sein soll, um Explorationen zu unterstützen, in unserem Fall Textexplorationen. TrajectoryLenses kann Antworten auf die Fragen liefern, wie man diese Lupen miteinander verbinden kann und diese miteinander zusammenarbeiten sollen. Ein wichtiger Aspekt hierbei ist eine leichte und schnelle Handhabung der Lupen, so dass der Nutzer sich möglichst wenig mit dem Werkzeug auseinandersetzen muss, sondern sich voll und ganz auf die Exploration und Analyse der eigentlichen Daten konzentieren kann. Um die Intuitivität bei der Benutzung sicherzustellen, werden bei TrajectoryLenses zum einen „logische“ Stützen verwendet, wie z. B. die Einteilung in einfache Gruppen, ohne dass der Nutzer sich mit der dahinterliegenden Theorie der Mengenlehre auseinandersetzen muss. Zum anderen gibt es auch visuelle Stützen, wie z. B. die Einfärbung der Lupen entsprechend ihrer Gruppen und die entsprechende Hervorhebung derer fokussierten Elemente. Hinzu kommt eine performante Umsetzung der Funktionen, um so eine flüssige Interaktion sicherzustellen, wie bereits weiter oben beschrieben. Diese Art von Stützen soll es auch bei uns geben.

Neben der beschriebenen Umsetzung und Handhabung der Lupenkombinationen gibt es noch Aspekte bei der anschließenden Filterung und Darstellung der entsprechenden Daten, die man beachten sollte. Wie bereits erwähnt unterscheidet sich die Art der Daten in TrajectoryLenses von der in unserer Arbeit, jedoch kann man sich auch hier an bestimmten Designprinzipien von TrajectoryLenses

orientieren. Hierbei geht es sowohl um die Daten der fokussierten Elemente selbst als auch um entsprechende Kontextinformationen. Eine gebündelte Visualisierung an einer sinnvollen Position, um eine flüssige Exploration nicht beeinträchtigen, so wie dies TrajectoryLenses mit kompakten Diagrammen bewerkstelligt, ist auch bei uns wichtig. Ebenso wichtig ist eine sinnvolle Auswahl und Visualisierung der Kontextinformationen, um so die Exploration mit den Lupen zu erleichtern, wie weiter oben bereits beschrieben.

Diese genannten Aspekte können und sollen ebenso in unsere Arbeit einfließen, um die Zielfunktionen optimal zu gestalten.

2.3 Jigsaw

Jigsaw [GLP+07] ist wie die bisher vorgestellten Arbeiten ein Ansatz, der zur Analyse von bestimmten Daten dient. In diesem Fall dienen als Daten Mengen von Dokumenten. Jigsaw soll hierbei ermöglichen, leichter solche Dokumentenmengen herauszufiltern, die eine enge Verbindung miteinander haben. So soll eine investigative Analyse der Dokumente erleichtert werden.

Dieses Ziel wird in Jigsaw mit Hilfe von Diagrammen und anderen Visualisierungen erreicht, den sog. Ansichten. Jigsaw extrahiert hierzu vorhandene Entitäten innerhalb der Dokumente und untersucht sie unter Einbeziehung verschiedener Aspekte. Daraus werden die genannten Ansichten erstellt, welche die verschiedenen Aspekte der Entitäten und deren Relation untereinander aufzeigen. Hierbei wird in den Ansichten zwar nur eine Teilmenge der Daten angezeigt, jedoch kann der Nutzer diese mit Hilfe von anpassbaren Anfragen einstellen und erweitern, so dass die Daten auf einfache Weise durchforstet werden können.

Weiteres Merkmal von Jigsaw ist die Verbindung der beschriebenen Ansichten untereinander. Diese Verbindung bewirkt, dass die Interaktion mit einer Ansicht sich direkt auf die anderen Ansichten auswirkt, so dass der Nutzer immer ein konsistentes Bild der Daten vorliegen hat. Zusätzlich kann der Nutzer auch mehrere Instanzen derselben Ansicht erzeugen und von den anderen Ansichten „abkoppeln“, so dass diese Ansicht unabhängig von den anderen Ansichten und den Interaktionen mit ihnen ist — so kann der Nutzer einen bestimmten Ausschnitt der Daten „festhalten“, falls er es wünscht.

In Jigsaw werden folgende Arten von Ansichten angeboten:

- Listenansicht: Hier werden listenartig die vorhandenen Entitäten angezeigt und zusätzlich Verbindungen farbig hervorgehoben.
- Graphansicht: In dieser Ansicht werden die Beziehungsverflechtungen der Entitäten in einem Graph dargestellt. Teile des Graphen können auch schrittweise aus- und eingeblendet werden, um einfacher einen Überblick zu erhalten bzw. Details einzusehen.
- Punktwolkenansicht: Hier werden Beziehungspaare von zwei gewählten Entitätstypen angezeigt, wobei die Größe der verwendeten Entitätenmengen eingestellt werden kann.
- Textansicht: Hier werden die Textstellen der Dokumente angezeigt und vorkommende Entitäten hervorgehoben.

- Zeitverlaufsansicht: Auf einer Zeitachse werden zeitbehaftete Entitäten markiert, wobei die Zeiträume zoombar sind.
- Kalendaransicht: Diese Ansicht gibt einen Überblick über die Dokumente und deren enthaltenen Entitäten, welche nach dem Veröffentlichungsdatum sortiert sind.

In Abbildung 2.4 [GLP+07] ist ein Ausschnitt von Jigsaw zu sehen: Dort sind die Listen-, Text-, Graph- und Kalendaransicht von Jigsaw dargestellt, welche dem Nutzer verschiedene Einblicke in die Daten ermöglichen.

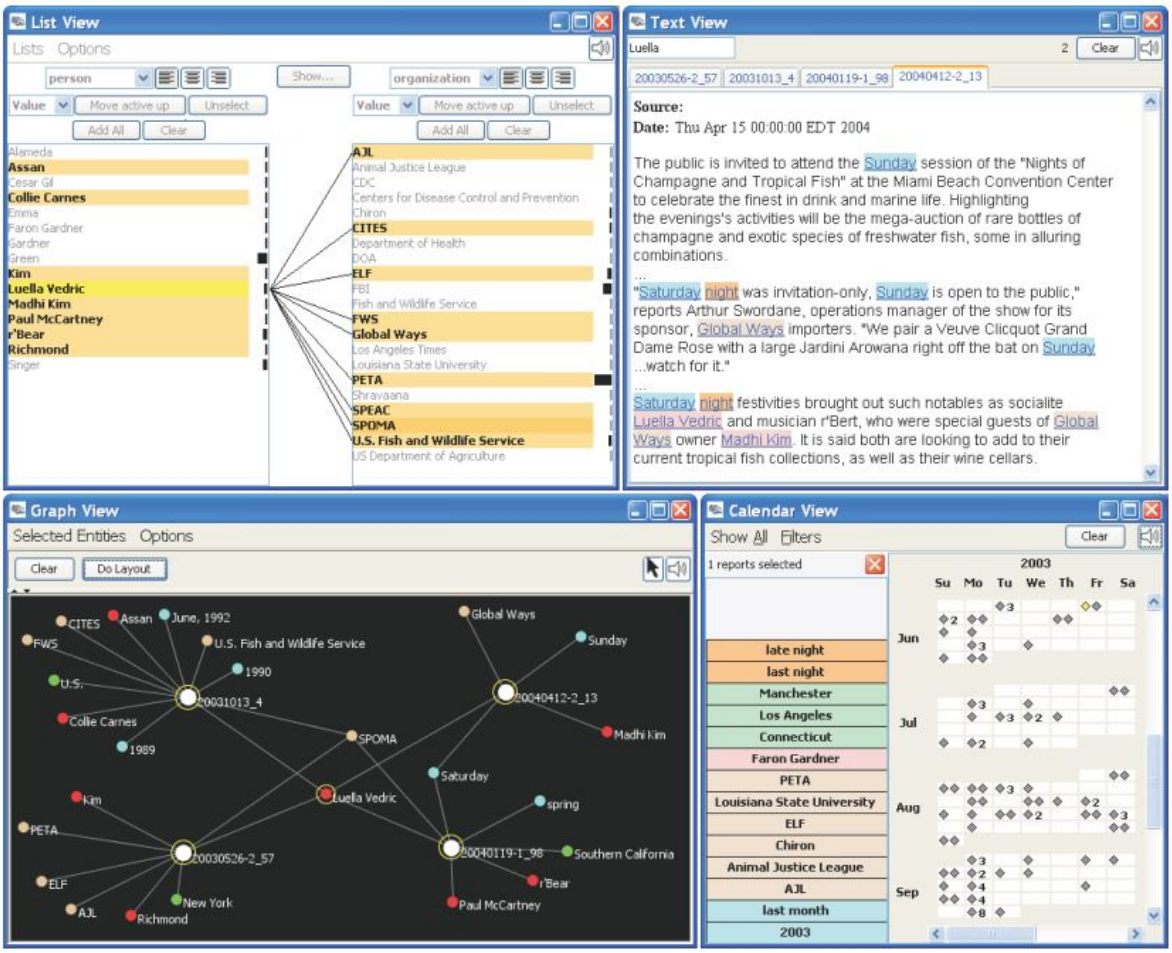


Abbildung 2.4: Ausschnitt aus Jigsaw: Zu sehen sind die Listen-, Text-, Graph- und Kalendaransicht von Jigsaw, welche dem Nutzer verschiedene Einblicke in die Daten ermöglichen [GLP+07].

Eine wichtige Eigenschaft von Jigsaw, welches eine Gemeinsamkeit zu unserer Arbeit darstellt, ist die Unterstützung zur Analyse von Dokumenten. Haupthilfsmittel in Jigsaw sind hierbei die genannten

Ansichten, welche wie beschrieben verschiedene Aspekte von Entitäten und deren Relationen untereinander darstellen. Solche Arten von Visualisierungen soll es ebenso in unserer Arbeit geben, wenn auch der Kontext nicht identisch ist — denn es fehlt hier die Betrachtung von mehreren Teilmengen von Daten, welche bei uns durch die Lupen eingeteilt werden. Auch die Darstellungsformen sind nicht pauschal übertragbar, denn sie müssen kompatibel mit unserer Dokumentenspatialisierung sein und intuitiv zusammen mit dieser bedienbar sein.

Dennoch helfen solche Ansichten in angepasster Form dabei, vergleichende Analysen zwischen den durch die Lupen fokussierten Mengen zu ermöglichen. Vor allem der Aspekt der Extraktion der vorkommenden Entitäten mit ihren Eigenschaften und Beziehungen soll uns als Vorbild dienen und in solche Vergleiche einfließen. Doch auch ohne den Kontext der Entitäten kann Jigsaw einen Anstoß geben, auf sinnvolle Weise grafische und textuelle Elemente zu kombinieren, um so anschauliche Darstellungen der Informationen der Dokumente anzubieten, so wie Jigsaw dies mit seinen verschiedenen Ansichten bewerkstelligt.

2.4 Weitere verwandte Arbeiten

In diesem Abschnitt werden einige weitere Arbeiten genannt, die ebenfalls mit dieser Arbeit verwandt sind, wobei ihr Zusammenhang jedoch nicht so ausgeprägt ist wie das der Arbeiten aus den vorigen Abschnitten. Trotzdem können uns diese Arbeiten als Orientierung dienen im Bezug auf verschiedene Aspekte, die sie beinhalten. Deshalb werden diese im Folgenden kurz beschrieben und ihre Bedeutung für uns aufgezeigt.

2.4.1 Word Cloud Explorer

Der Word Cloud Explorer [HLLE14] stellt einen Ansatz dar, womit sich Textanalysen durchführen lassen. Hierbei steht die Verwendung von sogenannten Word Clouds im Mittelpunkt. Word Clouds sind eine Visualisierungsmöglichkeit, um relevante Begriffe eines Textes anschaulich darzustellen. Beim Word Cloud Explorer wird diese Art der Visualisierung mit verschiedenen anderen Hilfsmitteln verbunden: Hierzu zählen Methoden der Computerlinguistik und verschiedene Filter- und Interaktionsmöglichkeiten. Diese Gesamtheit der Möglichkeiten in Kombination mit den Word Clouds ergeben einen effektiven Ansatz, um Analysen von Texten durchzuführen.

In Abbildung 2.5 [HLLE14] ist ein Ausschnitt aus dem Word Cloud Explorer zu sehen, einschließlich der Bedienelemente für die zuvor erwähnten Hilfsmittel, die Word Cloud Explorer in Verbindung mit den Word Clouds anbietet.

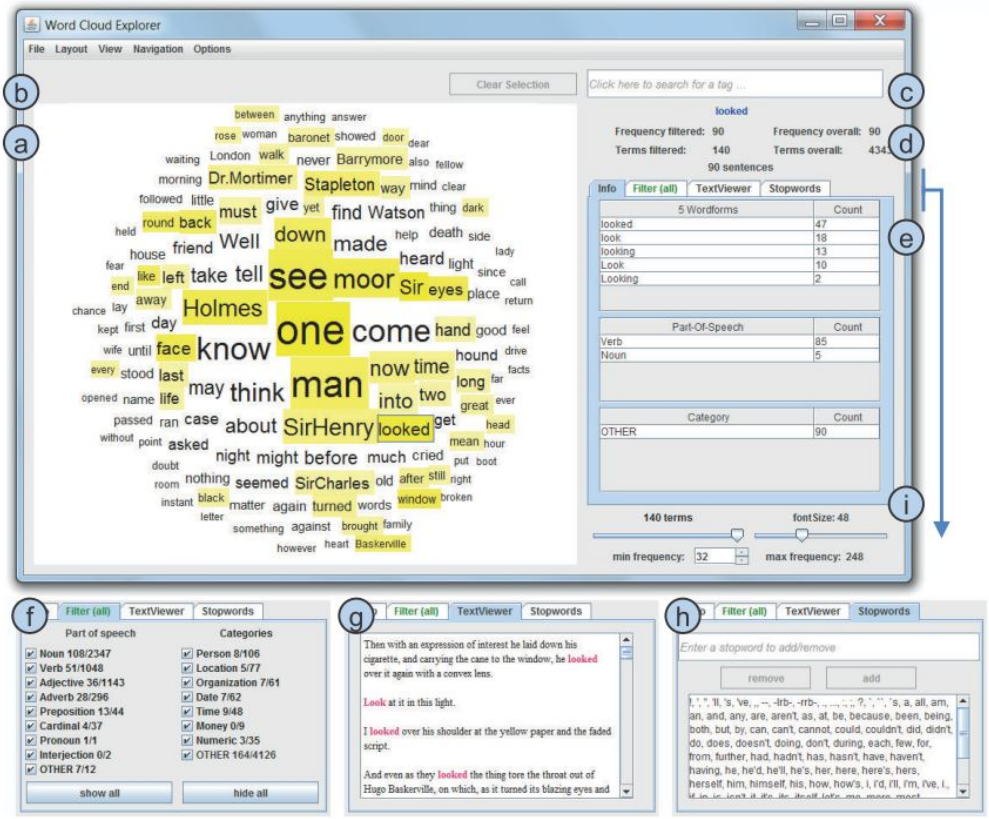


Abbildung 2.5: Ausschnitt aus dem Word Cloud Explorer: Dargestellt ist die Word Cloud, inklusive diverser Zusatzhilfsmittel, um Textanalysen zu unterstützen. Hierzu gehören Filter- und Suchfunktionen, Statistiken- und Informationsanzeigen, und Einstellungsmöglichkeiten für die Word Cloud [HLLE14].

Die Gemeinsamkeit vom Word Cloud Explorer und unserer Arbeit ist insofern offensichtlich, dass der Word Cloud Explorer ebenfalls ein Ansatz zur Analyse von Texten ist. Hierbei unterscheiden sich jedoch die verwendeten Hilfsmittel: Während wir mit Lupen arbeiten, setzt Word Cloud Explorer auf die Verwendung von Word Clouds als hauptsächliches Hilfsmittel.

Trotzdem kann uns Word Cloud Explorer als Orientierung dienen, denn für ausgewählte Bereiche kann es sich unter Umständen anbieten, ebenfalls Word Clouds als effektive Visualisierungsart von Informationen zu verwenden. Ebenso nützlich wäre eine ähnliche Miteinbeziehung von Interaktions- und Filtermöglichkeiten, so wie Word Cloud Explorer dies bewerkstelligt.

2.4.2 Phrase Nets

Phrase Nets [VHWV09] stellen ähnlich wie Word Clouds [HLLE14] eine Visualisierungsart für eine Menge von Wörtern dar. Hierbei spielen nicht nur die Darstellung der Wörter selbst und ihre Platzie-

rung eine Rolle, sondern auch deren Verbindungen, die sie zueinander haben. Die Bedingungen für die Verbindungen können bei Phrase Nets unterschiedlich sein, beispielsweise können sie auf einfacheren Mustern (englisch: pattern matching) aufbauen oder kompliziertere, syntaktische Analysen verwenden.

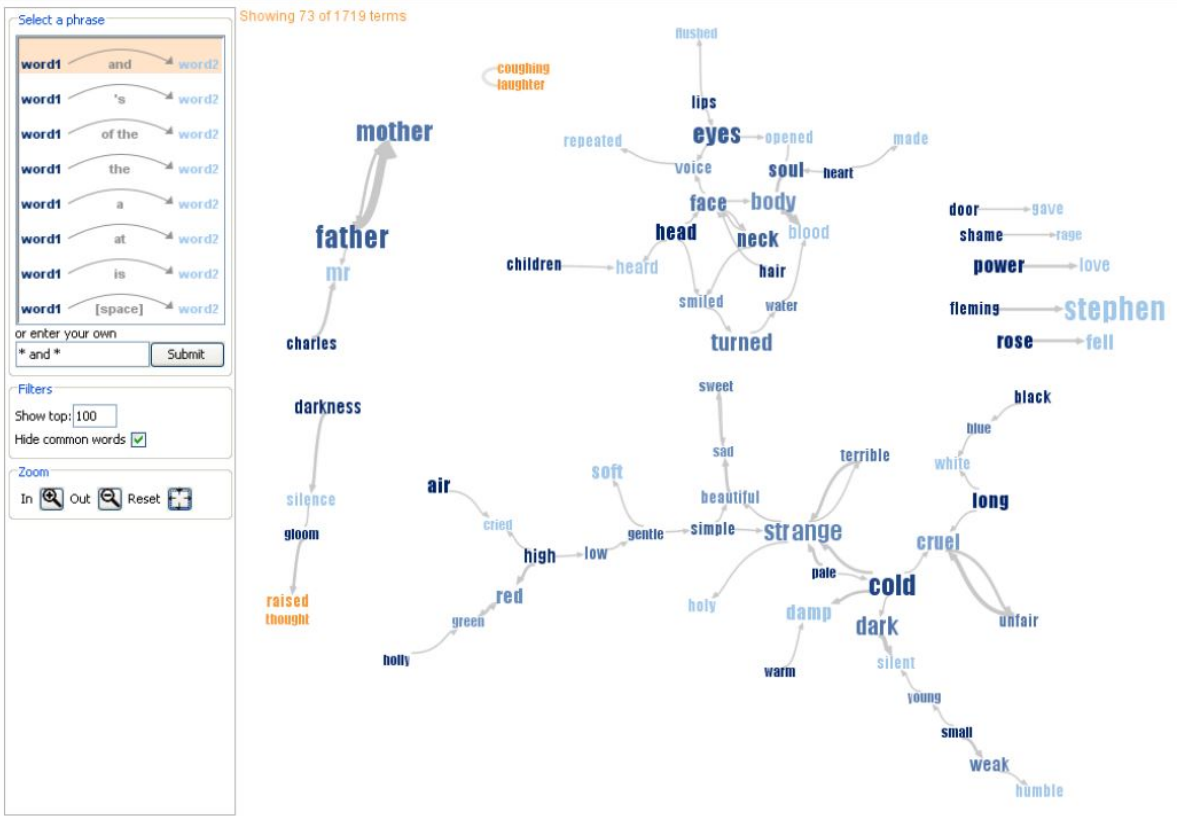


Abbildung 2.6: Beispiel aus Phrase Nets: Dargestellt ist ein Phrase Net für den Roman „Ein Porträt des Künstlers als junger Mann“. Links ist die Auswahl für die Art der Verbindungen zu sehen, rechts befindet sich das zugehörige Phrase Net [VHWV09].

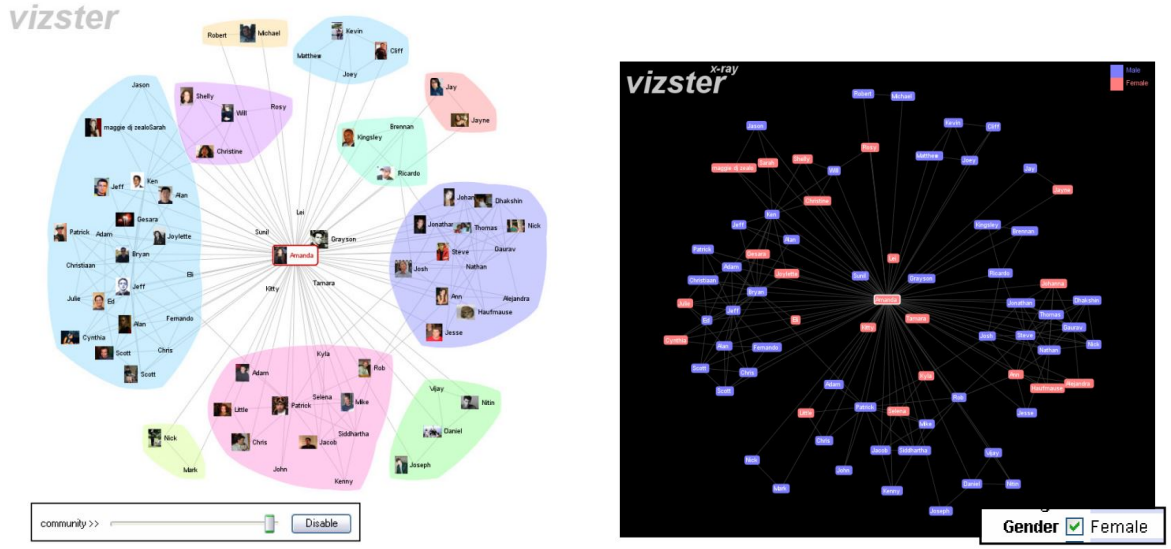
In Abbildung 2.6 [VHWV09] ist ein Beispiel für ein Phrase Net dargestellt: Zu sehen ist das Phrase Net für den Roman „Ein Porträt des Künstlers als junger Mann“ von James Joyce. Auf der linken Seite befindet sich das Bedienelement, in dem man die Art der Verbindung auswählen kann. Das entsprechende Phrase Net ist auf der rechten Seite zu sehen, dass die Wörter mit den zugehörigen Verbindungen enthält.

[VHWV09] sagt zwar nicht, dass Phrase Nets zur Verwendung in ausführlichen Textanalysen dient, sondern zur Überblicksbeschaffung bei unstrukturierten Texten. Dennoch kann Phrase Net einen Anstoß geben, wie man Informationen anschaulich darstellt. Vor allem die Darstellungen der Verbindungen steht hier im Vordergrund, da es auch in dieser Arbeit nützlich sein kann, Verbindungen zwischen bestimmten Elementen darzustellen, um so dem Nutzer einen tieferen Einblick in die Doku-

mente zu geben. Auch die damit verbundenen (syntaktischen) Analysen der Texte können für uns relevant sein, denn damit lassen sich nützliche Informationen herauskristallisieren, die als Grundlage der angebotenen Visualisierungen dienen können.

2.4.3 Vizster

Vizster [HB05] ist ein Ansatz zur Exploration von sozialen Netzwerken. Hierbei benutzt Vizster die Visualisierung der Freundschaftsnetzwerke per Node-Link-Diagramm und einige zugehörige Interaktionsmöglichkeiten, um so Analysen zu Personen, Beziehungen und Gemeinschaftskreisen innerhalb der Netzwerke zu ermöglichen. Zu diesen Interaktionsmöglichkeiten gehören beispielsweise das Selektieren von mehreren Personen und das Anzeigen von dazugehörigen Informationen, oder eine Suchfunktion. Auch sehr nützlich ist die optionale Visualisierung einer Vielzahl verschiedener Aspekte der Personen bzw. Personennetzwerke, wie z. B. das Geschlecht, die Anzahl der Freunde, direkte Freundschaftensverbindungen, Beziehungscluster usw.



(a) Hervorheben von Beziehungsclustern

(b) Farbige Markierung der Geschlechter

Abbildung 2.7: Beispiel aus Vizster: Neben dem eigentlichen Personennetzwerk sind außerdem zusätzliche Informationen durch die entsprechenden Visualisierungsdetails dargestellt [HB05].

In Abbildung 2.7 [HB05] ist dargestellt, wie die Verwendung dieser Funktionalitäten aussehen kann: Neben dem eigentlichen Personennetzwerk zeigt Abbildung 2.7a auch, wie die Beziehungscluster farbig hervorgehoben werden können. Abbildung 2.7b zeigt außerdem die farbliche Markierung der Geschlechter der Personen.

Auch wenn Vizster keinen direkten Bezug zu der Analyse von Texten besitzt, kann die Arbeit trotzdem für uns hilfreich sein. Denn in Dokumenten, in denen Personen und deren Beziehungen im Mittelpunkt stehen, kann es für den Nutzer hilfreich sein, auf entsprechende Visualisierungen zuzugreifen, wie

Vizster diese anbietet. Ebenso nützlich können die Optionen zur Visualisierung von verschiedenen Aspekten der Personen und Beziehungen sein, wobei hier differenziert werden muss, denn zu Personen aus Dokumenten existiert nicht zwingenderweise ein Profil eines sozialen Netzwerks, in denen Informationen abrufbar sind. Dennoch kann uns Vizster in den genannten Bereichen als Vorbild dienen.

3 Grundlagen

In dieser Arbeit spielt das Fachwissen aus mehreren Themenbereichen eine wichtige Rolle. Hierzu gehören sowohl der Bereich der Visualisierung als auch der Computerlinguistik. Da die weiteren Schritte dieser Arbeit auf verschiedenen Aspekten dieser Themengebiete basieren, sollen in diesem Kapitel einige Grundlagen daraus erläutert werden, um so das Verständnis der restlichen Arbeit zu erleichtern. Im Folgenden werden also einige ausgewählte Aspekte aus diesen Gebieten vermittelt und erläutert.

3.1 Visualisierung

Das Ziel dieser Arbeit ist es, eine Möglichkeit anzubieten, um die Exploration von großen Dokumentenmengen zu vereinfachen. Hierbei soll ein zentrales Hilfsmittel die Verwendung von Lupen sein, mit entsprechenden Dokumentenspatialisierungen als Grundlage. Diese Hilfsmittel basieren auf verschiedenen Techniken und Prinzipien der Visualisierung, welche in diesem Abschnitt erläutert werden sollen.

3.1.1 Das Visualisierungsreferenzmodell

Besondere Bedeutung hat in dieser Arbeit das Visualisierungsreferenzmodell (englisch: Visualization Reference Model) von Card et al. [CMS99]. Es beschreibt wie die Visualisierung von Informationen als „Mapping“ von Daten zu visuellen Darstellungsformen betrachtet werden kann, mit ihren einzelnen Zwischenschritten. Hierbei ist ein besonderes Merkmal die Möglichkeit per Interaktion des Nutzers in die einzelnen Zwischenschritte einzugreifen, um so gewünschte Anpassungen zu erreichen.

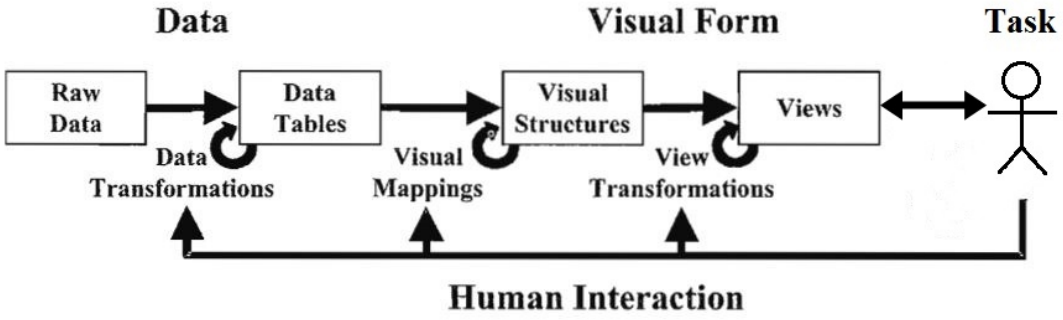


Abbildung 3.1: Schematische Darstellung des Visualisierungsreferenzmodells: Verschiedene Zwischenschritte führen von rohen Daten zu verwertbaren Visualisierungen, welche der Nutzer für Analysen benutzen kann. Besonderes Merkmal ist die Interaktionsmöglichkeit für den Nutzer, um so die Zwischenschritte zu beeinflussen [CMS99].

In Abbildung 3.1 ist das Visualisierungsreferenzmodell schematisch als Diagramm dargestellt. Der Knoten „Raw Data“ des Diagramms stellt die unbehandelte Datenmenge dar, welche typischerweise in ihrer unbelassenen Form idiosynkratisch für die weiteren Schritte ist. Deshalb findet zunächst eine „Data Transformation“ der Daten statt, welche beispielsweise aus einer Aggregation oder Filterung der Daten bestehen kann. Diese Daten werden durch die Transformation auf eine einheitliche Form, den Datentabellen (englisch: Data Tables), gebracht, welche durch den nächsten Knoten im Diagramm symbolisiert wird.

Im nun nächsten Schritt werden die Daten der Datentabellen auf den nächsten Knoten des Diagramms, den visuellen Strukturen (englisch: Visual Structures) gemappt. Hierbei werden bestimmte Werte der Datentabellen verwendet, um sie auf visuelle Elemente abzubilden, wie z. B. räumliche Positionen, besondere Markierungen und andere grafischen Eigenschaften.

Diese visuellen Strukturen können nun verwendet werden, um zum nächsten Knoten des Visualisierungsreferenzmodells zu gelangen, den eigentlichen Darstellungen (englisch: Views), welche das Ziel des Visualisierungsreferenzmodells sind und für den Nutzer erst verwertbar sind. Die sogenannte „View Transformation“ beinhaltet hierbei das Rendern der Darstellungen — sie kann aber auch aus weiteren Arten von Operationen bestehen, wie die Anpassung von Betrachtungsperspektiven.

In einem Beispiel könnte das Visualisierungsreferenzmodell wie folgt angewendet werden: Zunächst liegt eine Menge von Dokumenten vor, die Text beinhalten. Diese werden zu Dokumentenvektoren transformiert, deren Dimensionalität den enthaltenen Wörtern entspricht. Diese werden wiederum mit Hilfe einer Dimensionreduktion auf eine Tabelle mit x, y-Koordinaten abgebildet, welche schließlich verwendet werden kann, um eine Darstellung in einem Koordinatensystem umzusetzen. Diese konkrete Darstellung kann schließlich nochmal angepasst werden, indem der gezeigte Ausschnitt gezoomt oder verschoben wird.

Diese beschriebenen Übergänge (Data Transformation, Visual Mapping, View Transformation) können durch den Nutzer beeinflusst werden, in dem bestimmte Interaktionsmöglichkeiten angeboten werden.

Bei den Data Transformations können das anpassbare Filter für die Dokumente sein, bei Visual Mappings die Anspassungen von Farben, und bei View Transformations das Anpassen des betrachteten Ausschnitts. Mehr dazu folgt im Abschnitt 3.1.2.

Wie das Visualisierungsreferenzmodell konkret in dieser Arbeit verwendet wird, ist in den Kapiteln 4 und 5 zu lesen.

3.1.2 Shneiderman's Visual Information Seeking Mantra

Wichtiger Bestandteil der Visualisierung von Informationen ist die Art von Interaktionen, die in Verbindung mit der Visualisierung für den Nutzer zur Verfügung stehen. Diese sind von Nöten, wenn der Nutzer für bestimmte Analyseaufgaben die visuelle Darstellung ändern möchte, wie z. B. das Ändern von gezeigten Ausschnitten, Animationen usw. Shneiderman hat in [Shn96] hierzu ein Mantra formuliert, das eine Richtlinie darstellt, wie man Visualisierungen einschließlich ihrer Interaktionsmöglichkeiten optimal gestalten kann, damit der Nutzer sie für Analysezwecke verwenden kann. Shneiderman's Visual Information Seeking Mantra lautet wie folgt:

Overview first, zoom and filter, then details-on-demand. [Shn96]

- *Overview*: Dem Nutzer soll die Möglichkeit geboten werden, einen Überblick über die gesamte Menge von Daten zu erhalten. Hierbei wird typischerweise eine Ansicht verwendet, die die gesamte entsprechende Visualisierung beinhaltet. In Verbindung dazu wird oft eine Auswahlbox zur Verfügung gestellt, die der Nutzer verwenden kann, um den Ausschnitt zu wählen, der für eine parallel existierende Detailansicht verwendet werden soll.
- *Zoom*: Der Nutzer kann auf ein Artefakt der Visualisierung zoomen, das er näher untersuchen möchte. Hierbei wird der für ihn relevante Ausschnitt vergrößert. Oft ist es wünschenswert, dass das Zoomen möglichst stufenlos und gleichmäßig funktioniert, damit es einfacher für den Nutzer ist, den lokalen Kontext des Ausschnitts mit der Umgebung zu erfassen bzw. im Auge zu behalten.
- *Filter*: Der Nutzer kann die für ihn uninteressanten Artefakte herausfiltern. Durch das Ausblenden dieser Artefakte ist es für ihn einfacher sich auf relevante Inhalte zu konzentrieren. Typischerweise erstellt der Nutzer hierzu Filteranfragen, die auf verschiedene Weise angepasst werden können, wie z. B. Slidern, Buttons und/oder anderen ähnlichen grafischen Bedienelementen.
- *Details-on-demand*: Hier kann der Nutzer ein oder mehrere Artefakte auswählen und dazu Details anzeigen lassen, um diese genauer zu untersuchen. Typischerweise klickt der Nutzer hierzu das gewünschte Artefakt an und öffnet dadurch in einer zusätzlichen Ansicht die Detailinformationen dazu, die er nun einsehen kann.

Diese Teile des Mantras sollten allerdings nicht als „Einbahnstraße" verstanden werden, denn die Analyse ist typischerweise mit dem Untersuchen der Details allein nicht beendet. Hinzu kommen oft noch das in Relation setzen von untersuchten Artefakten und das Herstellen eines globalen Kontexts, weshalb normalerweise mehrere Iterationen der (Teil-)Schritte des Mantras benötigt werden.

Wie Shneiderman's Visual Information Seeking Mantra konkret in dieser Arbeit umgesetzt wird, wird in den Kapiteln 4 und 5 beschrieben.

3.1.3 Kräftebasierte Verfahren für das Layout von Graphen

Für den Entitätengraphen, der detaillierter in Kapitel 4 beschrieben wird, wird ein sogenannter kräftebasierter Graph benutzt (englisch: Force-directed Graph). In [Kob04] werden mehrere Algorithmen beschrieben, wie man solche kräftebasierten Graphen umsetzen kann.

Grundsätzlich lässt sich sagen, dass den Elementen von kräftebasierten Graphen Kräfte zugewiesen werden. Knoten bekommen dabei typischerweise abstoßende Kräfte, während Kanten anziehende Kräfte erhalten. Dabei sind die Stärken der Kräfte variabel und häufig von verschiedenen Attributen der Elemente abhängig. Durch diese Verteilung von Kräften ist es nun möglich in mehreren Iterationen eine Anordnung der Graphelemente zu erreichen, die auf den Ausgleich der abstoßenden und anziehenden Kräfte basiert.

Oft sind solche kräftebasierte Graphen dynamisch, d. h. der Nutzer kann Knoten verschieben, wodurch zunächst ein Kräfteungleichgewicht entsteht, das wieder in mehreren Iterationen ausgeglichen wird. Hieraus entsteht dann eine neue Anordnung der Graphelemente, bei der schließlich die Kräfte wieder ausgeglichen sind.

Ein Vorteil von solchen kräftebasierten Graphen ist, dass sich schnell Gruppen von Knoten identifizieren lassen, die semantisch zusammenhängen. Besonders bei einer großen Knotenmenge ist dies nützlich, da solche großen Mengen sonst zu unübersichtlich wären. Diese Gruppen bilden sich aufgrund der zugewiesenen Kräfte. Je ausgeprägter die Verbindungen zwischen den Knoten sind, desto höher ist die Dichte der Gruppe. In Abbildung 3.2 ist ein Beispiel für einen kräftebasierten Graphen gegeben. Dort sind die einzelnen Anhäufungen von Knoten zu erkennen, die wie beschrieben durch die Kräfte entstehen. Bei dem zuvor genannten Entitätengraphen ist dies besonders nützlich, da sich hierdurch Entitätengruppen, die eng miteinander in Verbindung stehen, leicht herausfinden lassen.

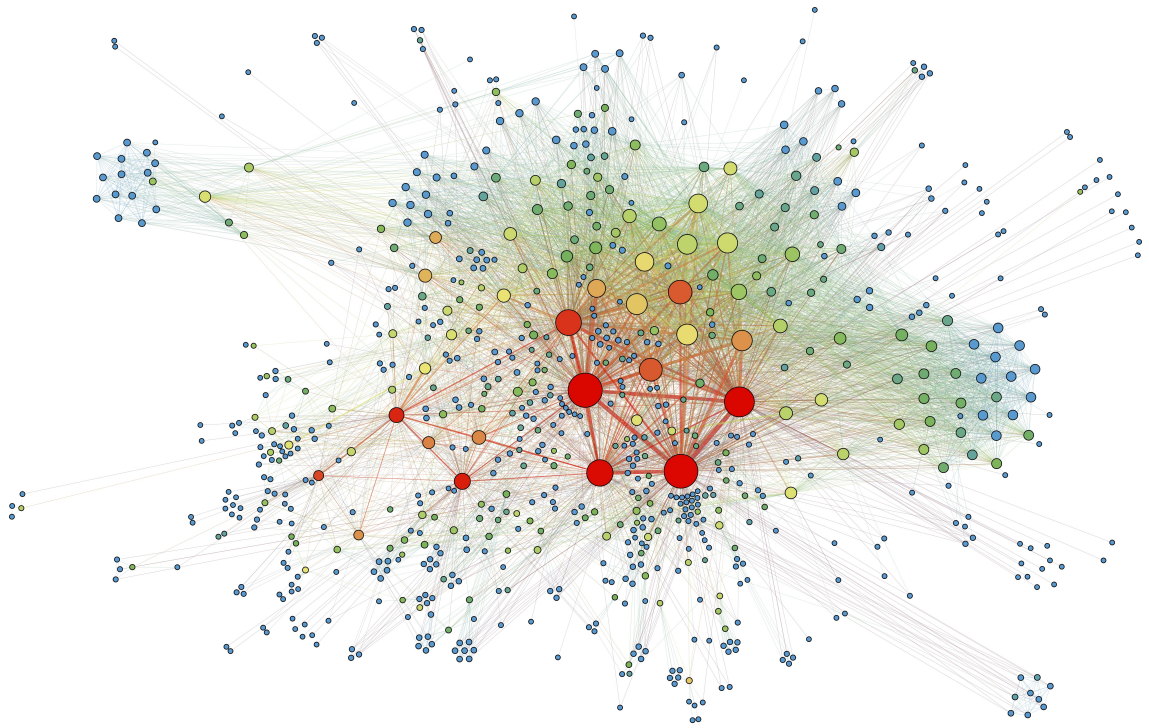


Abbildung 3.2: Beispiel für einen kräftebasierten Graphen: Erkennbar sind die einzelnen Anhäufungen von Knoten, die durch die Verteilung der Kräfte entstehen [Gra15].

3.2 Natürliche Sprachverarbeitung

Bei der Untersuchung von Textinhalten von gegebenen Dokumentenmengen ist es hilfreich, die Vorteile der natürlichen Sprachverarbeitung (oder auch Computerlinguistik) zu nutzen. In unserer Arbeit gibt es mehrere Bereiche, in denen die natürliche Sprachverarbeitung Anwendung findet und entsprechende Untersuchungen erleichtert. Deshalb soll im Folgenden auf das Gebiet der natürlichen Sprachverarbeitung genauer eingegangen werden und ausgewählte Aspekte daraus beschrieben werden.

3.2.1 Definition

Chowdhury gibt in [Cho03] folgende Definition für die natürliche Sprachverarbeitung an: Das Gebiet der natürlichen Sprachverarbeitung beschäftigt sich damit, wie man mit Hilfe von Computern automatisiert die natürliche, menschliche Sprache analysieren und verarbeiten kann. Oft wird dieses Gebiet auch mit NLP abgekürzt, was für die englische Bezeichnung, Natural Language Processing, steht. Bei

NLP handelt es sich um ein interdisziplinäres Gebiet, das mehrere Felder, wie die Linguistik, Informatik, Elektrotechnik, Robotik oder auch Psychologie einschließt, um so die komplexe menschliche Sprache korrekt zu erfassen.

3.2.2 Das Pipeline-Modell

Speziell für die Untersuchung und Verarbeitung von Texten gibt es ein grundlegendes Verfahren, das einer Pipeline gleicht [JJ00]: Es gibt eine Aufteilung in mehrere Schritte, in denen jeweils ein spezielles Ziel verfolgt wird und auf dem die weiteren Schritte aufbauen. Zu der grundlegenden Aufteilung gehören hierbei tokenbasierte Operationen, syntaktische Analysen und semantische Analysen. In Abbildung 3.3 [BKL09] ist diese Pipeline schematisch dargestellt.

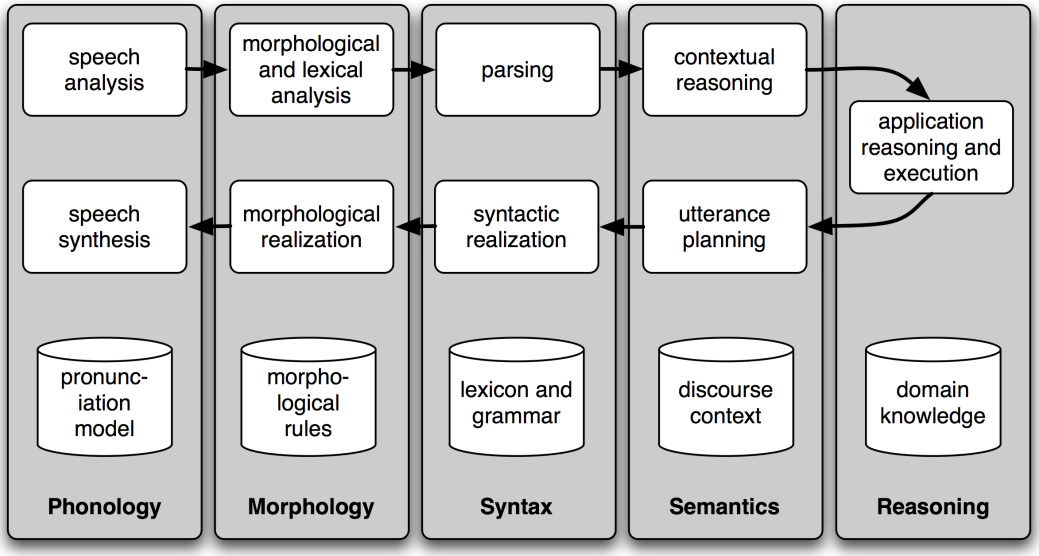


Abbildung 3.3: Schematische Darstellung der NLP-Pipeline: Es gibt eine Aufteilung in eigene Aufgabenbereiche, die jeweils von dem vorigen Teil der Pipeline abhängig sind. In unserer Arbeit ist besonders der obere Teil der Bereiche „Morphology“, „Syntax“ und „Semantics“ von Bedeutung [BKL09].

In unserer Arbeit gibt es Teile dieser Pipeline, die von besonderer Bedeutung sind, da sie Bestandteil bestimmter Funktionen sind. Deshalb sollen diese nun genauer betrachtet werden:

Tokenisierung

Dieser Schritt gehört zum morphologischen Teil der NLP-Pipeline (siehe Abbildung 3.3). Bei der Tokenisierung (englisch: Tokenization) werden aus einem zusammenhängenden Text einzelne Tokens extrahiert. Bei den Tokens handelt es sich typischerweise um einzelne Wörter, die im Text vorkommen.

In Abbildung 3.4 [MRS+08] ist ein Beispiel zu sehen, bei denen die resultierenden Tokens den jeweiligen Wörtern im Satz entsprechen.

Input: Friends, Romans, Countrymen, lend me your ears;
Output: [Friends] [Romans] [Countrymen] [lend] [me] [your] [ears]

Abbildung 3.4: Beispiel einer Tokenisierung: Aus einem Satz werden die einzelnen Tokens extrahiert [MRS+08].

In einfachen Fällen reicht eine Trennung nach Leerzeichen bzw. Satzzeichen. Es gibt jedoch auch kompliziertere Fällen, in denen dieser einfache Ansatz nicht ausreicht. Beispielsweise gibt es Wörter, die semantisch zusammenhängen, wie z. B. der Städtename *New York*. Hier wäre es wünschenswert, wenn man den vollständigen Namen als Token extrahiert, an Stelle von zwei separaten Tokens. Es gibt jedoch auch zusammengesetzte Wörter, bei denen man eventuell einzelne Wortteile als eigene Tokens behandeln möchte, wie z. B. beim Wort *Lebensversicherungsangesteller*. Weitere solcher Fälle erhält man, wenn man Sonderschreibweisen wie URLs betrachtet oder sogar Sprachen miteinbezieht, bei denen es keine Leerzeichen gibt, wie z. B. chinesisch oder japanisch.

Erwähnenswert sind an dieser Stelle auch die sogenannten Stop Words. Zu den Stop Words gehören solche Wörter, die für die weiteren Schritte nur wenig Bedeutung haben, wie *der, die, das, und, von, dass, oder* usw. Solche Wörter verwirft man in der Regel und erstellt dafür keine eigenen Tokens. Jedoch gilt auch hier, dass man mit Bedacht die Stop Words definieren sollte, da es Fälle gibt, in denen es Sinn machen kann, diese doch mit aufzunehmen, wie z. B. bei *Präsident der Vereinigten Staaten* an Stelle von *Präsident* und *Vereinigte Staaten*.

Eine einheitliche Lösung für die Tokenisierung zu finden kann sich als schwierig erweisen und kann fallabhängig sein. Jedoch sollte Wert darauf gelegt werden, da die weiteren Teile der Pipeline auf einer sinnvollen Tokenisierung aufbauen.

Segmentierung der Sätze

Dieser Schritt gehört ebenfalls zum morphologischen Teil der NLP-Pipeline (siehe Abbildung 3.3). Ähnlich wie bei der Tokenisierung teilt man einen zusammenhängenden Text auf. Hier sind die Aufteilungen jedoch nicht Tokens, sondern Sätze. Auch hier gibt es Regeln nach denen man die Sätze aufteilt, meist verwendet man die Satzzeichen, insbesondere Satzpunkte. Jedoch gilt es auch hier, Sonderfälle zu beachten: Beispielsweise sollte man Satzpunkte von Punkten unterscheiden, die für Abkürzungen benutzt werden.

Die Segmentierung von Sätzen (englisch: Sentence splitting) wird in dieser Arbeit vor allem verwendet, um den Zusammenhang von Personen und Orten in Dokumenten zu definieren: Je öfter Personen oder Orte im gleichen Satz vorkommen, desto größer ist ihr Zusammenhang. Diese Zusammenhänge und ihre unterschiedlichen Stärken werden unter anderem für den Entitätengraphen verwendet (mehr dazu in den Kapiteln 4 und 5).

Lemmatisierung

Die Lemmatisierung (englisch: Lemmatization) hat das Ziel, aus mehreren extrahierten Tokens, die semantisch die gleiche Bedeutung haben, eine gemeinsame einheitliche Form zu generieren. Hierzu versucht man bei der Lemmatisierung für die Tokens eine grammatikalisch korrekte Grundform zu finden. Beispielsweise wäre bei den Wörtern *gesagt* und *sagte* die Grundform *sagen* das Ziel. Ein weiteres Beispiel wäre die Vereinheitlichung von *U.S.A.* und *USA*. Auf diese Weise lassen sich die weiteren Schritte sinnvoller durchführen, wie die Named-Entity-Recognition. Auch das im nächsten Abschnitt beschriebene Part-of-Speech-Tagging basiert auf der Lemmatisierung. Die Lemmatisierung gehört zum morphologischen Teil der NLP-Pipeline (siehe Abbildung 3.3).

Part-of-Speech-Tagging

In diesem Schritt, welcher zum syntaktischen Bereich der NLP-Pipeline gehört (siehe Abbildung 3.3), versieht man die vorher beschriebenen Tokens mit Annotationen, den sogenannten Part-of-Speech-Tags, oder kurz POS-Tags. Diese enthalten grammatikalische Informationen zu dem jeweiligen Tokens, wie z. B. die Wortart (Substantiv, Verb, Adjektiv usw.), die Pluralität und einige andere Eigenschaften. Ein Beispiel für solche POS-Kategorien oder auch *Tagsets* ist das Penn Treebank POS Tagset für die englische Sprache, welches in Abbildung 3.5 [MMS93] dargestellt ist. Ein Beispiel für eine Anwendung des POS-Taggers könnte wie folgt aussehen [JJ00]:

- Eingabe: The grand jury commented on a number of other topics.
- Ergebnis: The/DT grand/JJ jury/NN commented/VBD on/IN a/DT number/NN of/IN other/JJ topics/NNS ./.

Table 2
The Penn Treebank POS tagset.

1. CC	Coordinating conjunction	25. TO	*to*
2. CD	Cardinal number	26. UH	Interjection
3. DT	Determiner	27. VB	Verb, base form
4. EX	Existential *there*	28. VBD	Verb, past tense
5. FW	Foreign word	29. VBG	Verb, gerund/present participle
6. IN	Preposition/subordinating conjunction	30. VBN	Verb, past participle
7. JJ	Adjective	31. VBP	Verb, non-3rd ps. sing. present
8. JJR	Adjective, comparative	32. VBZ	Verb, 3rd ps. sing. present
9. JJS	Adjective, superlative	33. WDT	*wh*-determiner
10. LS	List item marker	34. WP	*wh*-pronoun
11. MD	Modal	35. WP$	Possessive *wh*-pronoun
12. NN	Noun, singular or mass	36. WRB	*wh*-adverb
13. NNS	Noun, plural	37. #	Pound sign
14. NNP	Proper noun, singular	38. $	Dollar sign
15. NNPS	Proper noun, plural	39. .	Sentence-final punctuation
16. PDT	Predeterminer	40. ,	Comma
17. POS	Possessive ending	41. :	Colon, semi-colon
18. PRP	Personal pronoun	42. (	Left bracket character
19. PP$	Possessive pronoun	43.)	Right bracket character
20. RB	Adverb	44. "	Straight double quote
21. RBR	Adverb, comparative	45. ‘	Left open single quote
22. RBS	Adverb, superlative	46. “	Left open double quote
23. RP	Particle	47. ’	Right close single quote
24. SYM	Symbol (mathematical or scientific)	48. ”	Right close double quote

Abbildung 3.5: Penn Treebank POS Tagset: Tabellarisch aufgelistet sind die verschiedenen POS-Kategorien mit ihren zugehörigen Tags [MMS93].

Das Part-of-Speech-Tagging wird in dieser Arbeit verwendet, um Verben und Adjektive zu identifizieren, die im Zusammenhang mit bestimmten Personen und Orten vorkommen. So soll der Nutzer die Möglichkeit haben, zu erkennen, wie Personen oder Orte zueinander stehen, und die Möglichkeit diese Beziehungen zu bewerten (mehr dazu in den Kapiteln 4 und 5).

Named-Entity-Recognition

Dieser Schritt gehört zum semantischen Teil der NLP-Pipeline (siehe Abbildung 3.3). Die Named-Entity-Recognition, oder auch kurz NER, hat die Aufgabe, benamte Entitäten zu identifizieren. Beliebte Kategorien sind hierbei Personen, Organisationen, geopolitische Einheiten oder Orte, Zahlenwerte wie Jahreszahlen oder Geldbeträge usw. Ähnlich wie beim POS-Tagging werden die Tokens mit dem entsprechenden Tag annotiert. Beispielsweise würde *Artus* mit dem Tag „PERSON“ annotiert werden, oder *Britain* mit dem Tag „LOCATION“.

Wir verwenden in dieser Arbeit die NER für die Erkennung von Personen und Orten, welche für den Entitätengraphen benutzt werden (mehr dazu in den Kapiteln 4 und 5).

4 Konzept

In diesem Kapitel stehen die Konzepte im Mittelpunkt, die für die Lösung der gegebenen Problemstellung dieser Arbeit entwickelt wurden, nämlich der Textexploration von großen Dokumentenmengen mittels der visuellen Kombination von Lupen. Hierbei richtet sich der Fokus vor allem auf die enthaltenen Personen und Orte. Die hier vorgestellten Konzepte basieren auf Teilfunktionen des bereits vorhandenen Werkzeugs DocuCompass [HJH$^+$ar], das bereits in Kapitel 2.1 beschrieben wurde.

Wir gehen in diesem Kapitel wie folgt vor: Zuerst betrachten wir, welche bereits vorhandenen Funktionen von DocuCompass besonders relevant sind. Danach werden im Detail entwickelte Konzepte und Lösungsansätze und deren Funktionsweisen vorgestellt.

4.1 DocuCompass als Fundament

In diesem Abschnitt werden die Teile von DocuCompass (siehe Kapitel 2.1) beschrieben, die für die weiteren Konzepte und Lösungsansätze in diesem Kapitel von Relevanz sind. Sie sind insofern wichtig, da die entwickelten Funktionen diese Teile als Basis haben und darauf aufbauen.

4.1.1 Einlesen von Datensätzen mit anschließender Spatialisierung

Wichtiger Bestandteil der Exploration von Dokumentenmengen ist die initiale Verarbeitung der Dokumentenmengen selbst. Hierzu gehört nicht nur das Einlesen der Dokumente, sondern auch die anschließende Darstellung der Spatialisierung der Dokumente, da hierauf die Interaktionen mit den Lupen basieren.

DocuCompass stellt hierzu bereits fertige Funktionen bereit, bei denen es naheliegend ist, diese weiterzuverwenden. Es handelt sich hierbei um die Einlesefunktionen, welche die Texte der Dokumente tokenisieren und die resultierenden Tokens abspeichern (Tokenisierung siehe Kapitel 3.2.2). Diese Tokens sind die Grundlage für eine Vielzahl von Funktionen, welche im Verlauf des Kapitels noch vorgestellt werden. Weiterhin berechnet DocuCompass für die Dokumente die entsprechenden Spatialisierungen und stellt diese anschaulich dar, wobei hier unterschiedliche Spatialisierungsarten und dazugehörige Charakterisierungen der Dokumente möglich sind. In dieser Arbeit beschränken wir uns jedoch auf das sogenannte „t-distributed stochastic neighbor embedding“ (t-SNE) [MH08]. Ein Beispiel für eine Dokumentenspatialisierung ist in Abbildung 2.1 dargestellt: Jede kreisförmige Glyphe entspricht hierbei einem Dokument.

4.1.2 Lupenfunktionen

Die oben beschriebenen Dokumentenspatialisierungen dienen uns als Grundlage für die nächste wichtige Funktionskomponente: Die Lupen. DocuCompass bietet hierzu ebenfalls nützliche Funktionen an, die zu deren Weiterverwendung einladen.

Dies sind zum einen die essentiellen Funktionen, wie das Fokussieren und Erkennen von Untermengen der Dokumente. DocuCompass setzt dies ästhetisch ansprechend um: Der fokussierte Bereich wird aufgehellt und fokussierte Dokumente werden farbig hervorgehoben (siehe Abbildung 4.1). So ist es einfacher für den Nutzer, den fokussierten Bereich wahrzunehmen.

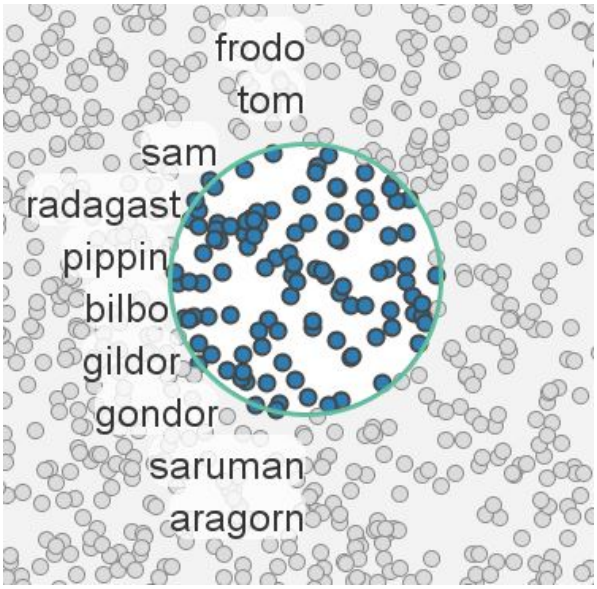
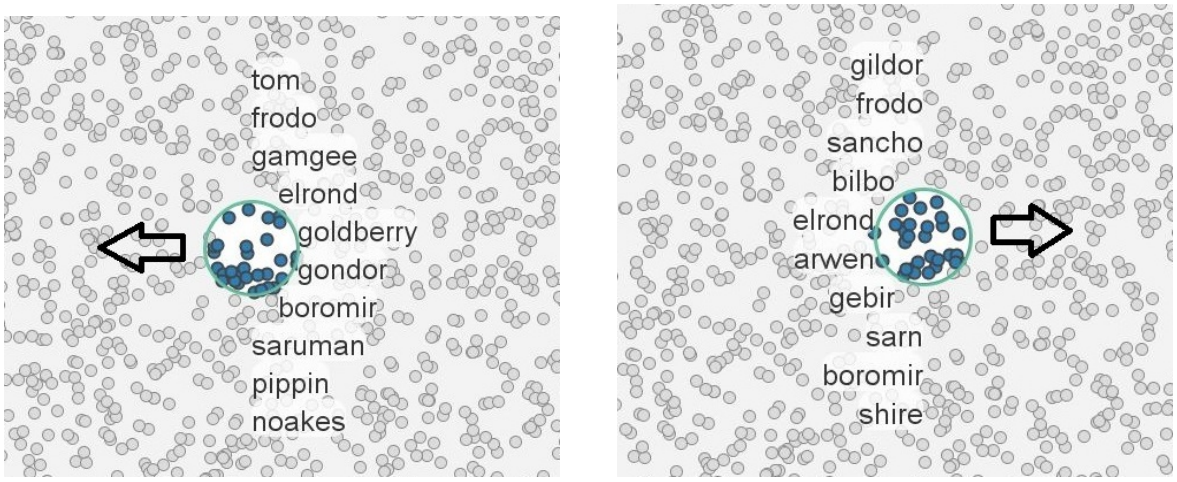


Abbildung 4.1: DocuCompass: Der fokussierte Bereich der Lupen wird aufgehellt und die entsprechenden Dokumente werden farbig hervorgehoben, wodurch der Nutzer die fokussierte Dokumentenmenge schneller identifizieren kann.

Ebenso ansprechend sind die Interaktionsmöglichkeiten mit den Lupen: Feinstufige Anpassungen der Lupengröße sowie flüssiges Bewegen der Lupen per Drag-and-Drop machen das Arbeiten mit den Lupen flexibel. Zudem unterstützt DocuCompass dynamisches Arbeiten mit den Lupen, indem die Position der angezeigten Zusatzinformationen angepasst wird, je nach Bewegungsrichtung der Lupe (siehe Abbildung 4.2). Auf diese Weise wird der Bereich vor der Lupe (in Bewegungsrichtung) nicht verdeckt und der Nutzer kann sich besser auf den Bereich der Spatialisierung konzentrieren, den er untersuchen möchte.

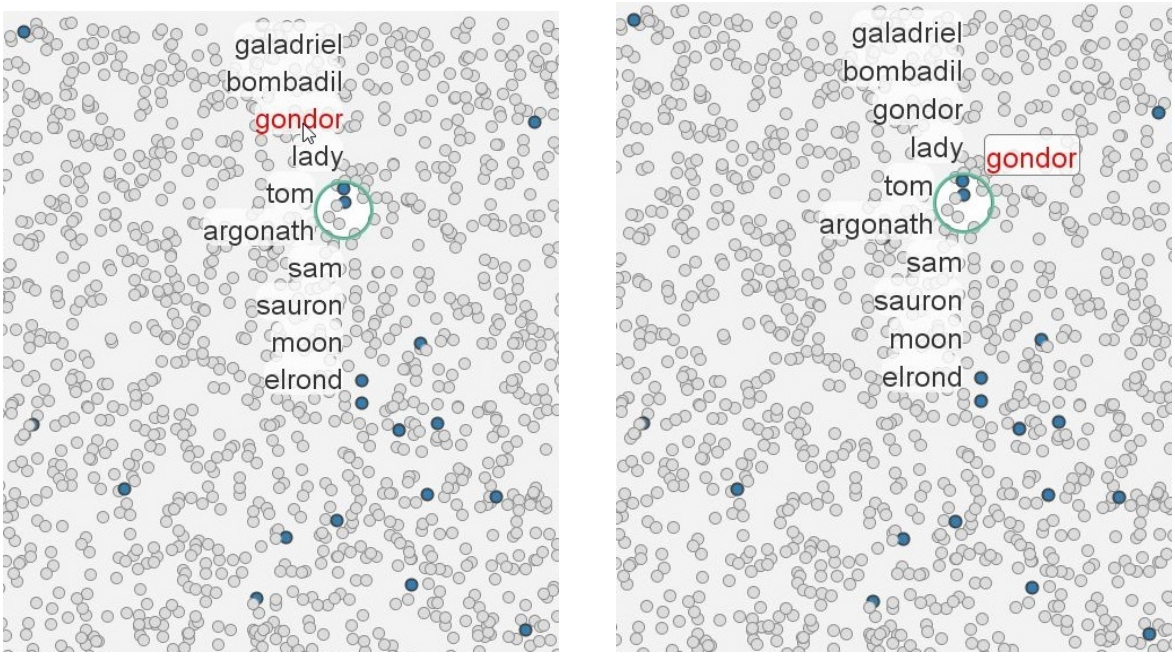


(a) Bewegungsrichtung der Lupe nach links, Informationen rechts von der Lupe

(b) Bewegungsrichtung der Lupe nach rechts, Informationen links von der Lupe

Abbildung 4.2: Dynamische Anpassung der Position für angezeigte Informationen in DocuCompass: Je nachdem in welche Richtung die Lupe bewegt wird, ändert sich die Position für die Informationen zu der fokussierten Menge. Auf diese Weise wird der Bereich vor der Lupe (in Bewegungsrichtung) nicht verdeckt und ist besser für den Nutzer einsehbar.

Eine weitere Funktionalität, auf die in dieser Arbeit aufgebaut wird, ist das Filtern nach Termen. DocuCompass bietet hierzu folgende Funktionen an: Beim Hovern eines Terms neben einer Lupe mit der Maus werden diejenigen Dokumente in der Spatialisierung hervorgehoben, die ebenfalls diesen Term beinhalten (siehe Abbildung 4.3a). Per Klick auf diesen Term kann dieser auch an die Lupe „angepinnt“ werden, so dass der Filter aktiv bleibt, auch wenn nicht mehr über den Term gehovert wird (siehe Abbildung 4.3b) — an dieser Stelle kann der Nutzer die Lupe (mit aktiven Filter) weiterbewegen, um so die hervorgehobenen Dokumente genauer zu untersuchen. Dieses Filtern von Dokumenten nach Termen wird für eine einige weitere Funktionen übernommen, welche im Laufe des Kapitels noch vorgestellt werden.



(a) Hovern des Terms mit der Maus

(b) Anpinnen des Terms an die Lupe per Mausklick

Abbildung 4.3: Termfilter im DocuCompass: Es lassen sich Dokumente hervorheben, die einen bestimmten Term enthalten — entweder per Hovern oder per Anpinnen des Terms. Erkennbar sind die eingefärbten Dokumentenglyphen, die den entsprechenden Term enthalten.

4.2 Konzepte und Ansätze zur Kombination der Lupen

In diesem Abschnitt des Kapitels werden die Konzepte und Lösungsansätze vorgestellt, die entwickelt wurden, um die Problemstellung dieser Arbeit zu lösen. Hierbei betrachten wir verschiedene Aspekte: Wir führen in DocuCompass neue Ansichten ein, die entsprechende Informationen zu den Dokumenten anzeigen sollen. Diese Informationen beziehen sich vor allem auf die enthaltenen Entitäten, wie Personen und Orte. Ebenso von Bedeutung sind die enthaltenen Verben und Adjektive. Hierbei wird auf verschiedene Darstellungsarten und Visualisierungen zurückgegriffen wie Graphen und Word-Clouds, und es werden verschiedene Interaktionsmöglichkeiten mit den Ansichten geboten. All diese Funktionen haben dabei die Spatialisierung der Dokumente mit der zugehörigen Konstellation von Lupen als Basis. Im Folgenden gehen wir also auf die einzelnen Funktionskonzepte ein und betrachten sie im Detail.

4.2.1 Unterscheidung der Lupen

Eines der angestrebten Ziele dieser Arbeit ist es, bei der Untersuchung von Dokumentenspatialisierungen die Arbeit mit mehreren Lupen gleichzeitig zu unterstützen. Hierbei ist zunächst eine intuitive

und für den Nutzer einfache Unterscheidung der Lupen notwendig. Hierzu wird ein einfaches Konzept verwendet: Die Lupen werden jeweils mit einer eigenen Farbe eingefärbt.

Die konkrete Auswahl der Farben ist dabei keineswegs willkürlich: Es muss darauf geachtet werden, dass Farben verwendet werden, die sich leicht voneinander unterscheiden lassen. Die Auswahl der Farben ist auch insofern wichtig, da verschiedene Funktionen, die später noch vorgestellt werden, sich auf die einzelnen Lupen beziehen — zur Kennzeichnung der Zugehörigkeit werden an diesen Stellen die entsprechenden Farben der Lupen verwendet. Weiterhin ist auch eine passende Sättigung der Farben, neben dem eigentlichen Farbton, notwendig: Die Farben müssen zum einen kräftig genug sein, um sich von einem weißen Hintergrund ausreichend abzuheben. Zum anderen müssen sie geeignet sein, um als Hintergrund für schwarzen Text zu dienen und dürfen daher nicht zu dunkel sein.

Die konkrete Auswahl der Farben für die Lupen wurde mit Hilfe von *ColorBrewer*[1] [HB03] getroffen. Dieses Werkzeug bietet vordefinierte Farbpaletten für die Darstellung von verschiedenen Arten von Datensätzen an. Für unseren Fall wurde eine Farbpalette für qualitative Datensätze benutzt (statt sequentiellen oder divergierenden Datensätzen), mit einer passenden Sättigung der Farben. Die Anzahl der benötigten Farben beläuft sich auf fünf, da der entwickelte Prototyp bisher eine Anzahl von fünf Lupen gleichzeitig unterstützt. In Abbildung 4.4 ist die konkrete Auswahl der Farben zu sehen.

Abbildung 4.4: Farbpalette für die Lupen: Jede der fünf unterstützten Lupen hat ihre eigene zugewiesene Farbe. Die Farben werden zur Kennzeichnung auch konsistent für andere Funktionen verwendet, die sich auf die jeweiligen Lupen beziehen.

Die anderen Funktionen der Lupen (siehe Abschnitt 4.1.2) werden größtenteils beibehalten und übernommen. Das Einfügen und Entfernen von Lupen funktioniert wie in DocuCompass bereits gegeben (wahlweise per Tastatur oder per Rechtsklick auf die Spatialisierung) und die Interaktionsmöglichkeiten wie das Anpassen der Größe und das Bewegen der Lupen funktionieren ebenfalls wie gehabt.

4.2.2 Anzeigen von Details zu den fokussierten Mengen

Am Anfang der Entwicklung des Konzepts warf sich die Frage auf, auf welche Weise Informationen und Details zu den fokussierten Mengen der Lupen angezeigt werden sollen. Hier bestand anfangs

[1] http://colorbrewer2.org

noch der Versuch, diese Informationen unmittelbar neben den Lupen anzuzeigen, also direkt in der Dokumentenspatialisierung. Eine konkrete Idee hierzu war, die Lupen als Eckpunkte zu nehmen und den dadurch „aufgespannten“ Bereich als Fläche zu benutzen, um die Informationen anzuzeigen.

Dieser Ansatz war jedoch insofern problematisch, da er für die geplante Word-Cloud (später Entitätengraph) weniger gut geeignet war. Zunächst wäre da die variable Anzahl der Lupen, welche bis zu fünf geplant ist. Hinzu kommen noch die vielen möglichen Anordnungen der Lupen, welche eine Vielzahl von möglichen Formen und Größen für die „aufgespannte“ Fläche bedeuten. Diese Faktoren erschweren es, eine konsistente und für den Nutzer intuitive Darstellung der Informationen zu gewährleisten. Ein weiteres Argument ist, dass die direkte Darstellung innerhalb der Dokumentenspatialisierung die Interaktion damit beeinträchtigt, da hierdurch größere Teile der Spatialisierung verdeckt werden würden.

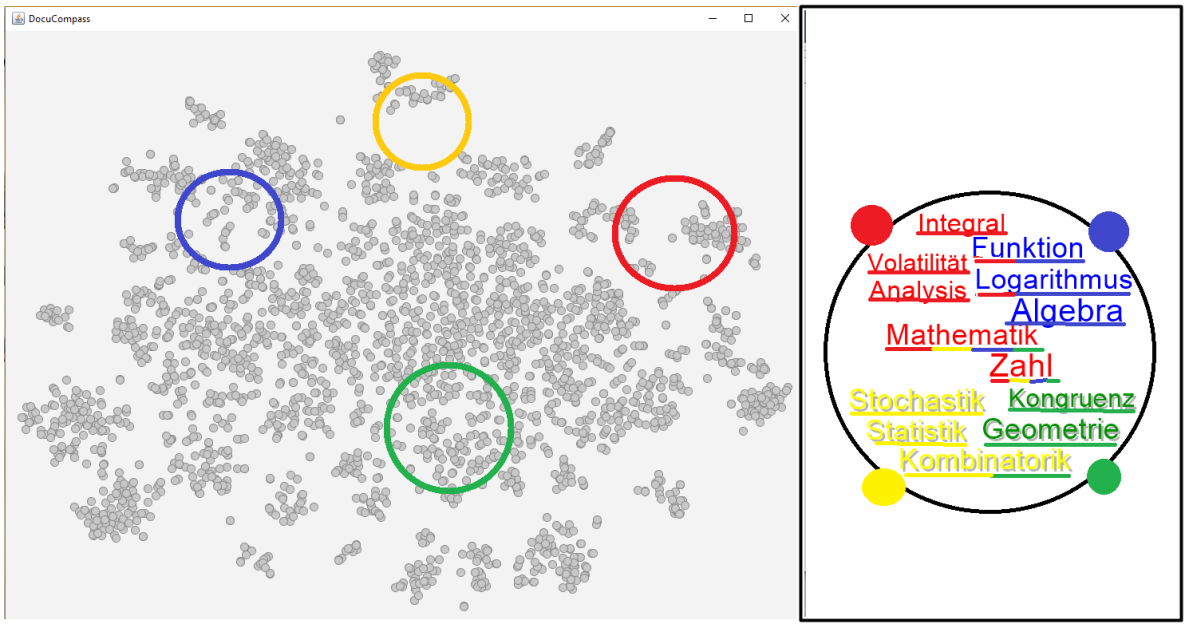


Abbildung 4.5: Separate Ansicht: Die Details und Informationen zu dem von den Lupen fokussierten Bereich werden in einem eigenen Bereich angezeigt — so wird die Konsistenz und Intuitivität sichergestellt und die Interaktion mit der Dokumentenspatialisierung weniger beeinträchtigt. Die Zugehörigkeit zu den Lupen soll mittels Farben angezeigt werden, wie z. B. bei den eingefärbten Balken unter den Begriffen (später Personen und Orte), die die Anteile der Vorkommen in den Lupen darstellen.

Aufgrund der genannten Punkte fiel die Entscheidung auf die Darstellung der Informationen in einem eigenen Bereich (siehe Abbildung 4.5). Um den Bezug zu den Lupen nicht zu verlieren, werden die Farben der Lupen verwendet, um die Informationen entsprechend zu kennzeichnen (siehe Abschnitt 4.2.1). Beispielsweise werden die Anteile der Vorkommen von Begriffen in den Lupen (später Personen und Orte) mit Hilfe von eingefärbten Balken angezeigt (siehe auch Abbildung 4.5).

(Es sei bemerkt, dass die verwendeten Farben in Abbildung 4.5 von den angegebenen Farben aus Abschnitt 4.2.1 abweichen, weil zu dem Entstehungszeitpunkt der Skizze, die Farben der Lupen noch nicht festlagen.)

4.2.3 Erkennung von enthaltenen Entitäten und Entitätengraph

Im Bezug auf die Exploration und Analyse der Dokumentenmengen soll der Fokus in dieser Arbeit vor allem auf den enthaltenen Personen und Orten, einschließlich ihrer Beziehungen zueinander, liegen. Aus diesem Grund sind wichtige Bestandteile dieser Arbeit die Erkennung und Extraktion dieser Entitäten und eine entsprechende Darstellung dessen.

Hierzu wird für die Visualisierung der Entitäten und deren Beziehungen ein kräftebasierter Graph (siehe Abschnitt 3.1.3) benutzt. Die Anziehungskraft zwischen zwei Entitäten entspricht hierbei deren inhaltlichen Zusammenhang: Je öfter zwei Entitäten (also Personen oder Orte) innerhalb einer bestimmten Spanne von Sätzen zusammen vorkommen, desto höher ist ihr Zusammenhang und desto höher ist folglich die Anziehungskraft zwischen ihnen. Um den Zusammenhang zusätzlich zu verdeutlichen, wird die Kantendicke der entsprechenden Kanten je nach Stärke angepasst. Auf diese Weise sollen dem Nutzer die Beziehungsnetzwerke von den enthaltenen Entitäten leichter zugänglich und einsehbar gemacht werden.

Der Entitätengraph wird hierbei aus den verwendeten Lupen in der Dokumentenspatialisierung generiert. Die fokussierten Dokumentenmengen der jeweiligen Lupen werden kombiniert und als Gesamtdokumentenmenge für die Extraktion der Personen und Orte, einschließlich ihrer jeweiligen „Zusammenhangsstärken“, verwendet. Hierbei soll die Generierung des Entitätengraphen möglichst flüssig verlaufen: Jedes Mal wenn der Nutzer die Position einer Lupe ändert, wird der Entitätengraph in kürzester Zeit „on-the-fly“ aktualisiert, so dass der Arbeitsfluss des Nutzers nicht durch lange Wartezeiten unterbrochen wird.

In Abbildung 4.6 ist eine Skizze dieses beschriebenen Ansatzes zu sehen: In der separaten Ansicht (siehe Abschnitt 4.2.2) ist der kräftebasierte Graph dargestellt, welcher die Zusammenhänge zwischen den Begriffen anzeigt.

(Es sei bemerkt, dass in Abbildung 4.6 als Beispiel noch allgemeine Begriffe statt Personen und Orte verwendet wurden, da zum Entstehungszeitpunkt der Skizze der Ansatz mit Personen und Orten noch nicht festgelegt war.)

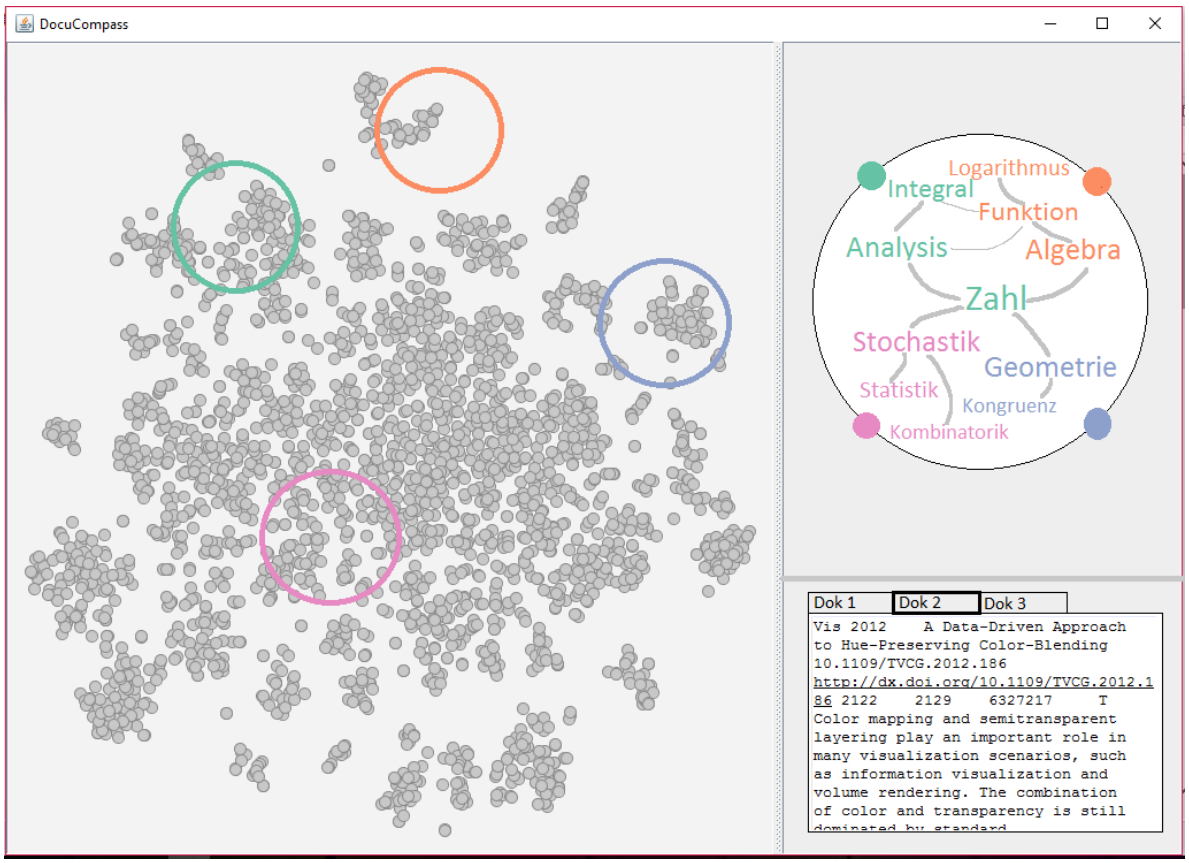


Abbildung 4.6: GUI-Mockup: In der separaten Ansicht sind sowohl der kräftebasierte Graph als auch der Textexplorer zu erkennen.

Neben diesen beschriebenen Funktionen bietet der Entitätengraph auch Interaktionsmöglichkeiten. Hierzu gehören zum einen das Zoomen des Graphen und das Verschieben des angezeigten Ausschnitts. So kann der Nutzer interessante Teile der Visualisierung genauer einsehen und begutachten bzw. einfacher einen Überblick über die Visualisierung bekommen. Weiterhin werden beim Hovern von Knoten oder Kanten, die direkten Nachbarknoten markiert. Dies ist in denjenigen Fällen hilfreich, in denen die enthaltenen Personen und Orte und ihre Beziehungen sehr zahlreich sind, wodurch der Entitätengraph sehr komplex werden kann. Auf diese Weise kann der Nutzer, trotz der Komplexität, die einzelnen Beziehungen identifizieren und Untersuchungen zu den Beziehungsnetzwerken durchführen.

Um weiterhin die Analyse der Personen und Orte zu unterstützen, gibt es noch weitere Funktionen, die hierzu konzipiert wurden. Parallel zum Entitätengraphen werden die meistvorkommenden Personen und Orte auch neben den jeweiligen Lupen angezeigt, ähnlich wie dies DocuCompass bereits für allgemeine Begriffe umgesetzt hat. Hier gelten auch die Vorteile, die bereits in Abschnitt 4.1.2 erläutert wurden: Beispielsweise kann hier auch die dynamische Anpassung der Positionen für die angezeigten Entitäten übernommen werden (siehe Abbildung 4.2).

Auf gleiche Weise kann insbesondere der Termfilter für die Personen und Orte übernommen werden (siehe Abbildung 4.3): Das Filtern und Hervorheben von Dokumenten nach entsprechenden Entitäten ermöglicht dem Nutzer, einfacher diejenigen Dokumente zu identifizieren, die für weitere Analysen in Frage kommen. Hierbei kann der Nutzer wahlweise die vorhandenen Termfilter-Funktionen oder auch den Entitätengraph verwenden: Per Hovern eines Knotens werden die entsprechenden Dokumente in der Spatialisierung hervorgehoben. Ebenso ist es möglich die Person bzw. den Ort per Mausklick an die Lupe anzupinnen, ähnlich wie dies für die vorhandenen Termfilter-Funktionen bereits funktioniert.

4.2.4 Explorer für Verben und Adjektive

In Abschnitt 4.2.3 wurden die Möglichkeiten beschrieben, die konzipiert wurden, um die enthaltenen Entitäten (Personen und Orte) und ihre Beziehungen zueinander zu identifizieren. Im Folgenden soll eine Möglichkeit vorgestellt werden, um diese Beziehungen näher zu untersuchen, einschätzen und bewerten zu können.

Hierzu führen wir eine weitere Ansicht ein, welche wir für die restliche Arbeit „Explorer“ nennen werden. In Abbildung 4.6 ist der hierfür vorgesehene Bereich in der Anwendung zu sehen, welcher sich unterhalb des im Abschnitt 4.2.3 beschriebenen Graphen befindet.

Der Explorer bietet die Möglichkeit bestimmte Verben und Adjektive anzeigen zu lassen, und zwar im Bezug auf die aktuell betrachteten Personen und Orte. Konkret bedeutet das, dass die Verben und Adjektive, die gemeinsam mit den jeweiligen Entitäten in einer bestimmten Spanne von Sätzen innerhalb der Dokumente vorkommen, extrahiert und abgespeichert werden.

Der Nutzer hat nun die Möglichkeit, sich diese Verben und Adjektive anzeigen zu lassen. Hierfür wird für deren Visualisierung auf Word-Clouds zurückgegriffen. Ein Beispiel dafür ist in Abbildung 4.7 zu sehen. Zunächst werden den Verben und Adjektiven eigene Farben zugewiesen, um sie schnell und einfach voneinander unterscheiden zu können. Je öfter ein Verb bzw. Adjektiv gemeinsam im Zusammenhang mit der entsprechenden Entität auftritt, desto größer wird dessen Tag in der Word-Cloud. So sollen relevante Verben und Adjektive hervorgehoben werden, damit der Nutzer diese leichter erkennen kann. Um bestimmte Verben bzw. Adjektive zu einem späteren Zeitpunkt leichter wiederfinden zu können, werden diese alphabetisch in der Word-Cloud angeordnet.

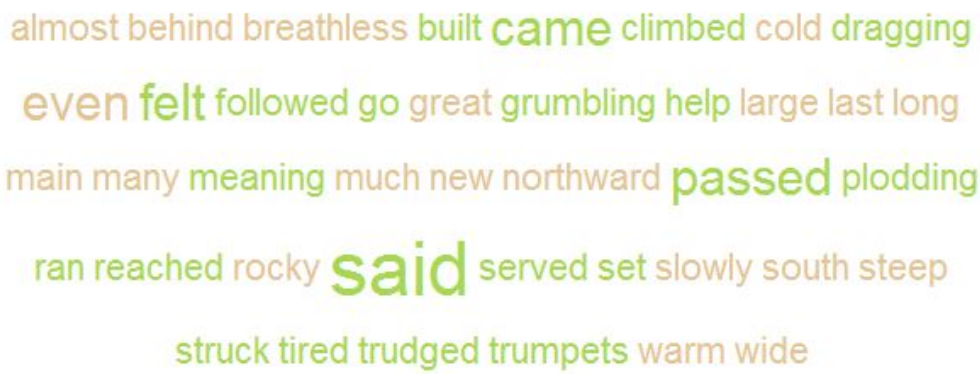


Abbildung 4.7: Die Word-Cloud lässt sich sowohl für einzelne Entitäten als auch für Entitätenpaare öffnen: Hier werden Verben und Adjektive angezeigt, die häufig im Zusammenhang mit den entsprechenden Personen und Orten vorkommen. Mittels dessen kann der Nutzer den Kontext der Entitäten und deren Beziehungen zueinander leichter verstehen und einschätzen.

Dieser Aufbau der Word-Cloud wird in der Exploreransicht angezeigt, indem der Nutzer auf einen Knoten des Entitätengraphen (siehe Abschnitt 4.2.3) doppelklickt, welcher einer Person oder einem Ort entspricht. Daraufhin werden die zuvor abgespeicherten Verben und Adjektive gefiltert, die im Zusammenhang mit der entsprechenden Person oder dem entsprechendem Ort vorkommen, einschließlich ihrer Anzahl der gemeinsamen Vorkommen. Aus diesen Daten wird die Word-Cloud generiert und in der Exploreransicht dargestellt.

Die Word-Cloud mit Verben und Adjektiven lässt sich nicht nur für einzelne Entitäten generieren, sondern auch für Paare von Entitäten. Hierzu doppelklickt der Nutzer statt auf einen Knoten auf die entsprechende Kante im Entitätengraphen, mit den entsprechenden angrenzenden Knoten, welche für die zu untersuchenden Personen bzw. Orte stehen. Das Kriterium für die Verben und Adjektiven, die in diesem Fall für die Word-Cloud verwendet werden, ist, dass die entsprechende Satzspanne beide zu untersuchenden Entitäten enthält. Die daraus resultierenden Verben und Adjektive werden für die Word-Cloud benutzt, welche wie vorher beschrieben aufgebaut wird.

Um die Bedienung weiterhin zu erleichtern, ist es auch möglich die Verben und Adjektive zu öffnen, indem der Nutzer auf die angezeigten Personen bzw. Orte unmittelbar neben den Lupen doppelklickt (siehe Abbildung 4.2). So muss der Nutzer nicht erst nach den entsprechenden Knoten im Entitätengraph suchen, was seinen Arbeitsfluss stören würde.

Um den Nutzer bei weiterführenden Analysen der Verben und Adjektive zu unterstützen, wird auch hier die Termfilterfunktion von DocuCompass (siehe Abbildung 4.3) übernommen. Hierzu kann der Nutzer mit der Maus die Tags der Word-Cloud hovern: In der Dokumentenspatialisierung werden die Dokumente entsprechend gefiltert und hervorgehoben, die das gehoverte Verb bzw. Adjektiv beinhalten. Ebenso funktioniert das Anpinnen des Verbs bzw. Adjektivs: Der Nutzer kann die Tags doppelklicken und damit den jeweiligen Tag an die Lupen anpinnen, so wie dies bisher funktionierte.

Nun kann der Nutzer die Lupen verwenden, um die durch den Termfilter markierten Dokumente näher zu untersuchen.

4.2.5 Explorer für Texte

Die bisher vorgestellten Methoden zielen darauf ab, dem Nutzer eine Möglichkeit zu bieten, schnell und effizient ein Verständnis der Dokumente mit ihren enthaltenen Entitäten und Beziehungen zu erlangen. Um tiefergehende Analysen zu ermöglichen, soll an dieser Stelle eine weitere Funktion vorgestellt werden: Der sogenannte „Textexplorer“. Hiermit kann der Nutzer die konkreten Texte der Dokumente öffnen und einsehen.

Hierfür teilt sich der Textexplorer den gleichen Bereich wie der Explorer für Verben und Adjektive (siehe Abbildung 4.6). Damit die Arbeit mit dem Explorer für Verben und Adjektive nicht beeinträchtigt wird, werden mehrere Tabs verwendet. Wenn der Textexplorer geöffnet wird, wird ein neuer Tab mit dem neuen Inhalt geöffnet, wobei jederzeit wieder auf vorige Tabs zurückgegriffen werden kann. Dieses Prinzip wird auch umgekehrt angewendet: Wird eine neue Word-Cloud mit Verben und Adjektiven geöffnet, dann wird eine neuer Tab verwendet, wobei auch hier jederzeit vorige Tabs mit Textexplorer-Inhalten aufgerufen werden können. Diese Tabs lassen sich bei Bedarf auch neu anordnen und schließen.

Die Dokumententexte lassen sich auf zahlreiche Weisen öffnen und besitzen je nach Kontext verschiedene Markierungen. Zunächst kann der Nutzer ein Dokument öffnen, indem er auf einen Knoten im Entitätengraph (siehe Abschnitt 4.2.3) doppelklickt. Um die Unterscheidung zu dem Explorer für Verben und Adjektive zu gewährleisten, gibt es verschiedene „Explorer-Modi“, die der Nutzer per Knopfdruck umschalten kann. Führt der Nutzer im „Textexplorer-Modus“ einen Doppelklick auf einen Knoten aus, werden diejenigen Dokumente angezeigt, die die Person oder den Ort enthalten, für den der Knoten steht. Hierbei werden die kombinierten Dokumentenmengen verwendet, die von den aktuell verwendeten Lupen fokussiert werden. Für jedes solche Dokument wird ein neuer Tab in der Explorer-Ansicht erstellt, der den entsprechenden Text des Dokuments enthält. Zusätzlich werden innerhalb des Textes die Vorkommen der ausgewählten Entität farbig markiert, damit der Nutzer relevante Textstellen schneller finden kann.

Dies funktioniert nicht nur für einzelne Entitäten, sondern auch für Entitätenpaare. Hierzu doppelklickt der Nutzer eine Kante im Entitätengraph. Die Dokumente, die das entsprechende Entitätenpaar enthalten, werden in neuen Tabs geöffnet und analog zu oben werden auch hier die entsprechenden Vorkommen der Personen bzw. Orte im Text farbig hervorgehoben, so dass die relevanten Textstellen einfacher aufzufinden sind.

Ähnlich wie beim Explorer für Verben und Adjektive (siehe Abschnitt 4.2.4) ist der Textexplorer auch aus der Dokumentenspatialisierung heraus aufrufbar. Hierzu kann der Nutzer die Personen bzw. Orte neben den Lupen (siehe Abbildung 4.2) doppelklicken und es öffnen sich analog zu oben die entsprechenden Textexplorer-Tabs mit den Markierungen.

Um auch aktuell betrachtete Verben und Adjektive näher untersuchen zu können, ist die Textexplorer-Funktion auch für die Word-Cloud der Verben und Adjektive verfügbar. Hierzu klickt der Nutzer doppelt auf das gewünschte Verb bzw. Adjektiv in der Word-Cloud. Es werden dann alle Dokumente

geöffnet, die die zugehörigen Entitäten bzw. Entitätenpaare enthalten *und* das ausgewählte Verb bzw. Adjektiv. Auch hier werden die Vorkommen im Text markiert, wobei Vorkommen von Verben und Adjektiven jeweils ihre eigene Farbe haben (die gleiche Farbe wie in der Word-Cloud für Verben und Adjektive), um die unterschiedlichen Arten der Vorkommen einfacher unterscheiden zu können.

4.3 Bezug zum Visualisierungsreferenzmodell

In diesem Abschnitt soll der Bezug zwischen den bisher vorgestellten Konzepten und dem Visualisierungsreferenzmodell hergestellt werden, das bereits in Abschnitt 3.1.1 erläutert wurde. Zusammengefasst beschreibt das Visualisierungsreferenzmodell wie die Visualisierung von Informationen als „Mapping“ von Daten zu visuellen Darstellungsformen betrachtet werden kann (siehe Abbildung 3.1). Im Folgenden wird beschrieben wie die einzelnen Teile des Modells durch unsere Konzepte umgesetzt werden.

Die unbehandelte Datenmenge (englisch: Raw Data) wird in unserem Fall durch den unbehandelten Dokumentensatz repräsentiert. Im Schritt der Datentransformation werden daraus die Datentabellen (englisch: Data Tables) erstellt: Für die einzelnen Dokumente wird jeweils der Dokumentenvektor berechnet, dessen Dimensionalität den enthaltenen Wörtern entspricht. Weiterhin werden mit Hilfe der vorgestellten NLP-Verfahren (siehe Abschnitt 3.2) sowohl die enthaltenen Entitäten als auch die dazugehörigen Verben und Adjektive extrahiert und gezählt. Diese Informationen werden in entsprechenden Datenstrukturen abgespeichert.

Mit Hilfe dieser Datentabellen werden die visuellen Strukturen (englisch: Visual Structures) erstellt: Aus den Dokumentenvektoren werden mittels einer Dimensionsreduktion x, y-Koordinaten berechnet, welche später für die Dokumentenspatialisierung verwendet werden. Außerdem werden für die gespeicherten Entitäten bzw. Entitätenpaare Knoten- und Kantentabellen erstellt, welche später für den Entitätengraph benötigt werden.

Aus diesen visuellen Strukturen werden schließlich die eigentlichen Darstellungen generiert (englisch: Views), in unserem Fall also die Dokumentenspatialisierung, die Entitätengraphen und die Word-Clouds, welche der Anwender für seine Analysen nutzen kann. Hierbei kann er mit diesen Visualisierungen interagieren, wie wir dies in den vorigen Abschnitten dieses Kapitels beschrieben haben. So kann er verschiedene Transformationen beeinflussen: Beispielsweise kann er die „View Transformation“ (zwischen den visuellen Strukturen und den eigentlichen Visualisierungen) beeinflussen, indem er Teile der Darstellungen zoomt oder verschiebt. Ebenso kann er die „Data Transformation“ (zwischen der unbehandelten Datenmenge und den Datentabellen) beeinflussen, indem er Filterfunktionen benutzt, wie die Lupen oder die vorgestellten Termfilter-Funktionen. Dieses Beeinflussen der Datentransformation beeinflusst wiederum die darauffolgenden Teile des Visualisierungsreferenzmodells: Durch die Lupen wird die Menge der Daten eingegrenzt, welche für die Visualisierungen verwendet wird. Durch die Termfilter-Funktionen können den Visualisierungen zusätzliche Markierungen hinzugefügt werden.

5 Implementierung

In diesem Kapitel der Ausarbeitung steht der Software-Prototyp im Mittelpunkt, der im Rahmen dieser Arbeit entwickelt wurde. Der Prototyp basiert hierbei auf den in Kapitel 4 vorgestellten Konzepten und setzt die dazugehörigen Funktionen um. Wir gehen dabei auf verschiedene Aspekte der Implementierung ein, wie eingesetzten Technologien, Struktur der Software und anderen Details der Implementierung.

5.1 Eingesetzte Technologien

Wie zuvor bereits erwähnt, wird in dieser Arbeit von DocuCompass als Fundament ausgegangen (siehe Abschnitt 2.1 und 4.1). Der Ansatz von DocuCompass ist dabei in der Programmiersprache **Java** geschrieben. Da dieser Ansatz mit den konzipierten Funktionen (siehe Abschnitt 4.2) erweitert wird, ist die Programmiersprache dieser Arbeit ebenfalls Java. Hierbei wird auf die Version Java JDK 8[1] zurückgegriffen.

Eine Besonderheit des Prototypen ist die Plattformunabhängigkeit, welche durch die Verwendung von Java ermöglicht wird. Da Java-Anwendungen in einer eigenen Laufzeit-Umgebung ausgeführt werden, ist die Ausführung weitgehend unabhängig vom vorliegenden Betriebssystem.

Für die Elemente der Bedienoberfläche, wie den Containern oder auch Bedienelementen wie Buttons usw., wird das GUI-Toolkit **Swing** für Java genutzt. Weiterhin wird ein Toolkit für die Visualisierung von verschiedenen Informationen verwendet: **Prefuse**[2] [HCL05]. Besonders beim Entitätengraphen, bei der Darstellung der Dokumentenspatialisierung und bei den Lupen kommt Prefuse mit seinen umfangreichen Frameworks zum Einsatz. Weiterhin wird speziell für die Berechnung der Dokumentenspatialisierung auf das Toolkit **Projection Explorer**[3] (PEx) [POM07] zurückgegriffen. Um die Interaktion mit den genannten Visualisierungen zu ermöglichen, wird das **Abstract Window Toolkit** (AWT) für Java benutzt. Ebenso kommt AWT stellenweise bei der Verwendung von grafischen Elementen wie Farben oder geometrischen Formen (Rechtecke usw.) zum Einsatz. Für die Verarbeitung der Dokumente und die dazugehörigen NLP-Aufgaben (siehe Kapitel 3.2) wird das **Stanford-CoreNLP-Toolkit**[4] [MSB+14] verwendet. Alle genannten Toolkits besitzen eine Java-API und sind somit gut für die Einbindung an das Fundament und für die umzusetzenden Funktionen geeignet.

[1] http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html
[2] http://prefuse.org/
[3] http://vicg.icmc.usp.br/vicg/tool/1/projection-explorer-pex
[4] https://stanfordnlp.github.io/CoreNLP/

Als Entwicklungsumgebung wurde **Eclipse**[5] verwendet. Weiterhin dient **Apache Subversion**[6] (SVN) zur Versionskontrolle.

Im Folgenden befindet sich eine Übersicht über die eingesetzten Werkzeuge:

- Abstract Window Toolkit (AWT)
- Apache Subversion (SVN)
- Eclipse
- Java JDK 8
- Prefuse
- Projection Explorer (PEx)
- Stanford CoreNLP
- Swing

5.2 Details zum Prototyp

Nachfolgend wird der implementierte Prototyp vorgestellt. Wir betrachten dabei die einzelnen Komponenten des Prototypen, gehen auf ihre Funktionsweisen ein, und betrachten Aspekte der Umsetzung.

5.2.1 Übersicht der Benutzeroberfläche

Zunächst soll ein Überblick über die Benutzeroberfläche gegeben werden. Generell werden verschiedenen Bereichen der Oberfläche bestimmte Funktionen zugewiesen. Diese Aufteilung richtet sich nach „Shneiderman's Visual Information Seeking Mantra" (siehe Abschnitt 3.1.2). In Abbildung 5.1 ist diese Aufteilung zu sehen und entsprechend markiert.

Im ersten Bereich (Abb. 5.1a) befindet sich die große Ansicht für die Dokumentenspatialisierung, wo der Nutzer mit den Lupen arbeitet. Diese Ansicht stellt zum einen den *Overview*-Teil von Shneiderman's Mantra dar, da hier der Nutzer eine komplette Übersicht über die Dokumente hat. Zum anderen setzt diese Ansicht auch den *Filter*-Aspekt des Mantras um, da hier der Nutzer mit Hilfe der Lupen die Dokumente filtert, welche für den Entitätengraphen und den Explorer verwendet werden.

Der zweite Bereich (Abb. 5.1b) enthält die Ansicht mit dem Entitätengraph. Dieser Graph wird aus den fokussierten Dokumenten der aktuell verwendeten Lupen generiert (siehe Abschnitt 4.2.3). In dieser Ansicht wird der *Zoom*-Aspekt von Shneiderman's Mantra umgesetzt: Hier kann der Nutzer zwar nicht direkt auf die Dokumente „zoomen", jedoch kann er hier die enthaltenen Entitäten untersuchen und entsprechende Teile des Graphen zoomen.

[5] https://www.eclipse.org/
[6] https://subversion.apache.org/

Der dritte Bereich (Abb. 5.1c) beherbergt die Exploreransicht, wo der Nutzer mehrere Tabs öffnen kann. Diese enthalten entweder Word-Clouds mit Verben und Adjektiven, oder die konkreten Texte der Dokumente (siehe Abschnitte 4.2.4 und 4.2.5). Diese Ansicht setzt den *Details-on-demand*-Aspekt von Shneiderman's Mantra um, denn hier kann der Nutzer Details zu gewünschten Entitäten aufrufen und diese genauer untersuchen.

Der vierte Bereich (Abb. 5.1d) bezieht sich auf die obere Leiste: Hier befindet sich die Toolbar, die verschiedene Hilfsfunktionen enthält, welche sich auf den Entitätengraphen und die Exploreransicht beziehen. Diese werden im Laufe des Kapitels noch beschrieben.

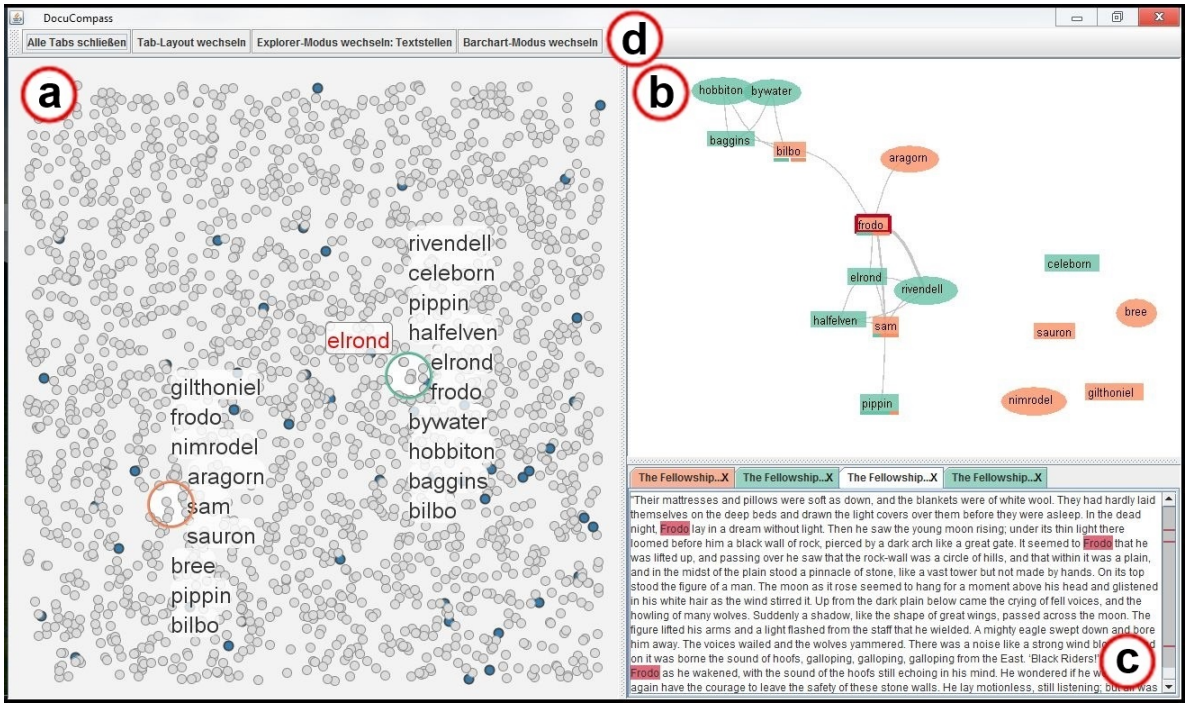


Abbildung 5.1: Überblick der Benutzeroberfläche: (a) Ansicht für die Dokumentenspatialisierung mit den Lupen, (b) Ansicht für den Entitätengraph, (c) Exploreransicht mit mehreren Tabs, (d) Toolbar mit verschiedenen Hilfsfunktionen

Diese beschriebenen Ansichten liegen in „Split-Panes", d. h. die Größe der jeweiligen Bereiche kann durch den entsprechenden „Divider" angepasst werden (ausgenommen der Toolbar).

5.2.2 Ansicht für die Dokumentenspatialisierung

Diese Ansicht ist eines der Herzstücke der Anwendung. Hier wird die Dokumentenspatialisierung dargestellt, welche Ausgangspunkt für die Exploration der Dokumente ist.

Für die Spatialisierung können TSV-Dateien verwendet werden (Abkürzung für „Tab-Separated Values", also Trennung durch Tabs), welche den Satz der Dokumente enthalten, die untersucht werden

sollen. Hierbei ist die Wahl der konkreten Dokumente dem Nutzer überlassen: Die Dokumente können eigenständig und unabhängig voneinander sein, oder sie können auch nur Teile eines größeren gemeinsamen Werks sein.

Beim Start der Anwendung wird dieser Satz an Dokumenten zunächst eingelesen und verarbeitet, d. h. es werden die verschiedenen Teile der NLP-Pipeline ausgeführt (siehe Abschnitt 3.2) und die entsprechenden Daten dazu werden abgespeichert. Hierzu gehört vor allem die Tokenisierung und die anschließende Extraktion von vorkommenden Entitäten (Personen und Orte). Ebenso findet auch die Segmentierung der Sätze statt, so dass die gemeinsamen Vorkommen der Entitäten innerhalb von bestimmten Satzspannen gezählt werden können. In dieser Arbeit wurde eine Spanne von 3 gewählt, d. h. es liegt ein gemeinsamer Zusammenhang vor, wenn die Entitäten innerhalb von 3 Sätzen zusammen vorkommen — diese Spanne ist jedoch auch anpassbar. Analog dazu werden auch die vorkommenden Verben und Adjektive extrahiert und ihre gemeinsamen Vorkommen mit anderen Entitäten bzw. Entitätenpaaren gezählt. All diese Daten werden beim Einlesevorgang berechnet und in passenden Datenstrukturen abgelegt. Hierbei gehören diese Daten immer zum jeweiligen Dokument, weshalb einfacherweise Hash-Maps zur Speicherung der Daten verwendet werden.

Dieser Einlesevorgang kann je nach Anzahl und Länge der Dokumente längere Wartezeiten verursachen. Das Einlesen von 150 Dokumenten, die jeweils die ungefähre Länge einer Kurzgeschichte besitzen, kann auf neueren Systemen ca. 30 Sekunden beanspruchen, und auf älteren 1-2 Minuten. Beim erstmaligen Einlesen des Datensatzes lässt sich dies nicht vermeiden. Es gibt jedoch auch die Möglichkeit der Serialisierung: Die berechneten Daten, die beim Einlesevorgang entstehen, können in Binärdateien abgespeichert werden. Wurde also ein Dokumentensatz einmal vollständig eingelesen und verarbeitet, ist es danach möglich den serialisierten Datensatz zu benutzen, ohne die Dokumente erneut einlesen zu müssen. Dieser Vorgang ist nahezu sofort fertig und ist eine deutliche Beschleunigung gegenüber dem erneuten Einlesen der Dokumentensätze.

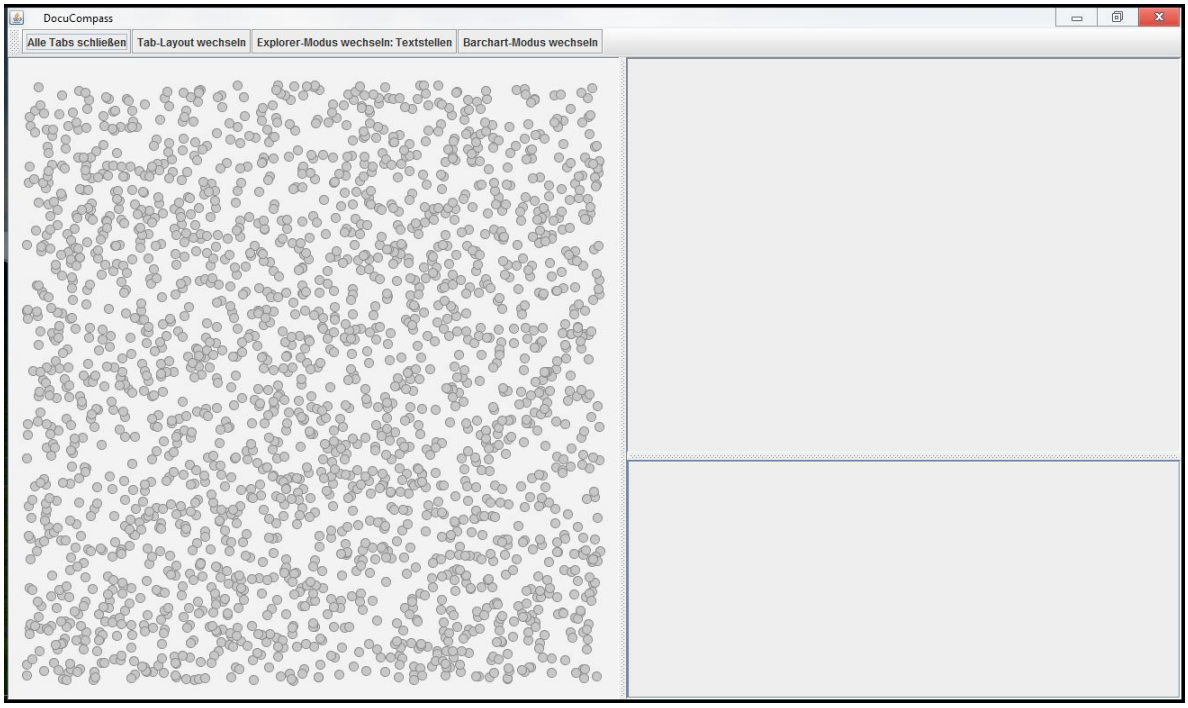
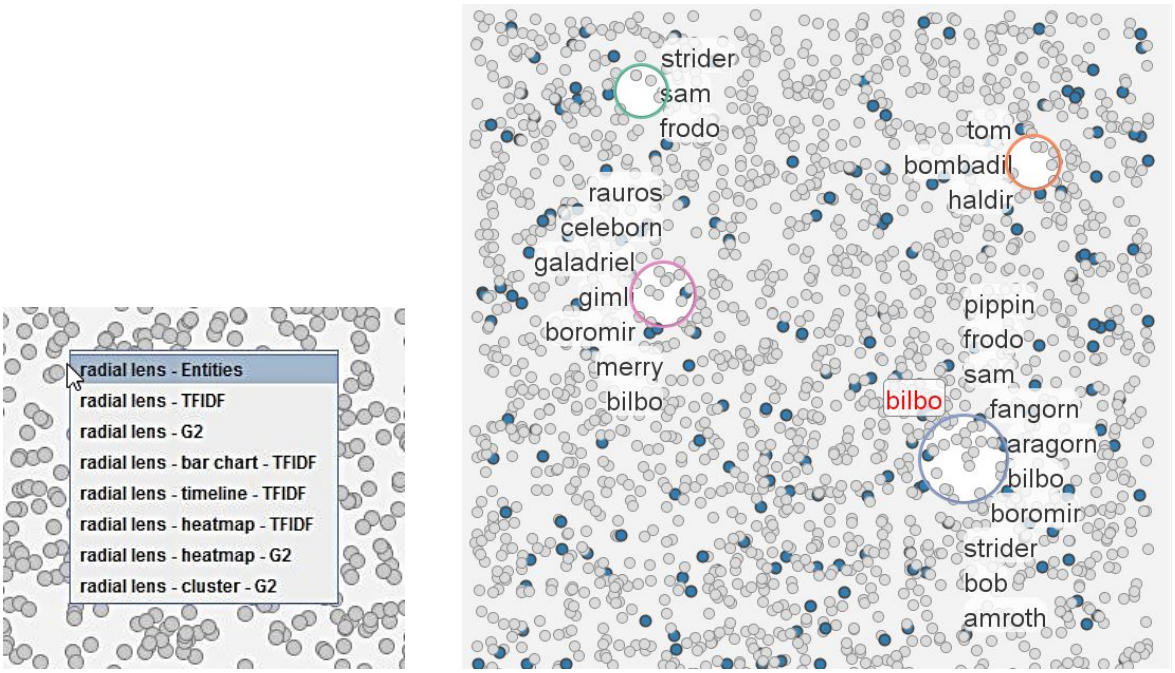


Abbildung 5.2: Start der Anwendung: Nach dem fertigen Einlesen der Dokumente wird die berechnete Dokumentenspatialisierung angezeigt. Die anderen Ansichten sind zunächst noch leer, da noch keine Lupen vorhanden sind.

Ist der Einlesevorgang der Dokumente abgeschlossen (bzw. das Einlesen der serialisierten Binärdatei), kann die Dokumentenspatialisierung berechnet werden und wird, wie in Abbildung 5.2 dargestellt, geöffnet. Hierbei sind die anderen Ansichten noch leer, da an dieser Stelle noch keine Lupen vorhanden sind. An dieser Stelle kann der Nutzer nun Lupen hinzufügen: Per Rechtsklick auf die Spatialisierung öffnet er eine Auswahl von verschiedenen Lupen, welche jeweils unterschiedliche Informationen und Details zu ihren fokussierten Mengen anzeigen (siehe Abbildung 5.3a). In dieser Arbeit neu, und im Folgenden genauer betrachtet, ist die Lupe „radial lens — Entities“: Diese zeigt die häufigsten Personen und Orte und wird außerdem für den Entitätengraphen verwendet. Die Funktionen der anderen Lupen werden im Rahmen dieser Arbeit nicht erklärt (jedoch nachzulesen in [HJH$^+$ar]). Die Lupe öffnet sich an der entsprechenden Stelle und kann nun für die Exploration der Dokumente verwendet werden. Der Nutzer kann auch weitere Lupen hinzufügen, oder bei Bedarf auch Löschen (per Tastendruck auf „Entfernen“). Jede Lupe hat hierbei ihre eigene Farbe (siehe Abschnitt 4.2.1). Die Lupen können per Mausrad vergrößert und verkleinert werden, und per Drag-and-Drop verschoben werden. Ein Beispiel mit mehreren Lupen ist in Abbildung 5.3b dargestellt.



(a) Das Auswahlmenü für die Lupen ist per Rechtsklick aufrufbar.

(b) Beispiel mit mehreren Lupen gleichzeitig (bis zu 5 werden unterstützt)

Abbildung 5.3: Arbeit mit der Dokumentenspatialisierung und den Lupen

5.2.3 Ansicht für den Entitätengraph

Nachdem der Nutzer, wie im vorigen Abschnitt beschrieben, die Lupen platziert hat, kann er nun in der nächsten Ansicht den Entitätengraphen nutzen, um enthaltene Personen und Orte und ihre Beziehungen zueinander zu erforschen (siehe Abschnitt 4.2.3). Hierbei platziert der Nutzer die Lupen über die gewünschte Menge an Dokumenten in der Spatialisierungsansicht, wobei ihn die nahezu beliebig anpassbaren Lupengrößen und -konstellationen unterstützen. Die so fokussierten Dokumentemengen werden in Kombination genutzt, um die entsprechenden Entitäten aus den gespeicherten Datenstrukturen zu filtern, die zuvor beim Einlesen der Dokumente erstellt wurden. Daraus wird schließlich der entsprechende Entitätengraph mit den Personen und Orten erzeugt. Dabei wird der Graph aktualisiert, sobald die Konstellation der Lupen verändert wird, d. h. beim Ändern der Position oder Größe einer Lupe, oder beim Hinzufügen oder Löschen einer Lupe. In Abbildung 5.4 ist ein Beispiel für einen solchen Entitätengraphen zu sehen, mit der entsprechenden Konstellation der Lupen in der Dokumentenspatialisierung.

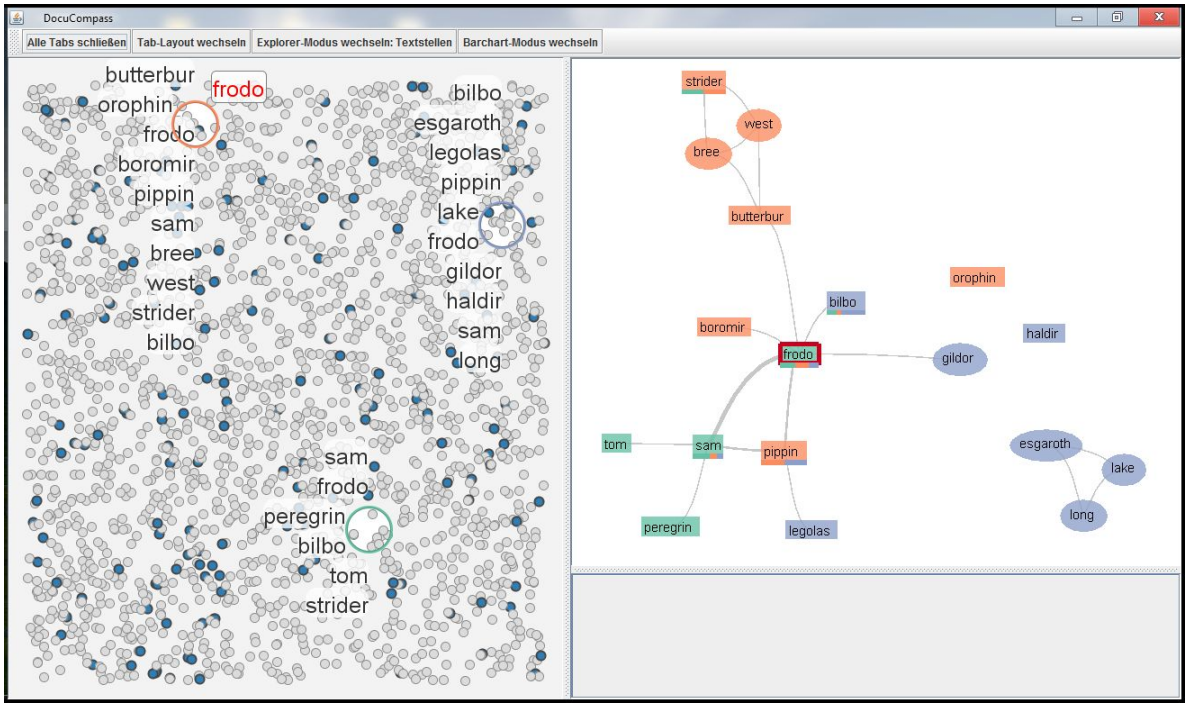


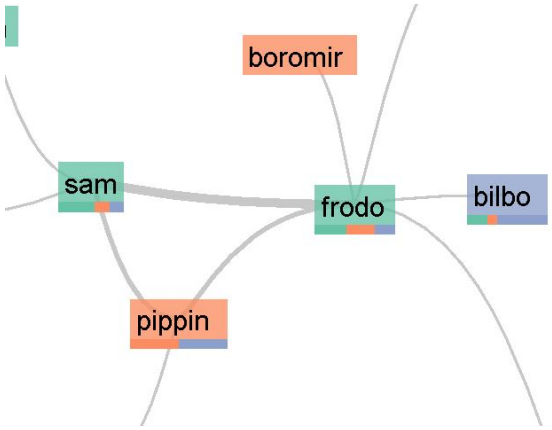
Abbildung 5.4: Beispiel der Ansicht für den Entitätengraphen: Der Entitätengraph wird aus den fokussierten Dokumentenmengen der aktuell verwendeten Lupen erzeugt.

Dabei haben die unterschiedlichen Entitätentypen jeweils eigene Glyphen: Personen sind als Rechtecke dargestellt und Orte als Ellipsen. So soll der Nutzer die verschiedenen Typen der Entitäten auf einen Blick schnell erkennen können.

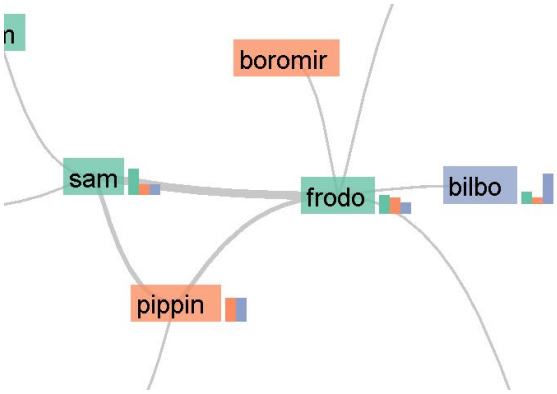
Die Knoten des Entitätengraphen werden zusätzlich eingefärbt: Je nachdem in welcher Lupe die jeweilige Entität vorkommt, wird die zugehörige Farbe der Lupe verwendet, um den Knoten im Graph einzufärben. Kommt beispielsweise die Person „Max Mustermann“ aus einem Dokument, das sich unter der orangenen Lupe befindet, so wird der Knoten im Graph entsprechend orange eingefärbt. So soll der Nutzer schnell einen Überblick über die Zusammensetzung des Entitätengraphen gewinnen können.

Für den Fall, dass eine Entität in mehreren Lupen gleichzeitig vorkommt, werden zusätzliche Balkendiagramme verwendet. Die eingefärbten Balken geben die Verteilung der Vorkommen in den verschiedenen Lupen an. Dabei kann der Nutzer zwischen zwei Optionen für die Position wählen, indem er den Button „Barchart-Modus wechseln“ in der Toolbar nutzt (Abbildung 5.5a). Zum einen können die Balkendiagramme unterhalb der Knoten angezeigt werden (Abbildung 5.5b). Zum anderen können sie auch neben den Knoten angezeigt werden (Abbildung 5.5c). Die erste Darstellungsart ist etwas kompakter, dafür ist bei der zweiten Darstellung der genaue Vergleich der Verteilungen einfacher. Das Umschalten der Darstellungen funktioniert dabei ohne weitere Verzögerungen, sobald der Button der Toolbar betätigt wird. Der Knoten selbst wird dabei entsprechend der Lupe eingefärbt, in der die Entität am meisten vorkommt.

(a) Der Button in der Toolbar zum Umschalten der Darstellung der Balkendiagramme



(b) Die Balkendiagramme können unterhalb der Knoten angezeigt werden…



(c) …oder neben den Knoten.

Abbildung 5.5: Balkendiagramme im Entitätengraph zur Darstellung der Verteilung der Vorkommen in den Lupen: Der Nutzer kann die Position der Darstellung bei Bedarf umschalten. Die Farbe der Knoten entspricht dabei den Lupen, in der sie am meisten vorkommen. Kommt eine Entität nur in einer Lupe vor, wird dessen Farbe verwendet.

Um weiterhin die Stärke der Zusammenhänge zwischen den Entitäten anzuzeigen, wird die Kantendicke angepasst: Je stärker der Zusammenhang ist (d. h. je öfter zwei Entitäten gemeinsam Vorkommen; siehe Abschnitt 4.2.3), desto dicker wird die entsprechende Kante im Graph (siehe Abbildung 5.6).

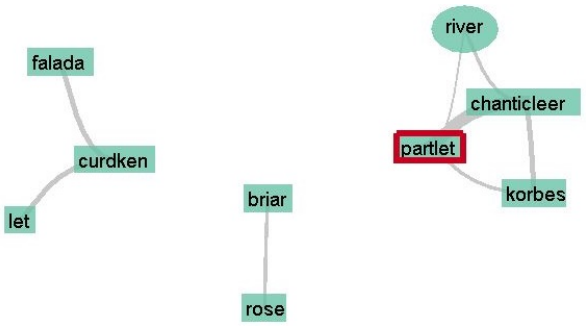
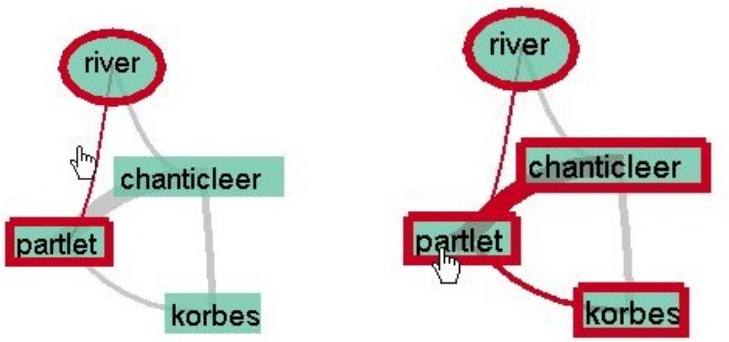


Abbildung 5.6: Die Kantendicke zeigt den Zusammenhang zwischen Entitäten an: Je dicker eine Kante ist, desto höher ist der Zusammenhang zwischen den Entitäten.

Der Nutzer hat außerdem die Möglichkeit, angrenzende Knoten hervorzuheben. Hierzu kann er entweder einen Knoten oder auch eine Kante mit der Maus hovern: Die zugehörige(n) Kante(n) werden rot markiert und die angrenzenden Knoten werden mit einer roten Umrandung hervorgehoben (siehe Abbildung 5.7). So soll der Nutzer auch in komplexeren Graphen auf einfache Weise die Zusammenhänge und Beziehungen zwischen den Entitäten identifizieren können. Um dem Nutzer einen ersten Anhaltspunkt für seine Analysen zu bieten, wird zusätzlich bei jedem Aktualisieren des Graphen diejenige Entität mit einer roten Umrandung hervorgehoben, die insgesamt am meisten in allen Lupen vorkommt.



(a) Hovern einer Kante **(b)** Hovern eines Knoten

Abbildung 5.7: Highlighting-Funktion des Entitätengraphen: Durch Hovern können die angrenzenden Knoten und Kanten hervorgehoben werden, wodurch der Nutzer auch bei komplexeren Graphen, die einzelnen Beziehungen der Entitäten erkennen kann.

Der Entitätengraph bietet auch mehrere Interaktionsmöglichkeiten an, die auch schon in Abschnitt 4.2.3 beschrieben wurden: Neben Zoomen und Verschieben des gezeigten Ausschnitts und den Termfilter-Funktionen, kann der Nutzer auch einzelne Knoten des Graphen per Drag-and-Drop verschieben. Die zunächst unausgeglichenen Kräfte des darunterliegenden kräftebasierten Graphen

gleichen sich daraufhin dynamisch wieder aus, so dass wieder ein neuer ausgeglichener Graph entsteht, der durch den Nutzer angepasst ist. So hat der Nutzer auch die Möglichkeit den Graphen, unter Beachtung der Kräfte des Force-Directed-Layouts, anzupassen, um so eventuell einfacher bestimmte Bereiche des Graphen zu untersuchen oder auch miteinander vergleichen zu können.

5.2.4 Ansicht für den Explorer

Die hier beschriebene Ansicht enthält sowohl den Explorer für Verben und Adjektive als auch den Explorer für Texte. Wie in den Abschnitten 4.2.4 und 4.2.5 bereits beschrieben, lassen sich diese sowohl für Entitäten als auch für Entitätenpaare auf verschiedene Weisen öffnen:

- In der Dokumentenspatialisierung:
  - Doppelklick auf den Namen der Person bzw. des Orts neben den Lupen
- Im Entitätengraph:
  - Doppelklick auf den entsprechenden Knoten der Entität
  - Doppelklick auf die entsprechende Kante des Entitätenpaars

Je nachdem welcher „Explorer-Modus“ eingestellt ist, öffnet sich ein neuer Tab bzw. mehrere neue Tabs in der Exploreransicht mit dem entsprechenden Inhalt. Der Explorer-Modus kann mit dem Button „Explorer-Modus wechseln:…“ in der Toolbar eingestellt werden. Hierbei ändert ein Klick des Buttons den Modus und zeigt den aktuellen Modus auch im Button selbst an. Im Modus „Verben und Adjektive“ (siehe Abbildung 5.8a) öffnet sich die Word-Cloud mit den entsprechenden Verben und Adjektiven. Im Modus „Textstellen“ (siehe Abbildung 5.8b) öffnen sich die Textinhalte der zugehörigen Dokumente unter den Lupen.

(a) In diesem Modus lassen sich die Word-Clouds mit Verben und Adjektiven öffnen.

(b) In diesem Modus lassen sich die Textinhalte öffnen.

Abbildung 5.8: Mit dem Button in der Toolbar lässt sich der Explorer-Modus umschalten. Je nach Modus öffnet sich die Word-Cloud mit Verben und Adjektiven, oder die Textinhalte der entsprechenden Dokumente.

Explorer für Verben und Adjektive

Öffnet der Nutzer für eine Entität bzw. für ein Entitätenpaar den Explorer für Verben und Adjektive, so wird in der Exploreransicht ein neuer Tab hinzugefügt. Dieser Tab wird nach der Entität benannt bzw. nach dem Entitätenpaar, so dass der Nutzer ihn anhand dessen wieder zuordnen kann. Zusätzlich wird der Tab eingefärbt:

- Bei einer Entität wird die Farbe der Lupe verwendet, in der die Entität am meisten vorkommt (entspricht also der Farbe des Knotens im Entitätengraph).
- Bei einem Entitätenpaar wird die Farbe der Lupe verwendet, in der die beiden Entitäten zusammen vorkommen.

Die Farbe des selektierten Tabs ist hierbei weiß. Ein Beispiel dafür ist in Abbildung 5.9 zu sehen.

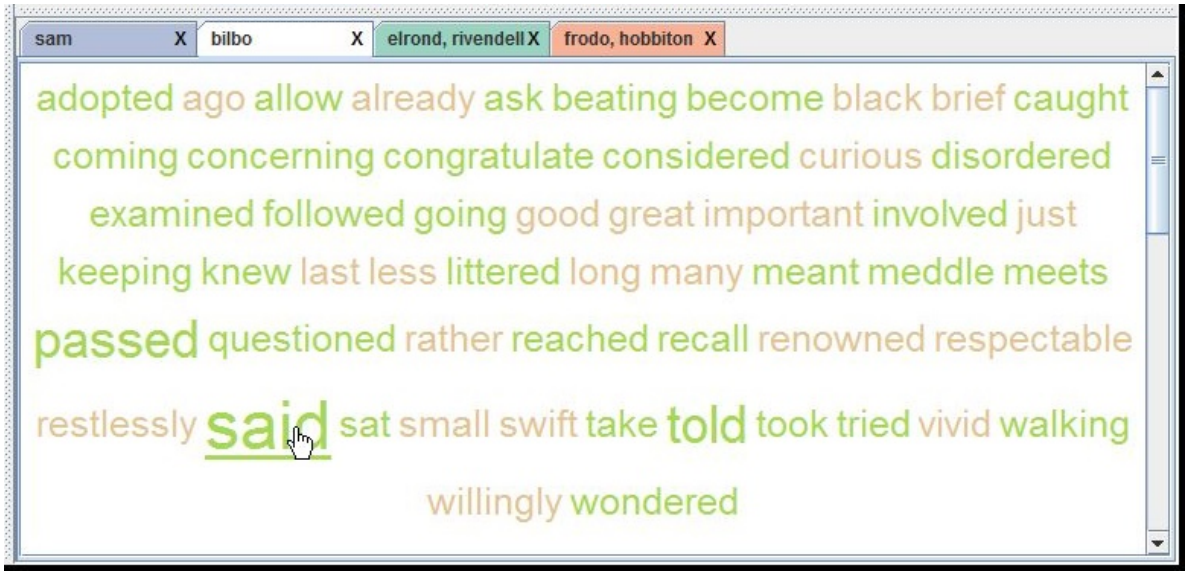


Abbildung 5.9: In der Exploreransicht lassen sich die Word-Clouds mit Verben und Adjektiven öffnen, welche sich in eigenen Tabs befinden. Die Tabs sind nach den zugehörigen Entitäten bzw. den Entitätenpaaren benannt und entsprechend der Vorkommen in den Lupen eingefärbt.

Die geöffnete Word-Cloud kann der Nutzer nun einsehen. Hierbei kann er für weiterführende Analysen auch die Termfilter-Funktionen nutzen, indem er die Tags entweder hovert oder auch doppelklickt (siehe Abschnitt 4.2.4).

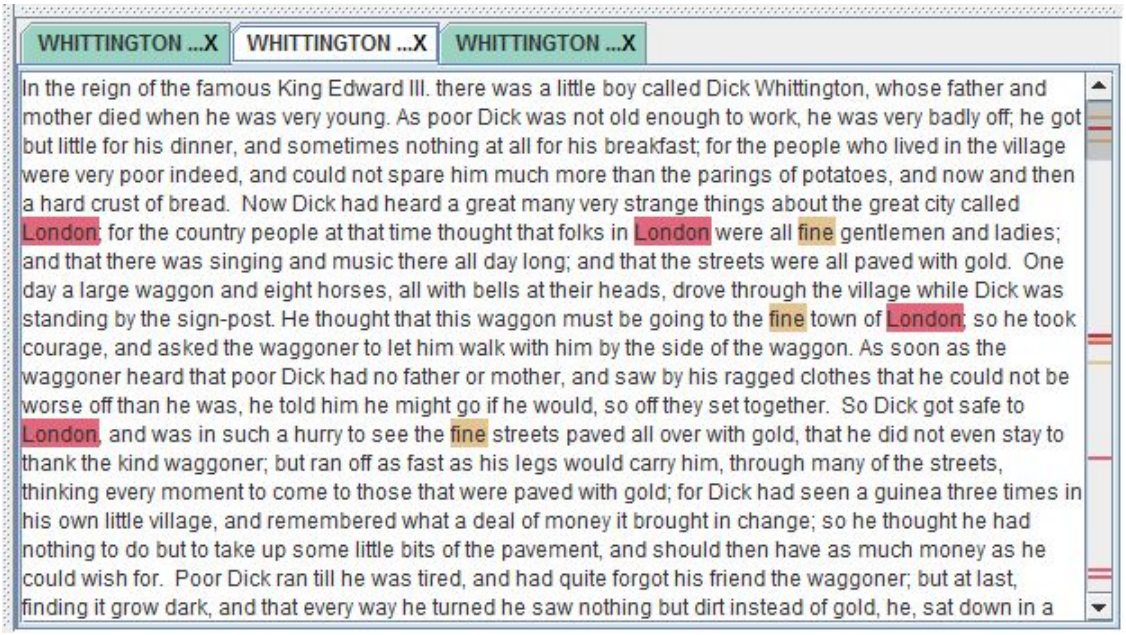
Explorer für Texte

Analog zu oben kann der Nutzer auch gewünschte Textinhalte öffnen. Hierbei unterscheidet sich der Vorgang jedoch leicht: Da möglicherweise eine Entität bzw. ein Entitätenpaar in mehreren Dokumenten vorkommt, werden eventuell mehrere Tabs gleichzeitig geöffnet. An dieser Stelle ist das Benennen der Tabs nach der Entität bzw. dem Entitätenpaar suboptimal, weshalb die Tabs stattdessen nach den jeweiligen Dokumenten benannt sind. Abhängig davon von welcher Lupe die jeweiligen Dokumente fokussiert sind, werden die entsprechenden Lupenfarben zum Einfärben der Tabs verwendet. Ansonsten werden Tabs mit Dokumententexten genauso hinzugefügt wie Tabs mit Verben und Adjektiven.

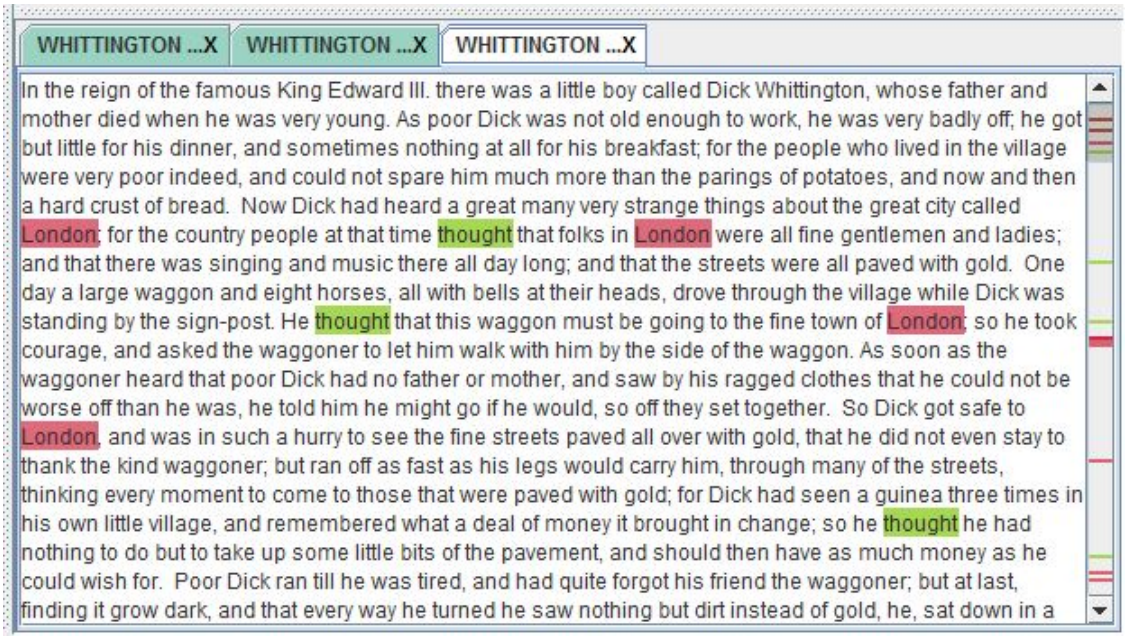
Der Textexplorer kann für Entitäten, Entitätenpaare, Verben oder Adjektive geöffnet werden (siehe Abschnitt 4.2.5). Hierbei werden in den geöffneten Dokumenten die Vorkommen der genannten Elemente mit unterschiedlichen Farben markiert:

- Entitäten werden rot markiert.
- Verben werden grün markiert (wie in der Word-Cloud).
- Adjektive werden beige markiert (wie in der Word-Cloud).

Neben den eigentlichen Textstellen werden auch Markierungen in der Scrollleiste mit der entsprechenden Farbe gesetzt, so dass die relevanten Textstellen auch bei langen Dokumenten schnell aufzufinden sind (siehe Abbildung 5.10). Ist der Text eines Dokuments so kurz, dass er vollständig in den Textexplorerbereich passt, wird keine Scrollleiste angezeigt und die genannten Markierungen in der Scrollleiste entfallen dementsprechend.



(a) Wird der Textexplorer für ein Adjektiv geöffnet, so wird das Adjektiv beige markiert und die dazugehörige Entität rot markiert.



(b) Wird der Textexplorer für ein Verb geöffnet, so wird das Verb grün markiert und die dazugehörige Entität rot markiert.

Abbildung 5.10: Je nachdem für welches Element der Textexplorer geöffnet wird, enthalten die geöffneten Dokumente unterschiedliche Markierungen, so dass der Nutzer die relevanten Textstellen schneller auffinden kann. Die Markierungen befinden sich ebenfalls in der Scrollleiste, falls eine vorhanden ist.

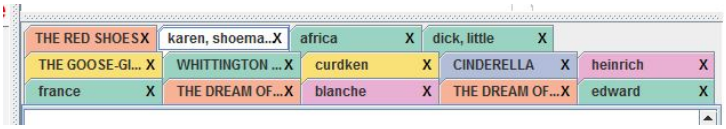
Organisation der Tabs

Im Falle, dass bei Analyse-Arbeiten viele Tabs anfallen, werden dem Nutzer einige Möglichkeiten geboten, um die Tabs in seinem Sinne zu organisieren. Zunächst sind die Tabs per Drag-and-Drop verschiebbar und können beliebig angeordnet werden. Weiterhin gibt es die Möglichkeit, die Tab-Leiste selbst anzupassen: Hierzu drückt der Nutzer den Button „Tab-Layout wechseln“ in der Toolbar (siehe Abbildung 5.11a). Der Nutzer kann damit entweder die vollständige Tab-Leiste mit allen Tabs anzeigen (Abbildung 5.11b) oder nur einen Ausschnitt davon (Abbildung 5.11c). Bei Letzterem werden zusätzliche Buttons angezeigt, damit der Nutzer innerhalb der Tab-Leiste durch die Tabs scrollen kann.

Ist die Breite eines Tabs nicht ausreichend, um den vollständigen Namen des Tabs anzuzeigen, so hat der Nutzer die Möglichkeit per Tooltip dessen vollständigen Namen zu erfahren (Abbildung 5.11d).

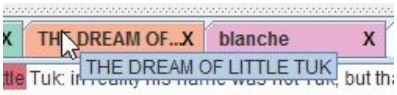
Schließlich kann der Nutzer den „X“-Button der Tabs nutzen, um einzelne Tabs zu schließen, oder auch den Button „Alle Tabs schließen“ in der Toolbar (Abbildung 5.11a), um alle Tabs zu schließen und so die Exploreransicht mit einem Klick zurückzusetzen.

(a) Buttons in der Toolbar zur Manipulation der Tab-Leiste in der Exploreransicht



(b) Tab-Layout mit allen Tabs der Tab-Leiste

(c) Tab-Layout mit einem Ausschnitt der Tab-Leiste und zusätzlichen Buttons zum Scrollen in der rechten Ecke



(d) Tooltip mit vollständigem Namen des Tabs

Abbildung 5.11: Der Nutzer hat verschiedene Möglichkeiten, um die Tab-Organisation der Exploreransicht nach seinem Wunsch anzupassen.

6 Anwendungsbeispiele und Diskussion

In diesem Kapitel werden einige Anwendungsfälle vorgestellt, um das Verständnis und die Verwendung der entwickelten Funktionen zu erleichtern. Anhand der Anwendungsfälle findet im Anschluss dazu eine Diskussion des entwickelten Ansatzes statt: Es wird auf einige Vor- und Nachteile eingegangen und im Zusammenhang dazu werden eventuelle Verbesserungsmöglichkeiten herausgearbeitet.

Bei den Anwendungsfällen gehen wir jeweils davon aus, dass ein fiktiver Literaturwissenschaftler eine bestimmte Aufgabe im Zusammenhang zu einem literarischen Werk bearbeiten möchte. Im Folgenden erklären wir, wie der entwickelte Ansatz dieser Arbeit ihn dabei unterstützt.

6.1 Literaturlandkarte von „Der Herr der Ringe“

In diesem ersten Anwendungsszenario betrachten wir den Roman „Der Herr der Ringe“ von J. R. R. Tolkien, welcher in den Jahren 1954/1955 in mehreren Bändern veröffentlicht wurde. In diesem Roman geht es um einen Ring, von dessen Vernichtung das Schicksal von Mittelerde abhängt, dem Schauplatz des Romans und ausgeklügelte Fantasiewelt von Tolkien. Mehrere Figuren sind in dieser Geschichte verwickelt, welche aus verschiedenen Völkern stammen, die in Mittelerde leben, darunter Menschen, Elben, Hobbits, Zwerge und Zauberer. Die Hauptfiguren sind hierbei vier Hobbits, denen die Aufgabe zugeteilt wird, den besagten Ring zu zerstören. Von mehreren Vertretern der oben genannten Völker begleitet, reisen sie also zum Schicksalsberg, denn nur dort kann der Ring zerstört werden. Während dieser Mission passieren sie unterschiedliche Schauplätze in Mittelerde und müssen zahlreiche Gefahren und Bedrohungen bewältigen.

Unseren fiktiven Literaturwissenschaftler interessiert es, wie verschiedene Autoren unterschiedliche literarische Räume erfinden. Hierzu möchte er literarische Werke explorieren und daraus Literaturlandkarten erstellen, die er für weitere Forschungen verwenden kann. Die Literaturlandkarten sollen dabei den imaginären Schauplätzen der Werke entsprechen und Informationen zu den verschiedenen Stationen des Handlungsverlaufs enthalten.

In unserem Beispiel möchte unser Literaturwissenschaftler also eine Literaturlandkarte von „Der Herr der Ringe“ mit seiner vielfältigen Welt erstellen. Er hat den Roman vor längerer Zeit schon gelesen und hat daher schon Vorwissen zu der Geschichte. Als Unterstützung verwendet er den in dieser Arbeit entwickelten Ansatz zur Textexploration.

Als Grundlage für die Exploration dient ein Dokumentensatz, der aus den einzelnen Abschnitten des Romans besteht. Hierbei entspricht jeder Abschnitt des Romans jeweils einem einzelnen Dokument. Die Exploration beginnt damit, dass unser Analyst eine Lupe auf der Dokumentenspatialisierung platziert. Nachdem er für kurze Zeit die Lupe auf der Spatialisierung bewegt, stößt er auf Rivendell

(deutsch: Bruchtal), welche als eine der häufigsten Entitäten neben der Lupe angezeigt wird (siehe Abbildung 6.1).

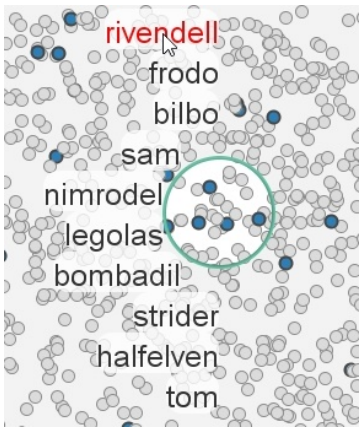


Abbildung 6.1: Mit Hilfe der Lupen erkennt der Analyst die häufigsten Entitäten innerhalb der fokussierten Menge.

Er erinnert sich zwar, dass Rivendell in der Geschichte vorkommt, jedoch weiß er nicht mehr, welche Rolle es in der Handlung spielt. Deshalb wählt er Rivendell neben der Lupe aus, damit der entsprechende Knoten im Entitätengraph selektiert wird. Beim Begutachten des Entitätengraphen bemerkt er, dass eine starke Relation zwischen Rivendell und Frodo besteht, welcher eine der Hauptpersonen im Roman ist (siehe Abbildung 6.2). Ebenso bemerkt unser Analyst eine starke Relation zwischen Frodo und Bilbo, welcher Frodos Onkel und Adoptivvater ist.

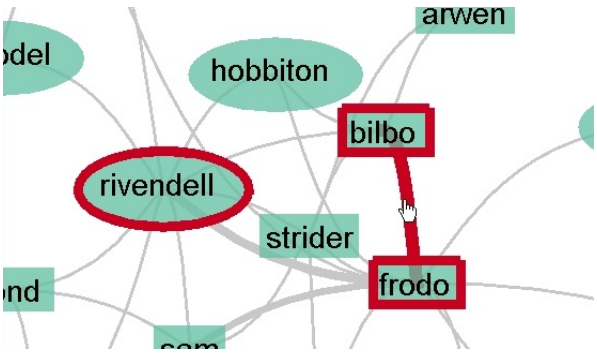


Abbildung 6.2: Im Entitätengraph kann der Anwender besonders starke Zusammenhänge zwischen den Personen und Orten feststellen.

Da ihn dies wundert, selektiert er die entsprechende Kante im Entitätengraph, während der Explorermodus für Textstellen aktiviert ist. Nun öffnen sich in der Exploreransicht die entsprechenden Textstellen mit den Vorkommen des Entitätenpaares. Hierbei entdeckt unser Analyst die Textstelle,

welche in Abbildung 6.3 dargestellt ist: Es handelt sich um die Stelle in der Geschichte, wo sich Frodo und Bilbo das erste Mal in Rivendell wiedersehen, seitdem sie aus ihrer Heimat aufgebrochen sind.

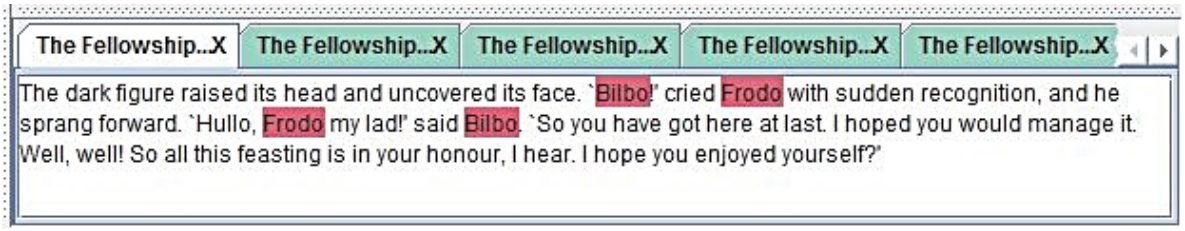


Abbildung 6.3: Der Anwender kann für tiefergehende Analysen den Textexplorer nutzen.

Daraufhin erinnert er sich, dass Rivendell auch der Ort ist, wohin Frodo gebracht wird, nachdem er von den Nazgûl, Dienern von Sauron, dem ursprünglichen Besitzer des Rings, beinahe getötet wird. Um dies zu verifizieren, untersucht unser Analyst weitere Textstellen: Zunächst aktiviert er den Termfilter, indem er Rivendell an die Lupe anpinnt (siehe Abbildung 6.4). Dann fügt er eine weitere Lupe hinzu, um die nun hervorgehobenen Dokumente in der Spatialisierung zu untersuchen.

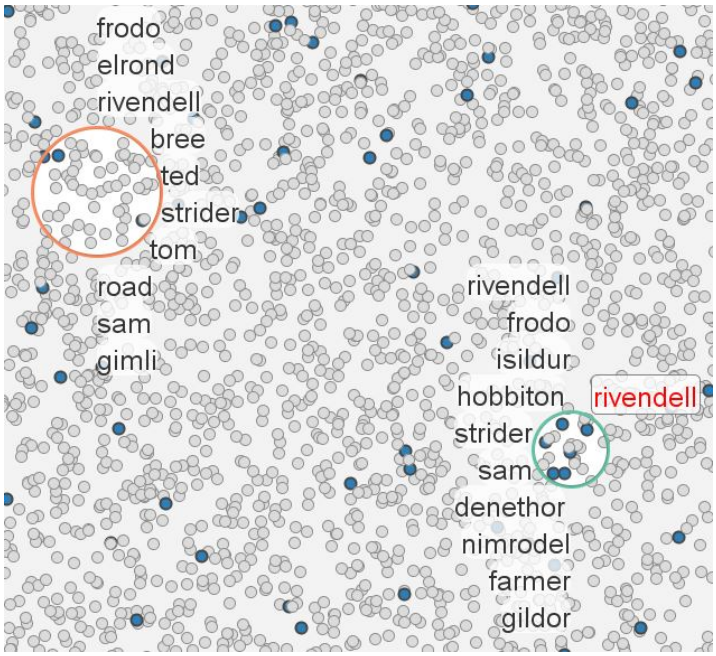
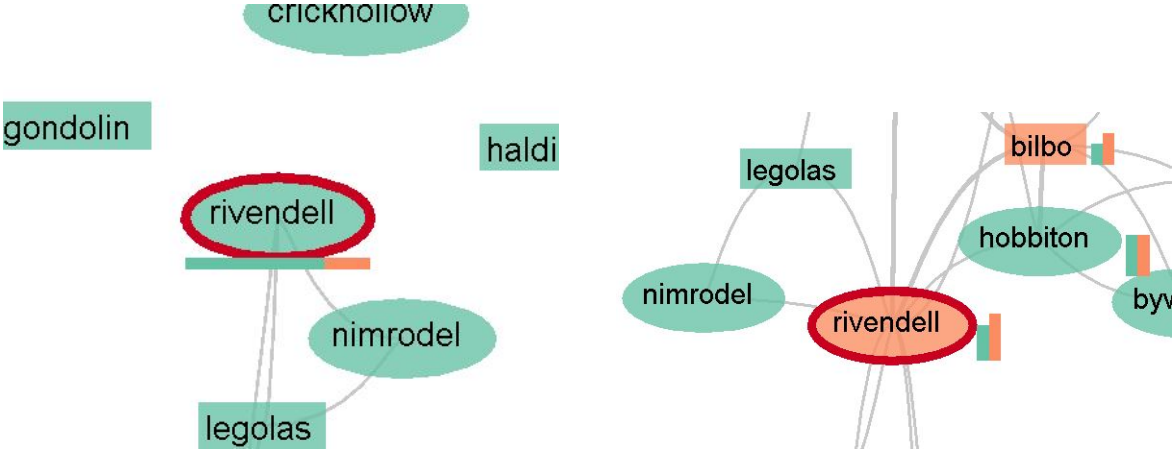


Abbildung 6.4: Um weiterführende Untersuchungen durchzuführen, können weitere Lupen und die Termfilter-Funktion genutzt werden.

Durch das Hinzufügen der neuen Lupe ändert sich zunächst der Entitätengraph. Unser Analyst sieht durch die Balkendiagramme unterhalb der Knoten, dass die Verteilung von Rivendell in der ersten

Lupe (grün) stärker ist, und dass in der zweiten Lupe (orange) noch wenig Vorkommen von Rivendell vorhanden sind (siehe Abbildung 6.5). Da er außerdem die Textstellen, die er zuvor gefunden hat, mit dem dazugehörigen Kontext im Entitätengraphen beibehalten möchte, belässt er die Position der ersten Lupe und verwendet nur die zweite Lupe für die weitere Untersuchung. Hierzu vergrößert er zusätzlich die zweite Lupe, um eine größere Anzahl von Dokumentenglyphen fokussieren zu können, welche durch den zuvor aktivierten Termfilter markiert sind. So kann unser Analyst die fokussierte Dokumentenmenge mit denjenigen Dokumenten erweitern, die für ihn relevant sind, ohne den vorher betrachteten Bereich zu verlieren.



(a) Die erste Lupe (grün) enthält zunächst noch mehr Vorkommen von Rivendell als die zweite Lupe (orange).

(b) Nachdem mehr relevante Dokumente durch die zweite Lupe fokussiert werden, erhöht sich dessen Anteil für Rivendell. Um die Verteilungen besser einschätzen zu können, kann unser Analyst auch den Barchart-Modus wechseln.

Abbildung 6.5: Die Balkendiagramme unterhalb bzw. neben den Knoten geben unserem Anwender Aufschluss über die Verteilung der Entitäten innerhalb der aktuellen Lupen. So kann er z. B. diejenige Lupe erkennen, die noch wenig Vorkommen von einer bestimmten Entität besitzt, und diese dann für dessen weitere Analyse verwenden.

Nun öffnet er die Textstellen, in denen Rivendell vorkommt, indem er den entsprechenden Knoten im Entitätengraph selektiert. Er findet nach einigem Suchen die in Abbildung 6.6 gezeigte Textstelle, welche seine Vermutung bestätigt. Rivendell ist tatsächlich der Ort, welcher Frodo zur Rettung verhilft, als er von den Nazgûl verfolgt wird.

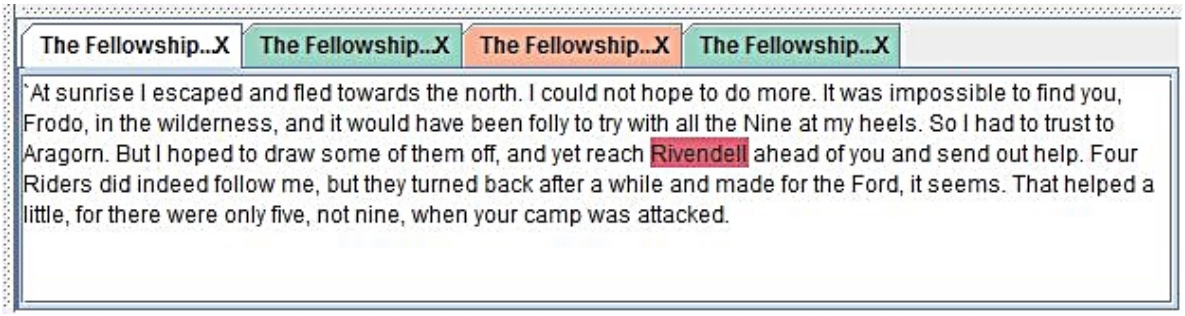


Abbildung 6.6: Der Analyst kann auch Vermutungen mit Hilfe des Textexplorers bestätigen.

Nach weiteren Untersuchungen entdeckt unser Analyst auch Textstellen, in der Informationen zu der Lage von Rivendell gegeben werden (siehe Abbildung 6.7). Mit Hilfe der genannten Informationen, welche er auf die bis hierher beschriebene Weise herausfindet, kann er Rivendell im Handlungsverlauf des Romans einordnen und auf der Literaturlandkarte eintragen.

Für die weitere Exploration kann unser Analyst mit Hilfe der Highlighting-Funktion des Entitätengraphen auch andere Orte identifizieren, die im Zusammenhang mit Rivendell auftreten (z. B. Bruinen, siehe Abbildung 6.8). So kann er die „Umgebung“ von Rivendell auf der Literaturlandkarte erforschen, welche eventuell weitere wichtige Schauplätze enthält, indem er für die hervorgehobenen Orte die Termfilter- und Explorer-Funktionen nutzt. Ihm steht jedoch auch frei, mit Hilfe der Lupen an einer anderen Stelle mit der Exploration fortzufahren. So gelangt er Schritt für Schritt zu einer vollständigen Literaturlandkarte, die die Schauplätze des Romans und ihre jeweiligen Rollen für die Handlung enthält.

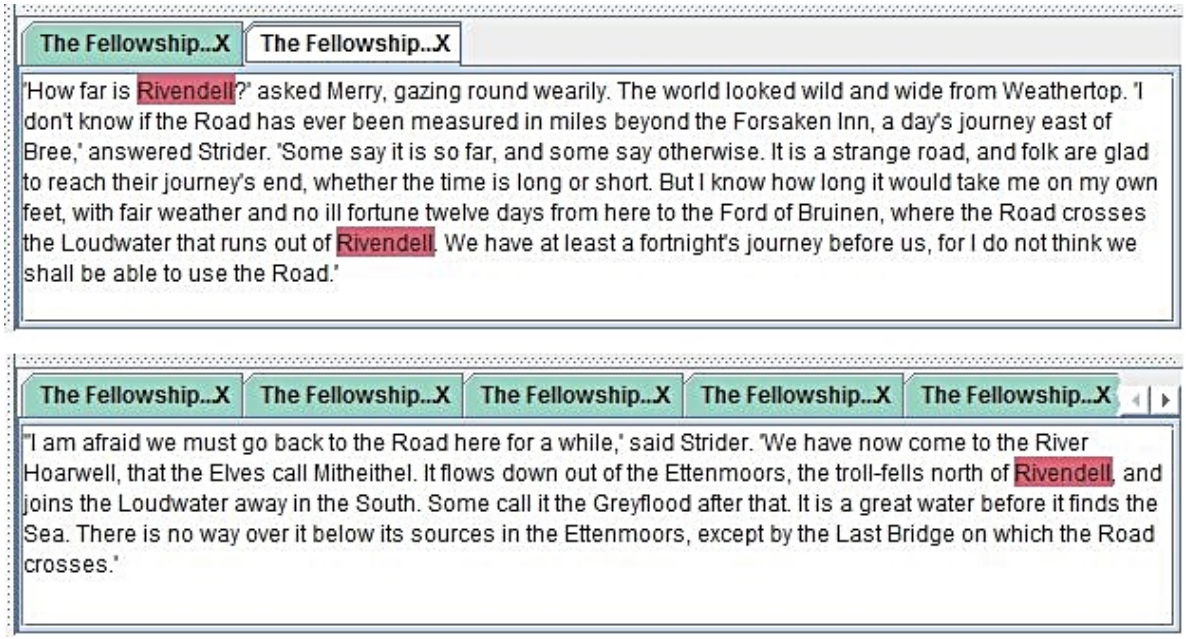


Abbildung 6.7: Die Textstellen können auch Angaben zu der Lage des jeweiligen Schauplatzes liefern, was nützlich für das Erstellen der Literaturlandkarte ist.

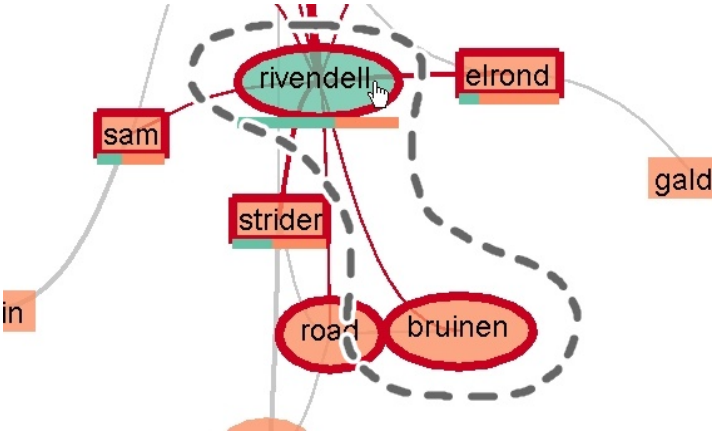


Abbildung 6.8: Mit Hilfe der Highlighting-Funktion des Entitätengraphen lassen sich die Orte identifizieren, die einen Zusammenhang mit der betrachteten Entität haben. Hier ist ein Beispiel Bruinen: Der Nutzer kann mit Hilfe des hervorgehobenen Knotens die Termfilter-Funktionen oder auch die Explorer-Funktionen nutzen, und so weiterführende Analysen zu Bruinen durchführen.

6.2 Personennetzwerk von „Harry Potter“

In dem zweiten Anwendungsszenario betrachten wir die Romanreihe „Harry Potter“ von J. K. Rowling, dessen sieben Bänder in den Jahren von 1997 bis 2007 veröffentlicht wurden. In der Reihe geht es um den jungen Harry Potter und sein Leben in der Zauberschule Hogwarts, welche er das erste Mal besucht als er 11 Jahre alt ist. Die sieben Bänder enthalten jeweils die Geschichte eines Schuljahres seines dortigen Aufenthalts und erzählen von den Bedrohungen und Gefahren, die er mit seinen Freunden, Ron Weasley und Hermione Granger, bewältigen muss. Dafür verantwortlich ist der mächtige dunkle Magier Lord Voldemort, welcher Harry Potter nach dem Leben trachtet. Lord Voldemorts Ziel ist es, unsterblich zu werden und alle Zauberer und „Muggles“ (Nicht-Zauberer) zu unterjochen. Die Schuljahre Harry Potters sind also gezeichnet von seinen Konfrontationen mit dem dunklen Lord, welchem er im Laufe der Geschichte immer wieder begegnet und bekämpfen muss.

Unser Analyst hat die Romanreihe vor längerer Zeit schon gelesen und möchte nun in diesem Fall die Charaktere näher untersuchen, die in der Geschichte vorkommen. Hierzu gehören Harry Potter selbst, seine Freunde und die vielen unterschiedlichen Mitschüler in Hogwarts. Weitere wichtige Personen sind sowohl der Schule zugehörig (wie die Lehrer, die Schulleitung und andere Bedienstete der Schule) als auch Personen außerhalb des schulischen Lebens (Verwandte von Harry, Verwandte der Freunde und andere unabhängige Personen). Um die zahlreichen Personen der Geschichte besser verstehen zu können, möchte unser Analyst ein Personennetzwerk erstellen, welches Profile zu den verschiedenen Charakteren enthält, mit ihren Bedeutungen für die Handlung. Weiterhin soll es auch die Beziehungen zwischen den Personen enthalten und wie sie innerhalb der Geschichte zueinander stehen. Zur Unterstützung wird auch hier unser entwickelter Ansatz zur Textexploration verwendet.

Als Erstes platziert unser Analyst eine Lupe auf der Dokumentenspatialisierung, bei der die einzelnen Abschnitte des Romans als Dokumente dienen, ähnlich wie im vorigen Beispiel in Abschnitt 6.1. Nachdem er die Lupe für kurze Zeit über die Spatialisierung bewegt, findet er einige Dokumente, in denen der Hauptprotagonist Harry vorkommt. Erkenntlich ist dies an dem Tag „Harry“, welcher neben der Lupe als eine der häufigsten Entitäten der fokussierten Menge angezeigt wird (siehe Abbildung 6.9).

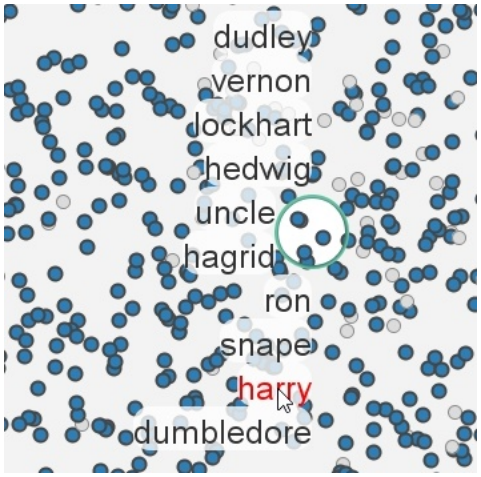


Abbildung 6.9: Mit Hilfe der Lupen findet der Analyst den Hauptprotagonisten der Geschichte, Harry.

Der Analyst begutachtet daraufhin den entsprechenden Entitätengraphen, welcher aus der aktuell fokussierten Menge erzeugt wird. Darin erkennt er eine Relation zwischen Harry und Vernon. Er kann sich nicht mehr erinnern, um welche Person es sich bei Vernon handelt, weshalb er die Ansicht im Entitätengraph auf den entsprechenden Teil des Graphen verschiebt und darauf zoomt, um diese Verbindung näher zu untersuchen (siehe Abbildung 6.10).

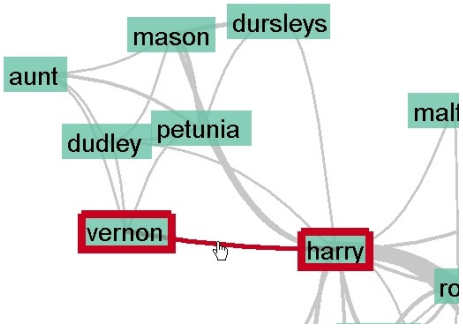
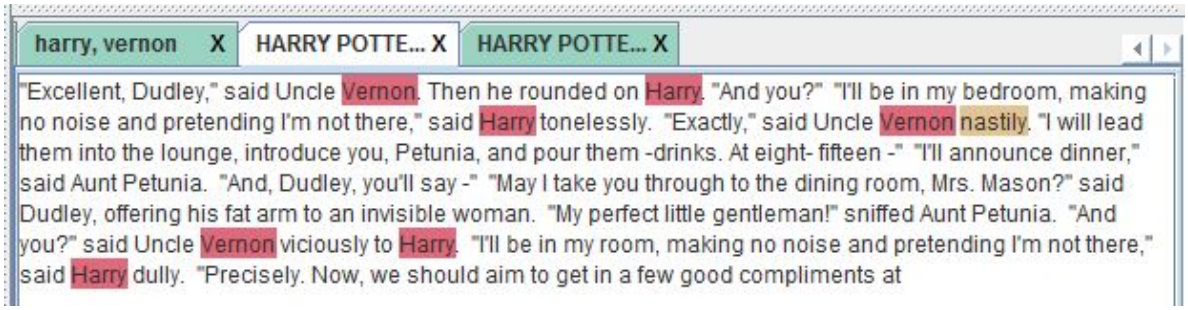


Abbildung 6.10: Im Entitätengraphen findet der Analyst ihm zunächst unbekannte Personen.

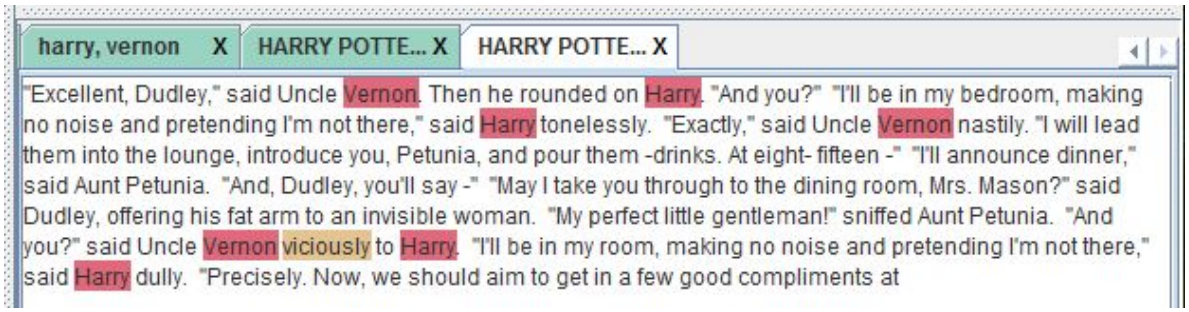
Da er mehr über Vernon herausfinden möchte, selektiert er nun die Verbindung zwischen Harry und Vernon, während der Explorermodus für Verben und Adjektive aktiviert ist. Es öffnet sich die Word-Cloud mit den Verben und Adjektiven, die in Zusammenhang mit Harry und Vernon vorkommen, welche in Abbildung 6.11 dargestellt ist.

Abbildung 6.11: Mit Hilfe des Explorers für Verben und Adjektive lässt sich auf schnelle Weise ein Eindruck über Personen und ihre Beziehungen zueinander gewinnen. Wörter wie „nastily“ und „viciously“ können Hinweise auf das Verhältnis zwischen Personen geben.

Beim Untersuchen der Word-Cloud findet er die Adjektive (bzw. Adverbien) „nastily“ und „viciously“, was er ungewöhnlich findet. Die Wörter „nasty“ (deutsch: gemein, scheußlich) und „vicious“ (deutsch: böse, boshaft), welche eher negativ behaftet sind, könnten auch auf eine negative Beziehung zwischen Harry und Vernon hinweisen. Um dies näher zu untersuchen, selektiert er deshalb jeweils die entsprechenden Tags in der Word-Cloud, woraufhin sich die Textstellen öffnen, in denen die Adjektive in Zusammenhang mit Harry und Vernon vorkommen (siehe Abbildung 6.12).



(a) Textexplorer für „nastily“



(b) Textexplorer für „viciously“

Abbildung 6.12: Der Textexplorer kann auch für Verben und Adjektive genutzt werden. Die entsprechenden Vorkommen der Entitäten, Verben und Adjektive werden mit ihren eigenen Farben hervorgehoben.

In den geöffneten Textstellen erkennt unser Analyst, dass Vernon boshaft mit Harry spricht und ihn herablassend behandelt. Dies bestätigt die Vermutung, dass Harry und Vernon ein schlechtes Verhältnis zueinander haben. Beim Analysieren der Textstellen findet unser Analyst weiterhin heraus, dass Vernon sein Onkel ist und dass Harry bei ihm wohnt.

Nachdem er diese Informationen gesammelt hat, wendet er sich wieder der Entitätenansicht zu, um auch mehr über die anderen Personen zu erfahren. Hierbei stellt er fest, dass der Entitätengraph „geteilt“ ist (siehe Abbildung 6.13): Es gibt Untermengen von Personen, die jeweils nur untereinander Verbindungen zueinander besitzen, so dass sich der Graph in nahezu unabhängige Teilgraphen unterteilen lässt (abgesehen von den Verbindungen zum Knoten „Harry“).

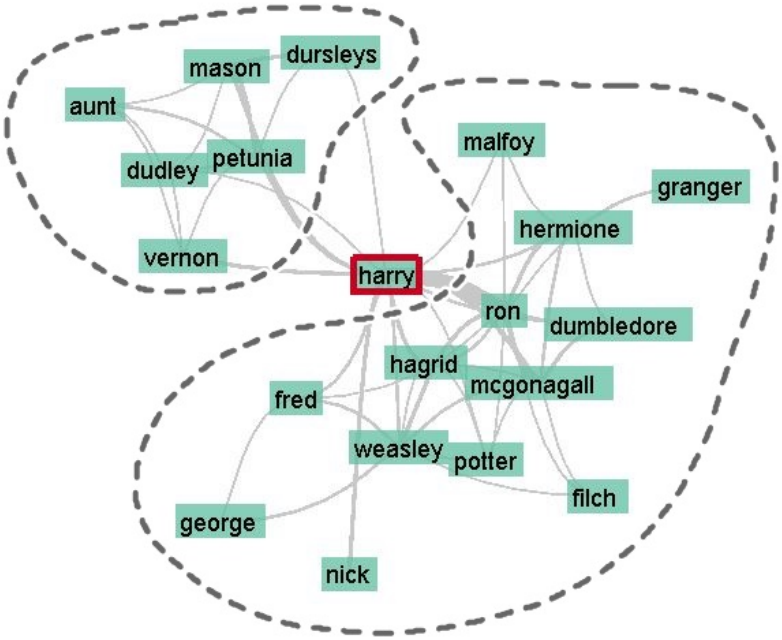


Abbildung 6.13: Im Entitätengraphen erkennt unser Analyst mehrere große Cluster, welche auf eigenständige Personengruppen der Geschichte hinweisen.

Dies deutet daraufhin, dass es sich bei den Teilgraphen um Untermengen von Personen handelt, die eigenständigen Personengruppen der Handlung entsprechen. Um dies genauer zu untersuchen, analysiert unser Anwender die Knoten der Teilgraphen, ähnlich wie zuvor bei Vernon. Hierbei findet er heraus, dass es sich dabei tatsächlich um eigene Personenkreise handelt: Er erfährt, dass Vernon das Familienoberhaupt der Dursleys ist, wozu auch Harrys Tante Petunia und sein Cousin Dudley gehören. Harry lebt bei den Dursleys, also in der Muggle-Welt (den Nicht-Zauberern), bis zu seinem elften Lebensjahr — bevor er das erste Mal von der Zauberschule Hogwarts erfährt. Seine Einschreibung in die Zauberschule stellt einen Einschnitt in der Geschichte dar: Hier entflieht er dem unglücklichen Leben bei den Dursleys und lernt die Zaubererwelt kennen, wozu die Charaktere des anderen Personenkreises gehören. Diesen Sachverhalt, also die Einteilung in Zauberer- und Muggle-Welt innerhalb der Geschichte, kann unser Analyst für die Erstellung des Personennetzwerks berücksichtigen.

So kann unser Analyst Schritt für Schritt die Profile der vorkommenden Charaktere erstellen, mit ihren Beziehungen zueinander und ihrer jeweiligen Bedeutung für die Handlung.

6.3 Diskussion

In diesem Abschnitt soll eine Diskussion des entwickelten Ansatzes anhand der vorgestellten Beispiele stattfinden. Es werden Vor- und Nachteile diskutiert und dadurch mögliche Verbesserungen herausgearbeitet.

Als vorteilhaft hat sich die Verwendung von mehreren Views herausgestellt. Durch die feste Zuweisung von Funktionen zu bestimmten Bereichen innerhalb der Benutzeroberfläche ist es einfacher, dem Nutzer eine intuitive und konsistente Bedienung der Anwendung zu ermöglichen. Der Nutzer muss sich beim Arbeiten mit dem Ansatz nicht unerwartet auf eine neue Darstellung der Informationen einstellen, sondern kann sich an ein durchgehendes Muster gewöhnen.

Für die Textexplorationen ist die Verlinkung der Views ebenfalls vorteilhaft: Einzelne Elemente einer Ansicht sind auf unterschiedliche Weise mit den anderen Ansichten verbunden. Wenn der Anwender ein Element in einer Darstellung entdeckt, das ihn interessiert oder verwundert, kann er dieses auswählen und näher untersuchen. Die Verlinkung der Ansichten unterstützt ihn dabei: Beispielsweise werden durch die Termfilter-Funktion, diejenigen Dokumente in der Dokumentenspatialisierung hervorgehoben, die eine bestimmte Entität enthalten. Diese Funktion ist sowohl in der Dokumentenspatialisierung als auch im Entitätengraph und ebenso in der Exploreransicht verfügbar. Hierdurch ist es auf einfache Weise möglich, die Analyse bezüglich eines bestimmten Tags (Entität, Verb, Adjektiv) zu vertiefen und weiterzuführen. Vor allem in einer offenen Exploration ist dies von Vorteil, weil so Anhaltspunkte bezüglich dieser Tags leichter innerhalb des verwendeten Dokumentensatzes verfolgt werden können. Die Verwendung von mehreren Lupen unterstützt diesen genannten Punkt zusätzlich: Hierdurch kann der Nutzer weiterführende Untersuchungen durchführen, ohne den Kontext der aktuell betrachteten Dokumente zu verlieren.

Speziell bei der Analyse von enthaltenen Personen und Orten erweist sich der Entitätengraph als effizientes Hilfsmittel: Hier lassen sich schnell starke Relationen zwischen Entitäten identifizieren. Durch das zugrunde liegende kräftebasierte Layout des Graphen lassen sich auch Cluster rasch erkennen. Weiterhin sind die Explorerfunktionen nützlich, wenn zu einer bestimmten Entität bzw. Entitätenrelation noch wenig Wissen vorhanden ist, und diese deshalb genauer betrachtet werden soll. Das Verwenden von entsprechenden Farben und Markierungen sowohl im Enitätengraph als auch in der Exploreransicht helfen dabei, die relevanten Informationen zu finden.

Der entwickelte Ansatz birgt jedoch auch einige Mängel, die nun beschrieben werden sollen. Beispielsweise fehlt in den Darstellungen der Informationen die zeitliche Komponente. Wenn dem Nutzer Informationen in den verschiedenen Ansichten angezeigt werden, kann er diese zunächst nicht ohne weiteres in ein zeitliches Verhältnis zueinander setzen. Bei der Exploration von Romanen und anderen ähnlichen Werken ist dies problematisch, da die korrekte zeitliche Zuordnung der Informationen wichtig für das Verständnis sein kann. Dieses Problem könnte im Ausmaß reduziert werden, indem die Dokumente entsprechend ihrer Vorkommen im literarischen Werk benannt werden, also mit Kapitel- und Abschnittsnummer, so dass zumindest bei der Verwendung des Textexplorers diese Information durch den Nutzer einsehbar ist und er sich ein grobes Bild des zeitlichen Verlaufs machen kann. Man könnte auch eine weitere Ansicht einführen, die die Zeitkomponente darstellt, wie z. B. mittels einem Zeitstrahl. Dieser Zeitstrahl könnte mit den anderen Views verlinkt werden und entsprechende

Markierungen enthalten, die dem Nutzer Aufschluss über die zeitliche Anordnung der Informationen geben.

Weiterhin ist die Erkennung der Entitäten verbesserungswürdig: Da die verwendeten Modelle ursprünglich für Zeitungsartikel und Journaltexte entwickelt wurden, funktioniert die Erkennung von Personen und Orten nicht immer einwandfrei. Vor allem wenn die Entitäten ungewöhnliche und fantasievolle Namen haben, kommt es vor, dass eine Person als Ort erkannt wird oder umgekehrt. Ebenso verbesserungswürdig ist die Erkennung von Koreferenzen, also verschiedenen Bezeichnungen, die sich auf dieselbe Entität beziehen. So werden in unserem Fall z. B. „Harry“ und „Potter“ fälschlicherweise als zwei eigenständige Entitäten behandelt und als solche im Entitätengraph dargestellt. Eine mögliche Lösung wäre hier, die Erkennung der Entitäten zu verbessern, indem ein anderes Toolkit für die NLP-Aufgaben verwendet wird, das hierfür besser geeignet ist. Da die falsche Erkennung jedoch trotzdem nicht vollständig vermieden werden kann, könnte man zusätzlich die Funktion integrieren, dass der Nutzer diese falschen Informationen selbst korrigiert. Wenn er z. B. entdeckt, dass im Entitätengraph eine Person vermeintlich als Ort dargestellt wird, könnte er diesen Knoten auswählen und dies korrigieren. Oder aber er entdeckt zwei Knoten im Entitätengraph, die eigentlich dieselbe Person darstellen. In diesem Fall würde er die beiden Knoten auswählen und als dieselbe Entität markieren. Diese Korrekturen würden von der Anwendung gespeichert werden und in Zukunft richtig angezeigt werden.

Unser Ansatz bietet viele Möglichkeiten, um weiterführende Analysen zu angezeigten Informationen durchzuführen. Wenn der Nutzer jedoch nach Informationen sucht, die keine Relation zu den aktuell angezeigten Informationen haben, stehen ihm hierzu keine unterstützenden Funktionen zur Verfügung. Dies könnte wichtig sein, wenn er Vermutungen oder Hypothesen überprüfen möchte, die er unabhängig von den aktuell angezeigten Informationen gebildet hat. An dieser Stelle wäre die Ergänzung einer globalen Suchfunktion sinnvoll: Hier könnte der Nutzer Begriffssuchen durchführen und so die Dokumente in der Spatialisierung filtern, die die gesuchten Begriffe beinhalten. Zusätzlich könnten entsprechende Vorkommen, sofern vorhanden, in den anderen Ansichten markiert werden.

7 Zusammenfassung und Ausblick

Ziel dieser Arbeit war es, einen lupenbasierten Ansatz zur Exploration von Dokumentenspatialisierungen zu entwickeln. Dabei sollte die Kombination von mehreren Lupen möglich sein. Grund hierfür war, dass bei vorigen existierenden Ansätzen vergleichende Analysen zwischen verschiedenen fokussierten Mengen nicht möglich waren, was auf eine begrenzte Anzahl der verwendeten Lupen zurückzuführen war. Deshalb sollten in dieser Arbeit bestehende lupenbasierte Ansätze um die Möglichkeit erweitert werden, mit mehreren Lupen gleichzeitig zu arbeiten.

Dazu wurde bei verwandten Arbeiten zunächst betrachtet, wie mit ähnlichen oder verwandten Problemstellungen umgegangen wird. Diese Arbeiten haben erste Ideen und Anregungen dazu gegeben, wie ein Lösungsansatz in dieser Arbeit aussehen kann. Weiterhin wurden die nötigen fachlichen Grundlagen erarbeitet, um einen entsprechenden Lösungsansatz entwickeln zu können. Hierzu gehörte das Fachwissen aus mehreren Themenbereichen, wie dem Bereich der Visualisierung und der natürlichen Sprachverarbeitung.

Daraufhin wurden die Konzepte betrachtet, die für die Lösung der gegebenen Problemstellung entwickelt wurden. Hierbei wurde von DocuCompass als Fundament ausgegangen, welches bereits Funktionen zur Textexploration mit einer Lupe enthielt. Im Rahmen dieser Arbeit wurde dieser vorhandene Ansatz um neue Funktionen erweitert. Hierzu gehörten, neben der Unterstützung von mehreren Lupen, das Einführen von neuen Ansichten, welche die Aufgabe hatten, Informationen zu den fokussierten Dokumenten der Lupenkombinationen anzuzeigen. Im Zusammenhang dazu wurde auch die entsprechende Verlinkung der Ansichten vorgestellt. Da der Fokus dieser Arbeit auf den Personen und Orten der Dokumente liegen sollte, war eine wichtige Visualisierungsart der Entitätengraph, welcher die Personen und Orte mit ihren zugehörigen Zusammenhängen darstellt. Um diese genauer untersuchen zu können, wurde eine weitere Funktion vorgestellt, der Explorer. Hier können sowohl Word-Clouds mit zugehörigen Verben und Adjektiven geöffnet werden als auch die konkreten Textstellen aufgerufen werden. Teil des Konzepts war hierbei auch die passende Verwendung von Farben und Markierungen, welche konsistent in den verschiedenen Ansichten verwendet wird. Außerdem standen auch verschiedene Interaktionsmöglichkeiten bezüglich der Ansichten im Vordergrund.

Dem Konzept entsprechend wurde im Rahmen dieser Arbeit ein Prototyp entwickelt, der die entwickelten Funktionen umsetzte. Dieser Prototyp wurde mit seinen verschiedenen Komponenten und Funktionsweisen näher betrachtet und erläutert. Um das Verständnis und auch die Bedienung des entwickelten Ansatzes zu veranschaulichen, wurden einige Anwendungsbeispiele vorgestellt, in denen gezeigt wurde, wie die einzelnen Funktionen zur Geltung kommen. Anschließend wurde anhand dessen eine Diskussion über Vor- und Nachteile durchgeführt und eventuelle Verbesserungsmöglichkeiten herausgearbeitet.

Ausblick

Diese Arbeit kann auf vielfältige Weise fortgeführt werden. Zum einen können vorhandene Teile des entwickelten Ansatzes verbessert werden, zum anderen können auch neue Aspekte miteinbezogen werden. Im Folgenden soll ein Ausblick auf die verschiedenen Möglichkeiten gegeben werden.

Zunächst könnte man die Ansichten und ihre Verlinkungen erweitern. Hierbei bieten sich mehrere Möglichkeiten an: Man könnte die Termfilter-Funktion auch für einzelne Wörter im Textexplorer umsetzen. Ebenso könnte man auch die Zugehörigkeit der Tabs im Explorer durch zusätzliche Interaktionsmöglichkeiten verdeutlichen, indem man das Auswählen eines Tabs mit dem gleichzeitigem Auswählen der zugehörigen Entität bzw. Entitätenpaares erweitert. Analog dazu könnte das Auswählen eines Vorkommens einer Entität im Textexplorer funktionieren. Den Entitätengraph selbst könnte man erweitern, indem man beim Highlighting der angrenzenden Nachbarn eines Knotens, die irrelevanten Knoten zusätzlich ausgraut, was die Anschaulichkeit verbessern würde.

Weiterhin könnte man dem Nutzer die Option bieten, verschiedene Teile der Anwendung seinen persönlichen Bedürfnissen anzupassen. Hierbei könnte der Nutzer z. B. die verwendeten Farben oder Schriftgrößen anpassen. Ebenso könnte man die Funktion integrieren, dass der Nutzer das Kriterium für den Zusammenhang von Entitäten einstellen kann: Hierbei könnte er die Spanne an Sätzen selbst definieren, in der die jeweiligen Entitäten vorkommen müssen.

Die Erkennung von Entitäten sollte verbessert werden, einschließlich der Erkennung von Koreferenzen. Wie in Abschnitt 6.3 bereits erläutert, könnte man dies zusätzlich um die Funktion ergänzen, dass der Nutzer von ihm erkannte Fehler verbessert und in das System eintragen kann. Weiterhin könnte auch eine semantische Analyse der Dokumententexte hinzugefügt werden, welche auf einen positiven oder negativen Inhalt verweisen würde. Dies könnte den Ansatz mit den Word-Clouds für Verben und Adjektive ergänzen oder sogar ersetzen.

Es wäre auch denkbar, weitere Ansichten für die Darstellung von Informationen einzuführen. Beispielsweise könnte man eine Ansicht für eine Kartendarstellung hinzufügen, welche Markierungen für reale Orte enthält, die in den Dokumenten vorkommen. Man könnte auch neue Ansichten einführen, welche Informationen zu anderen Aspekten der Dokumente darstellen, abgesehen von Personen und Orten. Ein Beispiel wäre ein Zeitstrahl, wie in Abschnitt 6.3 vorgestellt, welcher die zeitliche Zuordnung von Informationen beinhalten könnte.

Eine weitere Funktion, die die Bedienung erleichtern würde, wäre das automatische Einbinden von neuen Dokumenten, ohne den kompletten Dokumentensatz nochmal vollständig einlesen zu müssen. Ebenso könnte die Implementierung einer „Undo“-Funktion umgesetzt werden, welche einzelne Teilschritte rückgängig macht, oder eine Speicherfunktion, welche den aktuellen Stand der Lupen und Visualisierungen für ein späteres Abrufen festhält, so dass die Exploration auch ohne Probleme unterbrochen werden kann.

Literaturverzeichnis

[BKL09] S. Bird, E. Klein, E. Loper. *Natural language processing with Python: analyzing text with the natural language toolkit.* Ö'Reilly Media, Inc.", 2009. (Zitiert auf Seite 28)

[Cho03] G. G. Chowdhury. Natural language processing. *Annual review of information science and technology*, 37(1):51–89, 2003. (Zitiert auf Seite 27)

[CMS99] S. K. Card, J. D. Mackinlay, B. Shneiderman. *Readings in information visualization: using vision to think.* Morgan Kaufmann, 1999. (Zitiert auf den Seiten 23 und 24)

[GLP+07] C. Gorg, Z. Liu, N. Parekh, K. Singhal, J. Stasko. Visual analytics with Jigsaw. In *Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on*, S. 201–202. IEEE, 2007. (Zitiert auf den Seiten 16 und 17)

[Gra15] M. Grandjean. Introduction à la visualisation de données: l'analyse de réseau en histoire. *Geschichte und Informatik*, 18(19):109–128, 2015. (Zitiert auf Seite 27)

[HB03] M. Harrower, C. A. Brewer. ColorBrewer. org: an online tool for selecting colour schemes for maps. *The Cartographic Journal*, 40(1):27–37, 2003. (Zitiert auf Seite 37)

[HB05] J. Heer, D. Boyd. Vizster: Visualizing online social networks. In *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on*, S. 32–39. IEEE, 2005. (Zitiert auf Seite 21)

[HCL05] J. Heer, S. K. Card, J. A. Landay. Prefuse: a toolkit for interactive information visualization. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, S. 421–430. ACM, 2005. (Zitiert auf Seite 45)

[HJH+ar] F. Heimerl, M. John, Q. Han, S. Koch, T. Ertl. DocuCompass: Effective Exploration of Document Landscapes. In *2016 IEEE Conference on Visual Analytics Science and Technology (VAST)*, S. pp. to appear. (Zitiert auf den Seiten 11, 33 und 49)

[HLLE14] F. Heimerl, S. Lohmann, S. Lange, T. Ertl. Word cloud explorer: Text analytics based on word clouds. In *System Sciences (HICSS), 2014 47th Hawaii International Conference on*, S. 1833–1842. IEEE, 2014. (Zitiert auf den Seiten 18 und 19)

[JJ00] D. Jurafsky, H. James. Speech and language processing an introduction to natural language processing, computational linguistics, and speech. 2000. (Zitiert auf den Seiten 28 und 30)

[Kob04] S. G. Kobourov. Force-directed drawing algorithms. 2004. (Zitiert auf Seite 26)

[KTW+13] R. Krüger, D. Thom, M. Wörner, H. Bosch, T. Ertl. TrajectoryLenses–A Set-based Filtering and Exploration Technique for Long-term Trajectory Data. In *Computer Graphics Forum*, Band 32, S. 451–460. Wiley Online Library, 2013. (Zitiert auf den Seiten 13, 14 und 15)

[MH08] L. v. d. Maaten, G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(Nov):2579–2605, 2008. (Zitiert auf Seite 33)

[MMS93] M. P. Marcus, M. A. Marcinkiewicz, B. Santorini. Building a large annotated corpus of English: The Penn Treebank. *Computational linguistics*, 19(2):313–330, 1993. (Zitiert auf den Seiten 30 und 31)

[MRS+08] C. D. Manning, P. Raghavan, H. Schütze, et al. *Introduction to information retrieval*, Band 1. Cambridge university press Cambridge, 2008. (Zitiert auf Seite 29)

[MSB+14] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, D. McClosky. The stanford corenlp natural language processing toolkit. In *ACL (System Demonstrations)*, S. 55–60. 2014. (Zitiert auf Seite 45)

[POM07] F. V. Paulovich, M. C. F. Oliveira, R. Minghim. The projection explorer: A flexible tool for projection-based multidimensional visualization. In *Computer Graphics and Image Processing, 2007. SIBGRAPI 2007. XX Brazilian Symposium on*, S. 27–36. IEEE, 2007. (Zitiert auf Seite 45)

[Shn96] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Visual Languages, 1996. Proceedings., IEEE Symposium on*, S. 336–343. IEEE, 1996. (Zitiert auf Seite 25)

[VHWV09] F. Van Ham, M. Wattenberg, F. B. Viégas. Mapping text with phrase nets. *IEEE transactions on visualization and computer graphics*, 15(6), 2009. (Zitiert auf den Seiten 19 und 20)

Alle URLs wurden zuletzt am 28. 04. 2017 geprüft.

Erklärung

Ich versichere, diese Arbeit selbstständig verfasst zu haben. Ich habe keine anderen als die angegebenen Quellen benutzt und alle wörtlich oder sinngemäß aus anderen Werken übernommene Aussagen als solche gekennzeichnet. Weder diese Arbeit noch wesentliche Teile daraus waren bisher Gegenstand eines anderen Prüfungsverfahrens. Ich habe diese Arbeit bisher weder teilweise noch vollständig veröffentlicht. Das elektronische Exemplar stimmt mit allen eingereichten Exemplaren überein.

Ort, Datum, Unterschrift