Abschlussbericht
Studienprojekt: TaxoSearch

Lehrstuhl für Computerlinguistik
Ruprecht-Karls-Universität Heidelberg

vorgestellt von

Thorsten Beinhorn, Vesna Cvoro, Khaled Dhaoui und Christian Pretzsch

Was ist TaxoSearch?

- TaxoSearch ist ein Tool zur Unterstützung bei der Internetsuche. Unser Programm hilft dem Benutzer Suchbegriffe präziser zu wählen und unterstützt ihn bei der Sichtung der Suchergebnisse durch individuelle Darstellung des Rankings

Ziele vom Projekt

- Steigerung der Trefferquote durch automatische semantische Erweiterung der Suchanfrage durch Wordnet
- thematische Eingrenzung der Suchergebnisse mittels Ontologie
- Schnittstellen zu mehreren Suchmaschinen
- Information Retrieval zur Aufbereitung der Ergebnisse

Realisierung I

- Erweiterung der Suchanfrage
  - Wordnet als einzige Alternative
  - Manuell vs. Automatisch
    - Automatische Disambiguierung der Suchbegriffe nicht möglich

Realisierung II

- Schnittstellen zu verschiedenen Suchmaschinen
  - Praktische Gründe und strukturelle Ähnlichkeiten der Suchmaschinen => Google
  - Nachträgliche Erweiterung problemlos möglich

Realisierung III

- Information Retrieval
  - Ranking und indivuduelle Darstellung durch Selektion relevanter Begriffe
  - Vektormodell

Arbeiten mit TaxoSearch I

- Eingabe der Suchbegriffe
  - Unterstützung des Benutzers durch Thesaurus Funktion (WordNet)
  - Erweiterte GoogleSuche in GUI vereinfacht dargestellt

Arbeiten mit TaxoSearch II

- Information Retrieval:
  Ranking der Trefferseiten
  - Anpassung des Rankings durch individuelle Zusammenstellung relevanter Begriffe

Umsetzung

- Programmiersprache: Python
- Plattform: Unix, Mac OS, Windows
- GUI: TkInter
- Information Retrieval: Vektormodell

Vektormodell

- Berücksichtigung partieller Übereinstimmung zwischen Anfragetermen und Dokumenttermen durch nicht - binäre Werte für Termgewichtung
- Berechnung der Ähnlichkeit zwischen Anfragetermen und Dokumenttermen
- Sortierung von Dokumenten nach Grad der Ähnlichkeit

(Baeza-Yates/Ribeiro-Neto, 1999,27)

Vektormodell - Beispiel

Dokumentvektor: (4,8,0)
Vektor Query$_1$: (1,2,0)
Vektor Query$_2$: (3,6,0)

	Term 1 *Öl*	Term 2 *Preis*	Term 3 *Alaska*
	4	8	0
	1	2	0
	3	6	0

(vgl. Kowalski, 1997,153)

Externe Komponenten

- WordNet 1.7.1
- pywordnet 1.4
- pyGoogle 0.5.3
- MontyTagger 1.2

WordNet

- WordNet ist eine semantische (allgemeinorientierte) Wortdatenbank, die an der University of Princeton erstellt wurde
  - http://www.cogsci.princeton.edu/~wn/
- WordNet ist eine hierarchische Datenbank, die erfassten Konzepte werden in Baumform konzeptualisiert.
- WordNet ist ein erweitertes Lexikon, dessen Design an aktuelle psycholinguistische Theorien des menschlichen Wortgedächtnisses angelehnt ist.
- Englische Nomen, Verben und Adjektive werden in sogenannten *Synonym Sets* organisiert, von denen jeder ein zugrundeliegendes lexikalisches Konzept repräsentiert. Verschiedene Beziehungen verbinden diese Synonym Sets.

WordNet II

- WordNet speichert seine Informationen nicht nach alphabetischen, sondern nach konzeptuellen Gesichtspunkten.
  - Begriffe gleicher oder ähnlicher Bedeutung werden zusammen abgelegt.
  - die Datenbank von WordNet umfaßt im Gegensatz zu früheren Projekten ungefähr 95600 verschiedene Wortformen, welche in ca. 70100 Bedeutungsklassen (Synonym Sets) eingeordnet sind.
  - WordNet unterteilt das Lexikon in 5 Kategorien : Nomen, Verben, Adjektive, Adverben und Funktionswörter (nicht implementiert). Diese Einteilung beruht auf Untersuchungen über Wortassoziationen von Fillenbaum und Jones (1965)
    - Der Nachteil dieser Vorgehensweise besteht darin, daß mache Wörter in mehr als eine syntaktische Kategorie hineinfallen können und somit mehrmals gespeichert werden müssen.

Monty Tagger I

- Regelbasierter POS – Tagger
- Basiert auf Brills 1994 entwickelten Transformational – Based Learning POS Tagger
- In plattformunabhängigem Python und Java
- Benutzt als Grundlage die Penn Treebank

http://web.media.mit.edu/~hugo/research/montytagger.html

Monty Tagger II

- Tokenizer:
  - „Tokenization“ des Eingabetextes
    - children's --> children 's
    - parents' --> parents '
    - won't --> wo n't
    - I'm --> I 'm
  - ➢ Trennung von Wörtern und Interpunktion durch Leerzeichen
  - Ausnahme: Abkürzungen und Akronyme

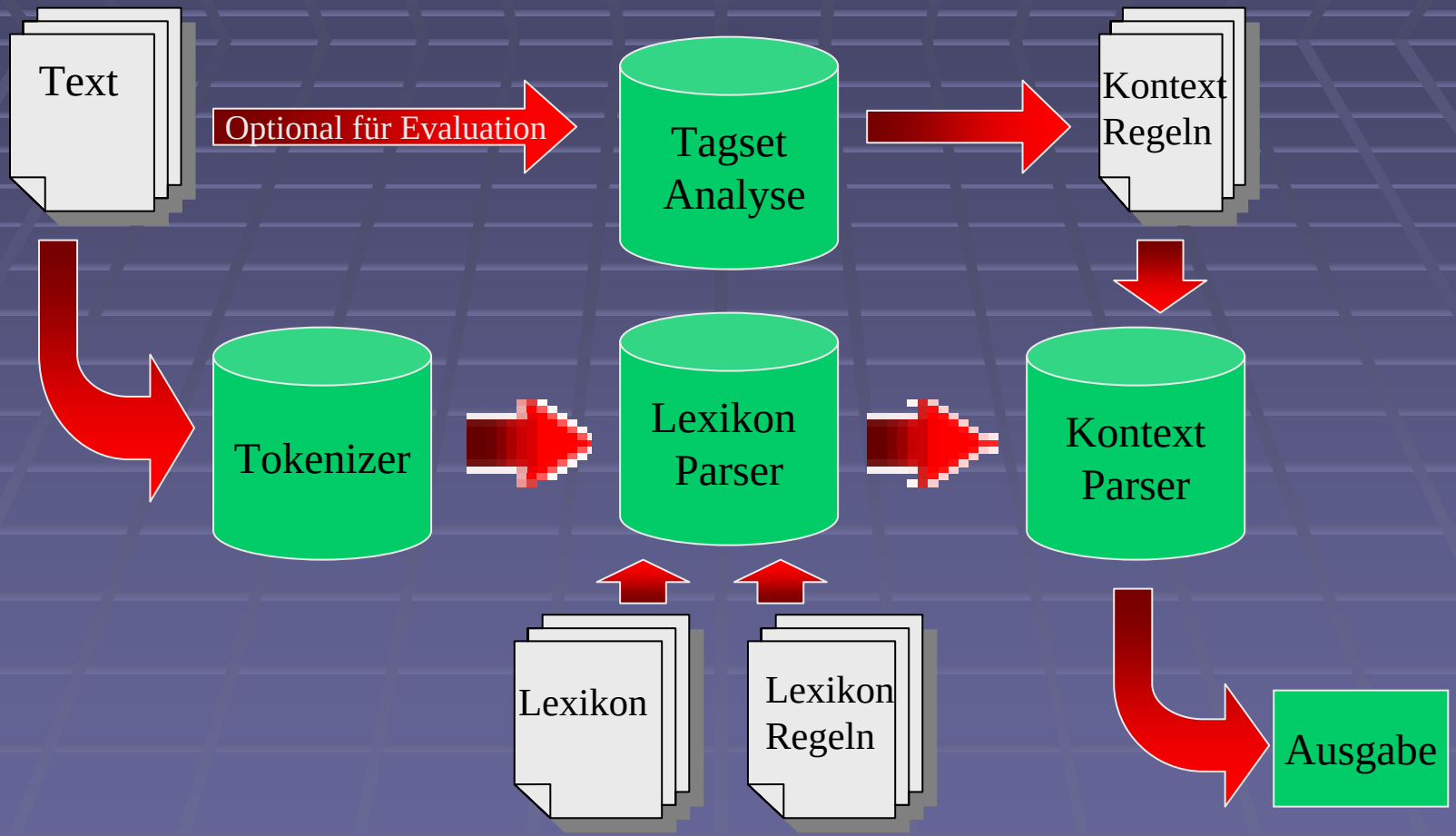
Monty Tagger III

- Lexikon und lexikalische Regeln:
  - Einbindung eines Lexikons und eines Regelsets („Brill94 lexical rule files“)
  - Morphosyntaktische Analyse
  - Zuordnung des „wahrscheinlichsten“ Tags
    - „golden gate“ -> /NNP
    - „race“ -> /NNS oder /VB ?

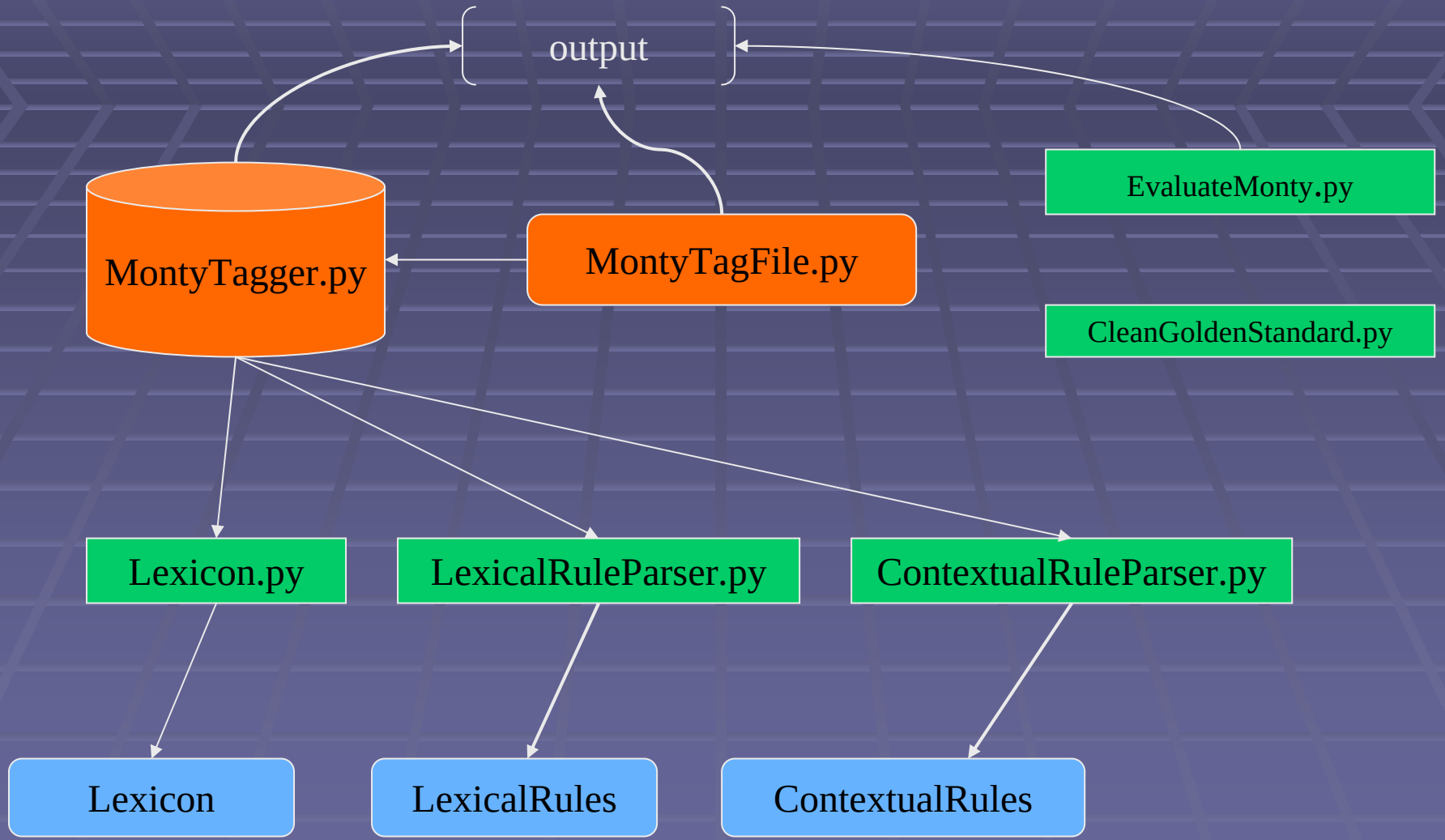
Monty Tagger IV

- Kontextregeln und Syntaxanalyse:
  - Einbindung der Kontextregeln („Brill94 context rule files“)
  - Syntaktische (Kontext-) Analyse: jede Regel wird für alle Wörter geprüft
  - Überprüfung und anschließende Zuordnung bzw. Auswahl der „wahrscheinlichsten“ Tags
    - „golden gate“ -> /NNP
    - „race“ -> /NNS oder /VB ? -> Entscheidung: race /NNS

Beispiel (Monty Tagger)



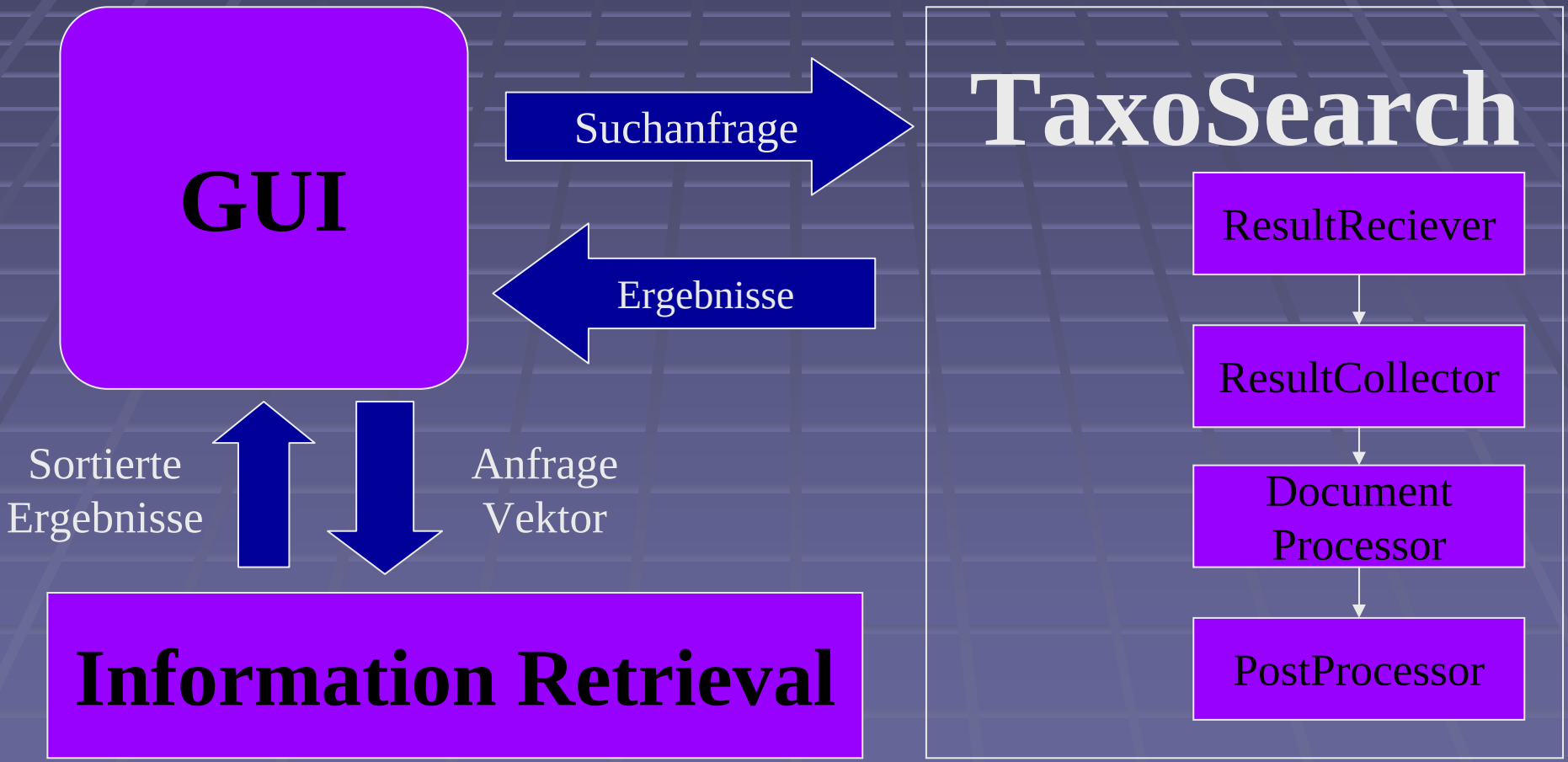
Architektur



Evaluation: Monty Tagger

- Performance
  - 500 Wörter /s
- Genauigkeit
  - ca. 96 – 97 %

TaxoSearch - Systemarchitektur



Modul: mod_DocumentObjects

- **Class DocumentCollection**
  - Speichert eine Sammlung an Dokumenten in einem Dictionary
- **Eigenschaften**
  - *dicDocumentCollection*
    - Dictonary zum Speichern der DocumentObjects (Key=URL, Value=DocumentObject)
- **Methoden**
  - *AddDocument(self, PageURL, Title)*
    - Fügt ein neues DocumentObject in die Auflistung ein.
  - *UpdateDocument(self, DocumentObject)*
    - Aktualisiert ein DocumentObject in der Auflistung mit einem extern veränderten DocumentObject

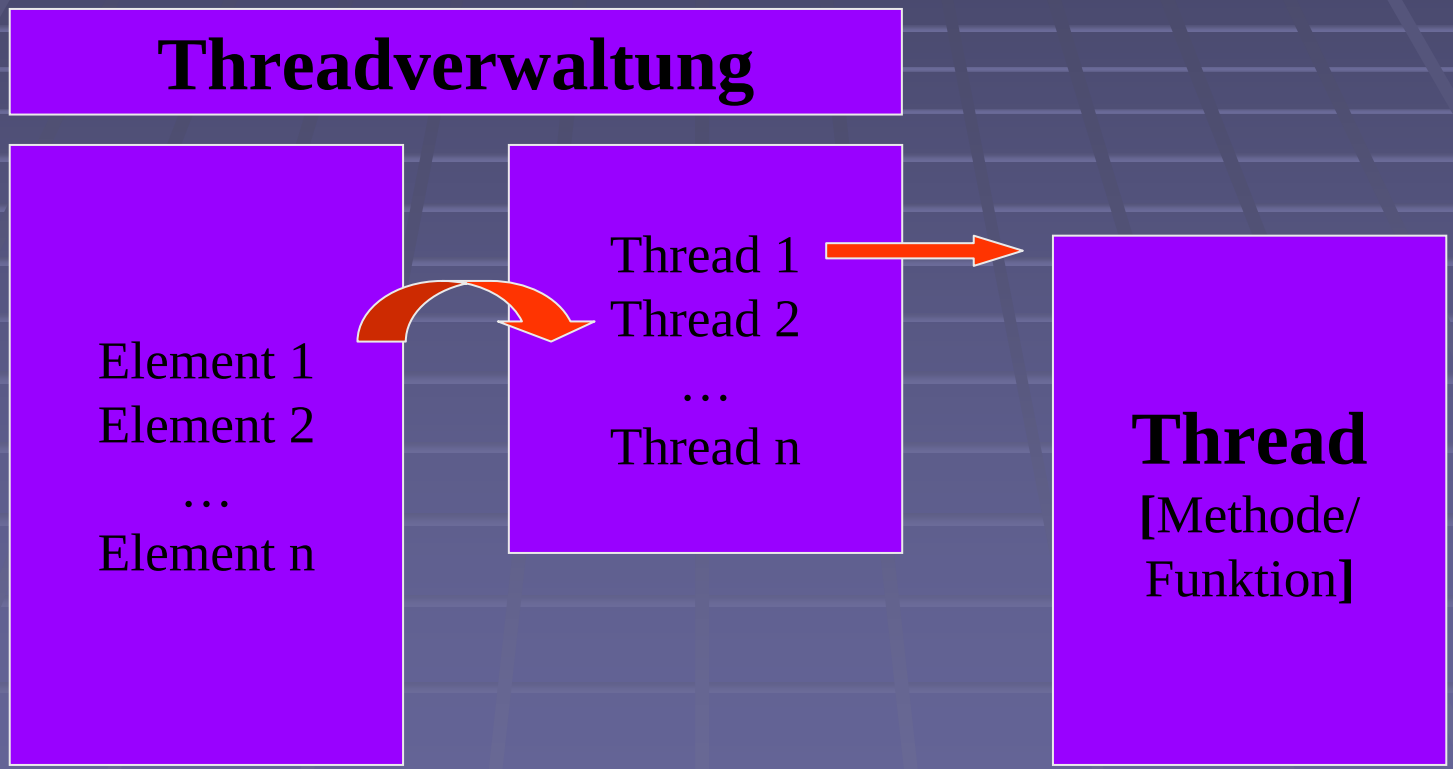
Modul: mod_DocumentObjects

Class DocumentObject

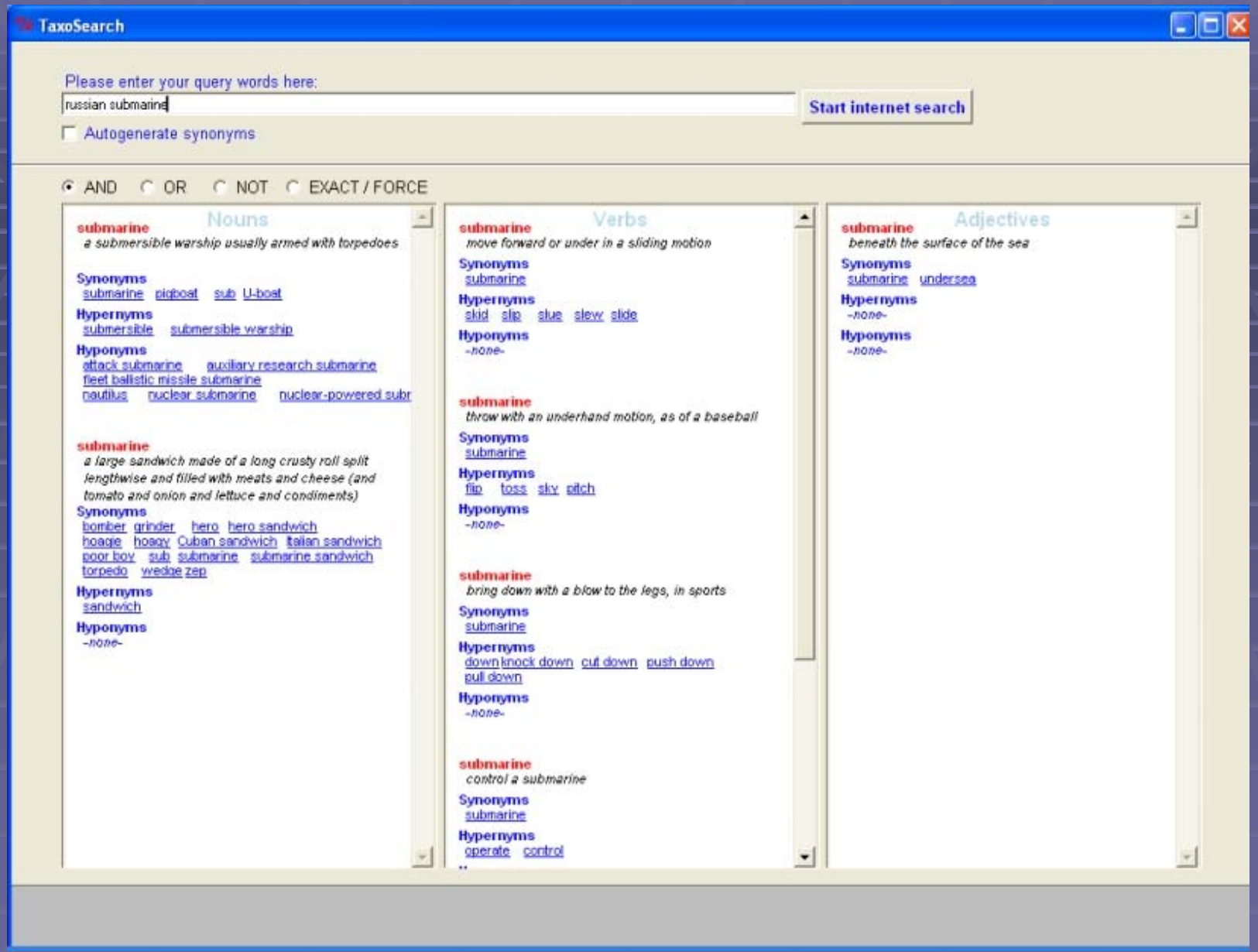
- Repräsentiert ein eine Trefferseite mit all ihren wichtigen Eigenschaften

- **Eigenschaften**
  - *URL*
    - Internet-Adresse der Trefferseite
- *Title*
  - Titel der Trefferseite (aus HTML-Header)
- *Page*
  - Kompletter, unveränderter Inhalt einer Trefferseite
- *PageContent*
  - Inhalt einer Trefferseite nach DeHTML, also ohne Tags
- *dicDocumentVector*
  - Dokumentvector einer Trefferseite, enthält alle Nomen und Verben Key=Wort, Value=Anzahl
- *numWordCount*
  - Anzahl aller Wörter im Dokumentvector
- *numDocumentRating*
  - PageRankingWert nach Information Retrieval

Multithreading



TaxoSearch in Aktion

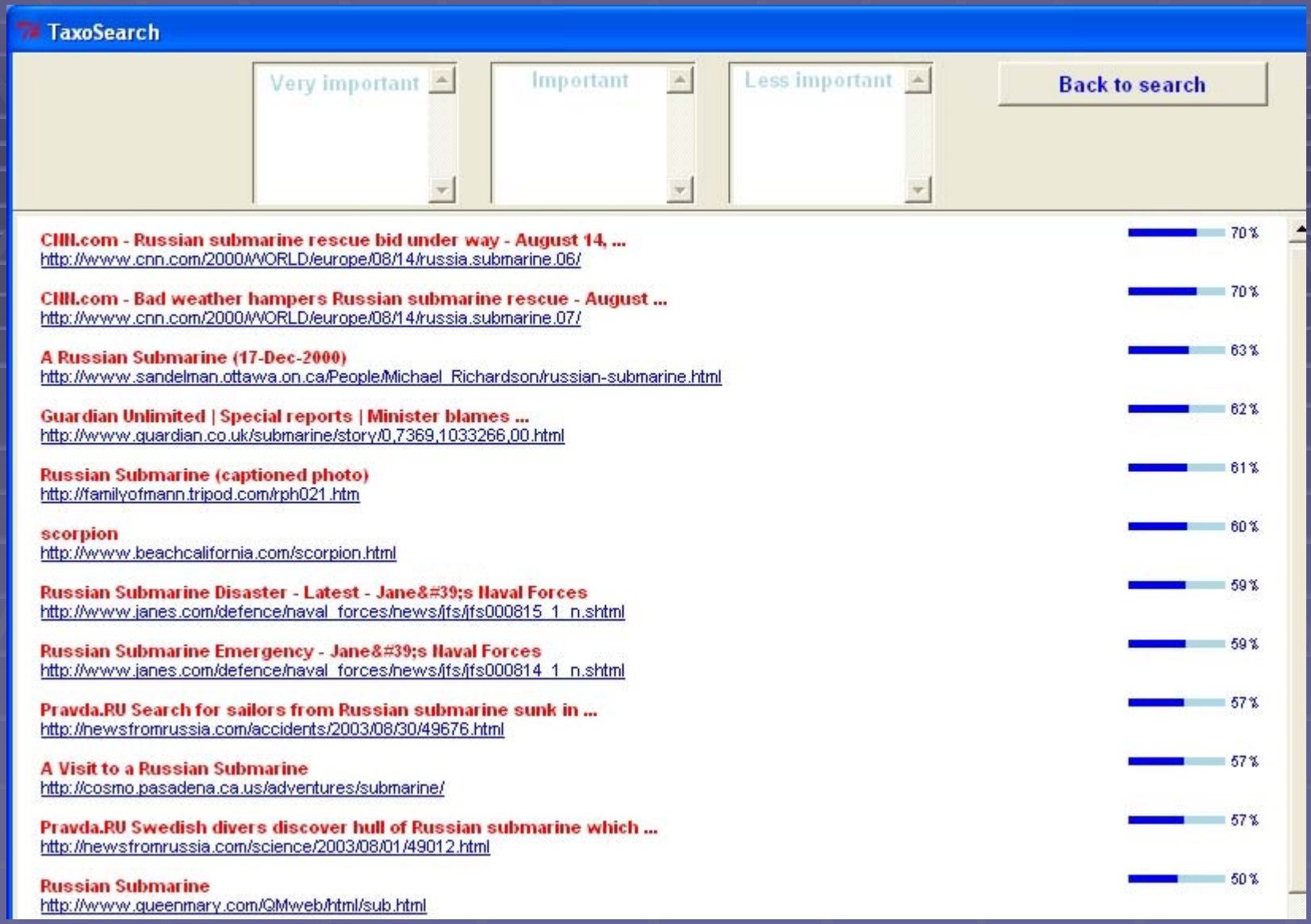

TaxoSearch
Please enter your query words here:
russian submarine
Start internet search
Autogenerate synonyms
AND
OR
NOT
EXACT / FORCE
Nouns
submarine
a submersible warship usually armed with torpedoes
Synonyms
submarine pigboat sub U-boat
Hypernyms
submersible submersible warship
Hyponyms
attack submarine auxiliary research submarine fleet ballistic missile submarine nautilus nuclear submarine nuclear-powered subr
submarine
a large sandwich made of a long crusty roll split lengthwise and filled with meats and cheese (and tomato and onion and lettuce and condiments)
Synonyms
bomber grinder hero hero sandwich hoagie hoagy Cuban sandwich Italian sandwich poor boy sub submarine submarine sandwich torpedo wedge zep
Hypernyms
sandwich
Hyponyms
-none-
Verbs
submarine
move forward or under in a sliding motion
Synonyms
submarine
Hypernyms
skid slip slue slew slide
Hyponyms
-none-
submarine
throw with an underhand motion, as of a baseball
Synonyms
submarine
Hypernyms
flip toss sky pitch
Hyponyms
-none-
submarine
bring down with a blow to the legs, in sports
Synonyms
submarine
Hypernyms
down knock down cut down push down pull down
Hyponyms
-none-
submarine
control a submarine
Synonyms
submarine
Hypernyms
operate control
Adjectives
submarine
beneath the surface of the sea
Synonyms
submarine undersea
Hypernyms
-none-
Hyponyms
-none-

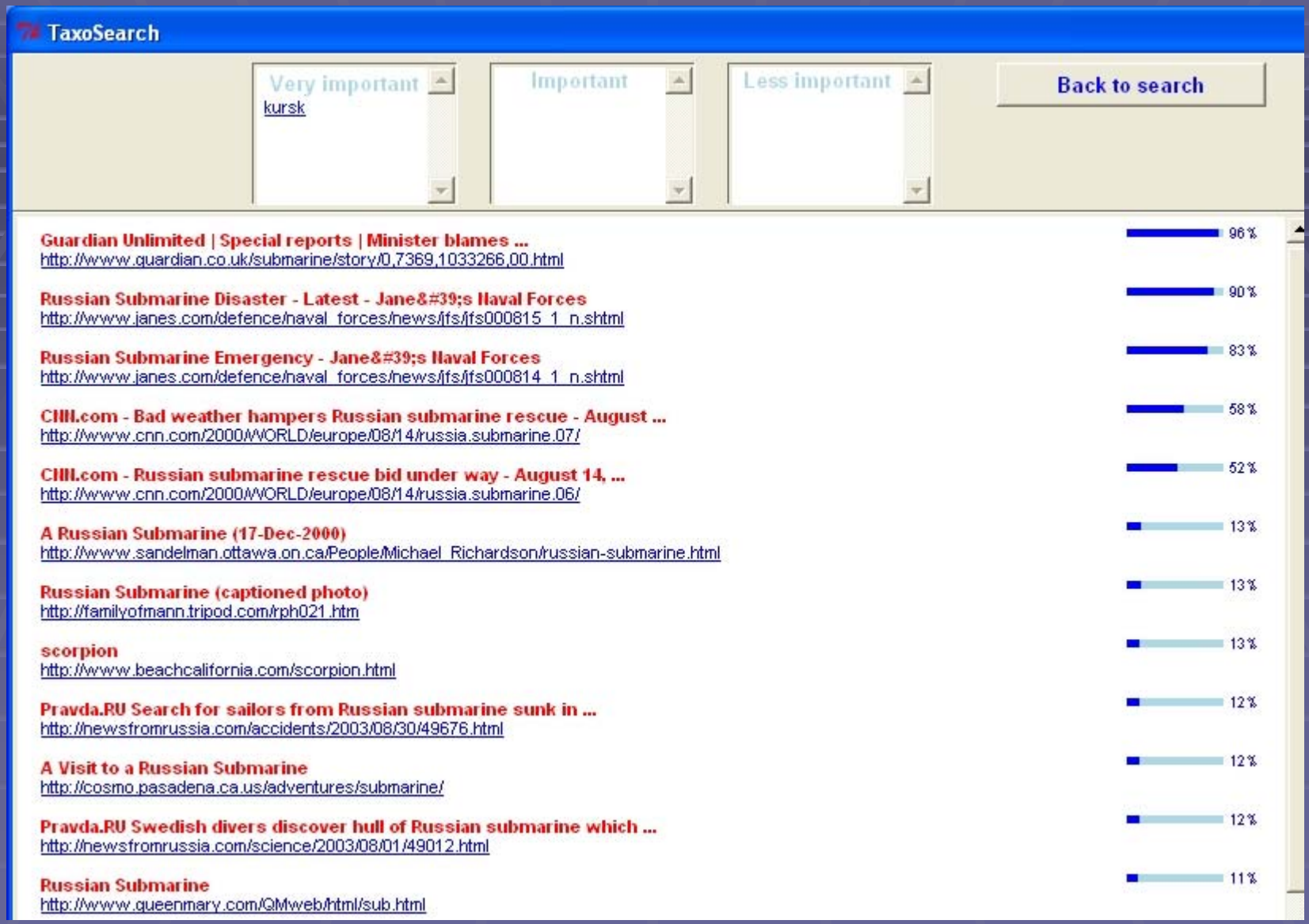
TaxoSearch

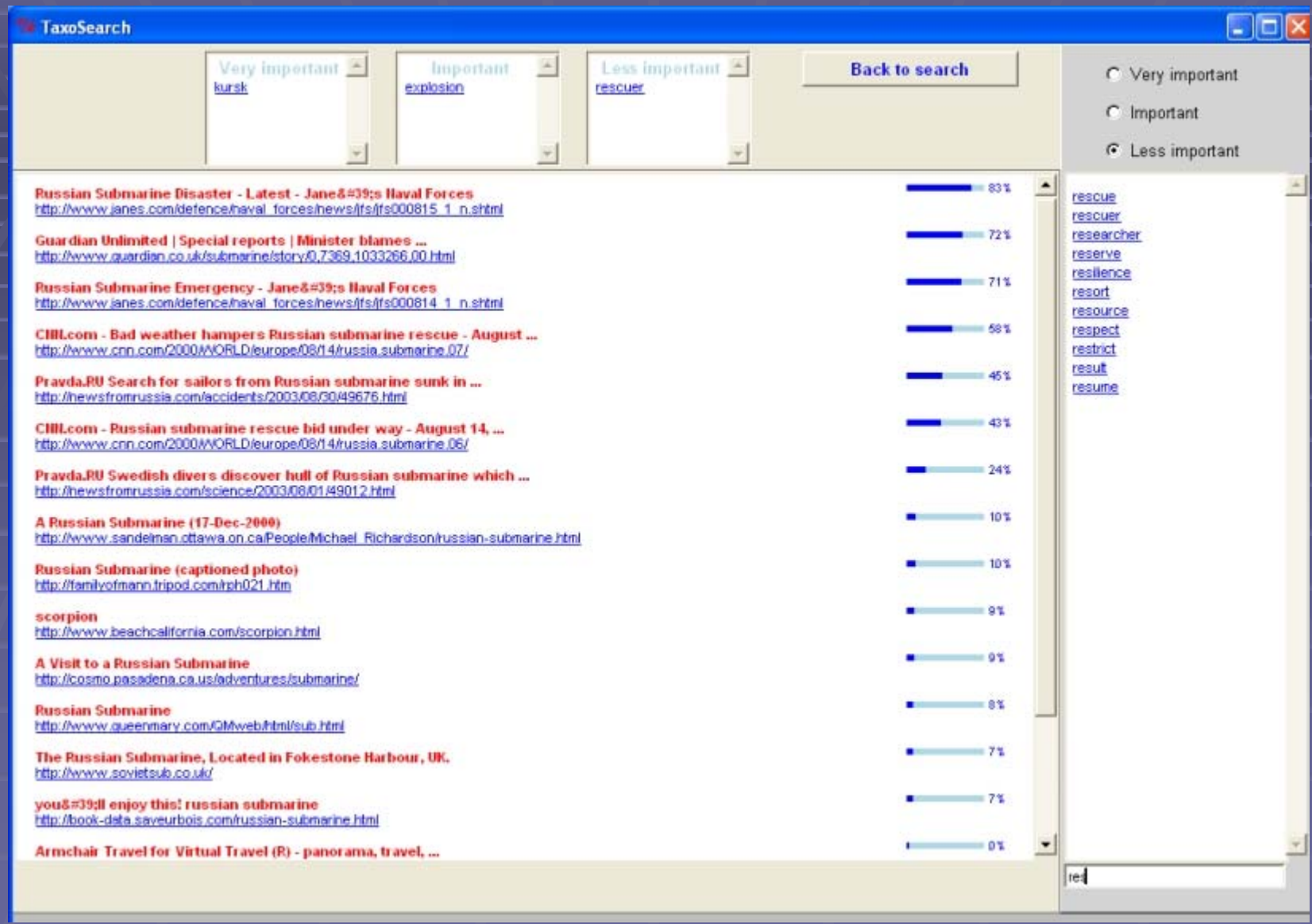
Performing google search with:

russian submarine

ResultReciever: google QueryString ready: russian submarine
ResultReciever: search complete in 0.117079 seconds
ResultReciever: IGNORED RESULT: http://news.bbc.co.uk/2/hi/europe/3193625.stm
ResultReciever: search complete in 0.0901 seconds
ResultReciever: search complete in 0.212211 seconds
ResultReciever: IGNORED RESULT: http://www.cdi.org/russia/273-17.cfm
ResultReciever: IGNORED RESULT: http://en.wikipedia.org/wiki/Russian_submarine_Kursk
ResultReciever: search complete in 0.207524 seconds
ResultReciever: IGNORED RESULT: http://www.cdi.org/russia/272-2.cfm
ResultReciever: IGNORED RESULT: http://news.bbc.co.uk/hi/english/world/europe/newsid_880000/880194.stm
ResultReciever: search complete in 0.224076 seconds
ResultCollector: Collecting page contents...
ResultCollector: page retrieved
ResultCollector: page retrieved
ResultCollector: page retrieved
ResultCollector: page retrieved
ResultCollector: page retrieved
ResultCollector: page retrieved
ResultCollector: page retrieved
ResultCollector: page retrieved
ResultCollector: page retrieved
ResultCollector: page retrieved
ResultCollector: page retrieved
ResultCollector: page retrieved
ResultCollector: page retrieved
ResultCollector: page retrieved
ResultCollector: page retrieved
ResultCollector: page retrieved
ResultCollector: page retrieved
ResultCollector: page retrieved
ResultCollector: page retrieved
ResultCollector: page retrieved


TaxoSearch
Very important
Important
Less important
Back to search
CNN.com - Russian submarine rescue bid under way - August 14, ...
http://www.cnn.com/2000/WORLD/europe/08/14/russia.submarine.06/
70 %
CNN.com - Bad weather hampers Russian submarine rescue - August ...
http://www.cnn.com/2000/WORLD/europe/08/14/russia.submarine.07/
70 %
A Russian Submarine (17-Dec-2000)
http://www.sandelman.ottawa.on.ca/People/Michael_Richardson/russian-submarine.html
63 %
Guardian Unlimited | Special reports | Minister blames ...
http://www.guardian.co.uk/submarine/story/0,7369,1033266,00.html
62 %
Russian Submarine (captioned photo)
http://familyofmann.tripod.com/rph021.htm
61 %
scorpion
http://www.beachcalifornia.com/scorpion.html
60 %
Russian Submarine Disaster - Latest - Jane's Naval Forces
http://www.janes.com/defence/naval_forces/news/jfs/jfs000815_1_n.shtml
59 %
Russian Submarine Emergency - Jane's Naval Forces
http://www.janes.com/defence/naval_forces/news/jfs/jfs000814_1_n.shtml
59 %
Pravda.RU Search for sailors from Russian submarine sunk in ...
http://newsfromrussia.com/accidents/2003/08/30/49676.html
57 %
A Visit to a Russian Submarine
http://cosmo.pasadena.ca.us/adventures/submarine/
57 %
Pravda.RU Swedish divers discover hull of Russian submarine which ...
http://newsfromrussia.com/science/2003/08/01/49012.html
57 %
Russian Submarine
http://www.queenmary.com/QMweb/html/sub.html
50 %


TaxoSearch
Very important
kursk
Important
Less important
Back to search
Guardian Unlimited | Special reports | Minister blames ...
http://www.guardian.co.uk/submarine/story/0,7369,1033266,00.html
96 %
Russian Submarine Disaster - Latest - Jane's Naval Forces
http://www.janes.com/defence/naval_forces/news/jfs/jfs000815_1_n.shtml
90 %
Russian Submarine Emergency - Jane's Naval Forces
http://www.janes.com/defence/naval_forces/news/jfs/jfs000814_1_n.shtml
83 %
CNN.com - Bad weather hampers Russian submarine rescue - August ...
http://www.cnn.com/2000/WORLD/europe/08/14/russia.submarine.07/
58 %
CNN.com - Russian submarine rescue bid under way - August 14, ...
http://www.cnn.com/2000/WORLD/europe/08/14/russia.submarine.06/
52 %
A Russian Submarine (17-Dec-2000)
http://www.sandelman.ottawa.on.ca/People/Michael_Richardson/russian-submarine.html
13 %
Russian Submarine (captioned photo)
http://familyofmann.tripod.com/rph021.htm
13 %
scorpion
http://www.beachcalifornia.com/scorpion.html
13 %
Pravda.RU Search for sailors from Russian submarine sunk in ...
http://newsfromrussia.com/accidents/2003/08/30/49676.html
12 %
A Visit to a Russian Submarine
http://cosmo.pasadena.ca.us/adventures/submarine/
12 %
Pravda.RU Swedish divers discover hull of Russian submarine which ...
http://newsfromrussia.com/science/2003/08/01/49012.html
12 %
Russian Submarine
http://www.queenmary.com/QMweb/html/sub.html
11 %


TaxoSearch
Very important
kursk
Important
explosion
Less important
rescuer
Back to search
Very important
Important
Less important
Russian Submarine Disaster - Latest - Jane's Naval Forces
http://www.janes.com/defence/naval_forces/news/jfs/jfs000815_1_n.shtml
83%
Guardian Unlimited | Special reports | Minister blames ...
http://www.guardian.co.uk/submarine/story/0,7369,1033266,00.html
72%
Russian Submarine Emergency - Jane's Naval Forces
http://www.janes.com/defence/naval_forces/news/jfs/jfs000814_1_n.shtml
71%
CNN.com - Bad weather hampers Russian submarine rescue - August ...
http://www.cnn.com/2000/WORLD/europe/08/14/russia.submarine.07/
58%
Pravda.RU Search for sailors from Russian submarine sunk in ...
http://newsfromrussia.com/accidents/2003/08/30/49676.html
45%
CNN.com - Russian submarine rescue bid under way - August 14, ...
http://www.cnn.com/2000/WORLD/europe/08/14/russia.submarine.06/
43%
Pravda.RU Swedish divers discover hull of Russian submarine which ...
http://newsfromrussia.com/science/2003/08/01/49012.html
24%
A Russian Submarine (17-Dec-2000)
http://www.sandelman.ottawa.on.ca/People/Michael_Richardson/russian-submarine.html
10%
Russian Submarine (captioned photo)
http://familyofmann.tripod.com/rph021.htm
10%
scorpion
http://www.beachcalifornia.com/scorpion.html
9%
A Visit to a Russian Submarine
http://cosmo.pasadena.ca.us/adventures/submarine/
9%
Russian Submarine
http://www.queenmary.com/QMweb/html/sub.html
8%
The Russian Submarine, Located in Fokestone Harbour, UK.
http://www.sovietsub.co.uk/
7%
you'll enjoy this! russian submarine
http://book-data.saveurbois.com/russian-submarine.html
7%
Armchair Travel for Virtual Travel (R) - panorama, travel, ...
0%
rescue
rescuer
researcher
reserve
resilience
resort
resource
respect
restrict
result
resume
res

Literatur

- Karin Haenelt
- **Alexander Valet / Christian Pretzsch / Vanessa Micelli**: *Ansätze des Tagging – Ein Seminarreferat*, „Hauptseminar: Parsing“, Universität Heidelberg, SS 2003 *kontext.fraunhofer.de/haenelt/kurs/Referate/Micelli_Pretzsch_Valet_SS03/MontyAnsatzRef.pdf*
- MontyTagger Version 1.2: http://web.media.mit.edu/~hugo/research/montytagger.html
- WordNet 1.7.1
- pyWordNet
- pyGoogle 0.5.3