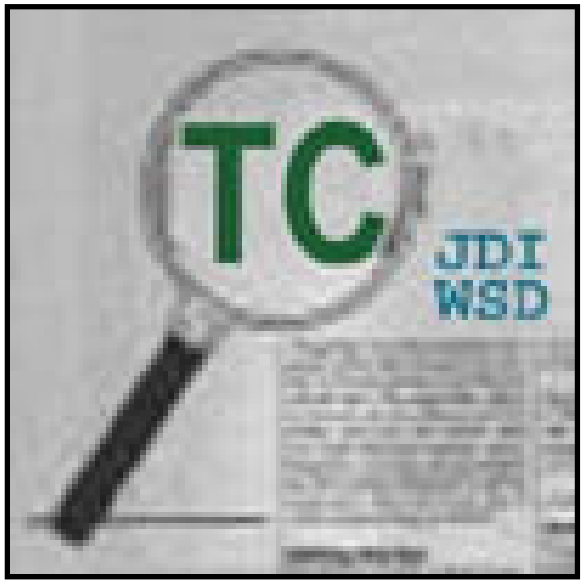
Text Categorization

Lexical Systems Group
National Library of Medicine
National Institutes of Health

Table of Contents

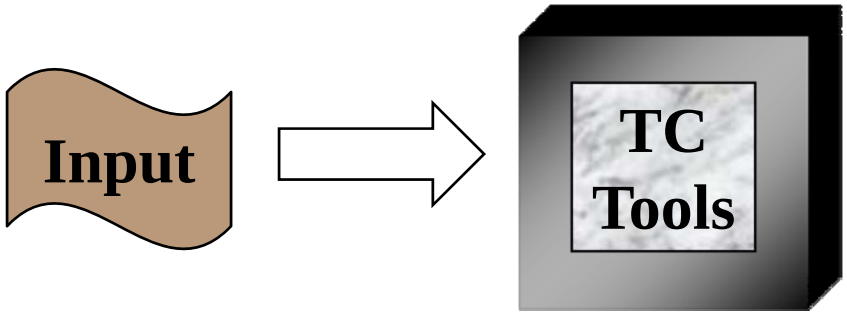
Introduction



Introduction - TC Tools

- A set of tools for
  - Text categorization
  - Indexing & retrieval
  - Document classification
  - Word sense disambiguation
  - etc..

Introduction - TC Tools



- A set of tools takes the given input
  - Free text: word, phrase, sentence, paragraph, etc..
  - MeSH terms

Introduction - TC Tools

- A set of tools that generates ranked JD or ST list with scores to categorize the input text

Introduction - Three Tools

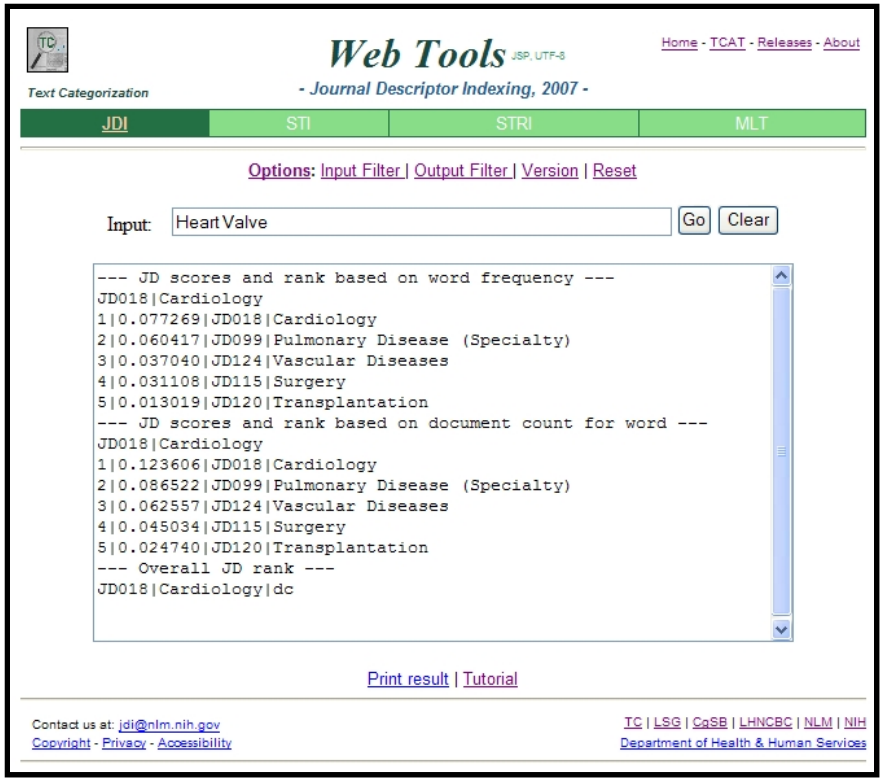
Introduction - Tool Types

- Command line tools
  - JDI (Journal Descriptor Indexing)
  - STI (Semantic Type Indexing)
  - STRI (Semantic Type Real-Time Indexing)
  - MLT (MEDLINE Tokenizer)
- Web Tools
- Java APIs

Introduction - Facts

- Free distributed with open source code
- 100% in Java
- Run on different platforms
- One complete package
- Documents & support
- Provides Java APIs, command line tools, and Web tools
- First release, TC 2007

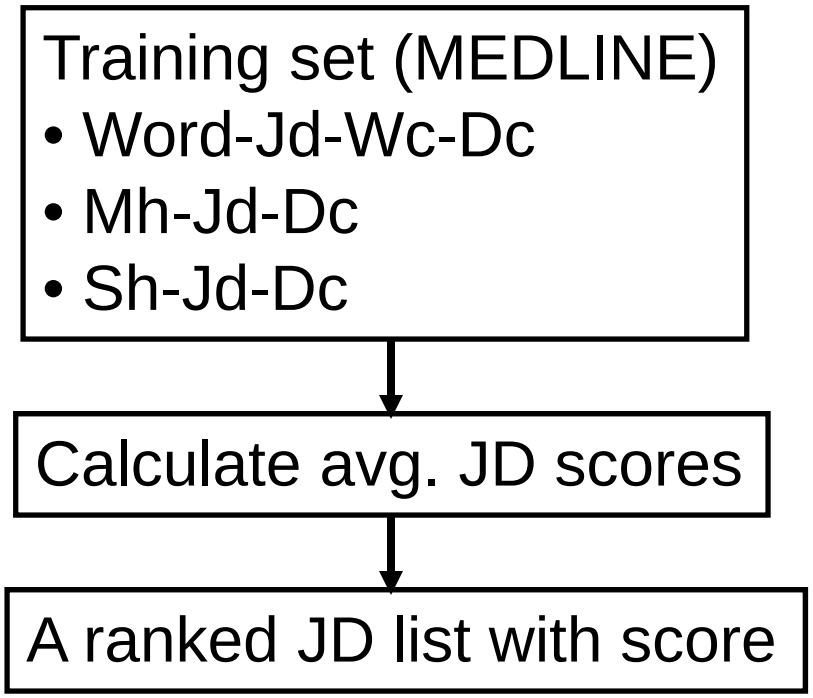
TC Tools


Text Categorization
Web Tools JSP, UTF-8
- Journal Descriptor Indexing, 2007 -
Home - TCAT - Releases - About
JDI
STI
STRI
MLT
Options: Input Filter | Output Filter | Version | Reset
Input: Heart Valve
Go
Clear
--- JD scores and rank based on word frequency ---
JD018|Cardiology
1|0.077269|JD018|Cardiology
2|0.060417|JD099|Pulmonary Disease (Specialty)
3|0.037040|JD124|Vascular Diseases
4|0.031108|JD115|Surgery
5|0.013019|JD120|Transplantation
--- JD scores and rank based on document count for word ---
JD018|Cardiology
1|0.123606|JD018|Cardiology
2|0.086522|JD099|Pulmonary Disease (Specialty)
3|0.062557|JD124|Vascular Diseases
4|0.045034|JD115|Surgery
5|0.024740|JD120|Transplantation
--- Overall JD rank ---
JD018|Cardiology|dc
Print result | Tutorial
Contact us at: jdi@nlm.nih.gov
Copyright - Privacy - Accessibility
TC | LSG | CgSB | LHNCBC | NLM | NIH
Department of Health & Human Services

JDI Methodology

- JDI is the core methodology of TC
- JDI categorizes text according to a set of JDs
- Journal Descriptors (JDs):
  - A set of 122 descriptors from MeSH (Medical Subject Headings) is used for indexing MEDLINE journals per se
  - For example, Journal of "*Pediatric Surgery*" is indexed and listed under both [Pediatrics] and [Surgery]

JDI Methodology



TC Tools - JDI

- Based on statistical word-JD associations from a training set:
  - 3 years MEDLINE (2002 ~ 2004)
  - 4093 journals
  - 12.4 M MEDLINE citations
- Word-Jd-Wc-Dc score table for text indexing
- Mh-Jd-Dc table score table for main-heading indexing
- Sh-Jd-Dc table score table for Sub-heading indexing

JDI: Word-Jd Table

- Get total word count and document count for all words from titles and abstracts in MEDLINE citations (testing set data)
- Get word count and document count for all words co-occurs with all JDs
- Generate Word-Jd-Wc-Dc table:

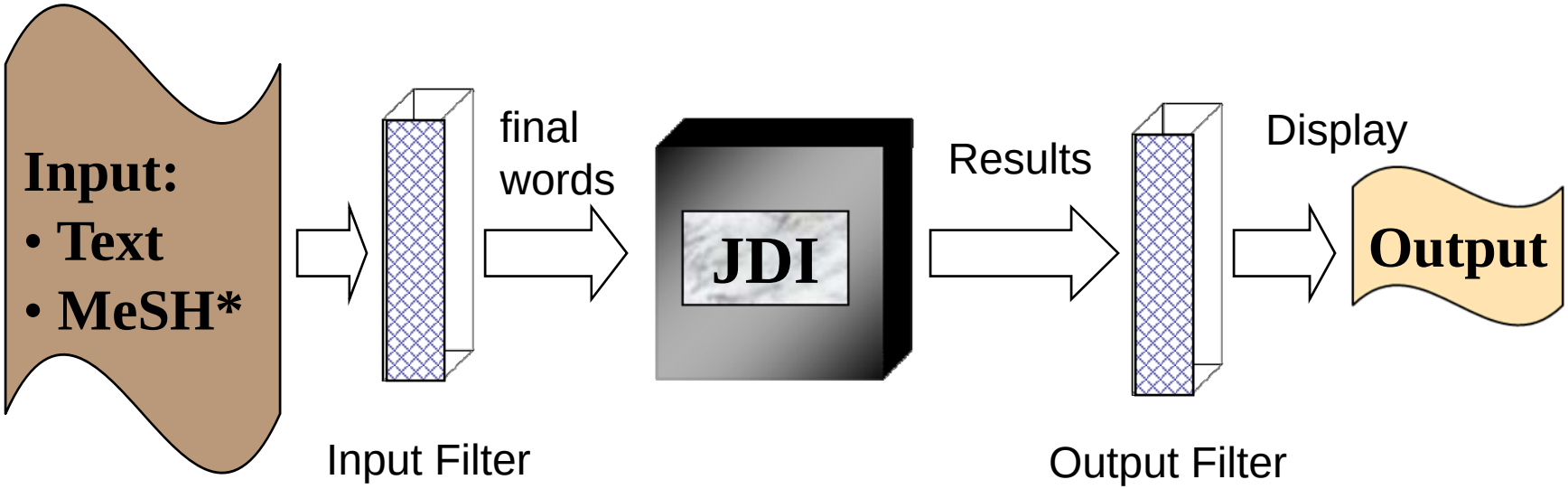
$$\text{word score} = \frac{\text{word - Jd co - occurs word count}}{\text{total word count}}$$

$$\text{document score} = \frac{\text{word - Jd co - occurs document count}}{\text{total document count}}$$

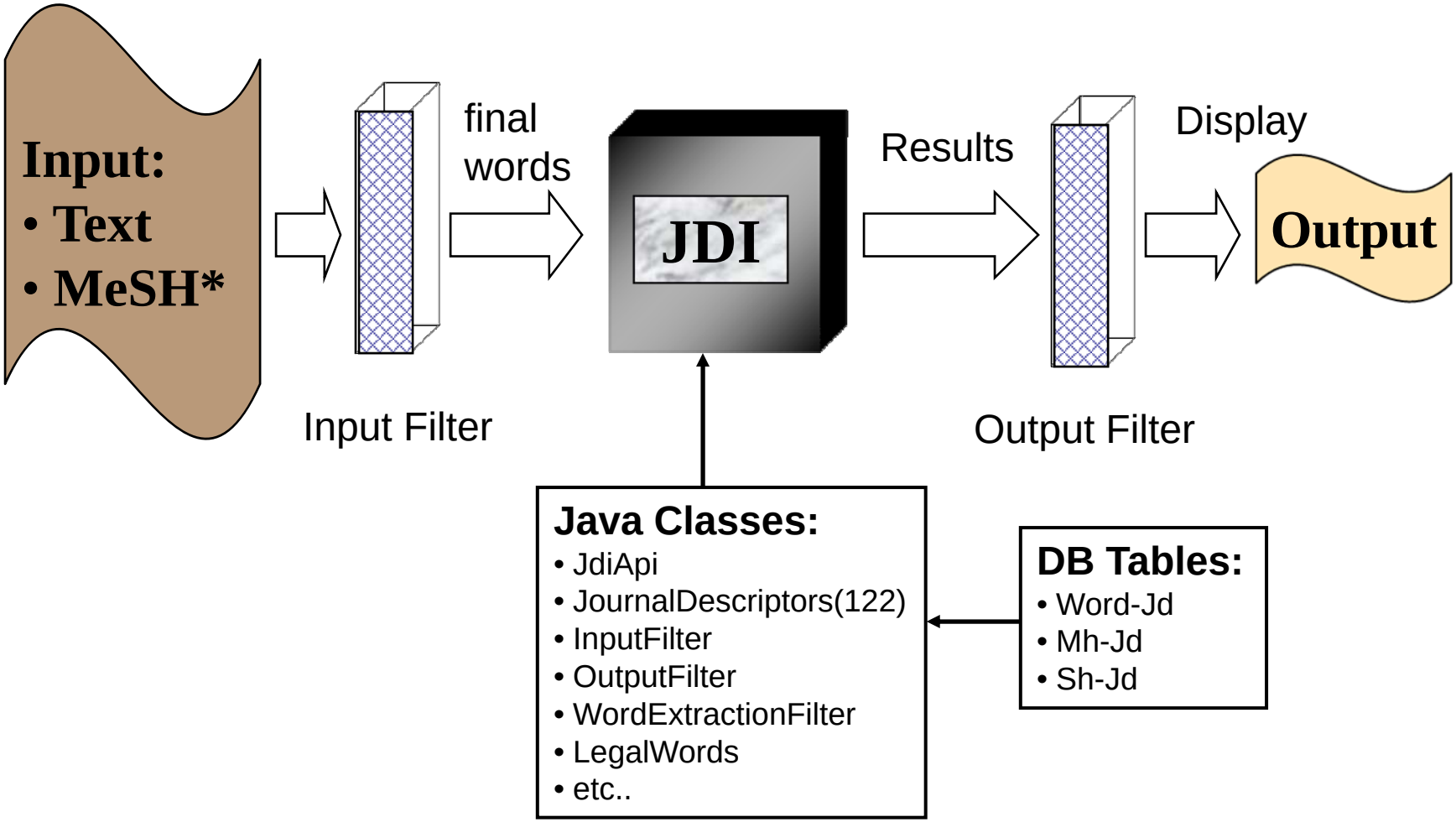
JDI: Word-Jd Table Cont.

- Derived files:
  - Journal Descriptors (from jid-Ta-Jds)
  - contractions (from Lisp)
  - stopwords (from Lisp)
  - Jid-Ta-Jds (from lsi.xml)
- MEDLINE citations:
  - retrieve words from TI, AB – pmid-words
  - word-wc-dc
  - pmid-jd
- Calculating scores:
  - word-Signal-Gt1 (normalized)
  - word-Wc-Dc-Scores
  - word-Jdid-Wc-Dc-Gt1
  - Jd-Dc
  - Jd-Dc-NFactor
  - word-Jd-Wc-Dc (file used for TC DB)
- Other files:
  - restricted words (from MRCON, contractions, word-wc-dc-gt1)

TC Tools - JDI


Input:
• Text
• MeSH*
final
words
JDI
Results
Display
Output
Input Filter
Output Filter

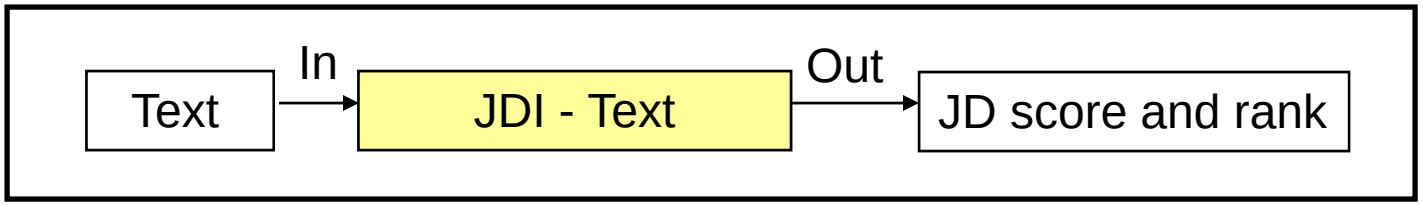
TC Tools - JDI


Input:
• Text
• MeSH*
final words
JDI
Results
Display
Output
Input Filter
Output Filter
Java Classes:
• JdiApi
• JournalDescriptors(122)
• InputFilter
• OutputFilter
• WordExtractionFilter
• LegalWords
• etc..
DB Tables:
• Word-Jd
• Mh-Jd
• Sh-Jd

TC Tools - JDI

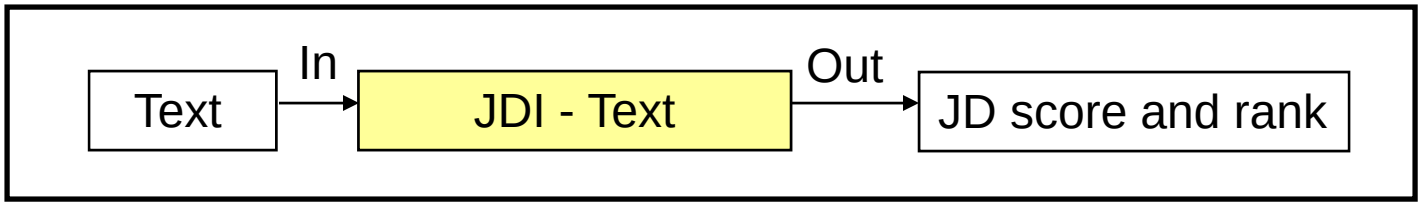
- Apply input filter to filter out irrelevant words
  - Word extraction filter
  - Unique word filter
  - Legal word filter: stopwords filter, restrictwords filter, word length, word count, document count, etc.
- Calculate JD score
  - Get JD scores from DB for each final word
  - Calculate average JD scores for the input
- Apply output filter on results
  - Ranked JD list with scores (0 ~ 1)
  - Cluster display
  - Display number
  - Candidate only display
  - etc..

JDI - Example



Input: **`The heart valve`**

JDI - Example

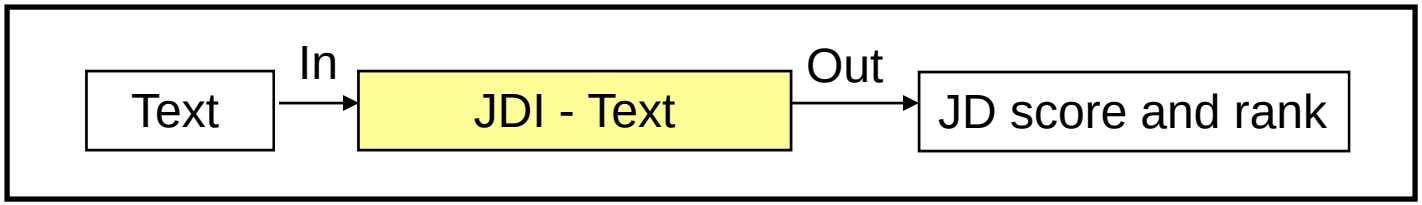


Input: **The heart valve**

Algorithm: Java Program

Filtered out (stop word)

JDI - Example



Input: **The heart valve**

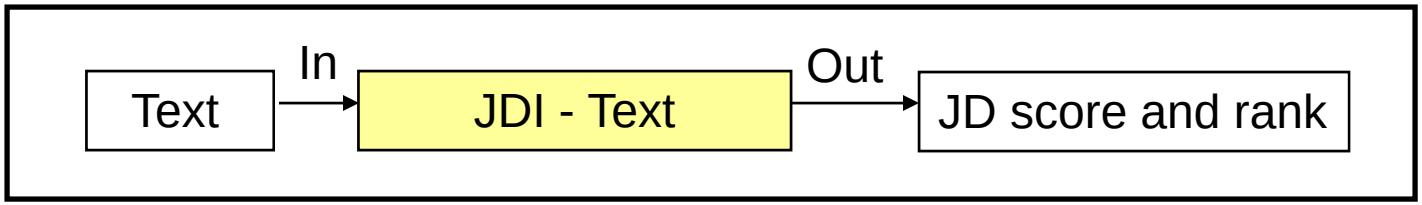
DB: Word-JD table

Heart	
JD001	0.0001787949
JD002	0.0015644556
…..	…..
JD018	0.09365937
…	…

Algorithm: Java Program

Filtered out (stop word)

JDI - Example



Input: **The heart valve**

DB: Word-JD table

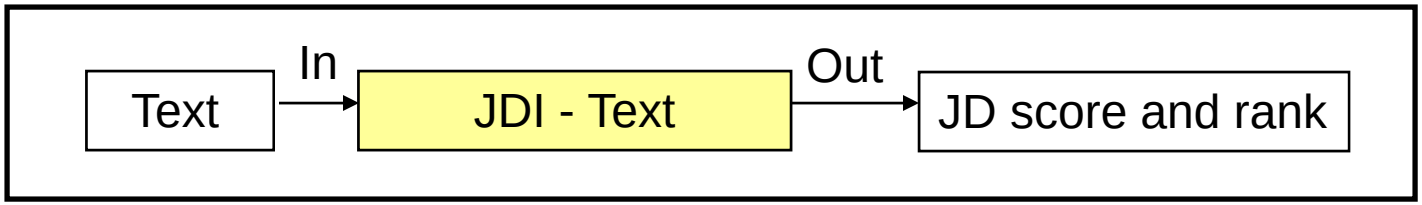
Heart	
JD001	0.0001787949
JD002	0.0015644556
…..	…..
JD018	0.09365937
…	…

Valve	
JD001	0
JD002	0.0008151288
…..	…..
JD018	0.15355311
…	…

Algorithm: Java Program

Filtered out (stop word)

JDI - Example



Input: The heart valve

DB: Word-JD table

Heart	
JD001	0.0001787949
JD002	0.0015644556
…..	…..
JD018	0.09365937
…	…

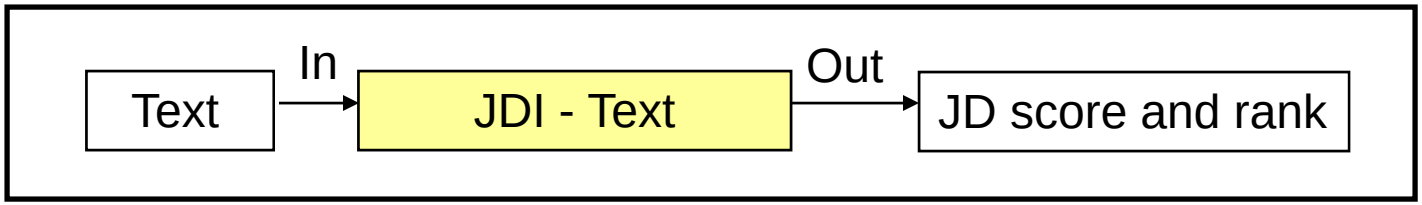
Valve	
JD001	0
JD002	0.0008151288
…..	…..
JD018	0.15355311
…	…

Algorithm: Java Program

Filtered out (stop word)

Calculate average scores:
(0.09365937 + 0.15355311)/2 = 0.1236062

JDI - Example



Input: The heart valve

DB: Word-JD table

Heart	
JD001	0.0001787949
JD002	0.0015644556
…..	…..
JD018	0.09365937
…	…

Valve	
JD001	0
JD002	0.0008151288
…..	…..
JD018	0.15355311
…	…

Algorithm: Java Program

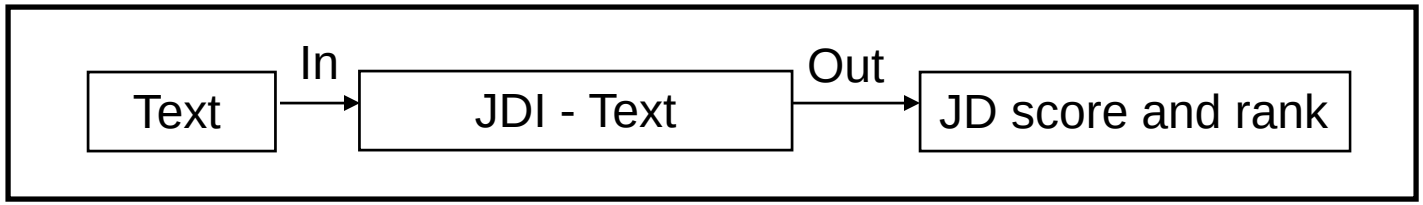
Filtered out (stop word)

Calculate average scores:
(0.09365937 + 0.15355311)/2 = 0.1236062

Output

Heart Valve	
JD001	0.0000894
JD002	0.0011898
…..	…..
JD018	0.1236062
…	…

JDI - Example



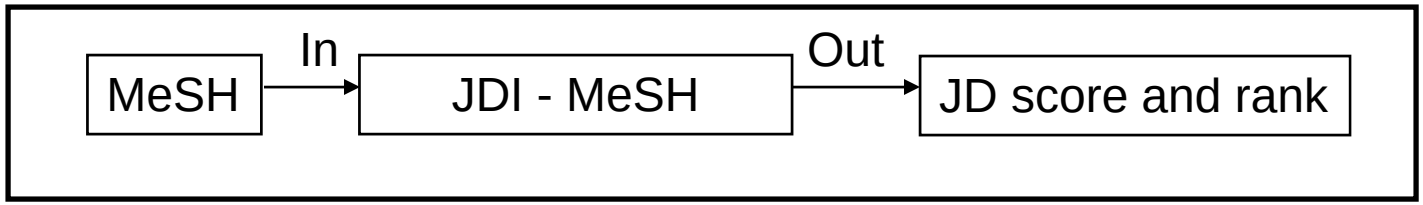
- **Inputs:** The Heart Valve
- **Outputs:**

```
--- JD scores and rank based on document count for word ---
JD018|Cardiology
1|0.123606|JD018|Cardiology
2|0.086522|JD099|Pulmonary Disease (Specialty)
3|0.062557|JD124|Vascular Diseases
4|0.045034|JD115|Surgery
5|0.024740|JD120|Transplantation
6|0.024412|JD005|Anesthesiology
7|0.023319|JD030|Diagnostic Imaging
8|0.016154|JD092|Physiology
9|0.012300|JD055|Internal Medicine
10|0.012124|JD086|Pediatrics
```

JDI: Mh-Jd & Sh-Jd Tables

- Get total document count for all MH* and SH* from MEDLINE citations
- Get document count for all MH* and SH* co-occurs with all JDs
- Get Mh-Jd-Dc table
  - $$\text{document score} = \frac{\text{MH*-JD co-occurs document count}}{\text{total document count}}$$
- Get Sh-Jd-Dc table
  - $$\text{document score} = \frac{\text{SH*-JD co-occurs document count}}{\text{total document count}}$$

JDI - Example



- **Inputs:** `Potassium Channels|ph`
- **Outputs:**

```
--- JD scores and rank based on document count for word ---
JD092|Physiology
1|0.068614|JD092|Physiology
2|0.033940|JD088|Pharmacology
3|0.029321|JD070|Neurology
4|0.027848|JD017|Brain
5|0.024574|JD018|Cardiology
6|0.024317|JD026|Cytology
7|0.023046|JD124|Vascular Diseases
8|0.022541|JD015|Biophysics
9|0.021362|JD106|Science
10|0.018340|JD035|Endocrinology
```

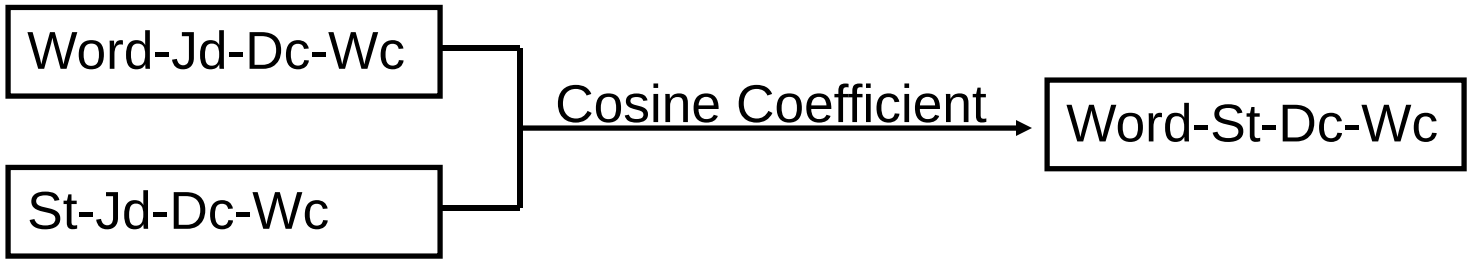
Tools – STI

- **Semantic Types:**
  - A set of 135 Semantic Types in the Semantic Network in NLM's UMLS (Unified Medical Language System) is used for STI.
  - Concepts in the UMLS Metathesaurus are assigned one or more STs which semantically characterize those concepts.
  - For example, concept ***Aspirin*** is assigned the STs [Pharmacologic Substance] and [Organic Chemical].

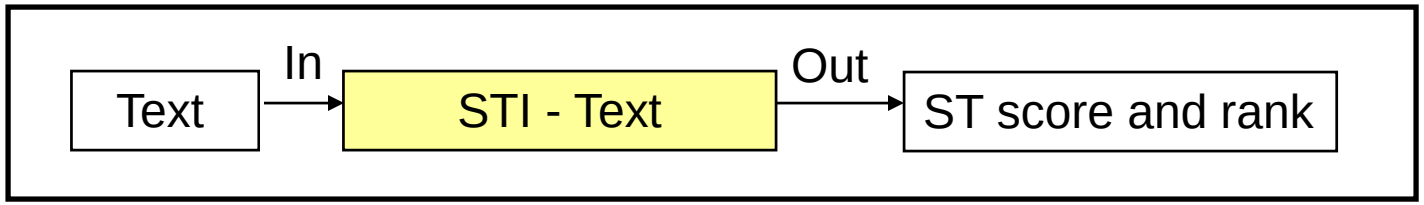
- **STI Tool:**
  - Use JDI methodology as basis
  - Calculate the average ST scores for input text from Word-St table
  - Print out ranked ST list with scores

STI: Word-St Table

- Generate ST-Documents (St-Words) from UMLS Metathesaurus MRCON (string) and MRSTY (Semantic Types) by associating with CUI
- Apply JDI on ST-Documents to generate St-Jd-Dc-Wc
- Calculate cosine coefficient on JDI of ST-Documents (St-Jd-Dc-Wc) and JDI on individual training set words (Word-Jd-Dc-Wc) to generate Word-St-Dc-Wc

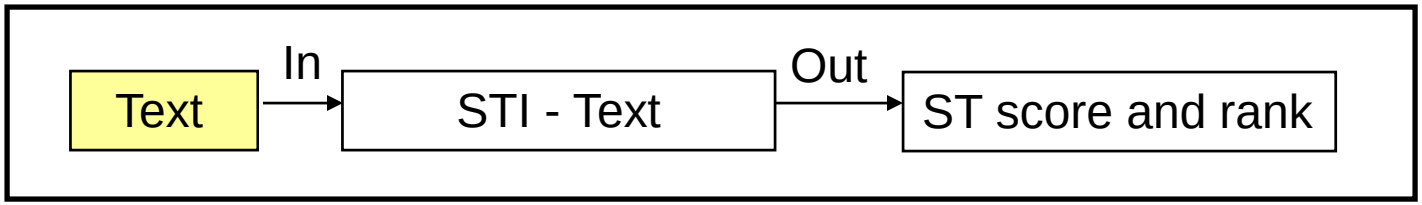


Tools - STI



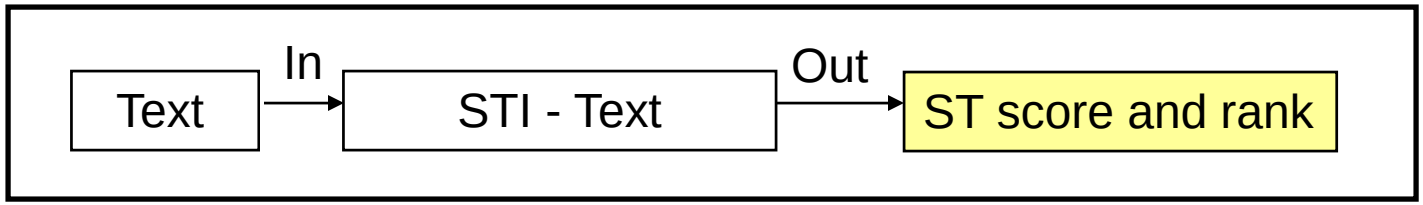
- **Input Filter:**
  - Tokenize and filter out words for processing
  - Apply Word Extraction Filter (used on MEDLINE citations)
  - Filter out illegal words
  - Filter out duplicated words (if unique option is specified)
- **Process:**
  - Get ST scores for each legal word from DB
  - Calculate average ST scores for the input term
- **Output filter:**
  - Apply the specified output filter options
  - Print out ranked ST list with scores

Tools – STI Inputs



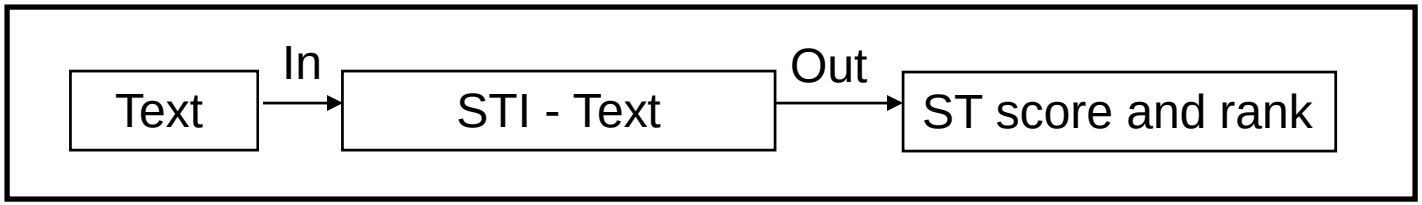
- Text
  - A word
  - A phrase
  - A sentence
  - A paragraph
  - etc..
- Application examples
  - A title from MEDLINE
  - An abstract from MEDLINE
  - Titles and abstracts from MEDLINE

Tools - STI Outputs



- Ranked ST list
- ST Score based on document count
- ST Score based on word frequency
- Input filter details
- Output filter details
- etc..

STI - Example



Inputs: Race, ethnicity, culture, and disparities in health care

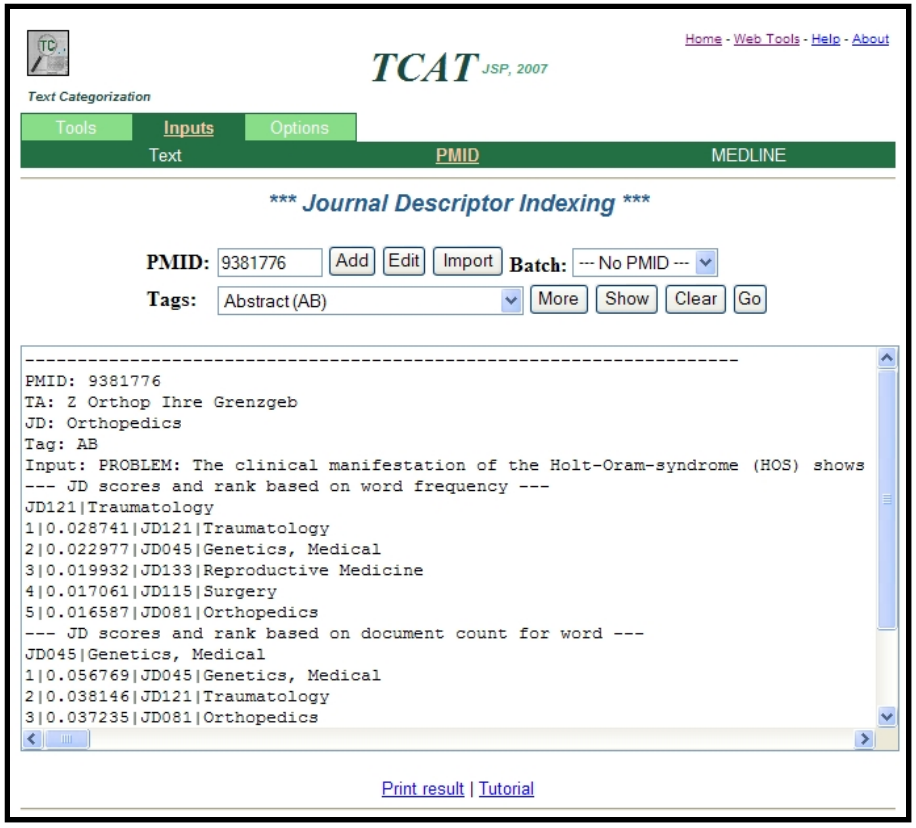
Outputs:

```
--- ST scores and rank based on document count for word ---
popg|Population Group
1|0.8373|popg|Population Group
2|0.7722|socb|Social Behavior
3|0.7591|aggp|Age Group
4|0.7385|idcn|Idea or Concept
5|0.7385|shro|Self-help or Relief Organization
6|0.7272|famg|Family Group
7|0.7232|orgt|Organization
8|0.7086|inbe|Individual Behavior
9|0.6965|gora|Governmental or Regulatory Activity
10|0.6905|edac|Educational Activity
```

TC Web Tools

- Web based tool
- Uses HTML forms as front end GUI
- Uses TC Java APIs as back end algorithm
- Same functions as command line tool
- Includes following tools:
  - JDI
  - STI
  - STRI
  - MLT
- Demo

Applications

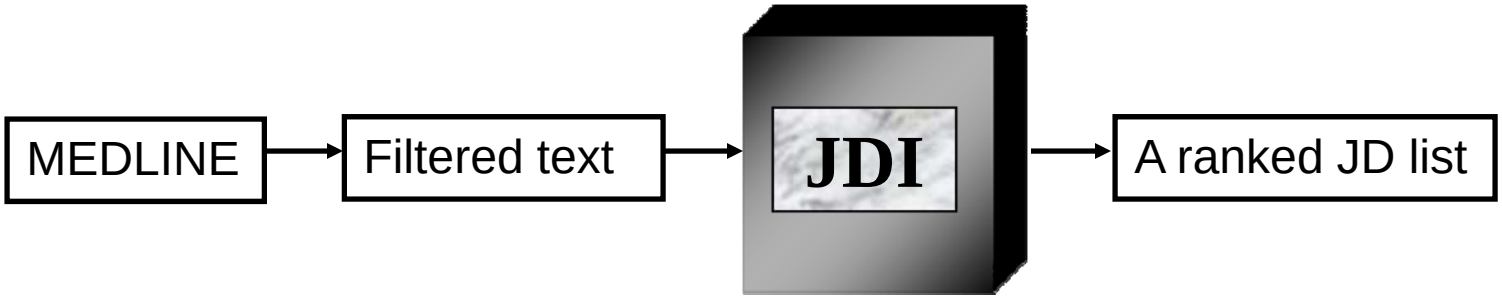


Application – JDI for TC

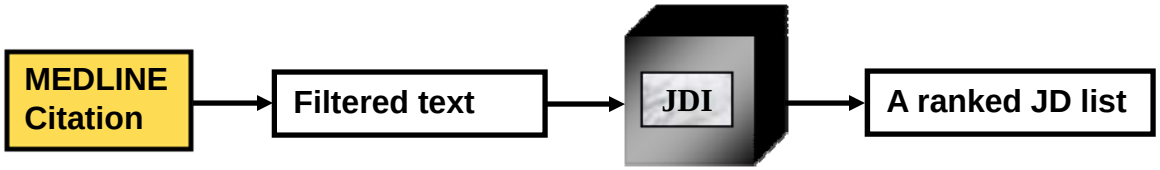
- JDI to index and categorize MEDLINE
- Inputs:
  - Title
  - Abstract
  - Title & abstract
  - Starred MeSH
  - Title, abstract, & starred MeSH
- Outputs:
  - A ranked JD list with scores

Application – JDI for TC

- Procedures:
  - Find the interested MEDLINE citation
  - Retrieve interested fields from MEDLINE citation
  - Filter out irrelevant characters and words
  - Apply JDI and get results



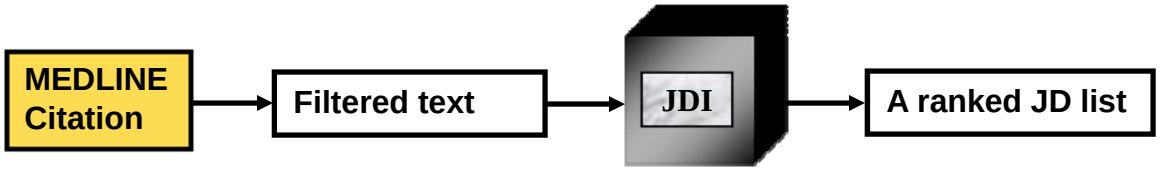
Application – JDI for TC



- MEDLINE Citation:

```
PMID- 15547873
OWN - NLM
STAT- MEDLINE
DA  - 20041119
DCOM- 20051108
PUBM- Print
IS  - 1531-5037 (Electronic)
VI  - 39
IP  - 11
DP  - 2004 Nov
TI  - Outcome and complications after resection of hepatoblastoma.
PG  - 1744-5; author reply 1745
FAU - Pritchard, Jon
AU  - Pritchard J
FAU - Stringer, Mark
AU  - Stringer M
LA  - eng
PT  - Comment
PT  - Letter
PL  - United States
TA  - J Pediatr Surg
JT  - Journal of pediatric surgery
JID – 0052631
…
```

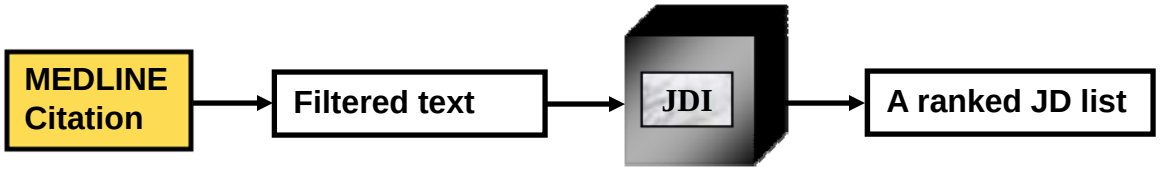
Application – JDI for TC



- MEDLINE Citation:

```
PMID- 15547873
OWN - NLM
STAT- MEDLINE
DA  - 20041119
DCOM- 20051108
PUBM- Print
IS  - 1531-5037 (Electronic)
VI  - 39
IP  - 11
DP  - 2004 Nov
TI  - Outcome and complications after resection of hepatoblastoma.
PG  - 1744-5; author reply 1745
FAU - Pritchard, Jon
AU  - Pritchard J
FAU - Stringer, Mark
AU  - Stringer M
LA  - eng
PT  - Comment
PT  - Letter
PL  - United States
TA  - J Pediatr Surg
JT  - Journal of pediatric surgery
JID – 0052631
…
```

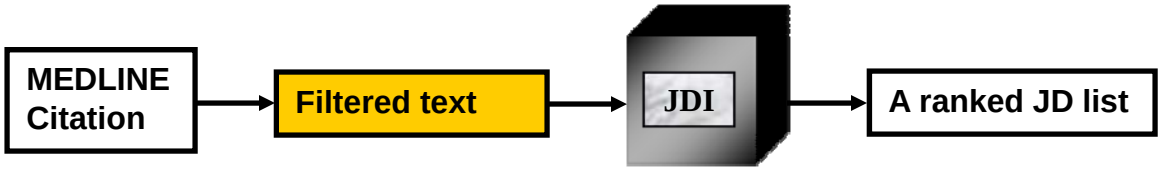
Application – JDI for TC



- MEDLINE Citation:

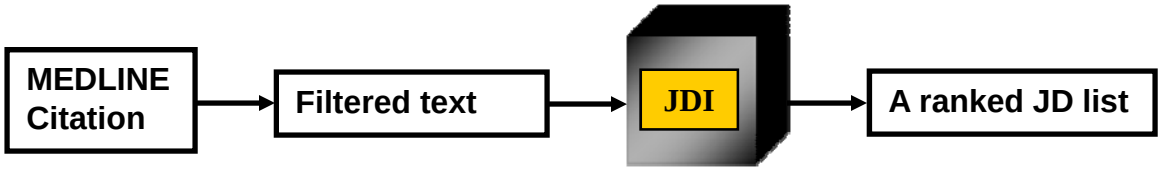
```
PMID- 15547873
OWN - NLM
STAT- MEDLINE
DA  - 20041119
DCOM- 20051108
PUBM- Print
IS  - 1531-5037 (Electronic)
VI  - 39
IP  - 11
DP  - 2004 Nov
TI  - Outcome and complications after resection of hepatoblastoma.
PG  - 1744-5; author reply 1745
FAU - Pritchard, Jon
AU  - Pritchard J
FAU - Stringer, Mark
AU  - Stringer M
LA  - eng
PT  - Comment
PT  - Letter
PL  - United States
TA  - J Pediatr Surg
JT  - Journal of pediatric surgery
JID – 0052631
…
```

Application – JDI for TC



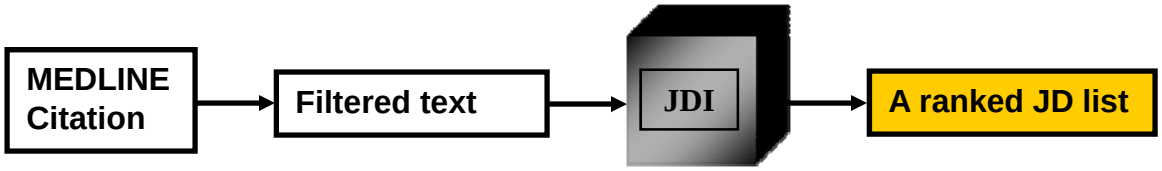
- MEDLINE Tokenizer (MLT) with tag TI:
  - `Outcome and complications after resection of hepatoblastoma.`

Application – JDI for TC



- Input Text (title):
  - `Outcome and complications after resection of hepatoblastoma.`
- Input Filter:
  - Word Extraction Filter:
    - `outcome and complications after resection hepatoblastoma`
  - Legal words Filter:
    - `resection hepatoblastoma`
  - Unique words Filter
    - `resection hepatoblastoma`
  - Final words
    - `resection hepatoblastoma`
- Get JD scores for both words from DB and calculate the average JD scores

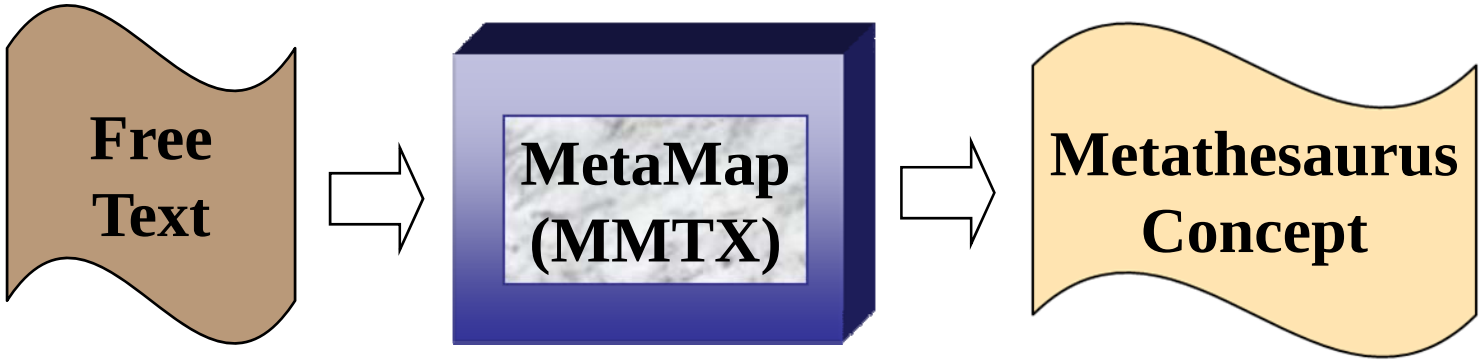
Application – JDI for TC



Input: Outcome and complications after resection of hepatoblastoma.
--- JD scores and rank based on word frequency ---
JD115|Surgery
1|0.031578|JD115|Surgery
2|0.028905|JD086|Pediatrics
3|0.024402|JD129|Neoplasms
4|0.023639|JD041|Gastroenterology
5|0.013307|JD045|Genetics, Medical
--- JD scores and rank based on document count for word ---
JD115|Surgery
1|0.060629|JD115|Surgery
2|0.050550|JD041|Gastroenterology
3|0.048146|JD129|Neoplasms
4|0.044012|JD086|Pediatrics
5|0.029093|JD123|Urology

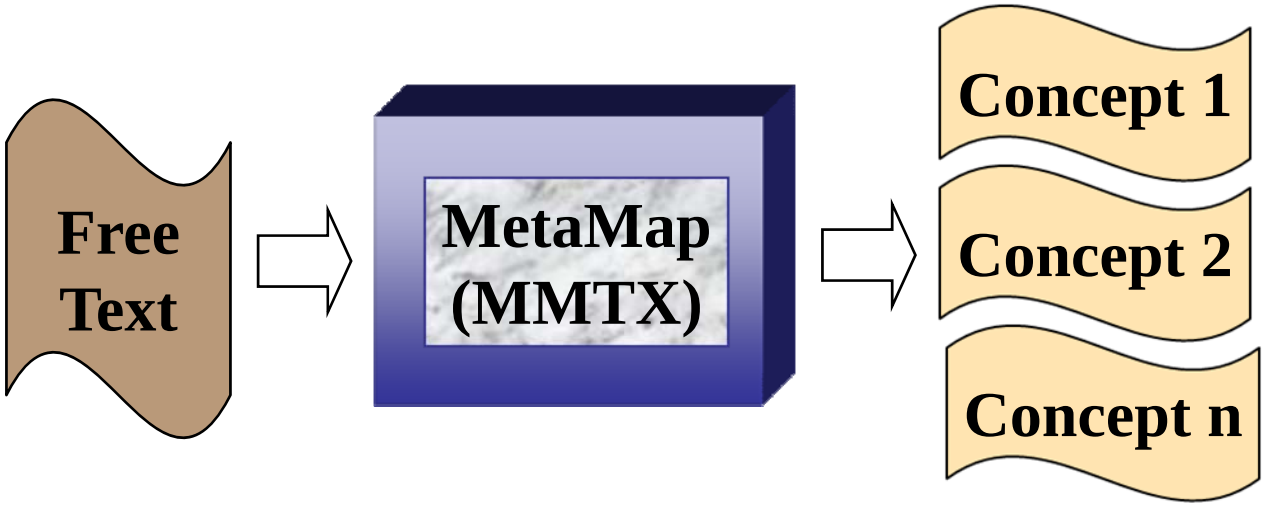
Application – STI for WSD

- NLP applications use MetaMap to map arbitrary text to concepts in the UMLS Metathesaurus



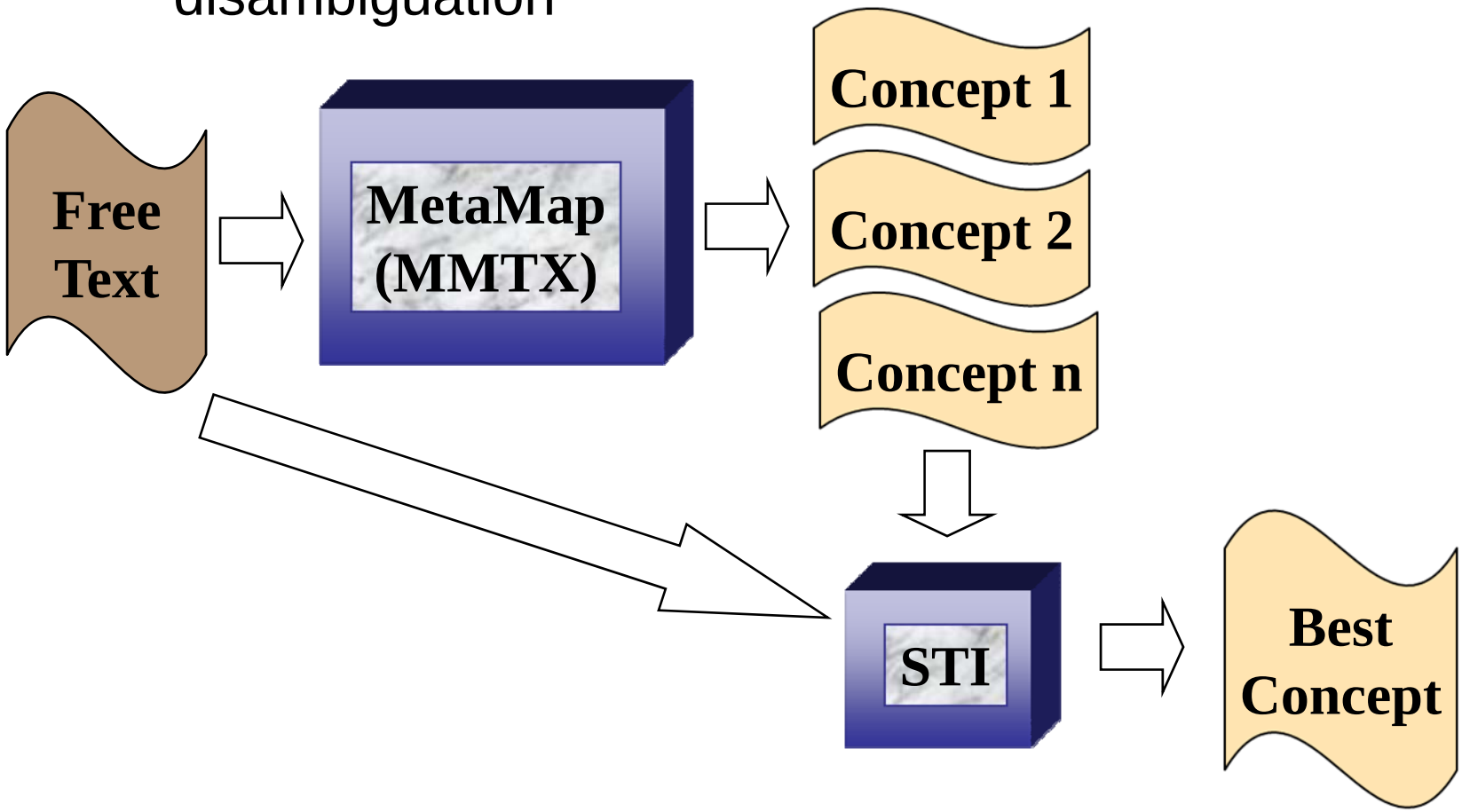
Application – STI for WSD

- Multiple mapped concepts with same confidence score generate ambiguity



Application – STI for WSD

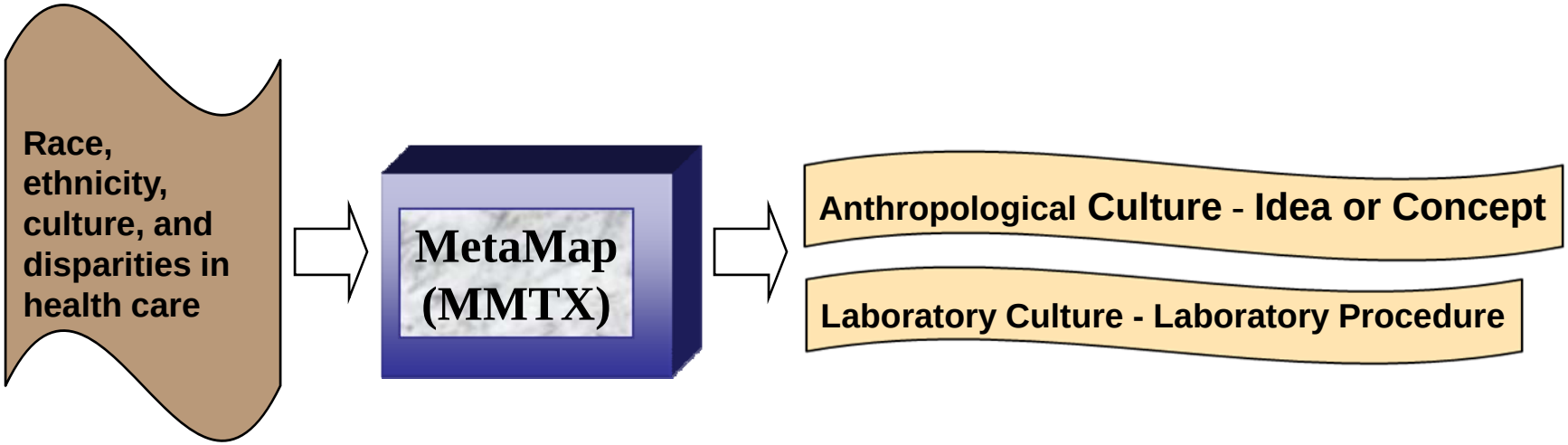
- Apply STI with candidate only option for disambiguation



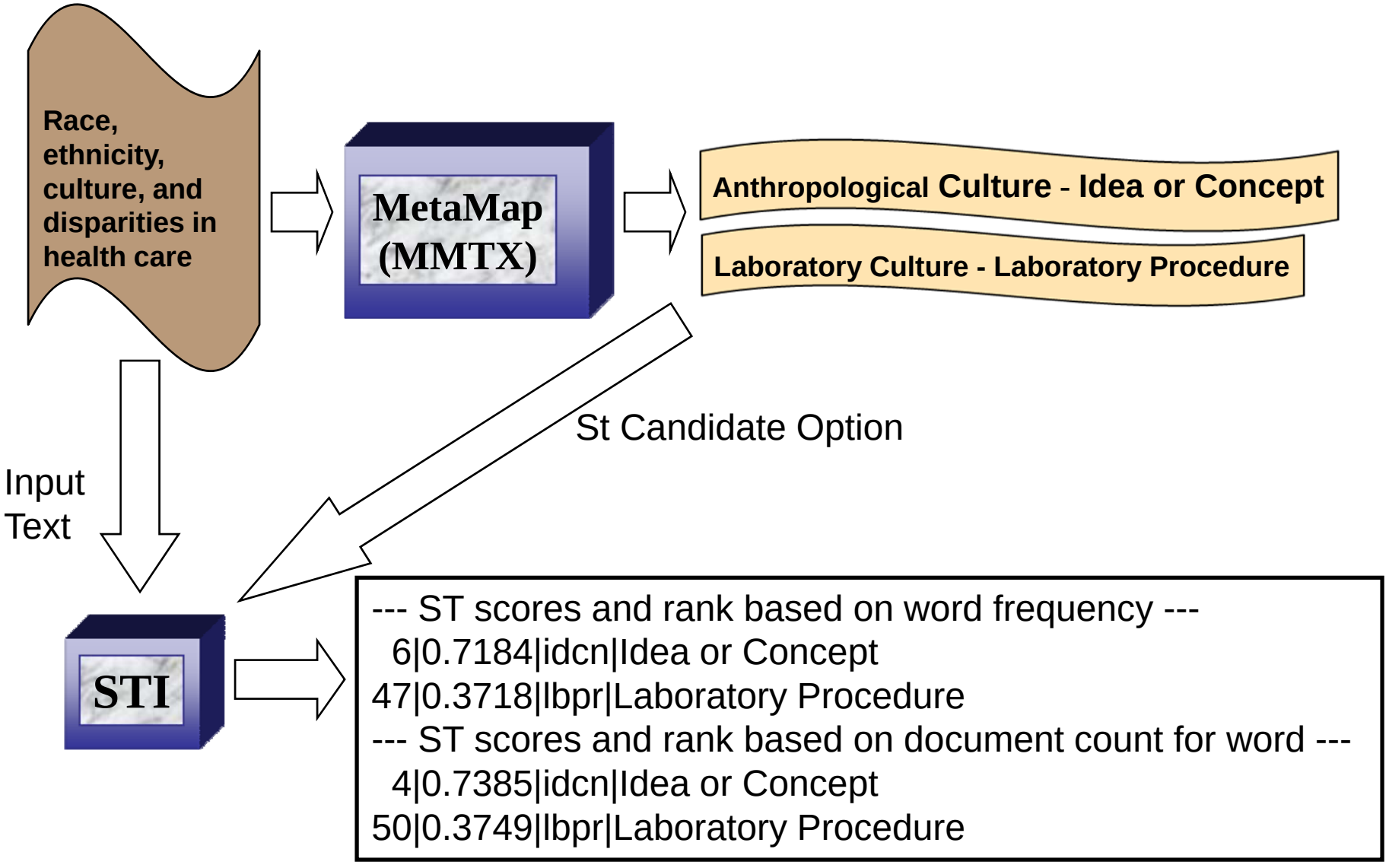
Application – STI for WSD

- **Example:**
  - Input: Race, ethnicity, culture, and disparities in health care
  - Where “culture” has two UMLS concepts/Semantic Types mapping from MetaMap/UMLSKS with same score:
    - Anthropological Culture - Idea or Concept
    - Laboratory Culture - Laboratory Procedure
  - Multiple mapping can cause ambiguity

Application – STI for WSD

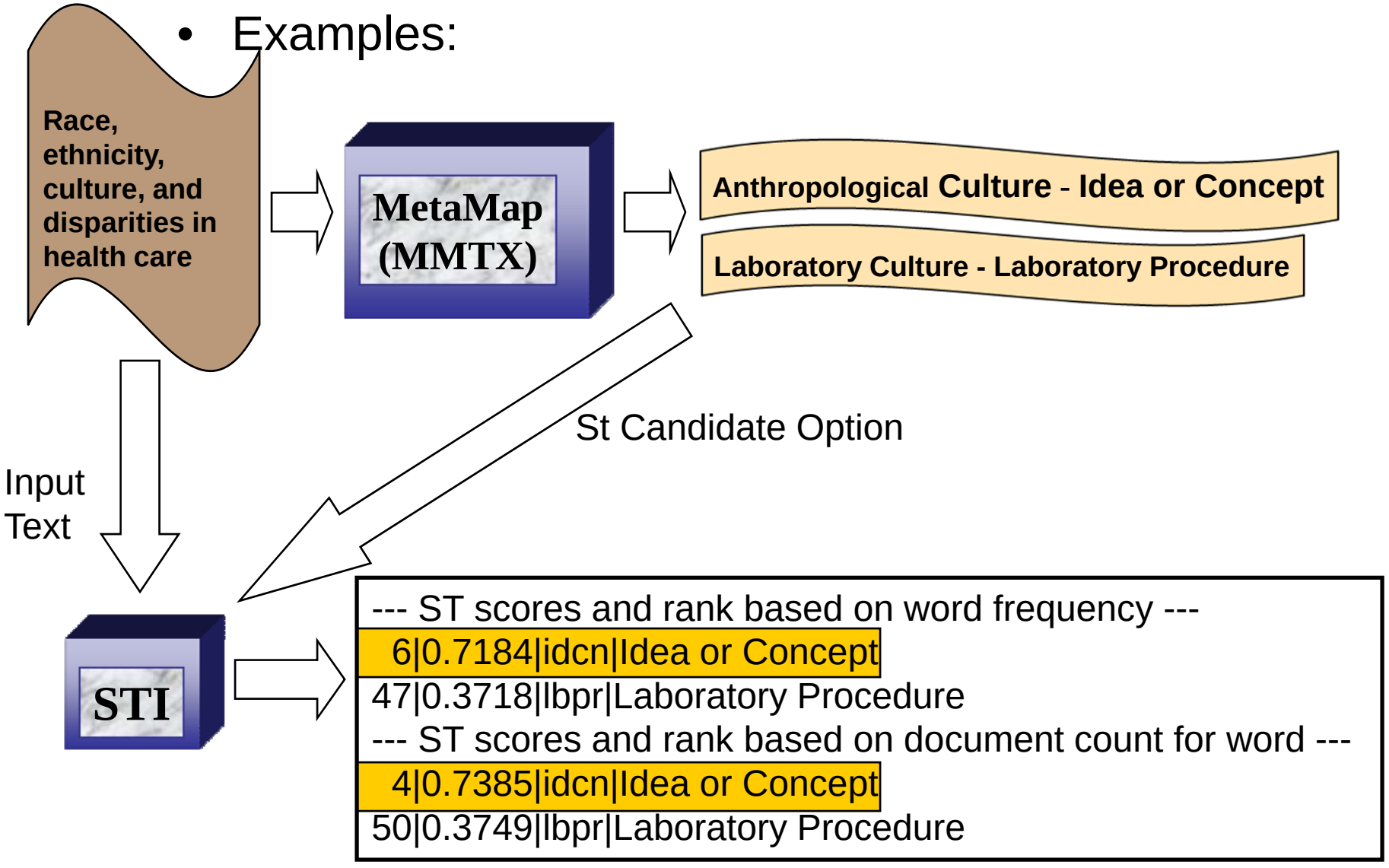


Application – STI for WSD



Application – STI for WSD

- Examples:



TCAT

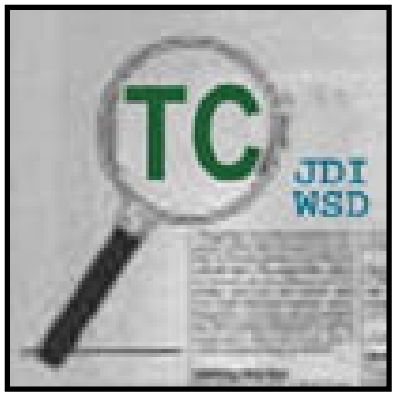
- Text Categorization Application Tools
- A showcase of applying TC package on our research projects
  - Use HTML as front end GUI
  - Use TC Java APIs as back end algorithm
  - Design to ease the processes of our existing research

- Demo
  - Apply JDI for TC on MEDLINE
    - PMID: 15547873
  - Apply STI for WSD
    - Input: Race, ethnicity, culture, and disparities in health care
    - ST candidate: idcn, lbpr

Future Work

- Tool package
  - Automated training set generation
  - Training set validation
  - Annual release with updated training set
- Research:
  - Use JD to index and retrieve MEDLINE
  - Apply TC tools on more medical databases
  - Apply TC tools on more WSD applications
  - Automatic stopwords determination
  - Enhance training set
  - Apply JDI methodology on Library of Congress (class number)

Thank You!



http://umlslex.nlm.nih.gov
http://umlslex.nlm.nih.gov/tc
jdi@nlm.nih.gov