# Gliederung Thesis: Determination of Document Similarity based on Bayesian Statistics for a Big-Data Information Retrieval Solution

1. Einleitung
   1. Motivation
   2. Problemstellung
   3. Zielsetzung
   4. Aufbau der Arbeit
2. Grundlagen, Analyse von Use Cases zur Dokumentähnlichkeitsbestimmung sowie auf Bayessche Statistik basierende Ansätze zur Dokumentähnlichkeitsbestimmung
   1. Informationsrückgewinnung
      1. **Definition**
      2. **Nutzen**
      3. **Mechanismus**
      4. **Informationsrückgewinnungsmodelle**
      5. Web Informationsrückgewinnung
   2. Bayes‘ sche Statistik
      1. **Definition und Hintergrund**
      2. **Anwendungsgebiete und Nutzen**
      3. **Modelle, Parameter und Überzeugungen**
      4. Die Wahrscheinlichkeit
      5. **Der Satz von Bayes**
   3. **Analyse von Use Cases für die Dokumentähnlichkeitsbestimmung**
      1. **Finden von Dokumenten mit ähnlichen Inhalten (Duplikaten Findung)**
      2. **Verwendung von a priori Wissen**
      3. **Systemübergreifende „Fremdschlüssel“**
      4. **Profil Matching**
      5. **Email Klassifikation (Spamfilter)**
      6. Anforderungsanalyse (Gucken J)
      7. Mögliche Bestimmungstechniken
   4. Analyse existierender Ansätze zur Dokumentähnlichkeitsbestimmung basierend auf Bayes' scher Statistik.
      1. Naive Bayes
      2. Bayes LSH
      3. More Like This

* Naive Bayes
* Bayes LSH (<http://vldb.org/pvldb/vol5/p430_venusatuluri_vldb2012.pdf> : Seite 440)

1. Die Informationsrückgewinnung-Middleware-Lösung
   1. Vorstellung
   2. Anforderungen
      1. Funktionelle Anforderungen
      2. Nicht funktionelle Anforderungen
      3. Plattformanforderungen
      4. Komponentenanforderungen
      5. Anforderungen an dem Ähnlichkeitsalgorithmus
   3. High-Level Architektur
2. Auswahl von Ansätzen zur Dokumentähnlichkeitsbestimmung in der Information-Retrieval-Middleware-Lösung
   1. Kriterien der Auswahl: sagen warum diese ausgewählt werden
   2. Funktionsweise Ansatz
3. Implementierung einer Teilmenge ausgewählten Ansätzen
   1. Programmierung usw.
4. Evaluierung der Skalierbarkeit von den realisierten Ansätzen basierend auf ausgewählten Probendaten
   1. Fähigkeiten und Grenzen (Erfüllung von Anforderung)
5. Zusammenfassung und Ausblick