Text Mining und Textzusammenfassung

Jürgen Kirkovits

Doris Rongitsch

Daniela Wagenhofer

Übersicht

1. Definition
2. Prozessablauf
3. Textzusammenfassung
4. Praxisbeispiel

Definition

"Text Mining is the art and technology to extract knowledge from text"

(Gao, Chang und Han)

- **automatisierte Entdeckung relevanter Informationen aus Textdaten**
- **nicht triviales Durchforsten von Daten mit unbekanntem Ergebnis**

Ziele

- **Wissenserweiterung**
  - **durch neue bzw. neu angeordnete Information**
- **Erkennen von Zusammenhängen**
  - **zwischen Texten und Textfragmenten**

Extraktion von „Wissen“ nicht trivial!

=>Ergebnisse und Anwendung nicht vorhersehbar

Text Mining vs. Data Mining

zentraler Unterschied: zugrundeliegende Datenbasis

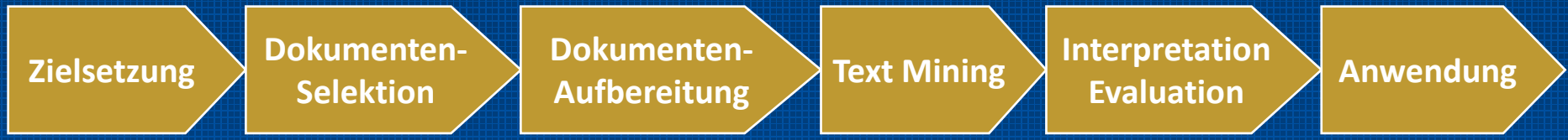
Data Mining:

- Zugriff auf klar definierte Daten/Attribute
- in einer Datenbank gespeichert

Text Mining:

- meist unstrukturierte Texte => Umwandlung in strukturiertes Format notwendig
- Aufbereitung mindestens so wichtig wie Analyse

Text-Mining-Prozess



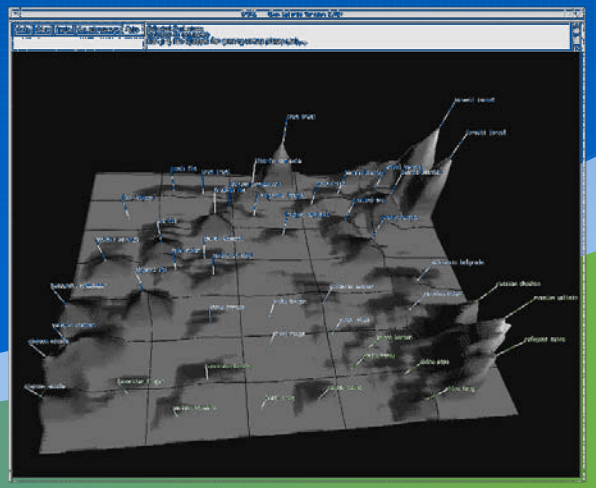
1.) Aufgabendefinition
    Marktforschung, Konkurrenzanalyse, ...
2.) Dokumentselektion
    Sammlung inhaltlich relevanter Texte
3.) Dokumentaufbereitung
    maschinelle Überführung in strukturierte Form
4.) Text-Mining-Methoden
    maschinelle Analysen (Mustererkennung => Klassifikation, Segmentierung, Abhängigkeitsbestimmungen)
5.) Interpretation und Evaluation der Ergebnisse
    Überprüfung auf Fehler und Verwendbarkeit
6.) Anwendung der Ergebnisse

Text-Mining-Aufgaben

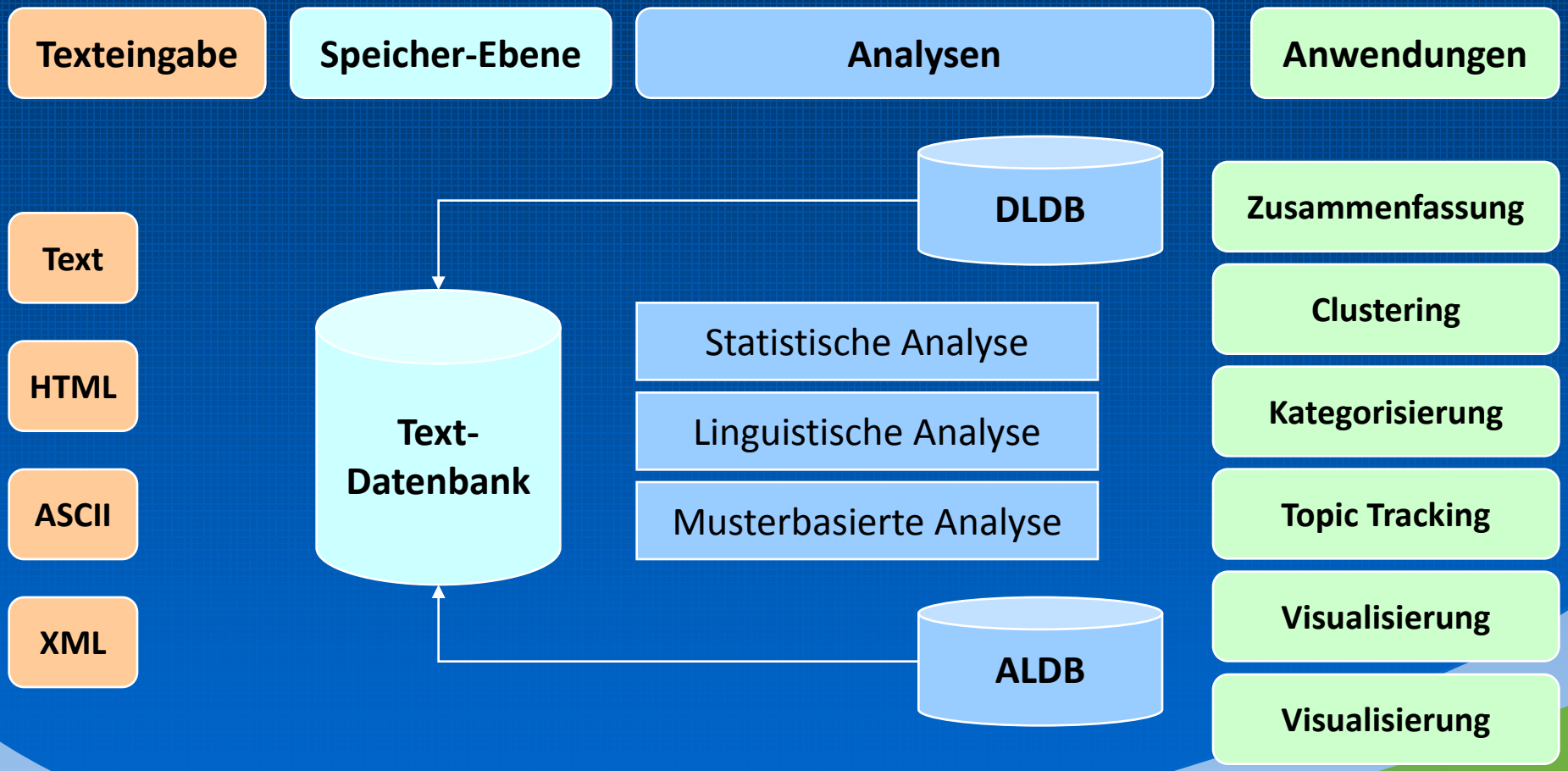
- **Informationsextraktion**
  - Auffinden und Extrahieren von Schlüsselinformationen (Zeit, Ort, Personen)
- **Topic Tracking**
  - Finden von Information anhand von Schlüsselwörtern (Tags)
- **Textzusammenfassung**
  - automatische Generierung von Abstracts

Text-Mining-Aufgaben

- Kategorisierung
  - Einteilung von Texten in vorbestimmte Kategorien
- Clusterbildung
  - Kategorisierung durch Gewichtung, fließende Grenzen
- Concept Linkage
  - Verknüpfen von Dokumenten durch Auffinden gemeinsamer Themen
- Informationsvisualisierung
- Frage-Antwort-Systeme



Dokumentaufbereitung



Morphologische Analysen

Untersuchung einzelner Wortformen und sinntragender Wortbestandteile

- **Tokenisierung**
  - Zerlegung in Wörter anhand von Satzzeichen
- **Stammformreduktion der Wörter („stemming“)**
  - „schrieb“ und „geschrieben“ werden zu „schreiben“
- **Finden von Satzgrenzen**
  - anhand von Interpunktionszeichen

Syntaktische Analysen

beschäftigen sich mit dem Satzbau bzw. mit den Beziehungen zwischen den Zeichen

- **Part-of-Speech-Tagging**
  - **ähnlich einer Wortartenbestimmung**
- **Phrase Recognition**
  - **identifiziert Wortgruppen und Phrasen**
- **Parsing (grammatikalische Analyse)**

  Bestimmung von Satzgliedern (Subjekt, Prädikat, Objekt, ...)

Semantische Analysen

beschäftigt sich mit Sinn und Bedeutung von Sprache

- Word Sense Disambiguation
  - Auflösung der Doppelsinnigkeit von Wörtern
- mit Hilfe von
  - Glossar (Liste von Begriffen mit zugehöriger Erklärung)
  - Taxonomie (Hierarchie von Begriffen)
  - Thesaurus (Erweiterung durch Ähnlichkeits- und Synonymrelation)
    - Bsp.: Thesaurus, Topic Map und Ontologie sind „ähnlich“, „Data Mining“ und „KDD“ sind Synonyme
  - Topic Map („Topics“, Assoziationen, Gültigkeitsbereiche)
  - Ontologie (Regelwerk für Zusammenhänge zwischen Objekten mittels „Wenn-Dann“-Beziehungen, Zuweisungen, logischen Verknüpfungen und weiteren Funktionen)

Textzusammenfassung

Aufgabe: Reduzierung großer Informationsmengen auf deren wichtigste Inhalte, um diese rasch auffassen zu können

Arten der Zusammenfassung:

1. **Abstract vs. Extract**
   - **Abstract: zusammenhängender Text, der den Quelltext ersetzt**
   - **Extract: Aneinanderreihung von Schlüsselwörtern**
2. **indikativ vs. informativ**
   - **indikativ: zeigt an, ob der Quelltext evtl. interessant ist**
   - **informativ: ersetzt den Quelltext**
3. **generisch vs. benutzerorientiert**
   - **generisch: eher allgemein gehalten**
   - **benutzerorientiert: Ergebnis an den jeweiligen Nutzer angepasst**

Statistische Verfahren

Oberflächenansatz:

- erstes System (Luhn)
  - berechnete Wichtigkeit der Wörter aufgrund ihrer Häufigkeit
- trainierbarer Zusammenfasser (Kupiec, Pedersen & Chen)
  - statistischer Ansatz, der manuell erstellte Abstracts zur Merkmalsbestimmung (Satzlänge, Indikatorphrasen, Absatzstruktur, Schlüsselwörter, Akronyme) heranzieht
  - errechnet für jeden Satz, mit welcher Wahrscheinlichkeit er aufgrund seiner Belegung mit Merkmalen von einem Menschen in ein Abstract aufgenommen werden würde

Wissensbasierende Verfahren

Entitätenebenenansatz:

Der Quelltext wird auf folgende Merkmale untersucht:

- **Ähnlichkeit von Ausdrücken**
- **Abstand zwischen zwei Wörtern, die miteinander verbunden sein könnten**
- **Betrachtung des Kontexts**
- **Untersuchung auf Synonyme, Hyperonyme, Homonyme**
- **Koreferenzen (sich aufeinander beziehende Ausdrücke)**
- **logische Beziehungen (in Widersprüchen oder Zustimmungen)**
- **syntaktische Analysen**

Wissensbasierende Verfahren

Diskursebenenansatz:

- Quelltext wird als Einheit gesehen und auf Handlungs- und Erzählstränge, Argumentationslinien und rhetorische Strukturen hin untersucht

Realisierung in der Praxis meist durch hybride Ansätze

- ein System kompensiert Schwächen des anderen
- linguistische Verfahren durch größeres „Hintergrundwissen“ bei komplexen Texten im Vorteil
- Tendenz hin zu benutzerorientierten Zusammenfassungen

Praxisbeispiel

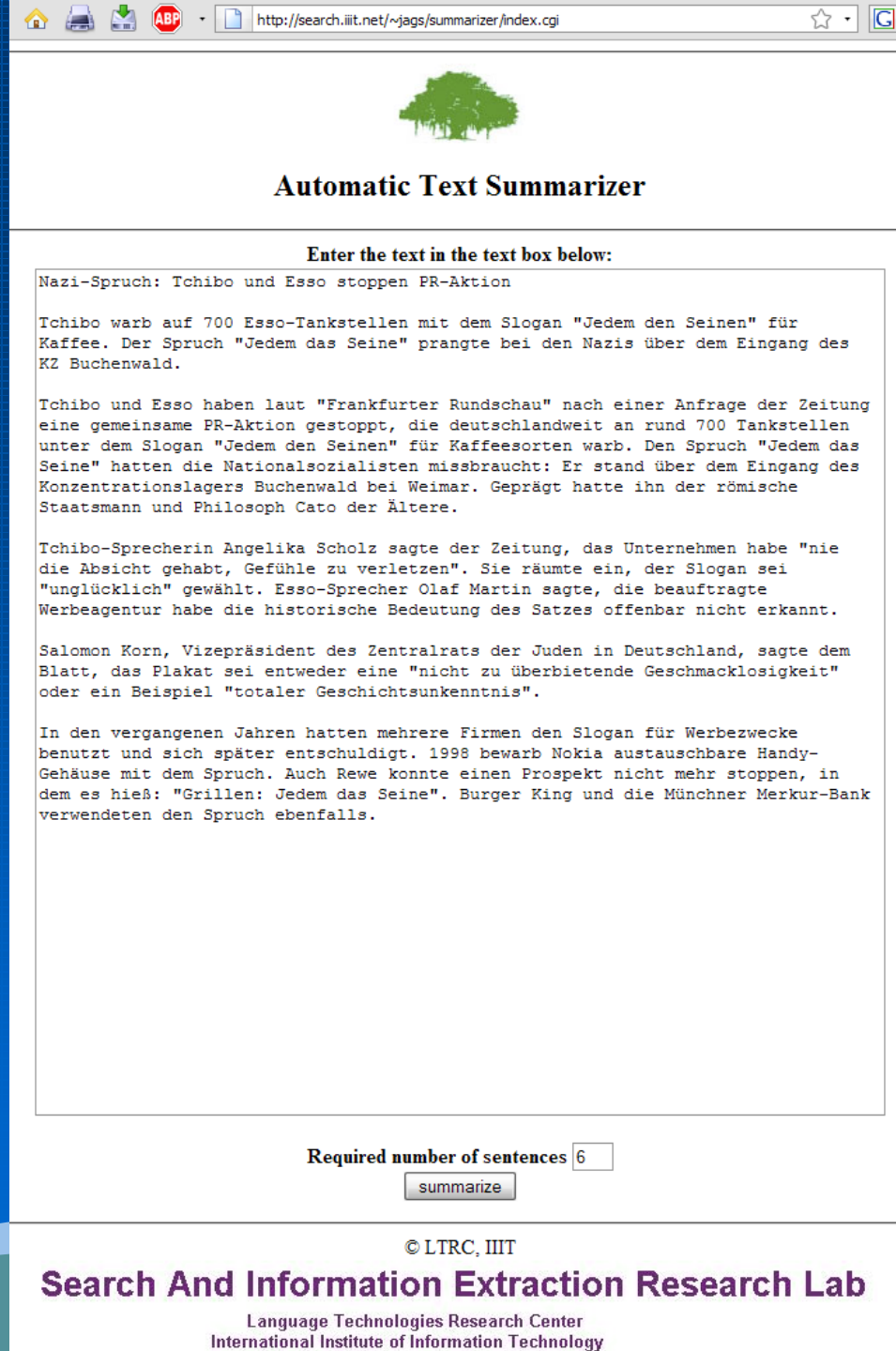
Anhand einer Pressemeldung wurden verschiedene Möglichkeiten zur Textzusammenfassung getestet

- Knowledge.Works Analyst Tool Box: Auto Summarizer (http://mskw.cipher-sys.com/Lectern/summary_submitter.asp)
- Automatic Text Summarizer (http://search.iiit.net/~jags/summarizer/index.cgi)
- AutoZusammenfassen-Funktion (Microsoft Word 2003)

Praxisbeispiel

Automatic Text Summarizer

eingestellt auf die Ausgabe von maximal 6 Sätzen



Ergebnis

Required number of sentences 6

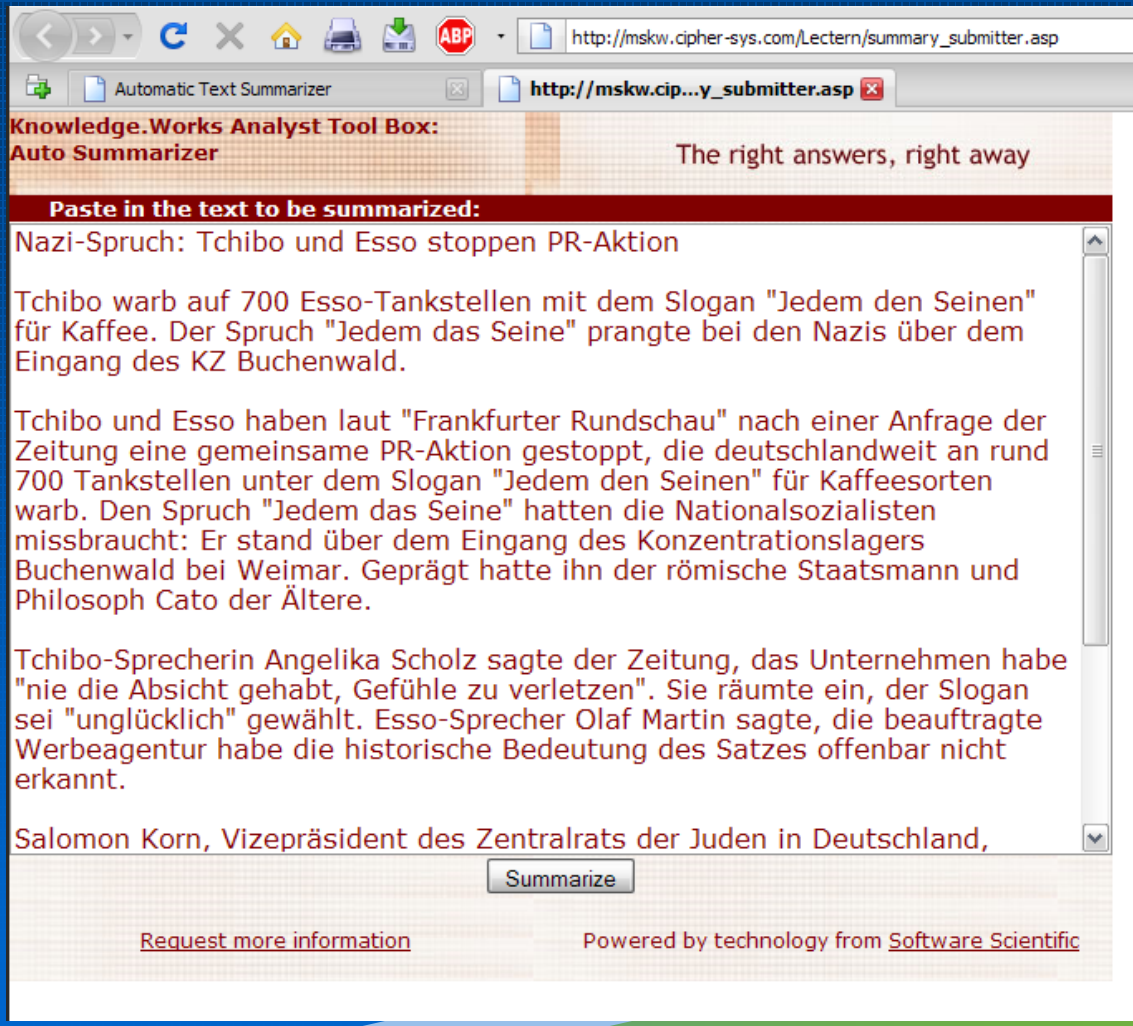
summarize

Nazi-Spruch: Tchibo und Esso stoppen PR-AktionTchibo warb auf 700 Esso-Tankstellen mit dem Slogan "Jedem den Seinen" für Kaffee.
Tchibo und Esso haben laut "Frankfurter Rundschau" nach einer Anfrage der Zeitung eine gemeinsame PR-Aktion gestoppt, die deutschlandweit an rund 700 Tankstellen unter dem Slogan "Jedem den Seinen" für Kaffeesorten warb.
Den Spruch "Jedem das Seine" hatten die Nationalsozialisten missbraucht: Er stand über dem Eingang des Konzentrationslagers Buchenwald bei Weimar.
Tchibo-Sprecherin Angelika Scholz sagte der Zeitung, das Unternehmen habe "nie die Absicht gehabt, Gefühle zu verletzen".
Sie räumte ein, der Slogan sei "unglücklich" gewählt.
Burger King und die Münchner Merkur-Bank verwendeten den Spruch ebenfalls.

Praxisbeispiel

Knowledge.Works Analyst Tool Box: Auto Summarizer

Slider auf Einstellung „2“

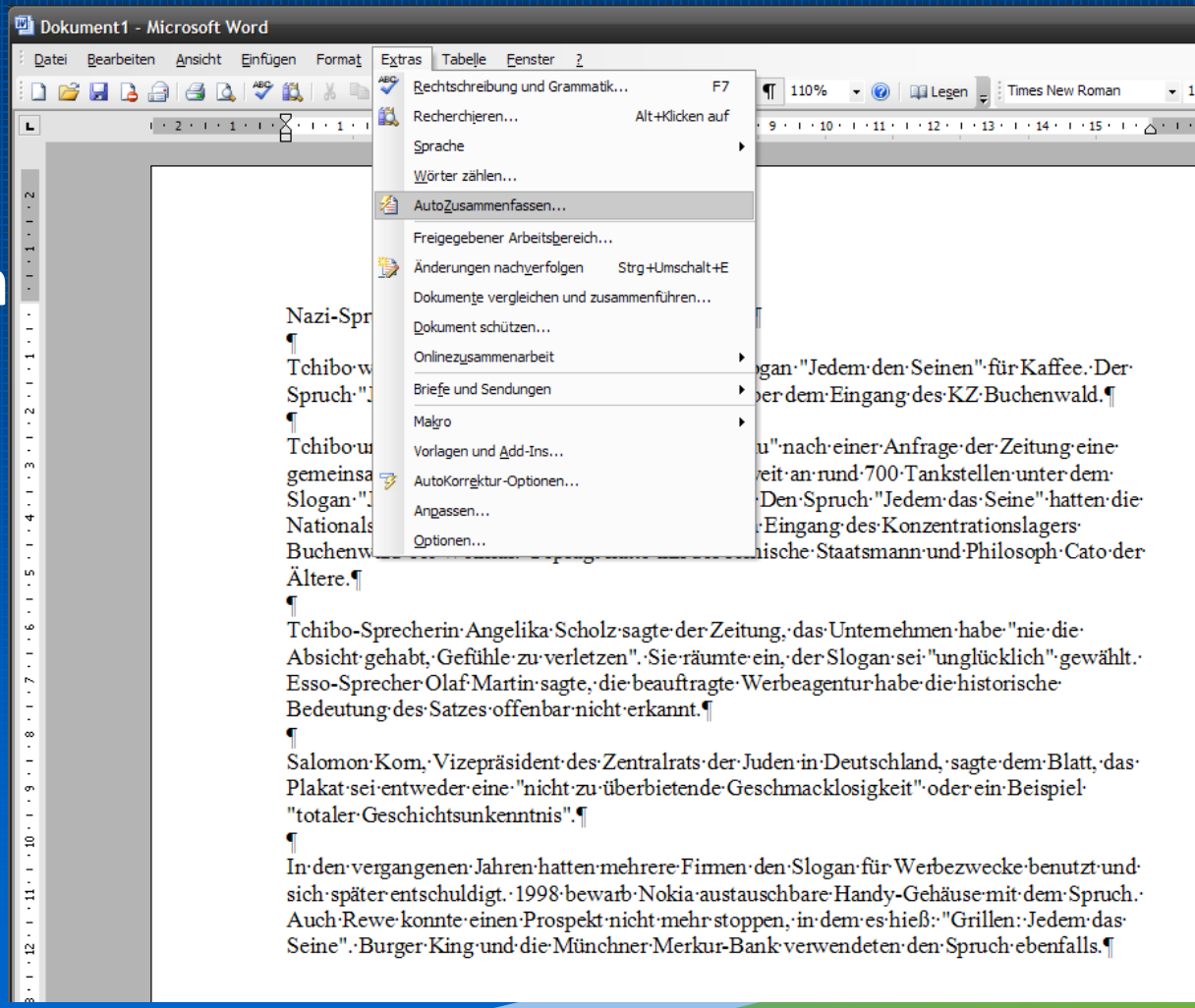


Ergebnis


http://mskw.cipher-sys.com/Lectern/summary_adjuster.
Automatic Text Summarizer
Slider
Knowledge.Works Analyst Tool Box:
Auto Summarizer
1 2 3 4
Use slider to adjust summary length:
Summarize Again
Close
Tchibo warb auf 700 Esso-Tankstellen mit dem Slogan "Jedem den Seinen" fr Kaffee. Der Spruch "Jedem das Seine" prangte bei den Nazis ber dem Eingang des KZ Buchenwald. Tchibo und Esso haben laut "Frankfurter Rundschau" nach einer Anfrage der Zeitung eine gemeinsame PR-Aktion gestoppt, die deutschlandweit an rund 700 Tankstellen unter dem Slogan "Jedem den Seinen" fr Kaffeesorten warb. Ltd

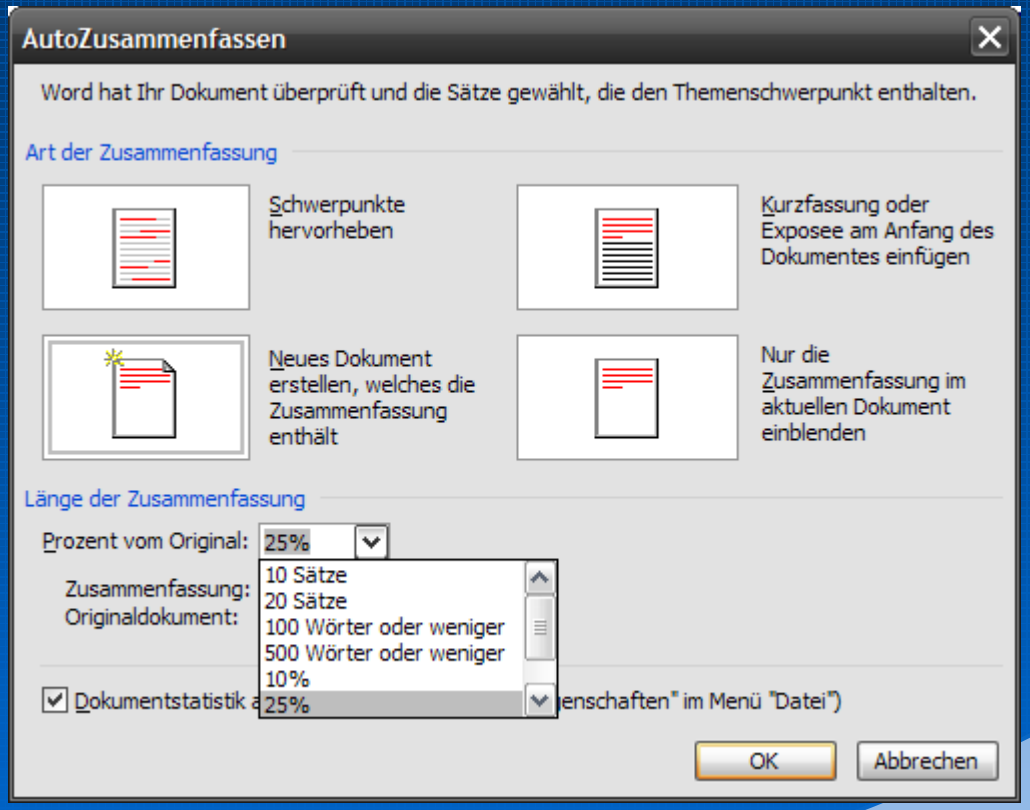
Praxisbeispiel

AutoZusammenfassen-Funktion (Microsoft Word 2003)

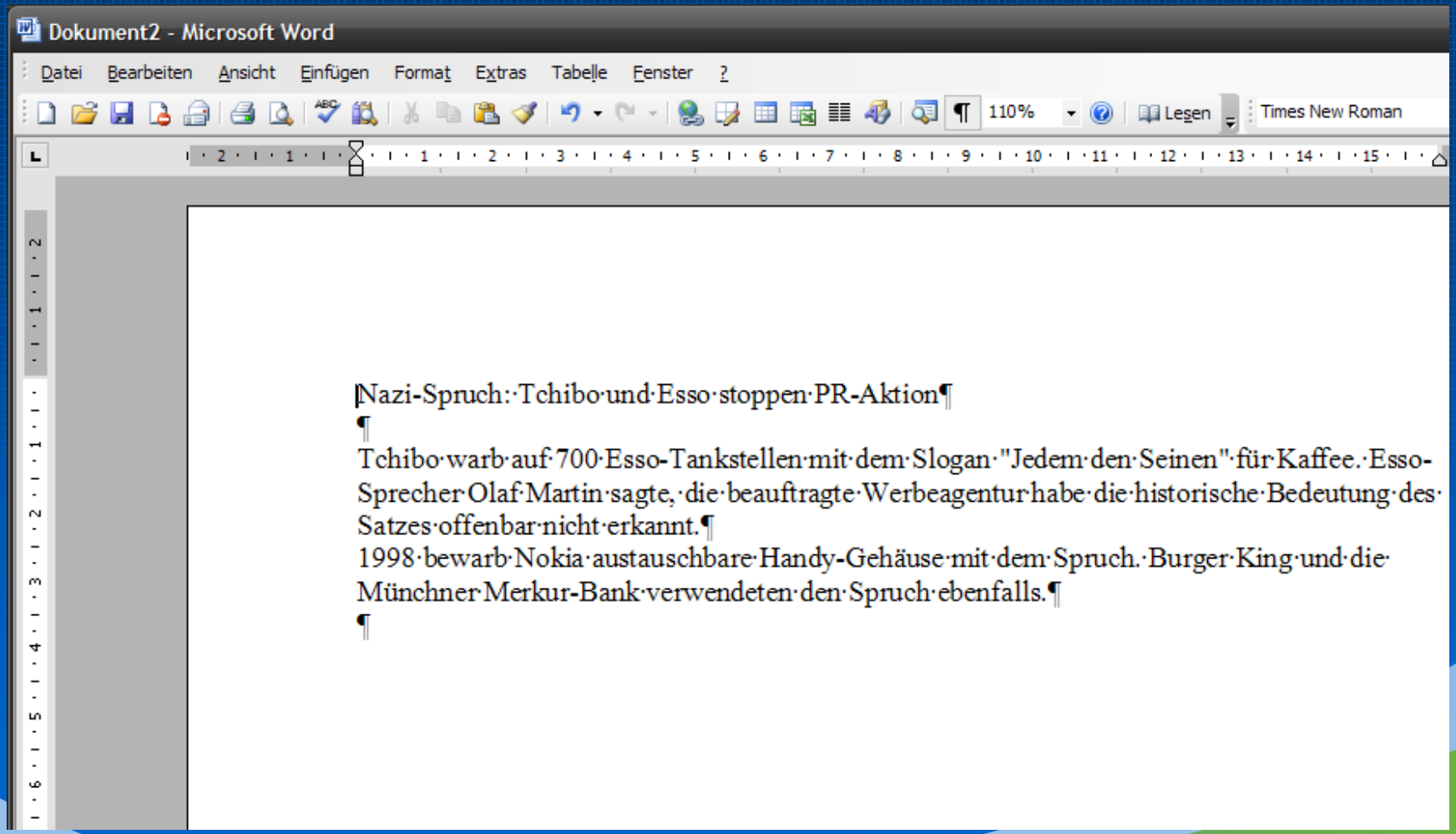


Praxisbeispiel

eingestellt auf “25% vom Original”



Ergebnis


Dokument2 - Microsoft Word
Datei Bearbeiten Ansicht Einfügen Format Extras Tabelle Fenster ?
110%
Lesen
Times New Roman
Nazi-Spruch: Tchibo und Esso stoppen PR-Aktion
Tchibo warb auf 700 Esso-Tankstellen mit dem Slogan "Jedem den Seinen" für Kaffee. Esso-Sprecher Olaf Martin sagte, die beauftragte Werbeagentur habe die historische Bedeutung des Satzes offenbar nicht erkannt.
1998 bewarb Nokia austauschbare Handy-Gehäuse mit dem Spruch. Burger King und die Münchner Merkur-Bank verwendeten den Spruch ebenfalls.