* 1. Versuchsaufbau für den Ansatz BayesLSH

Für die Implementierung des Ansatzes BayesLSH werden verschiedenen Elemente ins Spiel kommen, unter anderem ein Rechner, ein Algorithmus, eine Programmiersoftware und -sprache, Daten, den Code.

* + 1. Der Rechner

Der für diese Implementierung angewendete Rechner ist HP Pavilion Notebook, von der Firma HP hergestellt, das das Betriebssystem Windows 10 Home 64-Bit nutzt, mit einem Mikroprozessor Intel (R) Core(TM) i5-6400U CPU @2.30 GHz. Dieses Notebook verfügt auch über einen 8 GB fähigen Systemspeicher (RAM), eine Systemplatine 82OC 82.30 und ein System-BIOS F.25.

Das HP Pavilion Notebook verfügt auch über zwei Festplattenlaufwerke: Das Laufwerk 1 mit einer Speicherkapazität von 917,21 GB (davon 661,2 GB freier Speicherplatz besteht) und das Laufwerk 2 mit einer Speicherkapazität von 117,26 GB (davon 26,18 GB freier Speicherplatz besteht). Das Laufwerk 1 ist exklusiv für die Speicherung von Dateien (Partition 1) und für Recovery (Partition 2) versehen, während das Laufwerk 2 mit nur eine Partition exklusiv für Programme, Anwendungen und User versehen ist.

Nachdem den angewendeten Rechner beschrieben wurde, ist die nächste Aufgabe den entsprechenden Algorithmus für den BayesLSHs Ansatz zu erläutern.

* + 1. Der Algorithmus

Der für diesen Implementierungsteil angewendete Algorithmus ist der BayesLSH Algorithmus, deren Pseudocode beschrieben wurde im Kapitel 2. Es ist hinzuweisen, dass diesen Algorithmus als Leitfaden für die Implementierung des Ansatzes genommen wird, bzw. für das Schreiben des Codes. Weitere Details werden im Abschnitt (Code noch hinzufügen) erläutert.

Weiterhin gilt es zunächst anzugeben welche Programmiersprache und -software angewendet wurden.

* + 1. Die Programmiersoftware und -sprache

Die Firma Dibuco GmbH entwickelt normalerweise Softwarelösungen in die Programmiersprache Java oder Scala aber die verwendete Programmiersprache für diese Thesis ist R, und die Software dazu ist Rstudio. Da R Statistik mit Datenvisualisierung kombiniert und das zentrale Thema dieser Arbeit beruht auf Bayesscher Statistik wurde entschieden diese Programmiersprache anzuwenden. Dies hat unter anderem unterschiedliche Gründe. Es bestehen schon in R Bibliotheken und vorgefertigte Algorithmen, die nur zur Anwendung bereit sind. R kann unter anderem sowohl mit Spark (Big Data Framework für Analytik) als auch mit KNIME (Analytik Plattform) betrieben werden und fast alle Datenbanken können von der ausgelesen werden. Das Textreuse Paket, das später benutzt wird für die Implementierung des Ansatzes kommt vom MIT und es besteht eine große Community (Weltweit bestehend), die Pakete zur Verfügung stellt. Ein weiterer Vorteil von R wie schon im Abschnitt (1.1) erwähnt, ist dass, sie eine open Source ist. Außerdem viele wissenschaftliche Arbeiten können in R erstellt werden, bzw. Berechnungen und Graphiken können direkt in wissenschaftlichen Arbeiten gegeben werden

* + 1. Die Daten

Die in diese Implementierung angewendete Daten stammen ursprünglich aus Wikipedia, bzw. wurden als XML-Dateien von Wikipedia aufgenommen und dann mithilfe einer Software in JSON-Dateien umgewandelt um entsprechend für das Projekt BigData4Biz genutzt zu werden. Diese Daten entsprechen dann in der BigData4Biz Software Entitäten. Es wurden dann 90 Dateien im JSON Format genommen für diesen Fall.

* + 1. Die Bibliotheken/Pakete

* + 1. Der Code

Der Code dieser Implementierung wird in R geschrieben und anhand von dem existierender Packet Textreuse durchgeführt. Es werden auch ein Paar in R herunterzuladende Bibliothekpakete wie „pdftools“, zum Auslesen von PDF-Dateien, „textreuse“ zur Berechnung der Ähnlichkeit von Dokumente und zur Erkennung von wiederverwendete Passage.

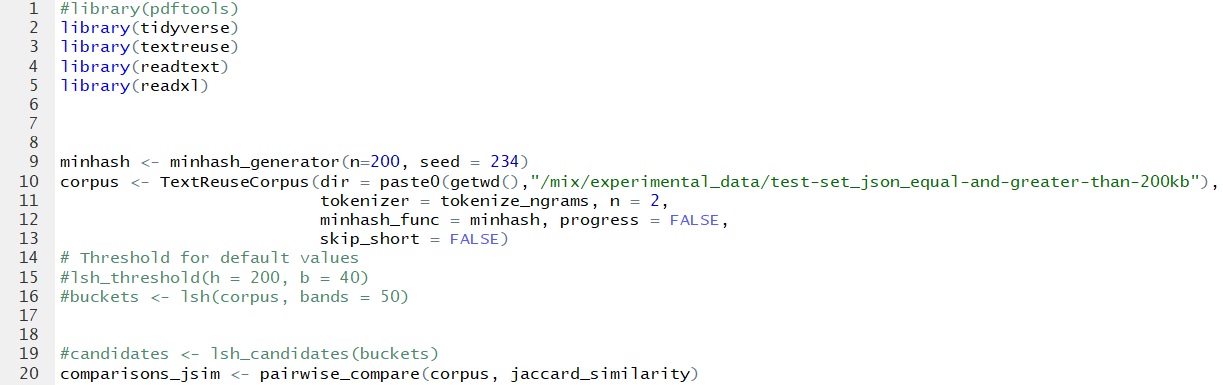


Abbildung 1: Der Code zur Implementierung des BayesLSH mit dem Jaccard Index

* + 1. Ergebnisse
  1. Versuchsaufbau für den Ansatz BayesLSH-Lite

[RPR18] <https://www.r-project.org/> (Letzter Abruf:21.06.2018)

[GNU18] <http://www.gnu.org/philosophy/free-sw> (Letzter Abruf:21.06.2018)

[HPS18] HP Support Assistant, integriertes Service in HP Pavilion Notebook