Masterprojekt: Topic Modeling im semantischen Information Retrieval – Implementation und Anwendung eines Software-Moduls

WiSe 2017/2018
Oberseminar: Technicum/Theoreticum
Dozent: Prof. Dr. Udo Hahn
Referent: Philipp Sieg
Datum: 01.12.2017

Gliederung

1. Einleitung: Semantische Suche
2. Topic Modeling
3. Schwächen von Topic Modeling
4. Herausforderungen bei der Umsetzung
5. Planung der Masterarbeit

1. Einleitung:
Was ist semantische Suche?

- Suchsysteme ohne semantische Komponente arbeiten lediglich auf der lexikalischen Oberfläche der Dokument- und Anfrageterme
- tieferes ‚Verständnis' für die Anfrage und die Daten
  - Semedico.org:

  "[…] semantic search engines are designed in such a way so that they automatically compensate for both the terminological as well as linguistic variety of natural language […]"
- Umsetzung dieses Ziels mit verschiedenen Mitteln:

  ‚Externe' Ressourcen vs. ‚interne' Ressourcen

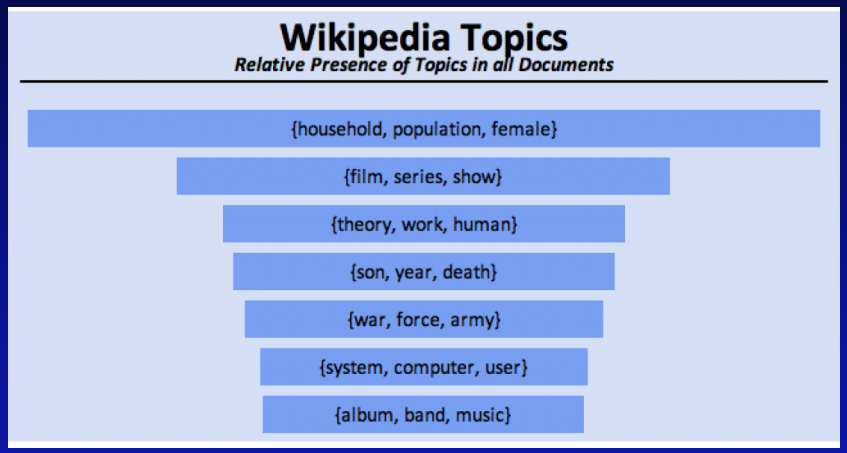
1. Einleitung: ‚Externe' Wissensrepräsentationen

- Hinzufügen von Wissen aus logisch-definitorisch aufgebauten Repräsentationen: Ontologien, Thesauri, digitale Terminologien …
- Probleme
  - Hoher Aufwand bei der Erstellung: manuelles Eintragen, Einarbeiten, Abstimmen, Expertenwissen …
  - Selbst bei automatisierten Verfahren bleiben Inkonsistenzen und definitorische Unsicherheiten

1. Einleitung:
Die Daten selbst als Wissensgrundlage

- Topic Modeling: automatische Abstraktion des Inhalts eines Dokuments anhand der Wahrscheinlichkeitsverteilungen seiner Wörter
  - Empirisches Datenfundament statt Definitionen (bottom-up statt top-down)
  - automatisiert errechenbar (maschinelles Lernen)
  - keine explizit definierten Relationen als Semantik, sondern Konzept der ‚Informativität' („informativeness", Manning/Schütze 2000:545)

2. Topic Modeling: Beispiele



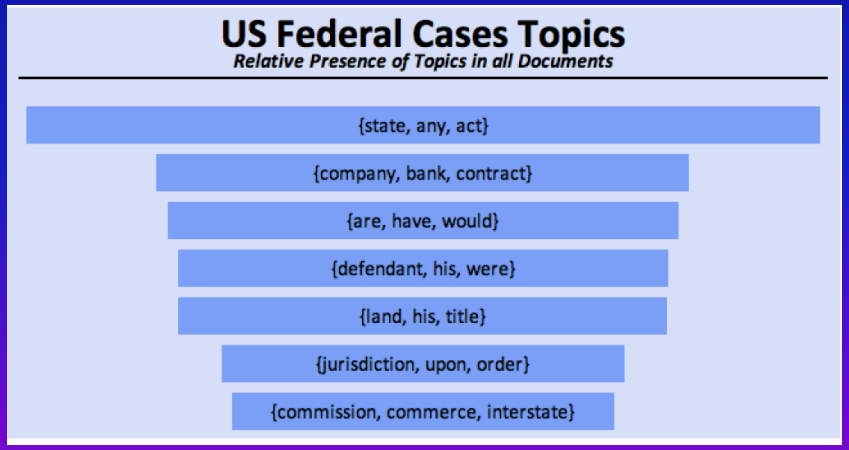


Abb. 1: Beispiele für Topics, aus Chaney/Blei (2012:422).

2. Topic Modeling: Sprachmodellierung aus ‚Topic-Perspektive'

- Was sind Topics im Information Retrieval?
  - Allgemein: für den Informationsbedarf des Nutzers abstrahierte Repräsentation von Dokumenten hinsichtlich deren Relevanz
  - Grundlegende These: Bestimmte Wörter haben eine höhere ‚Informativität' über den Inhalt des Textes und im Verhältnis zur Nutzeranfrage eine bestimmte Relevanz
  - Allerdings für jede Sprachmodellierung anders formalisiert
- Ursprüngliche Technik zur Formalisierung von Relevanz: tf-idf/Vektorraum-Modell (VSM)
- Weiterentwicklung: Singular Value Decomposition (SVD) in Latent Semantic Indexing/Analysis (LSI/LSA)
- Probabilistische Ansätze: pLSA (aspect model), LDA
  - Topics (original: aspects (Hofmann 1999)) sind diskrete Wahrscheinlichkeits-verteilungen von Wörtern
  - **Generativer** Ansatz bei LDA: ein Model generiert durch die semantische Verknüpfung/Konzepte die Wörter, die zu dem Thema passen (**informativ** sind)

2. Topic Modeling: Modellierungsannahmen (1)

- Turney/Pantel (2010:153)
  - *Statistical semantics hypothesis*:

    Je ähnlicher die statistische Verteilung der Wörter auf Texte ist, desto ähnlicher ist ihre Semantik.
  - *Bag of words hypothesis*:

    Je ähnlicher die statistische Verteilung der Wörter auf Dokumente ist, desto ähnlicher ist die Semantik der Dokumente.
  - *Distributional hypothesis*:

    Je ähnlicher die statistische Verteilung bestimmter Wörter auf Kontexte ist, desto ähnlicher ist die Semantik dieser Wörter.
  - *Extended distributional hypothesis*:

    Die *Distributional hypothesis* gilt auch für Wortpaare.
  - *Latent relation hypothesis*:

    Je ähnlicher sich die Kontexte sind, desto ähnlicher sind sich auch die Wortpaare in diesem Kontext.

2. Topic Modeling: Modellierungsannahmen (2)

Beispiel: grundlegendes Latent Dirchlet Allocation

- Probabilistische Modellierung:
  - Diskrete Verteilung:
    - (a) Wörter als diskrete Einheiten
    - (b) Topics als diskrete Einheiten
  - Dirichlet-Verteilung:
    - (a) Topics als **gleichmäßig seltene** Verteilung von Wörtern → Kohärenz der Topics; jedes Wort hat eine ‚Minimalwahrscheinlichkeit' (non-zero possibility)
    - (b) Dokumente als **gleichmäßig seltene** Verteilung von Topics → Kohärenz der Dokumente; jedes Topics hat eine ‚Minimalwahrscheinlichkeit' (non-zero possibility)

- Verarbeitungsschritte:
  - Generierung der Topics
  - Document Allocation
  - Kontextualisieren der Wörter
  - Inferenz der (‚hidden') Topics

Distribution	Density	Example Parameters	Example Draws
Gaussian	$\frac{1}{\sqrt{2\sigma^2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	$\mu = 2, \sigma^2 = 1.1$	$x = 2.21$
Discrete	$\prod_i \phi_i^{\mathbb{1}[w=i]}$	$\phi = \begin{bmatrix} 0.1 \\ 0.6 \\ 0.3 \end{bmatrix}$	$w = 2$
Dirichlet	$\frac{\prod_{i=1}^K \Gamma(\alpha_i)}{\Gamma(\sum_{i=1}^K \alpha_i)} \prod_{i=1}^K \theta_i^{\alpha_i - 1}$	$\alpha = \begin{bmatrix} 1.1 \\ 0.1 \\ 0.1 \end{bmatrix}$	$\theta = \begin{bmatrix} 0.8 \\ 0.15 \\ 0.05 \end{bmatrix}$

Abb. 2: Wahrscheinlichkeitsverteilungen, aus Boyd-Graber/Hu/Mimno (2016:151).

2. Topic Modeling: Generierung der Topics

	w_1	w_2	...	w_n	
t_1	$x_{1,1}$	$x_{1,2}$	...	$x_{1,n}$	φ_1
t_2	$x_{2,1}$	$x_{2,2}$	...	$x_{2,n}$	φ_2
...		...			
t_K	$x_{k,1}$	$x_{k,2}$	...	$x_{k,n}$	φ_K

Abb. 3: Topic-Wort-Matrix.

- *K* ist als Parameter vorher festgelegt; w_i ist Element aus dem Vokabular des Korpus
- für jedes Topic t_k wird ein Vektor gemäß der Dirichlet-Verteilung gezogen: $\varphi_k \sim \text{Dir}(\lambda \boldsymbol{u})$
  - Jedes Topic wird durch eine diskrete Verteilung der Wörter *u* generiert
  - Kontrolliert durch den Dirichlet-Parameter λ ist jedes Wort auf jedes Topic gleichmäßig selten und einer Minimalwahrscheinlichkeit verteilt

2. Topic Modeling: Document Allocation

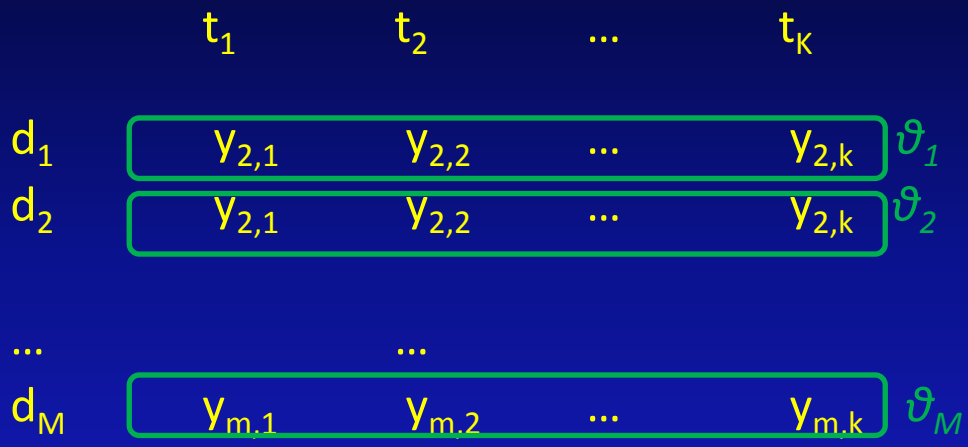


Abb. 4: Dokument-Topic-Matrix.

- für jedes Dokument *m* wird ein Vektor gemäß der Dirichlet-Verteilung gezogen: $\vartheta_m \sim \text{Dir}(\alpha \boldsymbol{u})$
  - Kontrolliert durch den Dirichlet-Parameter α ist jedes Topic auf jedes Dokument gleichmäßig selten und mit einer Minimalwahrscheinlichkeit verteilt
  - Jedes Dokument enthält Topics, die diskret verteilt sind

➔ Anhand der diskreten Wahrscheinlichkeitsverteilungen der Wörter für jedes Topic t_k können nun die Wörter für jedes Dokument generiert werden, indem ϑ_m als Parameter für die diskrete Verteilung genutzt wird

2. Topic Modeling: Kontextualisierung

- Wörter werden durch Topics der Dokumente generiert:
  - Topic Assignment:
    - Jedem Wort-Token *n* eines Dokumentvokabulars N_d wird ein Topic $z_{d,n}$ zugewiesen über die diskrete Topic-Verteilung dieses Dokuments ϑ_d
  - Generierung:
    - Aus der diskreten Wort-Verteilung dieses Topics $\varphi_{zd,n}$ wird nun ein Wort gezogen

➔ Generatives Modell, das Dokumente von der lexikalischen Oberfläche auf Topics abstrahiert

2. Topic Modeling: Inferenz

- Verschiedene Algorithmen
  - Bsp. Gibbs Sampling

    1. Für jedes Wort wird ein zufälliger Wert in der Topic-Wort-Matrix gesetzt, sodass jedem Wort-Token ein Topic zugeordnet wird (Kontextualisieren)

    2. Die zufälligen Ausgangswerte werden iterativ für jedes Token neu bestimmt

    a) Berechnung der Wahrscheinlichkeit, dass das Token topic *k* bekommt

$$p(z_{d,n} = i \mid \dots) = \theta_d \phi_j i = \left(\frac{N_{d,i} + \alpha_i}{\sum_k N_{d,k} + \alpha_k} \right) \left(\frac{V_{i,w_{d,n}} + \beta_v}{\sum_w V_{i,w} + \beta_w} \right)$$

Abb. 5: Produkt der Maximum Likelihood aus jeweils Wörter und Topics, aus Boyd-Graber/Hu/Mimno (2016:158).

    b) Ein Topic aus dem Wahrscheinlichkeitsraum wird dem Wort-Type zugeordnet; der Wahrscheinlichkeitsraum dieser Zuordnung vergrößert sich durch jede weitere Zuordnung

    3. Ende der Schleife beim ‚Steady state'

- Nutzbarer Output: Erlerntes Modell, dass Vektoren mit gewichteten Wort-Types enthält

2. Topic Modeling: Zusammenfassung und Anwendungen im IR

- Topic Modeling
  - Input: Vokabular, Dokumente, Token in Dokumenten
  - Hyperparameter: Anzahl der Topics *k*, Dirichlet-Parameter λ für Topics, Dirichlet-Parameter α für Dokumente, (Konzentrationsparameter der Topics über Dokumente α_0)
  - Output: trainierte *k* Topics, d.h. Wort-Vektoren mit Gewichtung

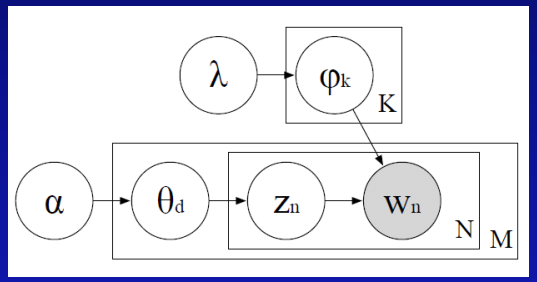


Abb. 6: LDA-Diagramm, aus Boyd-Graber/Hu/Mimno (2016:159).

- Anwendungen im ad hoc-Retrieval
  - Query Expansion: Erweiterung der Query-Terme durch Topics
  - Indexierung: breitere semantische Abdeckung durch Zusammenfassung der Wörter in Topic s als abstrahierte Zwischenschicht
  - Facettierung: Korpus (z. B.Query-Ergebnisse) können in Topics facettiert werden

2. Topic Modeling: Mögliche Ziel des Masterprojekts

- (A) Ad Hoc-Suche: Alternativer Index zu Ontologie
  - (1) alternative Suche
  - (2) kombinierte Suche
- (B) Zweistufige Suche: Query-Ergebnisse von Semedico durch Topic Modeling facettieren
  - Erst ab einer bestimmten Anzahl der Ergebnisse sinnvoll
  - In der Weise implementieren, dass die Suchergebnisse schon angezeigt werden, während die Facettierung vorbereitet wird (AJAX)
  - Mit einschlägigen Labels Facettierungsergebnisse repräsentieren

3. Schwächen von Topic Modeling

- Evaluation ist aufwendig, da implizite latente Semantik unüberwacht entstanden und daher potentiell korrekturbedürftig ist
  - Aufwand beim Modell-Design
  - Manuelles Evaluieren anhand von Visualisierungen
  - Perplexität als geeignetes Maß?

- Welche Aufgaben kann Topic Modeling überhaupt zuverlässig erfüllen? Ist es eine echte Alternative zu Wissensrepräsentationen?
  - Problem der Replizierbarkeit durch Zufallsvariablen bei der Inferenz
    ➔ Große Datenmengen nötig!

- Alternativen: Count-basierte Ansätze
  - Word Embeddings
  - *Nicht*-probabilistisches LSA

4. Herausforderungen bei der Umsetzung: Datengrundlage

- Kenntnisse des Korpus nötig, um Topic Modeling sinnvoll anzuwenden
  - Was ist ein Dokument?
  - Welche Wörter sind wichtig?
  - ➔ Wahl der Hyperparameter
- Welche Datenstrukturen werden benötigt?
- Welche Testdaten sollen gewählt werden?
  - *Für (B):* Jeweils unterschiedliche Medline-Teilmengen je nach Query-Ergebnis

4. Herausforderungen bei der Umsetzung: Implementation

- Trainingsorientierte Implementation von Topic Modeling
  - Testen verschiedener Parameter und Algorithmen
  - Bestehende Implementationen mit Vor- und Nachteilen
    - Zugriff auf Parameter
    - Evaluation der gefundenen Topics
    - Verwendeter Lernalgorithmus
    - Transparenz
    - …
- Programmatische Unabhängigkeit bzw. flexible Schnittstellen

4. Herausforderungen bei der Umsetzung: Anwendung in Semedico

- Schnittstellen
  - Anbindung an den Webservice
  - Schnittstellen in Tapestry
  - *Für (A)*: Einrichten eines eigenen Index
  - *Für (B):* Einbindung von AJAX
- Welche konkreten Datenstrukturen liegen vor?
- *Vor allem für (B):* Präsentation der Topics
  - Label
  - Die höchsten k Wörter
  - Verlinkung

4. Herausforderungen bei der Umsetzung: Evaluation

- Evaluation semantischer Funktionen bei der Suche
  - TREC Genomics Track Protocol 2005
  - Evaluationsmaß Mean Average Precision (MAP)

$$\mathrm{MAP}(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} \mathrm{Precision}(R_{jk})$$

Abb. 7: Mean Average Precision, aus Manning/Raghavan/Schütze (2008:160).

- Evaluation von Topic Modeling
  - Manuelle Evaluation durch Visualisierung und ändern der Parameter
  - Perplexität?

5. Planung der Masterarbeit: Zusammenfassung

- Datengrundlagen
  - Großes Korpus
- Unabhängige Implementation
  - Kernmodul
  - Training des Modells
  - Tapestry
  - AJAX
- Anwendung in Semedico
  - Anbindung an spezielle Datenstrukturen
- Evaluation
  - Funktionalität des Moduls: Durch manuelle Überprüfung mit Hilfe von Visualisierungen
  - Anwendung in Semedico: TREC 2005 Genomics Track Protocol

5. Planung der Masterarbeit: Zusammenfassung – Zwei Ansätze

(A) Indexierung

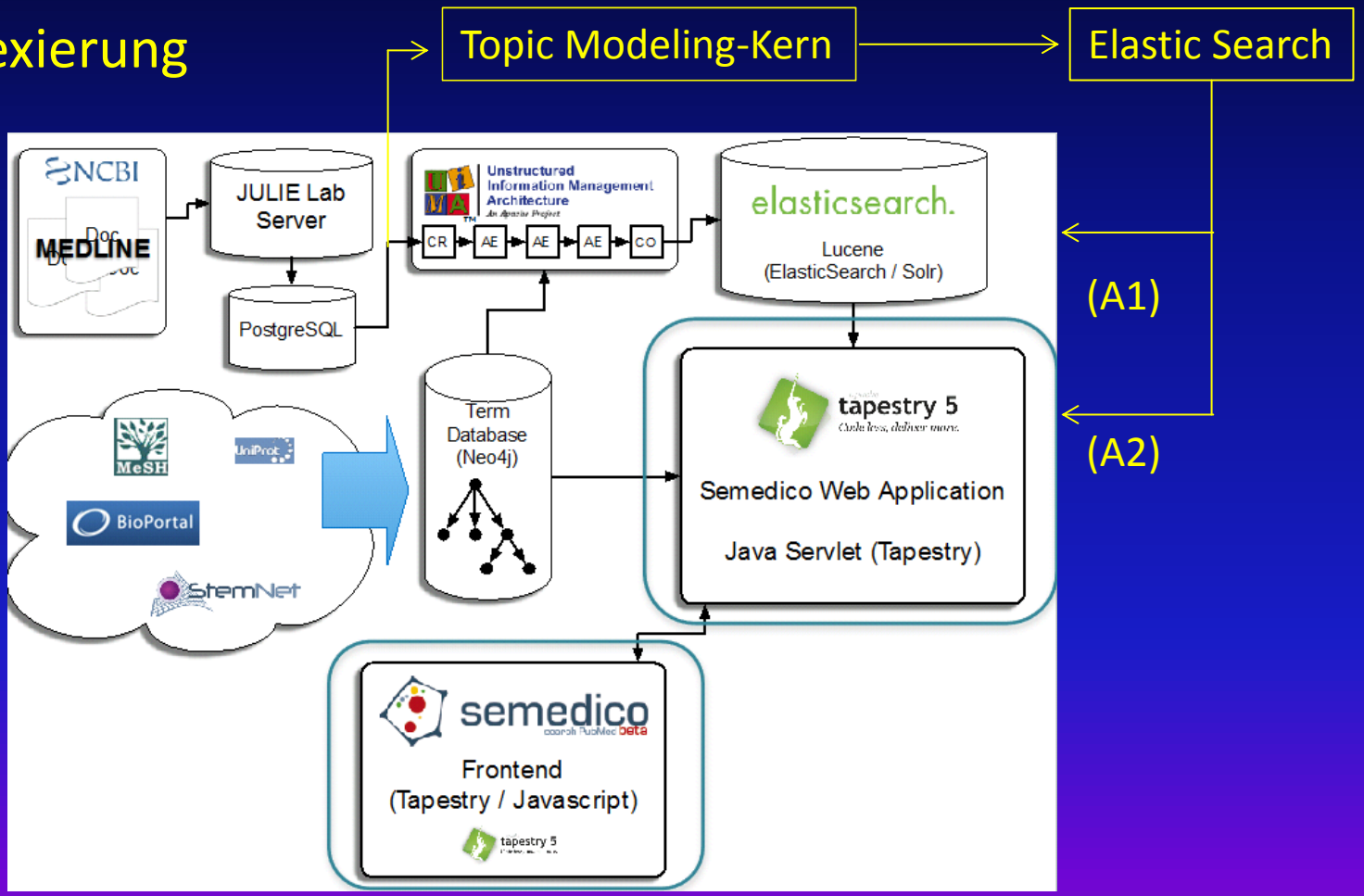


Abb. 8: Semedico Big picture, aus Präsentation von Erik Fäßler (2014): Semedico).

5. Planung der Masterarbeit: Zusammenfassung – Zwei Ansätze

(B) Nachverarbeitung des Ergebnis

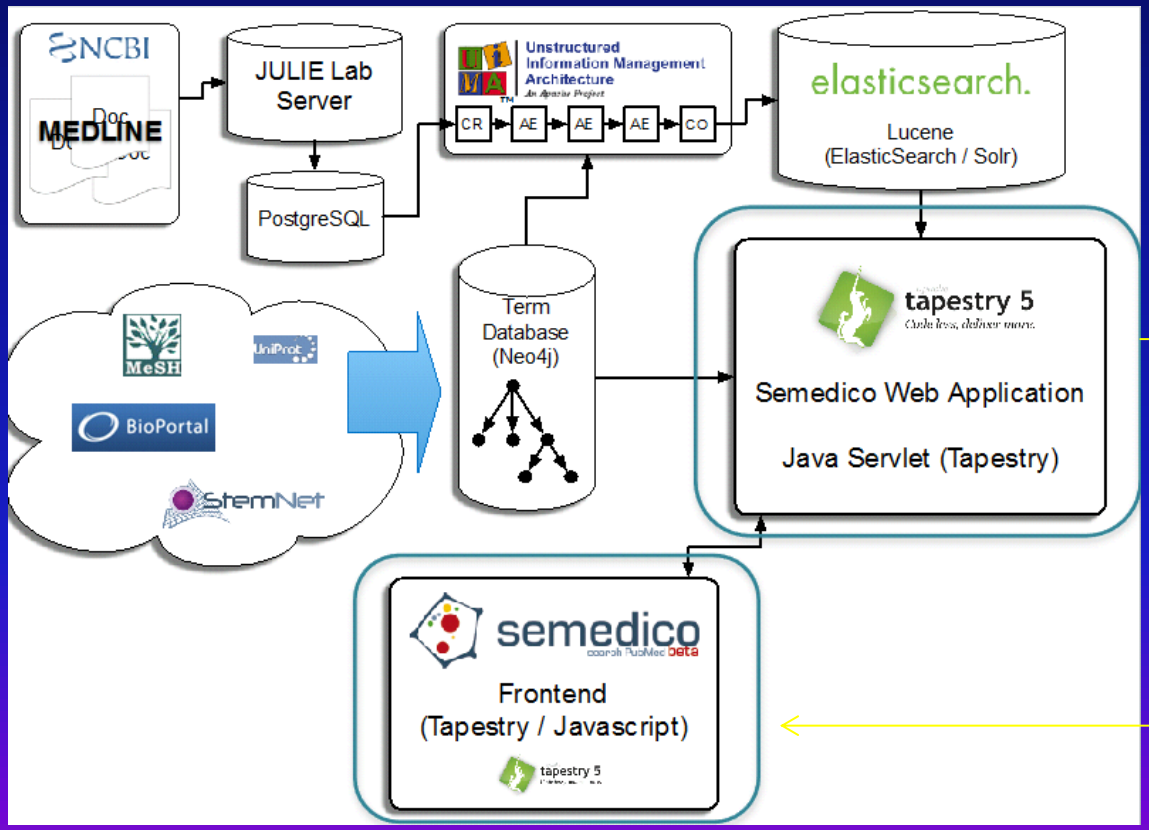


Abb. 8: Semedico Big picture, aus Präsentation von Erik Fäßler (2014): Semedico).

5. Planung der Masterarbeit: Inhaltsverzeichnis

5. Planung der Masterarbeit: Zeitplan

	1. Woche	2. Woche	3. Woche	4. Woche	5. Woche
November	**1.-5.11.** Treffen mit Erik und Johannes: Themenabsprache	**6.-12.11.** Grundlagen für 1. Vortrag zusammenfassen	**13.-19.11.** 1. Vortrag ausarbeiten	**20.-27.11.** Zusätzliche Infos für 1. Vortrag - Erik: Einführung Semedico bzw. andere Suchmaschinen/generelle Methodik	**28.-30.11.** 1. Vortrag ausarbeiten
Dezember	**1.-3.12.** 1. Vortrag im Oberseminar (01.12.'17)	**4.-10.12.** Neue Literatur sichten (v.a. Salton + die letzten 40 Jahre IR) (Anmeldung: 8.12.'17) Milestone I: Literatur und Konzeptionen gesichtet	**11.-17.12.** Software sichten -Suchmaschinen -Entwicklungs-umgebung: Tapestry, AJAX	**18.-24.12.** Programmieren: Modul-Kern	**25.-31.12.** ---
Januar	**1.-7.1.** Wiederholung Milestone II: methodisches Konzept muss stehen Programmieren: Modul-Kern	**8.-14.1.** Programmieren: Modul-Kern	**15.-21.1.** Programmieren: Modul-Kern; Tapestry (Standalone/gegen Semedico)	**22.-28.1.** Programmieren: Tapestry; AJAX (Standalone/gegen Semedico)	**29.-31.1.** Testen Milestone III: Programm muss stehen
Februar	**1.-4.2.** Evaluation	**5.-11.2.** Evaluation	**12.-18.2.** Evaluation 2. Vortrag ausarbeiten	**19.-25.2.** 2. Vortrag halten Milestone IV: Evaluation muss durchgeführt sein	**26.-27.2.** Schreiben
März	**1.-4.3.** Schreiben	**5.-11.3.** Schreiben	**12.-18.3.** Schreiben	**19.-25.3.** 3. Vortrag ausarbeiten und halten	**26.-31.3.** Schreiben
April	**1.4.** Schreiben	**2.-8.4.** Schreiben	**geplante Abgabe: 10.04.'18**	---	---
Mai	---	(offizielle Abgabe: 10.05.'18)	---	---	---

Referenzen

Hannah Bast, Björn Buchhold, Elmar Haussmann (2016). Semantic Search on Text and Knowledge Bases. Foundations and Trends® in Information Retrieval, 10(2–3),119–271. https://doi.org/10.1561/1500000032

David M. Blei, Andrew Y. Ng, Micheal I. Jordan (2003): Latent dirichlet allocation. In:Journal of Machine Learning Research, vol. 3, 993–1022.

Jordan Boyd-Graber, Yuening Hu, David Mimno (2017), "Applications of Topic Models", Foundations and Trends® in Information Retrieval: Vol. 11: No. 2-3, pp. 143-296. http://dx.doi.org/10.1561/1500000030

Allison Chaney and David M. Blei. Visualizing topic models. In International AAAI Conference on Weblogs and Social Media, 2012

Christopher D. Manning, Hinrich Schütze (2001): Foundations of Statistical Natural Language Processing. 4. Auflage. London, Cambridge: MIT Press.

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze (2008): *Introduction to Information Retrieval*, Cambridge University Press.

Gerard Salton (1971): The SMART retrieval system: Experiments in automatic document processing. Prentice-Hall, Upper Saddle River, NJ.

Peter D. Turney, Patrick Pantel (2010): From frequency to meaning: vector space models of semantics. *J. Artif. Int. Res.* 37, 1 (January 2010), pp. 141-188.

Internetquellen

- Topic Modeling Workshop, David Mimno, 2012. https://vimeo.com/53080123 (zuletzt aufgerufen am 29.11.2017)

- TREC 2005 Genomics Track Protocol http://skynet.ohsu.edu/trec-gen/2005protocol.html (zuletzt aufgerufen am 29.11.2017)