Bachelorarbeit

1. **Motivation:** Was ist das Problem, wie will ich das lösen ⇒ Begründung
2. **Konzept**, um Topic Modeling (Pipeline)Was muss ich machen um die Aufgabe zu erledigen, Feature engineering, etc.
3. **Algorithmus Model** aufbauen (theoretisch)
4. Implementierung
5. Evaluierung

**Was ist das Ziel:**

Mit dieser Bachelorarbeit werden verschiedene Einflussfaktoren und Parameter der Latent Dirichlet Allocation am Beispiel der ZVO untersucht. Das Ergebnis beschreibt, wie geeignet die implementierte LDA Methode von Gensim für das Klassifizieren von Emails zu Sachbearbeitern im ZVO ist.

**LDA:**

Bei der Latent Dirichlet Allocation werden Dokumente als eine Verteilung von Themen beschrieben. Diese Themen sind wiederum Verteilungen von verschiedenen Wörtern. Mehrere Dokumente ergeben zusammen einen Korpus, über den die **Wörter-Themen** Verteilung gleich ist, während die **Dokument-Themen** Verteilung variieren kann.

**LDA bei ZVO:**

Das Ziel ist, ein neues Dokument thematisch klassifizieren zu können, um es dem richtigen Sachbearbeiter weiterleiten zu können. Dafür ist es notwendig, zu wissen, welcher Sachbearbeiter welche Daten bereits bearbeitet hat, um die Ähnlichkeiten zu einem neuen Dokument herzustellen. Im Moment gibt es 18 Kategorien, denen händisch einkommende Emails zugeordnet werden. Das Ergebnis der Bachelorarbeit soll Aufschluss darüber geben, ob diese Kategorien die effizienteste Aufteilung der inhaltlichen Themen der ZVO in Bezug auf die verfügbaren Sachbearbeiter darstellen.

Titel

Topic-Modellierung für die Zuordnung von Nachrichten zu Sachnbearbeitern

Abstrakt

Einleitung

**Motivation**

Die digitalisierte Welt generiert täglich riesige Mengen an neuen Informationen. Die Kapazitäten, die ein Mensch aufbringen kann, um solche Massen an Daten zu organisieren und zu verstehen, sind schon lange übertroffen. Topic Modeling (dt. Themenmodellierung) beschreibt eine Gruppe von Verfahren, die es ermöglichen, große elektronische Datensammlungen automatisiert zu durchsuchen, organisieren und zu verstehen. Es können Muster innerhalb der Daten entdeckt und Themen extrahiert werden. Dabei stellen Themenmodelle statistische Modelle dar, die Verwendung in der Inferenz abstrakter Themen in unsortierten Datenmengen finden. In einer Welt von exponentiell wachsenden Datenmengen finden Methoden der Themenmodellierung stetig eine breitere Anwendung. Bereits heute wird Themenmodellierung in vielen Bereichen der Wirtschaft, Wissenschaft und Informationstechnologie verwendet. Um semantische Folgerungen aus Datenmengen zu generieren, gibt es verschiedene Ansätze – in dieser Arbeit wird es um die ‚Latent Dirichlet Allocation‘ gehen. Dabei werden ähnliche Wörter, die in ähnlichen Kontexten vorkommen in einem Cluster gruppiert.

**Ziele**

Diese Arbeit wird die Theorie der Themenmodellierung anhand des Beispiels des Zweckverband Ostholstein (ZVO) implementieren und die bezüglichen Parameter im Sinne der Auswertung bewerten. Der ZVO erhält jährlich ein große Menge an Kundenanfrage. Diese werden momentan händisch an die jeweils zuständige Abteilung weitergeleitet. Der Prozess soll zukünftig automatisch durch einen Klassifikationsmechanismus funktionieren. Nach der Implementation eines LDA Algorithmus zur Inferenz verschiedener Abteilungen aus den Kundenanfragen, kann die momentan händische Kategorisierung bewertet werden. Diese Arbeit beschäftigt sich mit der Vorhersage der Qualität des Klassifikators, indem die Qualität der manuell erstellten Kategorien und Kundenanfrage-Gruppen untersucht und mit den Ergebnissen verschiedener Themenmodellierung verglichen wird. Das Ergebnis einer Themenmodellierung hängt stark von der Qualität der Daten ab, die sie als Input bekommt. Diese Daten durchlaufen eine Reinigungsphase, bevor sie klassifiziert werden, um sie in eine gut zu verarbeitende Form zu bringen.

**Konzeption**

Themenmodellierung besteht aus vielen Methoden, die meist verbreitete ist die „Latent Dirichlet Allocation (LDA)“, was als Bag of Word modelliert ist, also keine Kontextinformationen beinhaltet. Dieses Verfahren ist eine Weiterentwicklung des ‚PLSI‘, das durch zwei Dirichlet-Priors ergänzt wurde. LDA liegt ein generierender Prozess zugrunde, den zwei Dirichlet Verteilungen maßgeblich beeinflussen: die Dokument-Themen Verteilung, die die Ausprägungen verschiedener Themen in einem Dokument beschreibt, und die Themen-Wörter Verteilung, die die Wahrscheinlichkeit beschreibt, dass ein bestimmtes Wort in einer gewissen Regularität in einem Themenbereich vorkommt. Dabei geht man davon aus, dass ein Dokument eine Verteilung von Themen ist, während ein Thema als eine Verteilung über Wörter betrachtet wird.

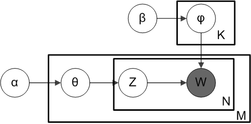
Die Wahrscheinlichkeit, dass ein bestimmtes Dokument generiert wird, ist das Produkt der Wahrscheinlichkeiten der beiden Verteilungen mit den Wahrscheinlichkeiten zweier multinomialen Verteilungen, die erst zufällig Topics, wie in der Dirichlet-Verteilung definiert, auswählen und aus diesen dann, mithilfe der zweiten Dirichlet-Verteilung, Wörter aus diesen Topics herleiten, wodurch das Enddokument entsteht. Das Enddokument wird durch höchstwahrscheinlich stark von dem gegebenen Dokument abweichen, jedoch kann durch anpassen der Dirichlet-Verteilungen eine Optimierungsproblem formuliert werden, nach dem die Dirchlet-Verteilungen gesucht werden, die ein möglichst ähnliches Dokument generieren.

Ein anderes verbreitetes Verfahren ist das „Latent Semantic Analysis“ (LSA), welches auf das Finden von sogenannten Hauptkomponenten in Dokumenten abzielt. Dadurch können sowohl ähnliche Wörter gefunden, als auch Textbereiche, die inhaltliche Überschneidungen mit einem bestimmten Begriff haben, aber das Wort selber nicht enthalten, gefunden werden. Die Methode basiert auf dem Prinzip der Singulärwertzerlegung(SVD). Als Ausgangslage wird aus einer Textsammlung eine Term-Dokument-Matrix erstellt. Diese Matrix wird in der SVD als Produkt von drei Matrizen dargestellt, von denen die mittlere eine Diagonalmatrix darstellt. Die Werte auf der Diagonalen lassen daraus die Topics der Textmenge ablesen. Auf das SVD Verfahren selbst hat der Entwickler wenig Einfluss. Um Rauschen zu verhindern, kann jedoch die Anfangsmatrix mithilfe der term-frequency und inverse-document-frequency verbessert werden, was sich auf das Gesamtergebnis auswirkt. LSA stellt sich als ein attraktives Verfahren heraus, da es Synonyme besser erkennen kann, als LDA und wird heutzutage unter anderem intensiv in dem Bereich des Digital Marketings genutzt.

Ein weiteres Verfahren, das auch mit Matrizen funktioniert, wird „Non-Negative Matrix Factorization“ (NMF) genannt. Dabei wird eine Matrix, die Wörter auf Dokumente abbildet, in zwei Teilmatrizen faktorisiert. Die erste Teilmatrix stellt die Topics in Dokumenten, die zweite die Wörter in Topics dar. Dadurch kann Speicherplatz gespart, und Themen aufgedeckt werden. Das Verfahren beginnt mit zwei möglichen faktorisierten Matrizen und verbessert sich durch die Errorfunktion iterativ, bis das Ergebnis gut genug ist. Dabei werden die errechneten Werte mit der gegebenen Matrix verglichen und angepasst.

Bevor eine Themenmodellierung auf Daten durchgeführt werden kann, müssen die Daten einem Prozess unterzogen werden. Dieser beginnt mit der Datenaquise, also der Akquirierung bestimmter relevanter Daten. Im Falle der ZVO bedeutet dies, dass es genügend Kundenanfragen gibt, die verarbeitet werden können. Wenn diese Daten bestehen werden sie auf die relevanten Wörter reduziert, aus denen eine bedeutsame Inferenz von Informationen möglich ist, sodass unter anderem die sogeneannten „Stop-Words“, also eine Menge von Verbindeungswörtern entfernt werden. Ein anderer Schritt der Datenreinigung ist das Transponieren aller Wörter in kleine Buchstaben, um eine Einheitlichkeit zu erlangen, da das Bag of Words Modell keine Reihenfolge mehr beachtet und somit große Satzanfänge irrelevant werden. Wenn die Daten in der geünschten Form vorliegen, beginnt der Schritt des Featureengineerings. Für einen Computer sind Wörter nicht so leicht zu verarbeiten, wie Zahlen, weshalb in diesem Schritt eine Quantisierung der Wörter und Überführung dieser in eine zahlenbasierte Form vorgenommen wird. Dies kann zum Beispiel in Form eines Bag-of-Words Modell, Dictionary oder TF-IDF, also einer relativen Vorkommensauflistung verschiedner Wörter über Dokumente umgesetzt werden. Nachdem die Daten in eine für den Computer kompatiblen Form gebracht wurden, kann das Themenmodell entwickelt werden.

**LDA Erklärung**



Das Prinzip der Themenmodellierung LDA ist erklärbar über die Abbildung. Der äußere Kasten beschreibt die Gesamtheit der |M| Dokumente, mit jeweils |N| Wörtern. Die beiden allein-stehenden Knoten sind die Dirichlet-Variablen, Alpha und Beta. Alpha definiert die Verteilungsstärke der Dokument-Themenverteilung, während Beta für die Wörter-Themenverteilung verantwortlich ist. W bezeichnet das Wort, das betrachtet wird, das hier als einziges Element dunkel eingefärbt ist. Das bedeutet, dass es die einzige Variable ist, die bekannt ist, nämlich als Element im „Bag of Words“ Modell. Z ist das Thema, dem das Wort zu einem bestimmten Anteil angehört. Z wird wiederum beeinflusst von Theta, das die Themenverteilung über das Dokument m definiert. Von oben wird das Wort w von Psi, das die Verteilung der Wörter über das gewählte Thema beschreibt beeinflusst.

**Der generative Prozess**

Mit LDA kann ein Dokument generiert werden, wenn die beiden Verteilungen Alpha und Beta bekannt sind. Zuerst wird eine Verteilung für die Themenausprägung in einem Dokument m ausgewählt. Folgend die Verteilung der Wörter auf die Themen. Wenn diese beiden Verteilungen festgelegt sind, kann für jede Position w im Dokument m die folgende Schleife durchlaufen werden:

1. Wähle ein Thema aus Theta
2. Wähle ein Wort dieses Themas aus Phi

Bei einem gegeben Dokument mit dem Ziel die Themenverteilung herauszufinden, wird angenommen, dass dieses Dokument wie oben beschrieben generiert wurde. Somit kann der Algorithmus nun neue Dokumente erstellen und das Ergebnis durch die Parameter, wie Alpha und Beta, anpassen, bis das Ergebnis ähnlich genug zu dem Anfangsdokument ist. Dann ist die Verteilung der Themen in diesem Dokument bekannt. Bei der Anwendung von LDA für praktische Problemstellungen, geht LDA das Prinzip rückwärts durch, d.h. für bestehende Gruppen an Dokumenten werden Verteilungen gesucht, durch die das Dokument generiert hätte werden können.

Hauptkapitel

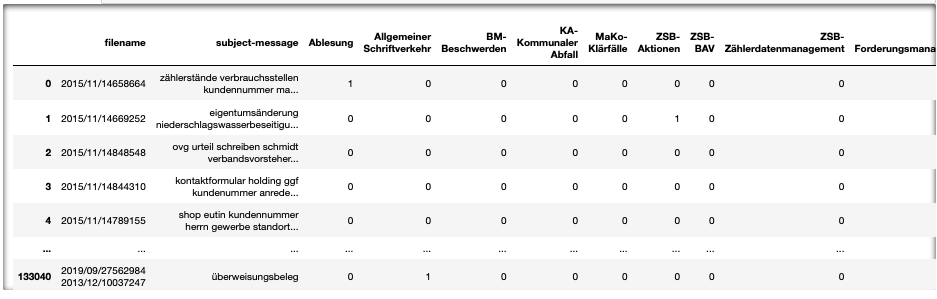
In diesem Abschnitt werden verschiedene Möglichkeiten implementiert und analysiert, wie LDA im Sinne der ZVO genutzt werden kann. Dabei ist die Zielfrage, wie am besten für ein unbekanntes Dokument die bestimmte Abteilung gefunden werden kann. Dafür müssen Korpora bestehen, die bereits durch Verteilung definiert sind, sodass die Dokument-Themen Verteilung für das neues Dokument inferiert werden kann. Wenn die Themenverteilung des neuen Dokuments gegeben ist, kann über Vergleiche der Verteilungen mit Dokumenten oder Durschnitten von Korpora eine Abteilung für das Dokument identifiziert werden.

Die Möglichkeiten, die sich für diesen Zweck ergeben sind folgende:

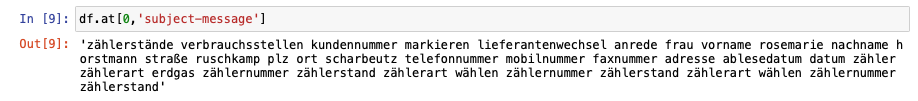
1. **Alle Dokumente ergeben ein Korpus**. Das neue Dokument wird mit dem gesamten Korpus verglichen und auf thematische Kompatibilität geprüft. Ein Gesamtbild der thematischen Aufteilung aller Dokumente kann eine mögliche effektivere Neuverteilung der Kategorien schlussfolgern.
2. **Jede Abteilung stellt einen Korpus da, deren Verteilung mit der des neuen Dokuments verglichen wird.** In diesem Fall hat jede der 18 Abteilungen eine Dokument-Themen Verteilung, die die Abteilung inhaltlich von den anderen unterscheidet. Die Wörter-Themen Verteilung ist für alle gleich, damit ein neues Dokument mit allen Abteilungsverteilungen verglichen werden kann. Dazu wird ein neues Dokument in jedem einzelnen Korpus integriert, um eine Dokument-Themen Verteilung für das neue Dokument auf Basis der gegebenen Wörter-Themen Verteilung zu inferieren.
3. **Jede Abteilung stellt einen Korpus dar, deren Dokumente alle einzeln mit der Verteilung des neuen Dokuments verglichen werden, das dann quantitativ einer Abteilung zugeordnet wird.** Die Wörter-Themen Verteilung ist gleich, während für jedes Dokument eine Dokument-Themen Verteilung errechnet wird. Wenn die Dokument-Themen Verteilung der bestehenden Dokumente und die des neuen Dokuments bereitstehen, können diese auf Ähnlichkeit überprüft werden. Da bereits bekannt ist, welcher Abteilung jedes Dokument angehört, stellen die Top X ähnlichsten Dokumente eine Verteilung der Abteilungen dar, denen das neue Dokument inhaltlich am ähnlichsten ist.

**Daten:**

Die Daten vom ZVO liegen in der folgenden Form vor:



Relevant für die Auswertung sind die subject-message und die jeweilige Abteilung. Die Tabelle verfügt über eine Matrix mit 18 Abteilungen, von denen pro subject-message eine mit einer 1 versehen ist. Dies beschreibt die Abteilung, der diese Anfrage manuell zugeordnet wurde. Die Daten in subject-message sind bereits bereinigt, also liegen vor, wie in diesem Beispiel, der Zeile 0:



Um die Einträge in eine computer-lesbare Form zu verwandeln, muss ein Dictionary erstellt werden, dass alle Wörter auf eine Anzahl ihrer Vorkommen abbildet. Dafür müssen die Wörter als alleinige Listeneinträge einlesbar sein:



**Möglichkeit 1:**

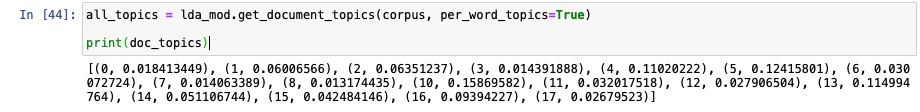
Das Ziel ist die Generierung einer Themenverteilung über alle Dokumente im Korpus. Diese könnte Aufschluss über die generelle Kategorien der Abteilungen geben. Dafür müssen zuerst alle Daten in einen großen Korpus vereinigt werden, das ist in der Methode „returnAll()“ implementiert. Danach wird diese in eine Liste aufgeteilt und in ein Dictionary verwandelt. Diese wird dann genutzt um mit der „gensim.models.LdaMulticore()“-Methode in einen LDA Modell verwandelt wird. Für dieses Modell wurde die Themenanzahl auf 18 gesetzt, da dies die aktuelle Anzahl der Themen bei dem ZVO ist.

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

Ein Bild, das Text enthält.

Automatisch generierte BeschreibungDer Output zeigt eine Liste mit 18 Elementen. Jedes Element beschreibt die Wortverteilung in einem der 18 Themen. Dabei kann LDA das Thema nicht semantisch benennen, sondern nur Cluster an häufig zusammen vorkommenden Wörtern formen. Daher kommt auch der „Latent“-Teil in der Benennung. Neben der Themen-Wortverteilung ist die Dokument-Themenverteilung von Interessen, also wie häufig kommt jedes Topic in der Gesamtmenge aller Anfragedaten vor. Diese Information errechnet man sich, indem man die folgende Methode aufruft:



Aus dieser Ausgabe wird nun ersichtlich, wie wahrscheinlich es ist, dass ein bestimmtes Thema in den Daten vorkommt. Das 10. Thema ist mit 15,9% das am häufigsten Vorkommende Thema und damit sind die Wörter „zweckverband“, „ostholstein“ und „sierksdorf“ die am stärksten vertretenen Wörter im Gesamtkorpus. Dieses Ergebnis überrascht nur bedingt, da dies der Name der Organisation ist und demensprechend häufig in Anfragen vorkommt.

Wenn man den gleichen Quelltext mit einer geringeren Themenanzahl aufruft, kommt man auf ein ähnliches Ergebnis, hier mit Beispiel 4:

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

Wieder hat das meistvertrene Thema die Wörter „Zweckverband“, „sierksdorf“ und „ostholstein“ als häufigste vorkommende Elemente. Als Einflussfaktor dient die Dirichlet Variable Alpha, die beeinflusst, wie stark die Wahrscheinlichkeitsunterschiede zwischen den einzelnen Werten ist. Im vorherigen Beispiel war Alpha 0,8 von 1.0. Im folgenden ist es 0,2 von 1.0: Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

**Möglichkeit 2:**

Hier wird ein neues Dokument einer Abteilung zugeordnet. Dabei wird über den Ähnlichkeitsoperator zwischen der Verteilung des neuen Dokuments und der des Abteilungskorpus‘ entschieden, wie passend das Dokument und die jeweilige Abteilung sind. Die Ähnlichkeit kann entweder vom durchschnittlichen Dokument-Themenverteilung des ganzen Korpus‘ sein, oder jedes Dokument einzeln mit dem neuen verglichen werden und die dann Ähnlichsten dann gezählt und abteilungsübergreifend verglichen werden. Dafür müssen zuerst alle Anfragen, die Abteilung X zugeordnet wurde in einen Text vereinigt werden, aus dem dann der Korpus entstehen kann. Dies wird mit der folgenden Methode implementiert:

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

Aus dem „Wenn“-Statement ist zu erkennen, dass nur die subject-messages zum „result0“ hinzugefügt werden, die in der Tabelle bei Abteilung 0 eine 1 haben. Das Ergebnis wird dann direkt in einzelne Listenelemente unterteilt, damit ein Dictionary erstellt werden kann.

Für die erste Abteilung wird dann, wie oben bereits durchgeführt, ein Korpus generiert und dessen durchschnittliche Verteilung errechnet:

Ein Bild, das Text enthält.

Automatisch generierte Beschreibung

Wenn die Verteilung der Themen in den Dokumenten des Korpus gegeben ist, kann nun ein neues Dokument in diesen Korpus integriert werden und auf Basis der Wort-Themenverteilung eine Dokument-Themenverteilung errechnet werden. Für die Inferenz dieser Themen im neuen Dokument verwenden wir die folgende Methode:

TODO – ein Dokument dem Korpus der Abteilung 0 also lda\_mod0 hinzufügen und Dokument-Themenverteilung für new0 inferieren.

Konklusion

Quellen