

# Verse des Alten Testaments clustern

## Belegarbeit

eingereicht am Fachbereich

### **Informatik**

der Hochschule Zittau/Görlitz (HAW)

als Prüfungsleistung im Fach

### **Data Mining**

vorgelegt von:

**Christof Ochmann (35989)**

**Ingo Körner (40586)**

Görlitz, 11. Juli 2012

Betreuer: Prof. ten Hagen

## **Abstract**

In diesem Projekt werden die Verse des Alten Testaments geclustert. Einander ähnliche Verse sollen durch Clusteralgorithmen in das selben Cluster gruppiert werden. Es wird untersucht, bei wievielen Attributen und wievielen Clustern die besten Resultate erzielt werden. Der Clusteralgorithmus soll dabei auf einem handelsüblichen Laptop ausgeführt werden. Verteiltes Clustern ist nicht Gegenstand dieser Arbeit.

# Inhaltsverzeichnis

<b>Literaturverzeichnis</b>	<b>VI</b>
<b>1 Theorie</b>	<b>1</b>
1.1 Einleitung . . . . .	1
1.2 Aufgabenstellung . . . . .	1
1.3 Relevanz des Forschungsgegenstandes . . . . .	1
1.4 Der aktuelle Wissensstand . . . . .	2
1.5 Testrechner . . . . .	2
1.6 Der Aufbau des Alten Testaments . . . . .	2
1.7 ARFF . . . . .	3
<b>2 Umsetzung</b>	<b>6</b>
2.1 Wie kann das AT in das ARFF-Format überführt werden? . . . . .	6
2.2 Analyse Text2ARFFConverter . . . . .	6
2.3 Entwurf Text2ARFFConverter . . . . .	6
2.4 Nach welchen Wörtern sollte geclustert werden? . . . . .	9
2.5 Hürden beim Einlesen der ARFF-Datei . . . . .	9
2.6 SimpleKMeans . . . . .	10
2.7 Weitere Cluster-Algorithmen . . . . .	11
2.8 geclusterte Instanzen wieder in Verse verwandeln . . . . .	11
2.9 AT mit SimpleKMeans geclustert . . . . .	12
2.10 Zusammenfassung . . . . .	14
2.11 Ausblick . . . . .	15
<b>A Arbeitsaufteilung</b>	<b>21</b>
<b>B Eigenständigkeitserklärung</b>	<b>22</b>

# Abbildungsverzeichnis

1.1	Aufbau des AT	3
2.1	Analyseklassendiagramm Text2ARFFConverter	7
2.2	Entwurfsklassendiagramm Text2ARFFConverter	8
2.3	Ausführungszeiten von SimpleKMeans	10
2.4	Analyseklassendiagramm Text2ClusterFile	12
2.5	Entwurfsklassendiagramm Text2ClusterFile	13
2.6	100 Attribute, 1024 Cluster, Cluster 15	14
2.7	100 Attribute, 1024 Cluster, Cluster 49	15
2.8	100 Attribute, 1024 Cluster, Cluster 61	16
2.9	100 Attribute, 1024 Cluster, Cluster 86	17
2.10	100 Attribute, 1024 Cluster, Cluster 554	17
2.11	1000 Attribute, 32 Cluster, Cluster 4	17
2.12	1000 Attribute, 32 Cluster, Cluster 5	18
2.13	1000 Attribute, 32 Cluster, Cluster 6	19
2.14	1000 Attribute, 32 Cluster, Cluster 15	20

# Listings

# Abkürzungsverzeichnis

<b>AT</b>	Altes Testament
<b>ARFF</b>	Attribute-Relation File Format
<b>JVM</b>	Java Virtual Machine
<b>EM</b>	Expectation Maximisation

# Literaturverzeichnis

- [1] Martin, Robert C. (2008): Clean Code: A Handbook of Agile Software Craftsmanship. Prentice Hall International
- [2] Freeman, Eric (2007): Entwurfsmuster von Kopf bis Fuß. O'REILLY
- [3] <http://www.cs.waikato.ac.nz/ml/weka/arff.html> (08.06.2012)
- [4] <http://wiki.pentaho.com/display/DATAMINING> (08.06.2012)

# **1 Theorie**

## **1.1 Einleitung**

Ziel dieses Projektes ist es, ähnliche Verse im Alten Testament, kurz AT, zu finden und zu gruppieren, d.h. zu clustern. Dazu gibt es verschiedene Clustering-Algorithmen. Diese Algorithmen bzw. die von ihnen erzeugten Cluster können miteinander verglichen werden.

## **1.2 Aufgabenstellung**

Durch Clustering werden Ähnlichkeiten in großen Datenbeständen gefunden. In diesem Projekt werden mit Hilfe von Clustering-Algorithmen Verse des AT geclustert. Dabei sollen einander ähnliche Verse in einem Clustern zusammenfasst werden, d.h. Datensätze, die sich ähneln, kommen in dasselbe Cluster. In diesem Projekt werden verschiedene Clusteralgorithmen auf das Alte Testament angewendet. In Werkzeugen wie Weka oder ELKI sind diese Algorithmen bereits implementiert und können genutzt werden. Es soll nur auf einer Maschine geclustert werden, d.h. skalierbares Datamining mit Apache Mahout wird in dieser Arbeit nicht behandelt.

## **1.3 Relevanz des Forschungsgegenstandes**

Der Forschungsgegenstand dieser Arbeit ist, die Verse des AT mit verschiedenen Clusteralgorithmen auf einem handelsüblichen Laptop zu clustern. Dafür wird aus einer Menge möglicher Clustersoftware eine passende ausgewählt. Der Forschungs-



gegenstand ist relevant, da bisher noch keine Ergebnisse für das Clustern von Versen des AT mit dem Testrechner vorliegen. Ziel der Forschung ist es, geeignete Clusteralgorithmen zu finden und mit verschiedenen Parametern auszuführen - wie Anzahl Cluster bzw. Anzahl Attribute nach denen geclustert werden soll. Es wird sich vertiefend in eine Clustersoftware und den Clusteralgorithmen eingearbeitet. Das geschieht z.B. unter Zuhilfenahme von Büchern und Online-Ressourcen. In diesen Medien ist der Forschungsstand zu Software und Algorithmen dokumentiert. Bei der Erstellung der passenden Eingabeformate und zur Auswertung der geclusterten Ergebnisse sind zudem Programme zu erstellen, bei denen technische Probleme gelöst werden müssen.

## 1.4 Der aktuelle Wissensstand

Noch nicht vorhandene Kenntnisse über die Anwendung von Clusteralgorithmen auf das AT werden hauptsächlich aus Onlinere Ressourcen bezogen. Primärliteratur zur gewählten Clustersoftware ist unter <http://www.cs.waikato.ac.nz/ml/weka/> zu finden. Unter dieser Adresse wird die Clustersoftware Weka vorgestellt. Auf der Seite wird auch auf das passende Eingabeformat ARFF eingegangen, auf dem dann die Clusteralgorithmen arbeiten. [4]

## 1.5 Testrechner

Der Testrechner besteht aus einem Laptop mit einem 64 Bit Dual Core Prozessor von AMD mit 2.1 GHz und 8GB RAM.

## 1.6 Der Aufbau des Alten Testaments

Jeder Vers des AT steht in einer eigenen Zeile, unabhängig, ob der Vers eher kurz ist, wie „Am Anfang schuf Gott Himmel und Erde.“

oder eher lang ist wie: „Da wurden berufen des Königs Schreiber zu der Zeit im dritten Monat, das ist der Monat Sivan, am dreiundzwanzigsten Tage, und wurde

geschrieben, wie Mardochai gebot, an die Juden und an die Fürsten, Landpfleger und Hauptleute in den Landen von Indien bis an das Mohrenland, nämlich hundert und siebenundzwanzig Länder, einem jeglichen Lande nach seiner Schrift, einem jeglichen Volk nach seiner Sprache, und den Juden nach ihrer Schrift und Sprache.“

Die Zeilennummer identifiziert somit einen Vers eindeutig.

Neben den Versen gibt es auch noch Kapitelüberschriften wie z.B. „Kapitel 1“ und Buchnamen wie z.B. „1. Buch Mose“, die auch eine eindeutige Zeilennummer haben. Zur Vereinfachung wird auch bei Buchnamen und Kapiteln von Versen gesprochen. Die Verse eines Buches im AT sind zwar auch durchnummeriert, diese Nummerierung wird aber ignoriert, da sie nur innerhalb eines Kapitels eindeutig ist.

```
1 1. Buch Mose
2 Kapitel 1
3 1. Am Anfang schuf Gott Himmel und Erde.
4 2. Und die Erde war wüst und leer, und es war finster
5 3. Und Gott sprach: Es werde Licht! und es ward Licht.
```

Abbildung 1.1: Aufbau des AT

## 1.7 ARFF

Damit die Clusteralgorithmen auf das AT angewandt werden können, muss das AT vorher in das ARFF-Format [3] (Attribute-Relation File Format) umgewandelt werden.

ARFF (Attribute-Relation File Format) ist eine Textdatei, die eine Liste von Instanzen beschreibt, die sich eine Menge von Attributen teilen. Eine Instanz wäre in unserem Beispiel ein Vers des AT. Die Attribute, die sich ein Vers mit anderen Versen teilt, sind die Wörter aus denen der Vers besteht. Der Vers „1. Am Anfang schuf Gott Himmel und Erde.“ besteht aus der Attribut-Menge {1, Am, Anfang, schuf, Gott, und, Erde, Himmel}

Die Elemente „1“, „Himmel“, „und“, „Erde“ teilt sich der Vers z.B. mit dem Vers „1. Also ward vollendet Himmel und Erde mit ihrem ganzen Heer.“

Man kann sagen, dass diese zwei Instanzen bzw. Verse deswegen eine gewissen Ähnlichkeit haben, da sie sich einige gemeinsame Attribute teilen.

Eine ARFF-Datei besteht aus zwei Teilen. Dem Header- und dem Data-Teil. Im Header-Teil befinden sich die Attribute. Zu beachten ist, dass es sich bei dem Wert eines Attributes um die Häufigkeit handelt, wie oft das Attribut im Vers vorkommt. Jedes Attribute hat einen Datentyp. Im vorliegenden Fall handelt es sich um Zahlen. Im Datenteil befinden sich die Instanzen, d.h. die Verse. Sie bestehen aus den Häufigkeiten, die mit Komma getrennt sind. Eine Instanz im Datenteil hat so viele mit Komma getrennte Zahlen, wie es Attribute im Header gibt. Jede Zahl steht für ein Attribut im Header. Die Position eines Attributes im Header stimmt mit der Position des Attributes in der Instanz überein. D.h. das die dritte Zahl einer Instanz für das Attribut „Anfang“ steht, da dieses auch an dritter Position im Header auftaucht.

Im folgenden sind drei Verse gegeben:

1. Buch Mose

Kapitel 1

1. Am Anfang schuf Gott Himmel und Erde.

Für diese drei Verse werden zehn Attribute definiert:

% 1. Title: AT @RELATION AT

@ATTRIBUTE 1 NUMERIC

@ATTRIBUTE Am NUMERIC

@ATTRIBUTE Anfang NUMERIC

@ATTRIBUTE schuf NUMERIC

@ATTRIBUTE Gott NUMERIC

@ATTRIBUTE Himmel NUMERIC

@ATTRIBUTE und NUMERIC

@ATTRIBUTE Erde NUMERIC

@ATTRIBUTE Fluss NUMERIC

@ATTRIBUTE Feld NUMERIC

Im Daten-Teil wird jeder Vers durch eine Instanz ausgedrückt. Die Instanz hat soviele Zahlen wie es Attribute im Header-Teil gibt. Eine 0 bedeutet, das Attribut kommt im Vers nicht vor. Eine 1 bedeutet, das Attribut kommt im Vers einmal vor.

@DATA

1,0,0,0,0,0,0,0,0,0

1,0,0,0,0,0,0,0,0,0

1,1,1,1,1,1,1,1,0,0

## 2 Umsetzung

### 2.1 Wie kann das AT in das ARFF-Format überführt werden?

Dazu wird ein Programm in Java namens Text2ARFFConverter entwickelt. Es liest alle Wörter die im AT vorkommen ein, filtert doppelte dabei aus. Dann zählt es, wie häufig die Wörter im AT vorkommen. Es sortiert dann die Wörter nach Häufigkeit. Mit einem Startparameter kann angegeben werden, wie viele Wörter bzw. Attribute in den Header der zu erzeugenden ARFF-Datei geschrieben werden sollen. Wird das Programm mit 100 gestartet, werden bei der Erzeugung der ARFF-Datei die 100 am häufigsten vorkommenden Wörter als Attribute in die ARFF geschrieben. Im Datenteil der ARFF-Datei werden die Instanzen erzeugt. Dazu wird das AT Zeilenweise eingelesen. Wenn sich ein Wort unter den z.B. häufigsten 100 Wörtern befindet, wird gezählt wie oft es im Satz vorkommt. Wurden alle Wörter eines Verses gezählt, kann die Instanz geschrieben werden.

### 2.2 Analyse Text2ARFFConverter

In Abbildung 2.1 auf Seite 7 ist das Analyseklassendiagramm des Text2ARFF-Converters zu sehen.

### 2.3 Entwurf Text2ARFFConverter

In Abbildung 2.2 auf Seite 8 ist das Entwurfklassendiagramm des Text2ARFF-Converters zu sehen.

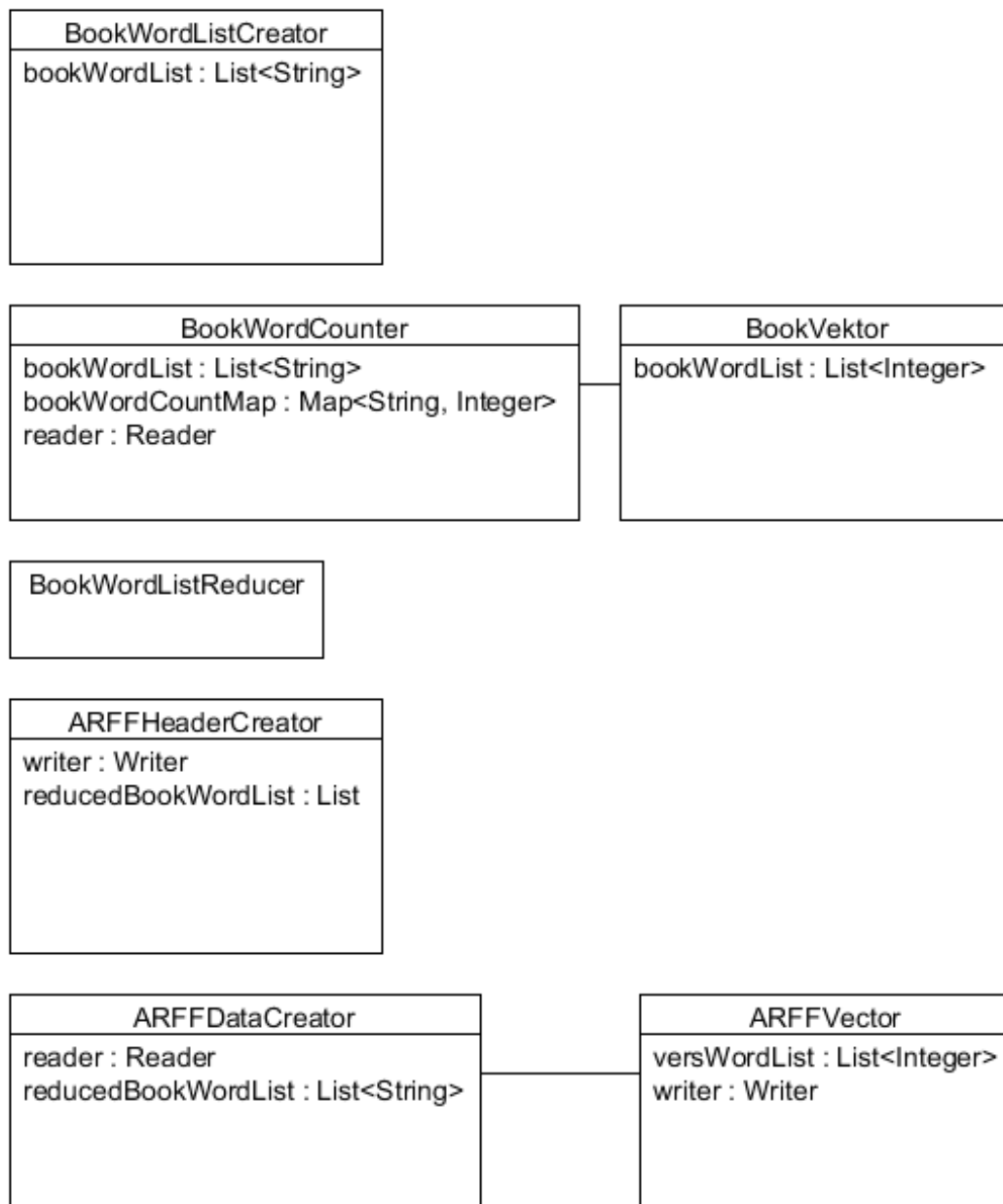


Abbildung 2.1: Analyseklassendiagramm Text2ARFFConverter

[\[1\]](#) [\[2\]](#)

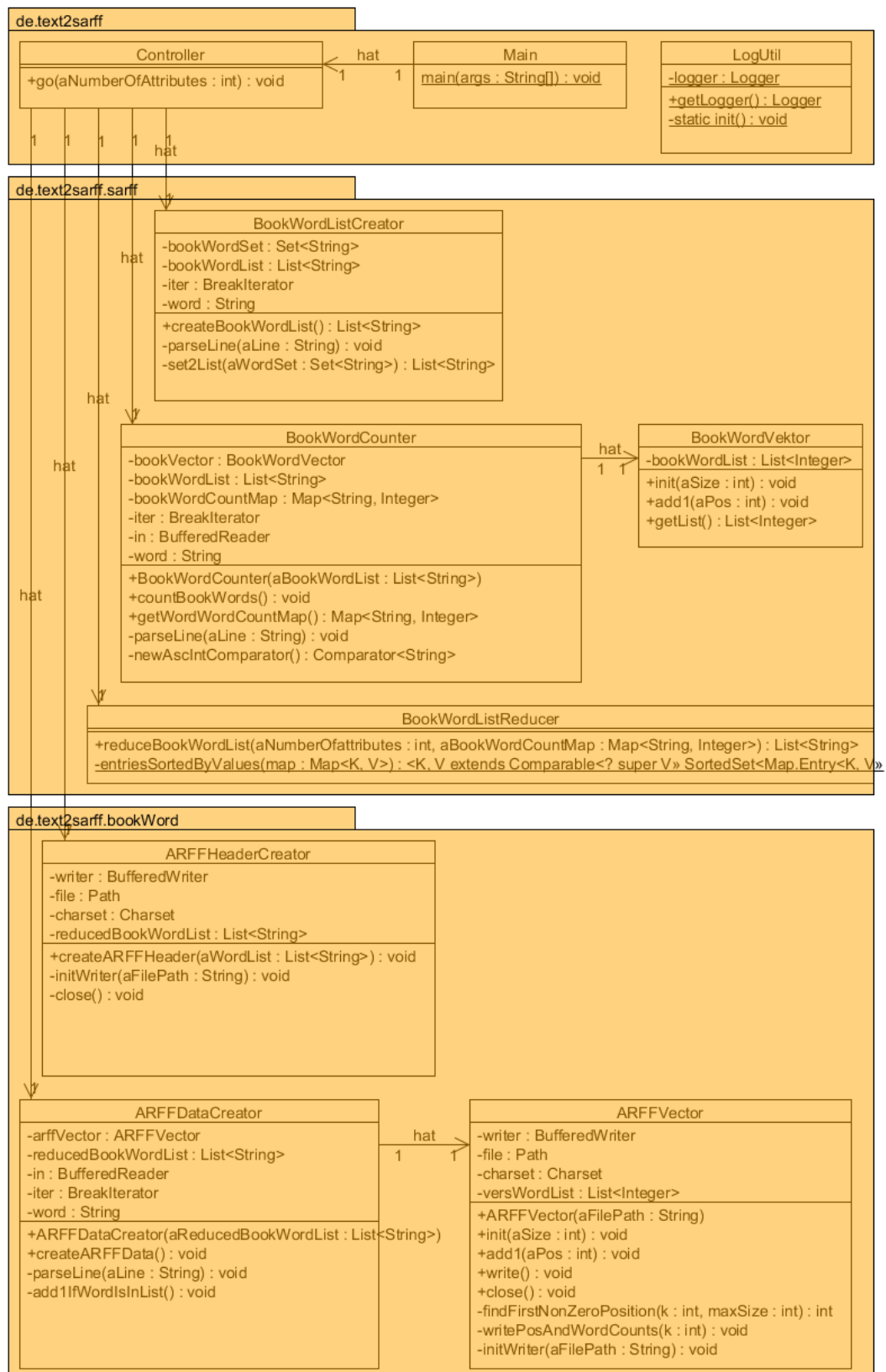


Abbildung 2.2: Entwurfsklassendiagramm Text2ARFFConverter

## 2.4 Nach welchen Wörtern sollte geclustert werden?

Bei Versuchen hat sich gezeigt, dass die Cluster ausgewogener gefüllt sind, wenn die häufigsten Wörter gewählt werden. Werden die seltensten Wörter gewählt, entstehen viele leere Instanzen und es sammeln sich alle Verse in einigen wenigen Clustern und die restlichen Cluster sind fast leer. Wahrscheinlich ist das damit zu begründen, dass beim Clustern ähnliche Verse in ein Cluster kommen. Wenn nun aber durch die seltensten Wörter die Verse alle sehr unterschiedlich sind, wird es für den Algorithmus schwierig, Ähnlichkeiten zu finden. Mit einem Wort, dass nur einmal vorkommt, kann man einen Vers gut von anderen Versen unterscheiden, aber es wird nicht helfen, zu sehen, wo die Ähnlichkeiten zu anderen Versen sind. Wenn dagegen Wörter gewählt werden, die häufig vorkommen, entstehen Cluster, die ausgewogener gefüllt sind. Es kommt mit höherer Wahrscheinlichkeit zwei Verse in dasselbe Cluster, wenn sie an den selben Positionen die gleichen Häufigkeiten haben.

## 2.5 Hürden beim Einlesen der ARFF-Datei

Die erzeugte ARFF-Datei wird für das AT 900 MB groß. In Weka kommt auf dem Testrechner beim Einlesen der 900MB großen ARFF-Datei die Fehlermeldung „OutOfMemory“. Und Weka hat sich darauf hin geschlossen. In der Konfigurationsdatei RunWeka.ini konnte der Parameter maxheap nur auf höchstens 1550m gesetzt werden. Wird mehr Platz für den heap space angegeben, kommt eine Fehlermeldung von der JVM. Darauf hin wurde von einem 32Bit-Java auf ein 64Bit-Java gewechselt. Damit verschwanden beide Fehlermeldungen und es konnte der maxheap auf 8000m gesetzt werden.

Wird die ARFF-Datei eingelesen, braucht Weka bzw. die JVM dafür 6,8 GB Arbeitsspeicher. Auf dem Entwicklungsrechner stehen aber nur 8GB zur Verfügung. Es wird davon ausgegangen, dass für die Anwendung einiger Clusteralgorithmen der restliche Arbeitsspeicher nicht ausreicht und die Geschwindigkeit durch Swapping ausgebremst wird.

Um die Größe der ARFF-Datei zu reduzieren, gibt es das Format Sparse ARFF, dass auch große Datenmengen kompakt speichern kann.



Sparse ARFF Dateien sind gewöhnlichen AARF Dateien ähnlich, außer das Attribut mit dem Wert 0 nicht repräsentiert werden. Nicht-Null Attribute werden durch die Attributnummer und den Wert angegeben.

Durch Sparse ARFF konnte die Dateigröße von 900MB auf 4 MB reduziert werden. Die Datei wird dadurch auch schneller von Weka eingelesen und Weka braucht weniger RAM. <http://wiki.pentaho.com/display/DATAMINING/>

## 2.6 SimpleKMeans

Beim Clustern des AT auf dem Testrechner wurden in Abhängigkeit von Attributanzahl und Anzahl der Cluster folgende Ausführungszeiten erzielt.

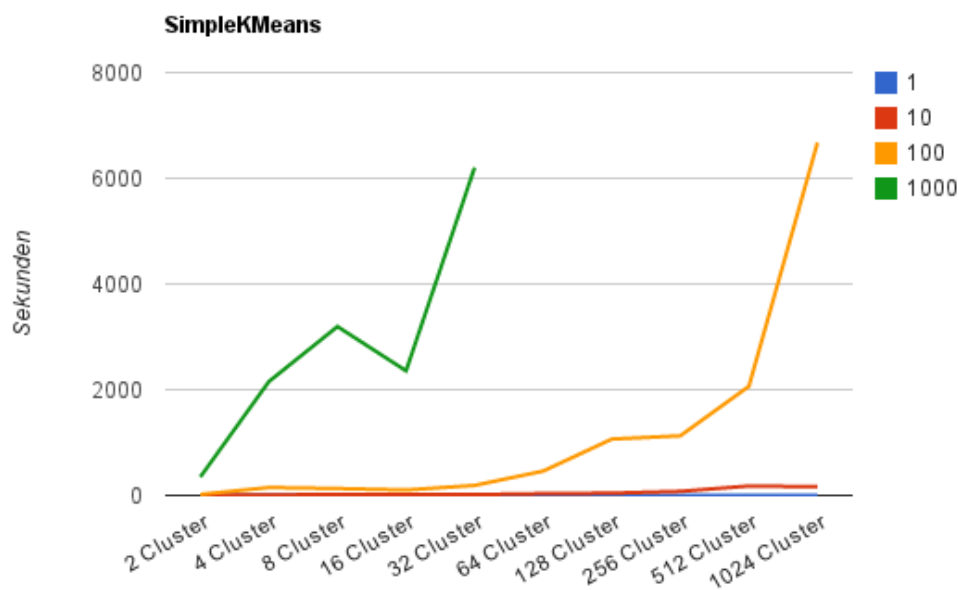


Abbildung 2.3: Ausführungszeiten von SimpleKMeans

In dem Diagramm ist erkennbar, dass die Ausführungszeit steigt, je mehr Attribute und je mehr Cluster es gibt. Auf einem einzelnen Rechner dauert das Clustern des AT für 1000 Attribute und 8 Clustern schon eine knappe Stunde. Für mehr Attribute oder mehr Cluster stößt der Testrechner an seine Grenzen.

## 2.7 Weitere Cluster-Algorithmen

Neben SimpleKMeans wurde auch XMeans auf das AT angewandt. Bei XMeans grenzt man die Cluster, die er finden könnte mit einer oberen und unteren Schranke ein. XMeans ist ein vergleichsweise schneller Clusteralgorithmus, der das AT selbst mit der maximalen Anzahl von 8950 Attributen in 77 Minuten clustert. Dabei findet er drei Cluster. Als obere Schranke wurden zwei und als obere Schranke wurden 16 Cluster angegeben.

Neben SimpleKMeans und XMeans wurden weitere Clustering-Algorithmen angewandt, die aus verschiedenen Gründen nicht weiter verfolgt wurden. Das Hierarchical Clustering braucht bei nur 5 Attributen mehr als 8GB Arbeitsspeicher um das AT zu clustern. Deswegen konnte dieser Algorithmus auf dem Testrechner nicht ausgeführt werden.

Der EM Algorithmus (expectation maximisation) benötigt auf dem Testrechner bei zehn Attributen 87 Minuten und findet dabei 7 Cluster. Auf ein clustern mit 100 oder gar 1000 Attributen wurde aus Zeitgründen verzichtet.

Der Clusteralgorithmus DBScan bringt beim starten die Fehlermeldung: „Problem evaluating cluster: null“. Auch er wurde nicht weiter verfolgt.

## 2.8 geclusterte Instanzen wieder in Verse verwandeln

Um aus den geclusterten Instanzen wieder Verse zu machen, die in die jeweils passende Cluster-Datei geschrieben werden, wird ein Programm namens Text2ClusterFile entwickelt. Das Programm gruppiert anhand der Clusternummer und der Zeilennummer die Verse. So wird erkenntlich, welche Verse wie geclustert wurden. In Abbildung 2.4 auf Seite 12 ist das Analyseklassendiagramm von Text2ClusterFile zu sehen.

In Abbildung 2.4 auf Seite 12 ist das Entwurfsklassendiagramm von Text2ClusterFile zu sehen.

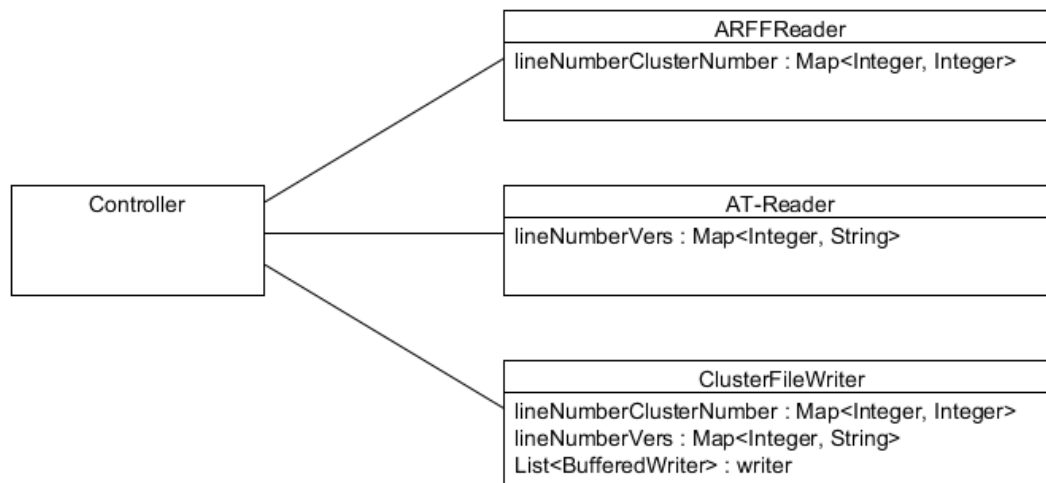


Abbildung 2.4: Analyseklassendiagramm Text2ClusterFile

## 2.9 AT mit SimpleKMeans geclustert

In Abbildungen 2.6, 2.7, 2.8, 2.9, 2.10, 2.11, 2.12 sowie 2.13 sind Verse zu sehen, die mit SimpleKMeans in ein Cluster gruppiert wurden.

Bei Abbildung 2.7 auf Seite 15 fällt auf, dass nicht nur in jedem Vers das Wort Herrn vorkommt, und dass auch alle mit der Versnummer 2 anfangen, sondern auch dass oft die Formulierung „Und er tat, was dem Herrn übel gefiel“ erscheint.

Bei Abbildung 2.10 auf Seite 17 fällt auf, dass nicht nur in jedem Vers das Wort „einen“ vorkommt, sondern dass auch alle mit der Versnummer 1 anfangen.

Auf Abbildung 2.9 auf Seite 17 ist zu sehen, dass hier alle Verse vorkommen, die die Formulierung „Was aber mehr von ... zu sagen ist und was er getan hat, und seine Macht, und wie er...“ beinhalten.

Wenn es eine hohe Anzahl von Attributen gibt, kann der Cluster-Algorithmus viele Wörter erfassen, mit denen Verse sich ähneln können. Bei vielen ähnlichen Versen braucht der Algorithmus dann aber auch viele Cluster auf die er die vielen ähnlichen Verse aufteilen kann.

Wenn der Algorithmus nur wenige Attribute aber viele Cluster hat, dann gruppiert er die wenigen Attribute, die ihm zur Verfügung stehen über alle Cluster auf. Wie

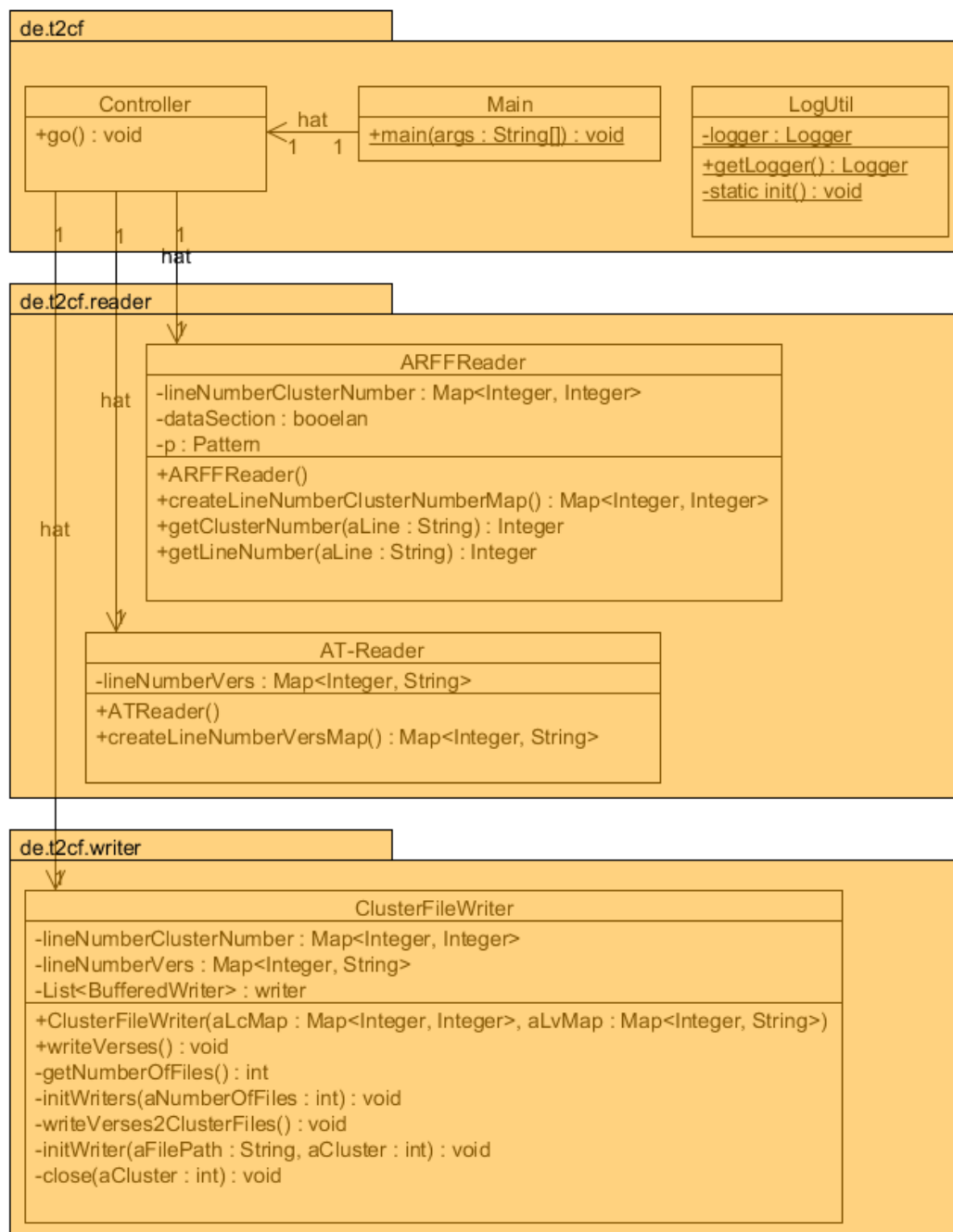


Abbildung 2.5: Entwurfsklassendiagramm Text2ClusterFile

gut er dabei ist, kann man bei wenigen Wörtern leicht übersehen, da die meisten Wörter der Verse gar nicht berücksichtigt wurden. In den Meisten der 1024 Cluster gibt es aber ein Wort, das alle Verse in diesem Cluster gemeinsam haben.

Es ist zu vermuten, dass wenn alle Attribute genommen werden, die mindestens zwei mal vorkommen, bei steigender Clusteranzahl, die Verse die in einem Cluster vorkommen, zwar im Mittel weniger werden, aber dafür immer ähnlicher.

```

1. Und der Herr redete mit Mose und sprach:
1. Und der Herr redete mit Mose und sprach:
1. Und der Herr redete mit Mose und sprach:
1. Und der Herr sprach zu Mose: Haue dir zwei steinerne Tafeln, wie die ersten waren, da
1. Und Mose versammelte die ganze Gemeinde der Kinder Israel und sprach zu ihnen: Das is
1. Da arbeiteten Bezaleel und Oholiab und alle weisen Männer, denen der Herr Weisheit un
1. Und der Herr redete mit Mose und sprach:
1. Und der Herr redete mit Mose und sprach:
1. [5:20] Und der Herr redete mit Mose und sprach:
8. [6:1] Und der Herr redete mit Mose und sprach:
1. Und der Herr redete mit Mose und sprach:
1. Und der Herr redete mit Mose und Aaron und sprach zu ihnen:
1. Und der Herr redete mit Mose und sprach:
1. Und der Herr redete mit Mose und Aaron und sprach:
1. Und der Herr redete mit Mose und sprach:
1. Und der Herr redete mit Mose und Aaron und sprach:
1. Und der Herr redete mit Mose, nachdem die zwei Söhne Aarons gestorben waren, da sie v
1. Und der Herr redete mit Mose und sprach:
1. Und der Herr redete mit Mose und sprach:
1. Und der Herr redete mit Mose und sprach:
1. Und der Herr redete mit Mose und sprach:
1. Und der Herr sprach zu Mose: Sage den Priestern, Aarons Söhnen, und sprich zu ihnen: I
1. Und der Herr redete mit Mose und sprach:
1. Und der Herr redete mit Mose und sprach:
1. Und der Herr redete mit Mose und sprach:
1. Und der Herr redete mit Mose auf dem Berge Sinai und sprach:
1. Und der Herr redete mit Mose und sprach:
1. Und der Herr redete mit Mose und Aaron und sprach:
1. Dies ist das Geschlecht Aarons und Mose's zu der Zeit, da der Herr mit Mose redete au
1. Und der Herr redete mit Mose und Aaron und sprach:
1. Und der Herr redete mit Mose und sprach:
1. Und der Herr redete mit Mose und sprach:
1. Und der Herr redete mit Mose und sprach:
1. Und der Herr redete mit Mose und sprach:
1. Und der Herr redet mit Mose und sprach:
1. Und der Herr redete mit Mose und sprach:

```

Abbildung 2.6: 100 Attribute, 1024 Cluster, Cluster 15

## 2.10 Zusammenfassung

Beim Clustern des AT nach Versen wurde je nach Algorithmus viel RAM oder viel CPU Zeit benötigt. Der Testrechner ist dabei relativ schnell an seine Grenzen gestoßen. Wird eine größere Menge von Daten geclustert, sollte das nicht nur auf einem einzelnen Rechner erfolgen, sondern verteilt.

```

11 2. Und die Philister riefen ihre Priester und Weissager und sprachen: Was sollen wir mit der Lade des He
12 2. Und von dem Tage an, da die Lade des Herrn zu Kirjath-Jearim blieb, verzog sich die Zeit so lange, bi
13 2. Der Geist des Herrn hat durch mich geredet, und seine Rede ist auf meiner Zunge.
14 2. Aber das Volk opferte noch auf den Höhen; denn es war noch kein Haus gebaut dem Namen des Herrn bis :
15 2. Und das Wort des Herrn kam zu ihm und sprach:
16 2. [3] Und Joas tat, was recht war und dem Herrn wohl gefiel, solange ihn der Priester Jojada lehrte,
17 2. und er tat, was dem Herrn übel gefiel, und wandelte nach den Sünden Jerobeams, des Sohnes Nebats, der
18 2. und er tat, was dem Herrn übel gefiel, doch nicht wie die Könige Israels, die vor ihm waren.
19 2. Und er tat, was dem Herrn übel gefiel, nach den Greueln der Heiden, die der Herr vor den Kinder Israe
20 2. Und er tat was dem Herrn wohl gefiel, und wandelte in allem Wege seines Vaters David und wich nicht,
21 2. Und da David die Brandopfer und Dankopfer ausgerichtet hatte, segnete er das Volk im Namen des Herrn
22 2. Aber das Wort des Herrn kam zu Semaja, dem Mann Gottes, und sprach:
23 2. Und es gingen ihm entgegen hinaus Jehu, der Sohn Hananis, der Seher, und sprach zum König Josaphat: :
24 2. Und Joas tat, was dem Herrn wohl gefiel, solange der Priester Jojada lebte.
25 2. Und er tat, was dem Herrn wohl gefiel, doch nicht von ganzem Herzen.
26 2. Und er tat, was dem Herrn wohl gefiel, ganz wie sein Vater Usia getan hatte, nur ging er nicht in der
27 2. Und er tat, was dem Herrn wohl gefiel, wie sein Vater David.
28 2. und tat, was dem Herrn wohl gefiel, und wandelte in den Wegen seines Vaters David und wich weder zur
29 2. Und er bestellte die Priester zu ihrem Dienst und stärkte sie zu ihrem Amt im Hause des Herrn
30 2. Die Könige der Erde lehnen sich auf, und die Herren ratschlagen miteinander wider den Herrn und seine
31 2. Bringet dem Herrn die Ehre seines Namens: betet an den Herrn im heiligen Schmuck!

```

Abbildung 2.7: 100 Attribute, 1024 Cluster, Cluster 49

## 2.11 Ausblick

In einem Folgeprojekt könnten Werkzeuge wie Apache Mahout verwendet werden, mit dem auch verteilt geclustert werden kann.

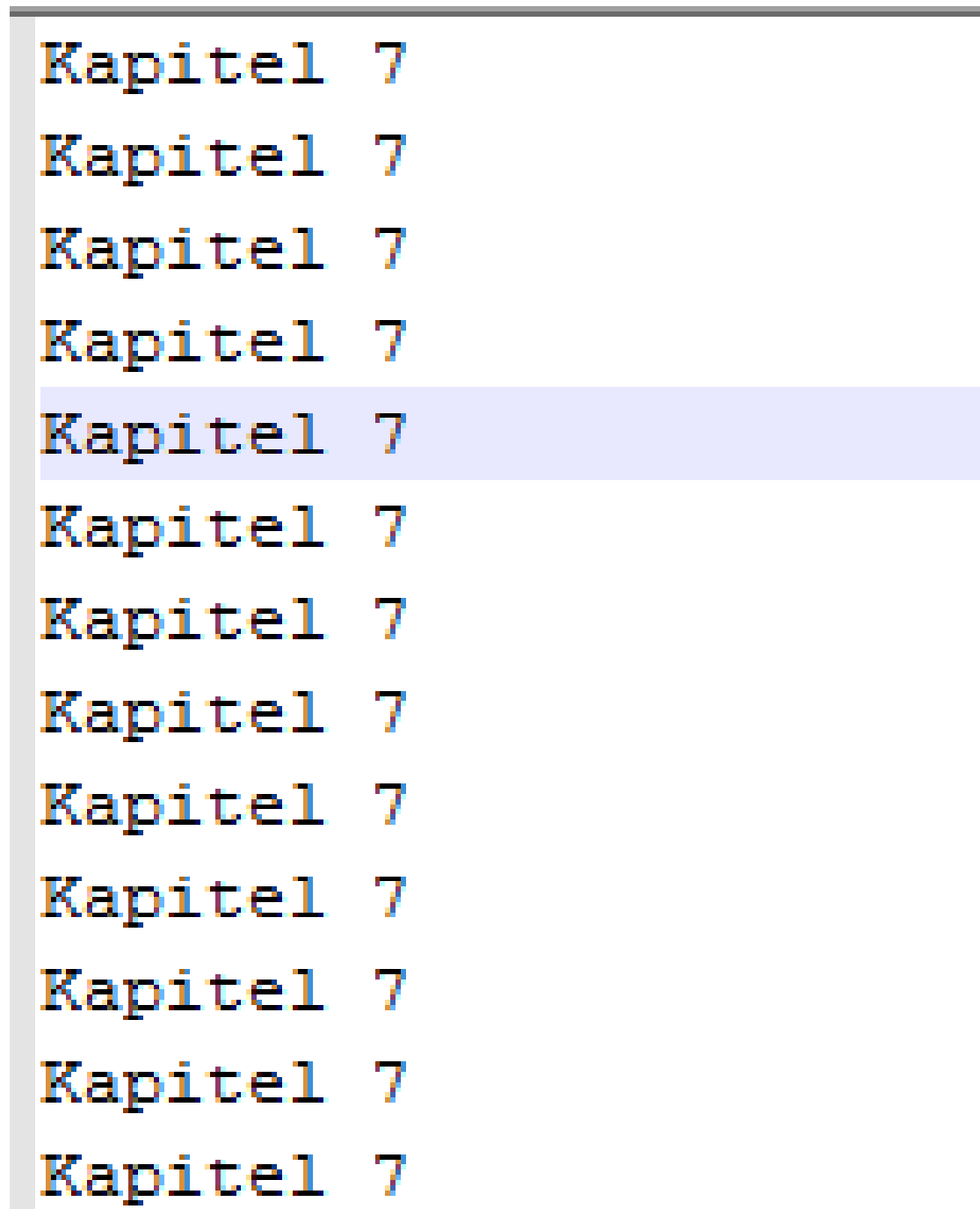


Abbildung 2.8: 100 Attribute, 1024 Cluster, Cluster 61

1 23. Was aber mehr von Asa zu sagen ist und alle seine Macht und alles, was er getan hat, und die Städte, 1  
 2 27. Was aber mehr von Omri zu sagen ist und alles, was er getan hat, und seine Macht, die er geübt hat, s.  
 3 39. Was mehr von Ahab zu sagen ist und alles, was er getan hat, und das elfenbeinerne Haus, das er baute,  
 4 12. Was aber mehr von Joas zu sagen ist und was er getan hat und seine Macht, wie er mit Amazja, dem Köni  
 5 15. Was aber mehr von Joas zu sagen ist, was er getan hat, und seine Macht, und wie er mit Amazja, dem Kö  
 6 28. Was aber mehr von Jerobeam zu sagen ist und alles, was er getan hat, und seine Macht, wie er gestrit  
 7 20. Was mehr von Hiskia zu sagen ist und alle seine Macht und was er getan hat und der Teich und die Wass

Abbildung 2.9: 100 Attribute, 1024 Cluster, Cluster 86

1. Und Adam erkannte sein Weib Eva, und sie ward schwanger und gebar den Kain und sprach: Ich habe **einen** Mann gewo  
 1. Und Gott sprach zu Jakob: Mache dich auf und ziehe gen Beth-El und wohne daselbst und mache daselbst **einen** Alta  
 1. Und nach zwei Jahren hatte Pharao **einen** Traum, wie er stünde am Nil  
 1. Du sollst **einen** Altar machen von Akazienholz, fünf Ellen lang und breit, daß er gleich viereckig sei, und drei  
 1. Das ist's auch, was du ihnen tun sollst, daß sie mir zu Priestern geweiht werden. Nimm **einen** jungen Farren und  
 1. Du sollst auch **einen** Räuchaltar machen, zu räuchern, von Akazienholz,  
 1. Und es begab sich darnach, daß Absalom ließ sich machen **einen** Wagen und Rosse und fünfzig Mann, die seine Traba  
 1. Nach diesen Geschichten begab sich's, daß Naboth, ein Jesreeliter, **einen** Weinberg hatte zu Jesreel, bei dem Pal  
 1. Und David hielt **einen** Rat mit den Hauptleuten über tausend und über hundert und mit allen Fürsten  
 1. Er machte auch **einen** ehernen Altar, zwanzig Ellen lang und breit und zehn Ellen hoch.  
 1. Ich habe **einen** Bund gemacht mit meinen Augen, daß ich nicht achtete auf eine Jungfrau.  
 1. Wohlan, ich will meinem Lieben singen, ein Lied meines Geliebten von seinem Weinberge: Mein Lieber hat **einen** We  
 1. Und er spricht: Wenn sich ein Mann von seinem Weibe scheidet, und sie zieht von ihm und nimmt **einen** andern Mann  
 1. So spricht der Herr zu mir: Gehe hin und kaufe dir **einen** leinenen Gürtel und gürte damit deine Lenden und mache  
 1. So spricht nun der Herr: Gehe hin und kaufe dir **einen** irdenen Krug vom Töpfer, samt etlichen von den Ältesten d  
 1. So spricht der Herr: Siehe, ich will **einen** scharfen Wind erwecken wider Babel und wider ihre Einwohner, die sich  
 1. Im zweiten Jahr des Reiches Nebukadnezars hatte Nebukadnezar **einen** Traum, davon er erschrak, daß er aufwachte.  
 1. Im ersten Jahr Belsazers, des Königs zu Babel, hatte Daniel **einen** Traum und Gesichte auf seinem Bett; und er schi

Abbildung 2.10: 100 Attribute, 1024 Cluster, Cluster 554

16. [32:17] und tat sie unter die Hand seiner Knechte, je eine Herde besonders, und sprach zu ihnen: **Gehet** vor mir hin und lasset Raum zwischen e  
 55. Da nun das ganze Ägyptenland auch Hunger litt, schrie das Volk zu Pharao um Brot. Aber Pharao sprach zu allen Ägyptern: **Gehet** hin zu Joseph:  
 4. Da sprach der König in Ägypten zu ihnen: Du Mose und Aaron, warum wollt ihr das Volk von seiner Arbeit frei machen? **Gehet** hin an eure Dienste!  
 25. [8:21] Da forderte Pharao Mose und Aaron und sprach: **Gehet** hin, opfert eurem Gott hier im Lande.  
 8. Mose und Aaron wurden wieder zu Pharao gebracht; der sprach zu ihnen: **Gehet** hin und dienet dem Herrn, eurem Gott. Welche sind es aber, die hin  
 13. Da stand Bileam des Morgens auf und sprach zu den Fürsten Balaks: **Gehet** hin in euer Land; denn der Herr will's nicht gestatten, daß ich mit e  
 30. [5:27] Gehe hin und sage ihnen: **Gehet** heim in eure Hütten.  
 23. Und da er euch aus Kades-Barnea sandte und sprach: **Gehet** hinauf und nehmet das Land ein, das ich euch gegeben habe! wart ihr ungehorsam dem M  
 8. Da machten sich die Männer auf, daß sie hingingen; und Josua gebot ihnen, da sie hin wollten gehen, das Land aufzuschreiben, und sprach: **Gehet**  
 20. Und sie geboten den Kindern Benjamin und sprachen: **Gehet** hin und lauert in den Weinbergen.  
 5. sandte er aus zehn Jünglinge und sprach zu ihnen: **Gehet** hinauf gen Karmel; und wenn ihr zu Nabal kommt, so grüßet ihn von mir freundlich  
 2. von solchen Völkern, davon der Herr gesagt hatte den Kindern Israel: **Gehet** nicht zu ihnen und laßt sie nicht zu euch kommen; sie werden gewiß  
 5. Er aber sprach zu ihnen: **Gehet** hin bis an den dritten Tag, dann kommt wieder zu mir. Und das Volk ging hin.  
 6. Sie sprachen zu ihm: Es kam ein Mann herauf uns entgegen und sprach zu uns: **Gehet** wiederum hin zu dem König, der euch gesandt hat, und sprecht  
 2. Laß uns an den Jordan gehen und einen jeglichen daselbst Holz holen, daß wir uns daselbst eine Stätte bauen, da wir wohnen. Er sprach: **Gehet** h  
 13. **Gehet** hin und fraget den Herrn für mich, für dies Volk und für ganz Juda um die Worte dieses Buches, das gefunden ist; denn es ist ein großer  
 2. Und David sprach zu Joab und zu des Volkes Obersten: **Gehet** hin, zählt Israel von Beer-Seba an bis gen Dan und bring es zu mir, daß ich wisse,  
 21. **Gehet** hin und fraget den Herrn für mich und für die übrigen in Israel und Juda über die Worte des Buches, das gefunden ist; denn der Grimm de  
 11. **Gehet** heraus und schauet an, ihr Töchter Zions, den König Salomo in der Krone, damit ihn seine Mutter gekrönt hat am Tage seiner Hochzeit und  
 2. das Botschafter auf dem Meer sendet und in Rohrschiffen auf den Wassern fährt! **Gehet** hin, ihr schnellen Boten, zum Volk, das hochgewachsen und  
 20. **Gehet** aus von Babel, fliehet von den Chaldäern mit fröhlichem Schall: verkündiget und lasset solches hören; bringt es aus bis an der Welt End  
 10. **Gehet** hin, **Gehet** hin durch die Tore! bereitet dem Volk den Weg! machet Bahn, machet Bahn! räumt die Steine hinweg! werft ein Panier auf über  
 10. **Gehet** hin in die Inseln Chittim und schauet, und sendet nach Kedar und merket mit Fleiß und schauet, ob's daselbst so zugeht!  
 1. **Gehet** durch die Gassen zu Jerusalem und schauet und erfahret und suchet auf ihrer Straße, ob ihr jemand findet, der recht tue und nach dem Gla  
 12. **Gehet** hin an meinen Ort zu Silo, da vormals mein Name gewohnt hat, und schauet, was ich daselbst getan habe um der Bosheit willen meines Volk  
 5. Zu jenen aber sprach er, daß ich's hörte: **Gehet** diesem nach durch die Stadt und schlaget drein; eure Augen sollen nicht schonen noch übersehen  
 2. **Gehet** hin gen Kalne und schauet, und von da gen Hamath, die große Stadt, und zieht hinab gen Gath der Philister, welche bessere Königreiche ge  
 8. **Gehet** hin auf das Gebirge und holet Holz und bauet das Haus; das soll mir angenehm sein, und ich will meine Ehre erzeigen, spricht der Herr.  
 7. Die starken gingen und zogen um, daß sie alle Lande durchzögen. Und er sprach: **Gehet** hin und durchziehet die Erde! Und sie durchzogen die Erde

Abbildung 2.11: 1000 Attribute, 32 Cluster, Cluster 4



1 18. Dornen und Disteln soll er dir tragen, und sollst das Kraut auf dem Felde essen.  
2 6. [22:5] Wenn ein Feuer auskommt und ergreift die Dornen und verbrennt die Garben oder Getreide, das noch steht, oder den Acker, so soll der wie  
3 55. Werdet ihr aber die Einwohner des Landes nicht vertreiben vor eurem Angesicht, so werden euch die, so ihr überbleiben laßt, zu Dornen werden.  
4 7. Gideon sprach: Wohlan, wenn der Herr Sebah und Zalmuna in meine Hand gibt, will ich euer Fleisch mit Dornen aus der Wüste und mit Hecken zerdr  
5 16. Und er nahm die Ältesten der Stadt und Dornen aus der Wüste und Hecken und ließ es die Leute zu Sukkoth fühlen.  
6 40. so mögen mir Disteln wachsen für Weizen und Dornen für Gerste. Die Worte Hiobs haben ein Ende.  
7 9. [10] Ehe eure Dornen reif werden am Dornstrauch, wird sie ein Zorn so frisch wegreißen.  
8 12. Sie umgeben mich wie Bienen; aber sie erlöschen wie Feuer in Dornen; im Namen des Herrn will ich sie zerhauen.  
9 6. [7] Denn das Lachen der Narren ist wie das Krachen der Dornen unter den Töpfen; und das ist auch eitel.  
10 2. Wie eine Rose unter den Dornen, so ist meine Freundin unter den Töchtern.  
11 6. Ich will ihn wüst liegen lassen, daß er nicht beschnitten noch gehackt werde, sondern Disteln und Dornen darauf wachsen, und will den Wolken g  
12 23. Denn es wird jetzt zu der Zeit geschehen, daß wo jetzt tausend Weinstöcke stehen, tausend Silberlinge wert, da werden Dornen und Hecken sein,  
13 24. daß man mit Pfeilen und Bogen dahingehen muß. Denn im ganzen Lande werden Dornen und Hecken sein,  
14 25. daß man auch zu allen den Bergen, die man mit Hauen pflegt umzuhacken, nicht kann kommen vor Scheu der Dornen und Hecken; sondern man wird Ocl  
15 18. [17] Denn das gottlose Wesen ist angezündet wie Feuer und verzehrt Dornen und Hecken und brennt wie im dicken Wald und gibt hohen Rauch.  
16 17. Und das Licht Israels wird ein Feuer sein, und sein Heiliger wird eine Flamme sein, und sie wird seine Dornen und Hecken anzünden und verzehr  
17 4. Gott zürnt nicht mit mir. Ach, daß ich möchte mit den Hecken und Dornen kriegen! so wollte ich unter sie fallen und sie auf einen Haufen anste  
18 13. Denn es werden auf dem Acker meines Volkes Dornen und Hecken wachsen, dazu über allen Häusern der Freude in der fröhlichen Stadt.  
19 12. Und die Völker werden zu Kalk verbrannt werden, wie man abgehaue Dornen mit Feuer ansteckt.  
20 13. und werden Dornen wachsen in seinen Palästen, Nesseln und Disteln in seinen Schlössern; und es wird eine Behausung sein der Schakale und Weid  
21 6. Und du Menschenkind, sollst dich vor ihnen nicht fürchten noch vor ihren Worten fürchten. Es sind wohl widerspenstige und stachelige Dornen bei  
22 24. Und forthin sollen allenthalben um das Haus Israel, da ihre Feinde sind, keine Dornen, die da strechen, noch Stacheln, die da wehe tun, bleibe  
23 6. [8] Darum siehe, ich will deinen Weg mit Dornen vermachen und eine Wand davorziehen, daß sie ihren Steig nicht finden soll;  
24 6. Siehe, sie müssen weg vor dem Verstörer. Ägypten wird sie sammeln, und Moph wird sie begraben. Nesseln werden wachsen, da jetzt ihr liebes Göt  
25 8. Die Höhen zu Aven sind vertilgt, durch die sich Israel versündigte; Disteln und Dornen wachsen auf ihren Altären. Und sie werden sagen: Ihr Be  
26 10. Denn wenn sie gleich sind wie die Dornen, die noch ineinanderwachsen und im besten Saft sind, so sollen sie doch verbrannt werden wie dürres :

Abbildung 2.12: 1000 Attribute, 32 Cluster, Cluster 5

---

25. Denn was ich gefürchtet  
6. Denn Mühsal aus der Erde  
18. Denn er verletzt und ve  
3. Denn nun ist es schwerer  
4. Denn die Pfeile des Allm  
21. Und warum vergibst du m  
8. Denn frage die vorigen G  
14. Denn seine Zuversicht v  
17. Denn er fährt über mich  
32. Denn er ist nicht meine  
6. und zeigte dir die heiml  
11. Denn er kennt die losen  
19. Wer ist, der mit mir re  
26. Denn du schreibst mir B  
25. Denn er hat seine Hand  
34. Denn der Heuchler Versa  
22. Denn die bestimmten Jah

Abbildung 2.13: 1000 Attribute, 32 Cluster, Cluster 6

1. Und mich hob ein Wind auf und brachte mich zum Tor am Hause des Herrn, da  
 1. Und des Herrn Wort geschah zu mir und sprach:  
 1. Und des Herrn Wort geschah zu mir und sprach:  
 1. Und es kamen etliche von den Ältesten Israels zu mir und setzten sich vor  
 1. Und des Herrn Wort geschah zu mir und sprach:  
 1. Und des Herrn Wort geschah zu mir und sprach:  
 1. Und des Herrn Wort geschah zu mir und sprach:  
 1. Und des Herrn Wort geschah zu mir und sprach:  
 1. Du aber mache eine Wehklage über die Fürsten Israels  
 1. Und es begab sich im siebenten Jahr, am zehnten Tage des fünften Monats,  
 45. [21:1] Und des Herrn Wort geschah zu mir und sprach:  
 1. [6] Und des Herrn Wort geschah zu mir und sprach:  
 1. Und des Herrn Wort geschah zu mir und sprach:  
 1. Und des Herrn Wort geschah zu mir und sprach:  
 1. Und es geschah das Wort des Herrn zu mir im neunten Jahr, am zehnten Tage  
 1. Und des Herrn Wort geschah zu mir und sprach:  
 1. Und es begab sich im elften Jahr, am ersten Tage des ersten Monats, gesch  
 1. Und des Herrn Wort geschah zu mir und sprach:  
 1. Und des Herrn Wort geschah zu mir und sprach:  
 1. Im zehnten Jahr, am zwölften Tage des zehnten Monats, geschah des Herrn W  
 1. Und des Herrn Wort geschah zu mir und sprach:  
 1. Und es begab sich im elften Jahr, am ersten Tage des dritten Monats, gesc  
 1. Und es begab sich im zwölften Jahr, am ersten Tage des zwölften Monats, g  
 1. Und des Herrn Wort geschah zu mir und sprach:  
 1. Und des Herrn Wort geschah zu mir und sprach:  
 1. Und des Herrn Wort geschah zu mir und sprach:  
 1. Und du, Menschenkind, weissage den Bergen Israels und sprich: Höret des F  
 1. Und des Herrn Wort kam über mich, und er führte mich hinaus im Geist des  
 1. Und des Herrn Wort geschah zu mir und sprach:

Abbildung 2.14: 1000 Attribute, 32 Cluster, Cluster 15

# A Arbeitsaufteilung

Arbeit	C. Ochmann	I. Körner
Abstract		0
Einleitung		1.1
Aufgabenstellung		1.2
Forschungsgegenstand		1.3
akt. Wissensstand		1.4
Testrechner		1.5
Der Aufbau des AT		1.6
ARFF		1.7
Wie AT in das ARFF-Format überführen?		2.1
Analyse Text2ARFFConverter		2.2
Entwurf Text2ARFFConverter		2.3
Nach welchen Wörtern clustern?		2.4
Hürden beim Einlesen der ARFF-Datei		2.5
SimpleKMeans		2.6
Weitere Cluster-Algorithmen		2.7
Instanzen in Verse verwandeln		2.8
AT mit SimpleKMeans geclustert		2.9
Zusammenfassung		2.10
Ausblick		2.11

Tabelle A.1: Aufteilung

## **B Eigenständigkeitserklärung**

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst habe. Mir ist bekannt, dass jede Form des Plagiats mit der Note 5 (Betrugsversuch) bewertet wird.

**Ochmann, Christof**

Unterschrift:

**Körner, Ingo**

Unterschrift: