<Bachelorarbeit> am Institut für Informatik der Freien Universität Berlin

Human-Centered Computing (HCC), AG NBI

Topic Modeling mittels Spreading Activation auf Wikidata Knowledge Graph

– Exposé –

Luka Stärk

Matrikelnummer: 374532

luka.staerk@campus.tu-berlin.de

Betreuerin: Prof. Dr. C. Müller-Birn

Berlin, 18. August 2019

1 Struktur des Exposé

1.1 Motivation der Arbeit

Im Prozess einer kollaborativen Ideen-Entwicklung ist es von Vorteil, die einzelnen Ideen in Beziehung zu setzten, um Ideen nach Themen und Ähnlichkeit zu sortieren und somit im Brainstorming Prozess zielgerichteter von den Ideen andere zu profitieren. Teil des kreativen Prozesses in Ideation ist es bestehende Ideen in Clustern zusammenzufügen und sich auf diese Weise den Lösungsraum zu erschließen und die Ideen darin zu sortieren und zu organisieren. In dieser Phase ist es von Nutzen, nach Aspekten wie Ähnlichkeit von Ideen und Themenbereich, effektiv einzelne Ideen abzufragen. Zudem können dann verschiedene Schritte automatisiert und verkürzt werden. Beispielsweise durch das Vorschlagen von Clustern, das Anzeigen von Ideen nach Ähnlichkeit oder aus einem bestimmten Themenbereich.

Mittels Topic Modeling werden Textdokumente auf ihre abstrakten Themen reduziert, zudem lassen sich versteckte und implizit erwähnte Themen extrahieren. Dadurch sind Inhalte besser zugänglich und es ist möglich diese nach Topics zu sortieren.

Topic Modeling ist ein gut erforschtes statistisches Probleme im Bereich Mashine Learning und NLP, wobei die Frequenz mit der Wörter gemeinsam in Texten vorkommen entscheidet, welche Wörter ein Topic bilden.

Nachteil dieses Ansatzes ist, dass relevante Topics fehlen, weil Beispielsweise die Daten bestimmte Zusammenhänge nicht abbilden und dass die Topics in Form als *bag of words* für Menschen oft schwer Interpretierbar sind.

Diese Arbeit fokussiert sich auf einen semantischen Ansatz relevante Topics von Ideen Beschreibungen zu extrahieren. Dazu wird ein Knowledge Graph, der aus den semantischen Verbindung von Konzepten in Winkidata besteht erstellt, sodass die Konzepte der Ideen im Knowledge Graph definiert sind. Notwendige Bedingung dafür ist, dass die *word sense disambuigutie* für die Ideen gelöst ist. Durch Spreading Activation, eine Information Retrival Methode in Informations-Netzwerken, werden die Konzepte der Ideen im Knowledge Graph in Superklassen beziehungsweise Topics zusammen geführt. Das ermöglicht es Winkidata-Konzepte den Topics zuzuordnen und eine logische Aussage über die Beziehung von Topic und Text zu treffen.

Die Arbeit bewegt sich im Rahmen von automatisierter Text und Dokument Klassifizierung, durch semantisches Wissen aus Wikidata.

1.2 Thematische Einordnung der Arbeit

- Welche Artikel/Literatur sind/ist relevant für diese Arbeit?
  - APPLICATION OF SPREADING ACTIVATION TECHNIQUES IN INFORMATION RETRIEVAL [Cre97]
  - Node Similarities from Spreading Activation [TB10]
  - SaskNet: A Spreading Activation Based Semantic Network [Har16]
  - TODO: Artikel für den probabilistischen Ansatzes.

- Bitte geben Sie die relevanten Inhalte der Artikel kurz wieder. TODO:

1.2.1 Spreading Activation

Spreading Activation ist eine Methode um in Netzwerken Information zu suchen. Für ein oder mehrere Knoten können Verbindungen abgefragt und nach Knoten, Kanten und Subgraphen gesucht werden. Dazu werden die Abfrage Konten aktiviert. Die Aktivierung wird jeweils abgeschwächt und an die Nachbarknoten weiter gegeben, bis ein Terminierungskriterium erreicht ist und kein Knoten mehr aktiviert wird. Das Ergebnis ist ein durch die Aktivierung induzierter Subgraph und dessen Knoten mit Aktivierungsgrad. Konten werden als Ähnlich erkannt desto mehr direkte und indirekte Nachbarn sie teilen [TB10].

1.2.2 Simimlarity of Concepts

The paper of G. Zhu and C. A Iglesias[ZI17] discusses different metrics for concepts similarity on Knowledge Graphs and compairing them and their own aproach *wpath* with gold standard datasets where humans juge the similarity of words in meaning. Their metric *wpath* for similarity measure considers path lenght between of two concepts in the Knowledge Graph and the Information Content of their Least Common Subsumer.

1.2.3 probabilistisches Modell???

Hier ist noch die Entscheidung zu treffen welcher probabilistische oder Mashine Learning Ansatz zum Abgleich der Spreading Activation benutzt wird.

Entweder die Ergebnisse des LDA [Ble03] mit Anreicherung der *bag of words* der Ideen mit dessen direkten Superklassen aus dem Knowledge Graph.

Oder ein Vektor Knowledge Graph Embedding, indem die Ideen in ein Vektor Raum transformiert werden und mittels des Skalarprodukts vergleichbar sind, zum Beispiel durch deep averaging networks (DAN).

1.3 Zielstellung

Ziel ist es eine Anwendung zu entwickeln die Texte, insbesondere Beschreibungen und Erläuterung von Ideen, ein oder mehrere Topics zuteilt. Das Label eines Topics soll ein definiertes Wikidata-Konzept sein.

Die Topics zu den Texten sollen für Menschen intuitive sinnvoll und verständlich sein und sollen eine gewisse Abstraktionsebene haben, sodass sie nicht zu speziell sind und nur für einzelne Ideen gelten, gleichzeitig nicht zu abstrakt und für sehr viel gelten.

Für alle Konzepte der Ideen, die in einem extrahierten Topic enthalten sind, soll die Nähe zu diesem Topic als $weight \in [0, 1]$ angegeben werden. Somit können Ideen und Topics über die *weight* der Konzepte auch in qualitative Beziehung gesetzt werden.

1.4 Geplante Vorgehensweise

Voraussetzung ist, dass die vorkommenden Begriffe in den Texten definierten Wikidata-Konzepten zuortenbar sind. Folglich die *word sense disambiguation* für diese gelöst ist, sodass die Konzepte in *Wikidata*-Instanzen übertragen werden können.

Zur Anwendung des entwickelten Verfahrens wird ein Goldstandart von Ideen mit annotierten Konzepten verwendet. Über das Set an Konzepten wird ein Knowledge Graph aus Wikidata extrahiert. Welche Prädikate für den Graph dieser Anwendung relevant sind und bis zu welcher Tiefe der Graph erschlossen wird, muss evaluiert werden. Zunächst sind die relevanten Prädikate in Wikidata *subclass of (P279), instance of (P31), part of (P361)*

Auf dem erstellten Graph für eine Menge von Ideen werden relevante Superklassen mittels *Spreading Activation* gefunden. Diese Superklassen werden den einzelnen Ideen zugeordnet, falls sie Konzept enthalten, die im *Subtree* der Superklasse liegen.

Wie die genau Umgesetzt von Spreading Activation auf dem Knowledge Graph aussieht, wird nach weiterer Auseinandersetzung bestimmt.

Um Superklassen die weniger dem intuitiven Verständnis von relevanten Topics und Cluster entsprechen und Ideen die falsch zugeordnet wurden, zu entdecken und sie aus dem Topic-Modell zu entfernen, folgt im letzten Schritt ein Abgleich der Superklasse mit der Lösung eines probabilistischen Ansatzes. Wenn ein Topic mit hoher Überschneidung der enthaltenen Ideen gefunden wird, bleibt die Superklasse erhalten, im anderen Fall nicht. Dabei gilt es empirisch einen guten Schwellenwert der Übereinstimmung s festzulegen, ab dem Superklassen erhalten bleiben.

Seien $A = \{i_1, \dots i_n\}$ und $B = \{i_1, \dots i_m\}$, wobei i_j jeweils die zugehörigen Ideen sind. Der Schnittmengen-Quotient s ergibt sich dann:

$$s = \frac{\|A \cap B\|}{\|A \cup B\|}$$

(Mehr eine einfache Idee)

1.5 Technische Umsetzung

Die Software wird als Jupyter Notebook in Python realisiert. Das Knowledge Graph Modell, wird aus Wikidata mit SPARQL querys extrahiert. Für den probabilistischen Ansatz werden die entsprechenden Software-Pakete importiert.

1.6 Erster Terminplan

- 12. Juli Proposel an Michael
- 12.-16. August Vortrag: Vorstellung des Bachelor Themas
- 19. August 1. Betreuer*in der TU-Berlin finden.
- 21. August Anmeldung des Bachelorarbeit an der TU Berlin

- 21. September Die Technische Umsetzung ist fertig und die Ergebnisse aufbereitet (Visuell). Experten Analyse des Ergebnisses Mit Prof. Dr. Claudia Müller-Birn, Michael Tebbe und Maximilian Mackeprang.
- 31. September ist die Bachelorarbeit fertig geschrieben, Zeit für Korrekturlesungen.
- Abgabe am 15. Oktober 2019

2 Wie geht es nach dem Exposé weiter?

Nachdem die Phase der Exposé-Erstellung abgeschlossen ist (das kann bis zu drei Iterationen dauern), können Sie mit der Erstellung der eigentlichen Abschlussarbeit beginnen. Bitte nutzen Sie die Inhalte des Exposés gleich als inhaltlichen Rahmen für die Arbeit (vor allem in Kapitel 1). Ihnen wird wieder eine LaTeXVorlage zur Verfügung gestellt. In dieser Vorlage finden Sie wieder viele Informationen und Hilfestellungen zur Erstellung der Arbeit. Sie sollten nun Ihre Arbeit anmelden. Das entsprechende Formular finden Sie auf den Institutsseiten (Link).Bitte bringen Sie das ausgefüllte Formular zu einer unserer Sitzungen mit. Ich unterschreibe es und leite es weiter. Nun sollten Sie auch bald darüber nachdenken, wer der Zweitgutachter Ihrer Arbeit sein könnte. Ich berate Sie dabei gern.

Viele weitere, nützliche Informationen finden Sie in der Prüfungsordnung Ihres Studiengangs. Bitte lesen Sie den Sie betreffenden Absatz im Anhang (2).

Literatur

[Ble03] David M Blei. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3(Journal of Machine Learning Research 3):30, 2003.

[Cre97] Fabio Crestani. Application of Spreading Activation Techniques in Information Retrieval. *Artificial Intelligence Review*, 11:453–482, December 1997.

[Har16] Brian Harrington. SaskNet: A Spreading Activation Based Semantic Network. page 48, 2016.

[TB10] Kilian Thiel and Michael R. Berthold. Node Similarities from Spreading Activation. In *2010 IEEE International Conference on Data Mining*, pages 1085–1090, Sydney, Australia, December 2010. IEEE.

[ZI17] G. Zhu and C. A. Iglesias. Computing Semantic Similarity of Concepts in Knowledge Graphs. *IEEE Transactions on Knowledge and Data Engineering*, 29(1):72–85, January 2017.

Anhang I: Auszug Prüfungsordnung Bachelor

FU-Mitteilungen

§ 5
Bachelorarbeit und mündliche Abschlussprüfung

(1) Die Bachelorarbeit soll zeigen, dass die Studentin oder der Student in der Lage ist, ein Thema aus dem Bereich der Informatik unter Anleitung nach wissenschaftlichen Methoden in einer vorgegebenen Zeit zu bearbeiten und seine Arbeit und die Ergebnisse selbständig darzustellen, wissenschaftlich einzuordnen und zu dokumentieren.

(2) Die Bearbeitungsdauer einer Bachelorarbeit beträgt zwölf Wochen.

(3) Studierende werden auf Antrag zur Bachelorarbeit zugelassen, wenn sie

1. die Module
   - Datenstrukturen und Datenabstraktion
   - Grundlagen der Theoretischen Informatik
   - Logik und Diskrete Mathematik
   - Analysis oder Analysis I
   - Lineare Algebra oder Lineare Algebra I sowie
   - Rechnerarchitektur

   erfolgreich absolviert haben,
2. im Bachelorstudiengang Informatik zuletzt an der Freien Universität Berlin immatrikuliert gewesen sind.

(4) Dem Antrag auf Zulassung zur Bachelorarbeit sind Nachweise über das Vorliegen der Voraussetzungen gemäß Abs. 3 beizufügen, ferner die Bescheinigung einer prüfungsberechtigten Lehrkraft über die Bereitschaft zur Übernahme der Betreuung der Bachelorarbeit sowie eine Erklärung, dass die oder der Studierende nicht an einer anderen Hochschule im gleichen Studiengang, im gleichen Fach oder in einem Modul, welches einem der im Bachelorstudiengang Informatik studierten Modulen vergleichbar ist, Leistungsnachweise endgültig nicht erbracht oder Prüfungsleistungen endgültig nicht bestanden hat oder sich in einem schwebenden Prüfungsverfahren befindet. Der zuständige Prüfungsausschuss entscheidet über den Antrag.

(5) Der Prüfungsausschuss gibt in Abstimmung mit der Betreuerin bzw. dem Betreuer das Thema der Bachelorarbeit aus. Thema und Aufgabenstellung müssen so beschaffen sein, dass die Bearbeitung innerhalb der Bearbeitungsfrist abgeschlossen werden kann. Ausgabe und Fristeinhaltung sind aktenkundig zu machen.

(6) Als Beginn der Bearbeitungszeit gilt das Datum der Ausgabe des Themas durch den Prüfungsausschuss. Das Thema kann einmalig innerhalb der ersten drei Wochen zurückgegeben werden und gilt dann als nicht ausgegeben. Ausnahmsweise kann der Prüfungsausschuss auf begründeten Antrag im Einvernehmen mit der Betreuerin bzw. dem Betreuer die Bearbeitungszeit der Bachelorarbeit um bis zu vier Wochen verlängern. Bei der Abgabe hat die bzw. der Studierende schriftlich zu versichern, dass sie bzw. er die Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt hat.

(7) Die Bachelorarbeit ist von zwei Prüfungsberechtigten zu bewerten, die vom Prüfungsausschuss bestellt werden. Einer der beiden Prüfer soll die Betreuerin bzw. der Betreuer der Bachelorarbeit sein. Mindestens einer der beiden Prüfer muss dem Kreis der Professorinnen und Professoren des Instituts für Informatik angehören.

(8) Die Ergebnisse der Bachelorarbeit werden im Rahmen einer mündlichen Abschlussprüfung, bestehend aus einem etwa 15-minütigen Vortrag mit anschließender etwa 15-minütiger Diskussion und Prüfungsgespräch, vorgestellt und wissenschaftlich eingeordnet und verteidigt.

(9) Voraussetzung für die Teilnahme an der mündlichen Abschlussprüfung ist die Abgabe der Bachelorarbeit. Der Prüfungstermin wird rechtzeitig in geeigneter Form bekannt gegeben.

(10) Die mündliche Abschlussprüfung wird von denjenigen Prüfungsberechtigten, welche die Bachelorarbeit bewertet haben, abgenommen.

(11) Ist die Note der Bachelorarbeit oder die Note der mündlichen Abschlussprüfung nicht mindestens „ausreichend" (4,0), so dürfen Bachelorarbeit und mündliche Abschlussprüfung einmal wiederholt werden.

84 FU-Mitteilungen 6/2007 vom 08.02.2007

Abbildung 1: Auszug Prüfungsordnung Bachelor