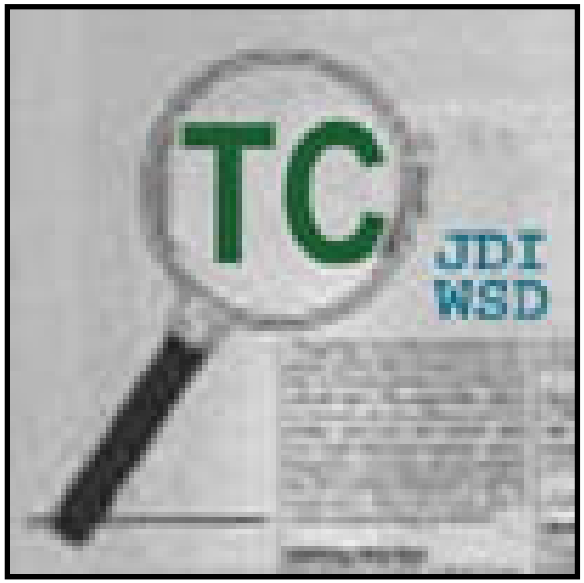
Text Categorization

Lexical Systems Group
National Library of Medicine
National Institutes of Health

Table of Contents

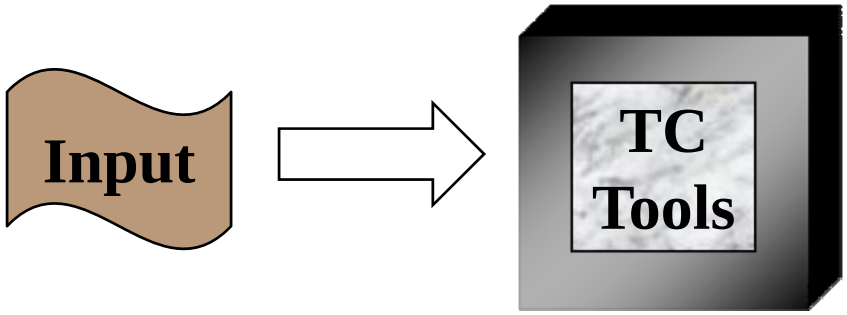
Introduction



Introduction:
Text Categorization Tools

- A set of tools for
  - Document classification
  - Identify text in a domain of interest
  - Word sense disambiguation
  - Indexing & retrieval
  - etc..

Introduction - TC Tools



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

Introduction - TC Tools

- A set of tools that generates ranked JD (biomedical disciplines) or ST (semantic categories) lists with scores to categorize the input text

What are JD & ST?

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

- **Semantic Types (STs):**
  - A set of 135 Semantic Types in the Semantic Network in NLM's UMLS (Unified Medical Language System)
  - 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]

Introduction - Example

- **Inputs:** 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
…

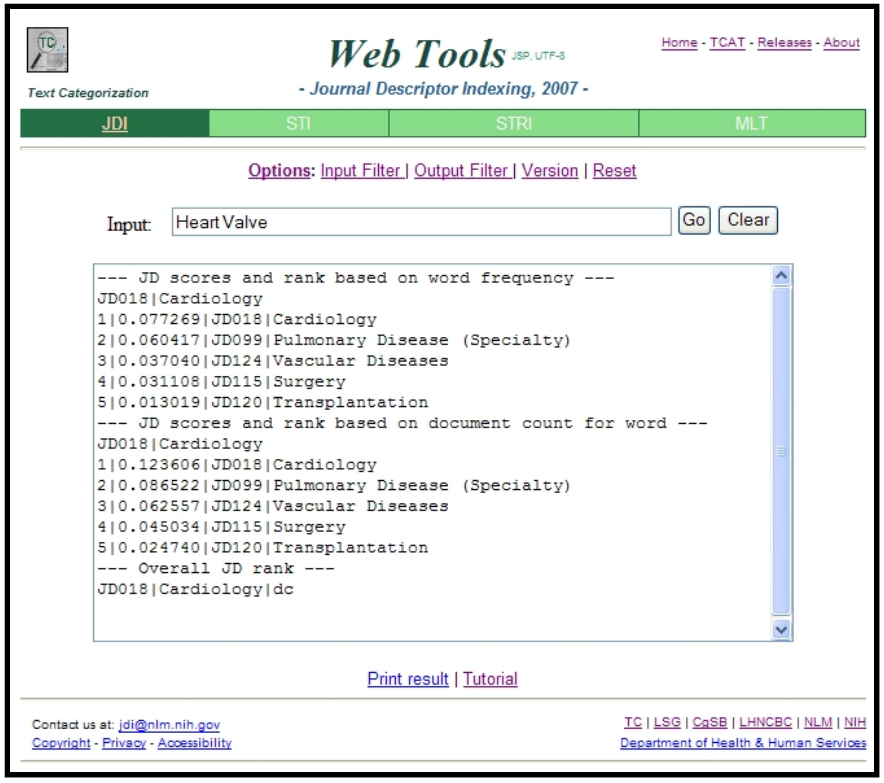
Introduction - Tool & Types

- Tools:
  - JDI (Journal Descriptor Indexing)
  - STI (Semantic Type Indexing)
  - STRI (Semantic Type Real-Time Indexing)
  - MLT (MEDLINE Tokenizer)
- Tool Types
  - Command line tools
  - 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 (Humphrey, 1998)
- 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]

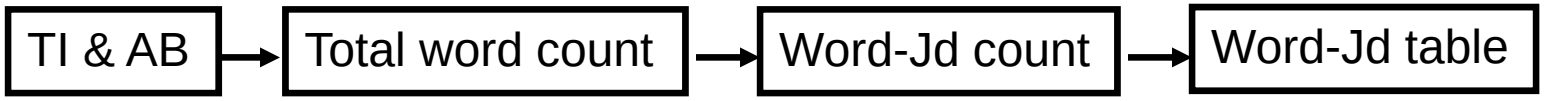
Tools – Journal Descriptor Indexing

- Based on statistical Word-Jd associations from a training set:
  - 3 years MEDLINE (2002 ~ 2004)
  - 4093 journals
  - 1.38 M MEDLINE citations
- Word-Jd score table:
  - Word count
  - Document count

JDI: Word-Jd Table

- **Word Count Score:**
  - Titles and abstracts in MEDLINE citations (training set)
  - Get total word count for all words from titles and abstracts
  - Get word count for all words co-occurs with all JDs
  - Calculate word score to generate Word-Jd table (WC):

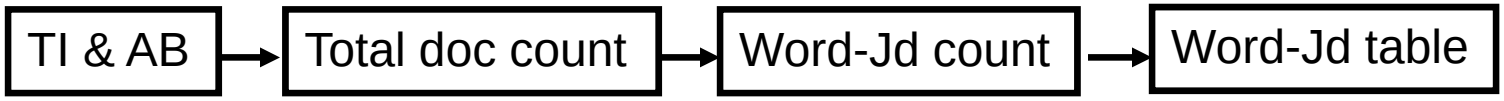
$$\text{word score} = \frac{\text{word count of word-Jd co-occurs}}{\text{total word count}}$$



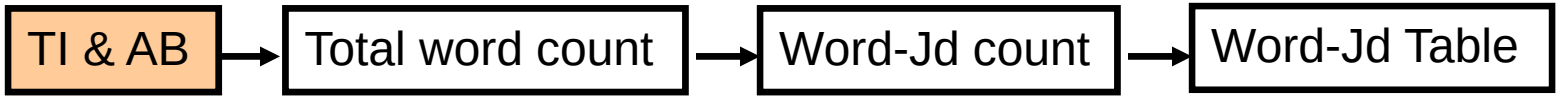
JDI: Word-Jd Table

- **Document count score:**
  - Titles and abstracts in MEDLINE citations (training set)
  - Get total document count for all words from titles and abstracts
  - Get document count for all words co-occurs with all JDs
  - Calculate document score to generate Word-Jd table (DC):

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

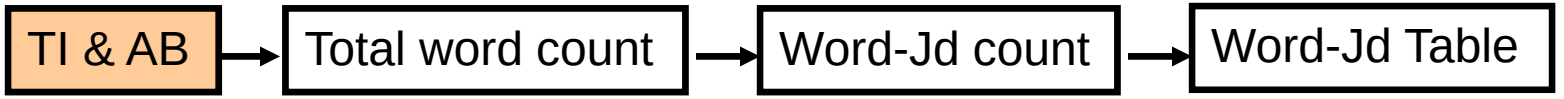


JDI: Word-Jd Table



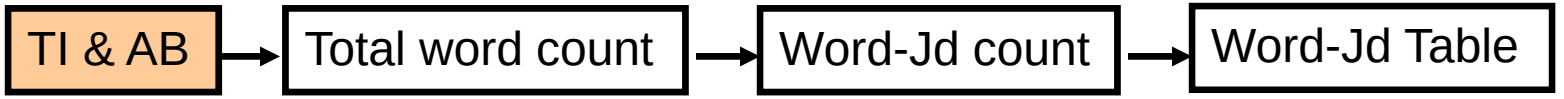
PMID- 961031
OWN - NLM
STAT- MEDLINE
DA - 19761020
DCOM- 19761020
LR - 20041117
PUBM- Print
IS - 0042-2835 (Print)
VI - 10
IP - 1
DP - 1976 Jan-Feb
TI - Postoperative arrhythmias in open-heart surgery, A study on fifty cases.
PG - 30-7
AB - 50 consecutive patients undergone open heart surgery were analyzed regarding postoperative arrhythmias in the first postoperative 3 days. Disturbances of rhythm occurred in each case of our group, serious or not serious (100%). Ventricular premature beats were the most frequent type of arrhythmia in the first and second postoperative days (80%). Two cases expired postoperatively. In one of them complete atrioventricular block developed after double valvular replacements (mitral and tricuspid). The other died of low cardiac output syndrome. Etiology of the arrhythmias
...

JDI: Word-Jd Table



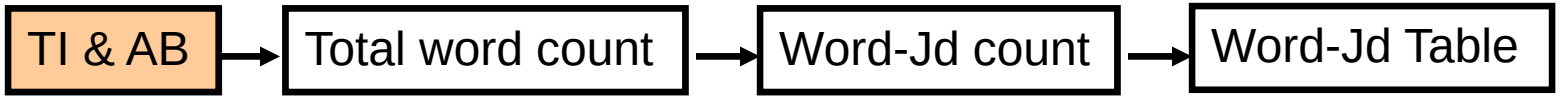
PMID- 961031
OWN - NLM
STAT- MEDLINE
…
DP - 1976 Jan-Feb
TI - Postoperative arrhythmias in open-heart surgery, A study on fifty cases.
PG - 30-7
AB - 50 consecutive patients undergone open heart surgery were analyzed regarding postoperative arrhythmias in the first postoperative 3 days. Disturbances of rhythm occurred in each case of our group, …
…
JT - Vascular surgery
JID - 0103277

JDI: Word-Jd Table



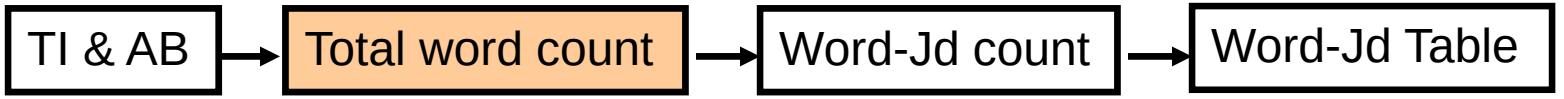
PMID- 961031
OWN - NLM
STAT- MEDLINE
…
DP - 1976 Jan-Feb
TI - Postoperative arrhythmias in open-heart surgery, A study on fifty cases.
PG - 30-7
AB - 50 consecutive patients undergone open heart surgery were analyzed regarding postoperative arrhythmias in the first postoperative 3 days. Disturbances of rhythm occurred in each case of our group, …
…
JT - Vascular surgery
JID - 0103277

JDI: Word-Jd Table



PMID- 961031
OWN - NLM
STAT- MEDLINE
…
DP - 1976 Jan-Feb
TI - Postoperative arrhythmias in open-heart surgery, A study on fifty cases.
PG - 30-7
AB - 50 consecutive patients undergone open heart surgery were analyzed regarding postoperative arrhythmias in the first postoperative 3 days. Disturbances of rhythm occurred in each case of our group, …
…
JT - Vascular surgery
JID - 0103277

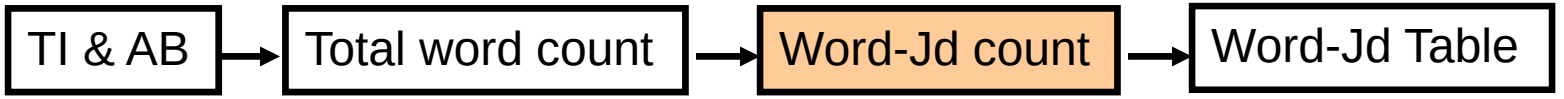
JDI: Word-Jd Table



TI - Postoperative arrhythmias in open-heart surgery, A study on fifty cases.
AB - 50 consecutive patients undergone open heart surgery were analyzed regarding postoperative arrhythmias in the first postoperative 3 days. Disturbances of rhythm occurred in each case of our group, …

Word	Count
postoperative	3
arrhythmias	2
in	3
open	3
heart	2
surgery	2
a	1
…	…

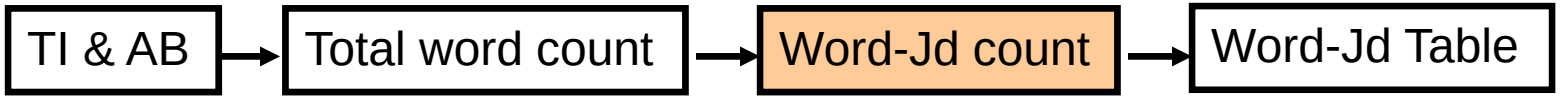
JDI: Word-Jd Table



PMID- 961031
OWN - NLM
STAT- MEDLINE
…
DP - 1976 Jan-Feb
TI - Postoperative arrhythmias in open-heart surgery, A study on fifty cases.
PG - 30-7
AB - 50 consecutive patients undergone open heart surgery were analyzed regarding postoperative arrhythmias in the first postoperative 3 days. Disturbances of rhythm occurred in each case of our group, …
…
JT - Vascular surgery
JID - 0103277

Cardiology|JD018
Surgery|JD115

JDI: Word-Jd Table



TI - Postoperative arrhythmias in open-heart surgery, A study on fifty cases.
AB - 50 consecutive patients undergone open heart surgery were analyzed regarding postoperative arrhythmias in the first postoperative 3 days. Disturbances of rhythm occurred in each case of our group, …
JDs – Cardiology|Surgery

- Cardiology|JD018

Word	Count
study	412
rhythm	304
…	…
heart	4248
…	…

- Surgery|JD115

Word	Count
group	128
surgery	5043
…	…
heart	605
…	…

JDI: Word-Jd Table

TI & AB → Total word count → Word-Jd count → Word-Jd Table

- Cardiology|JD018

Word	Count
study	412
rhythm	304
…	…
heart	4248
…	…

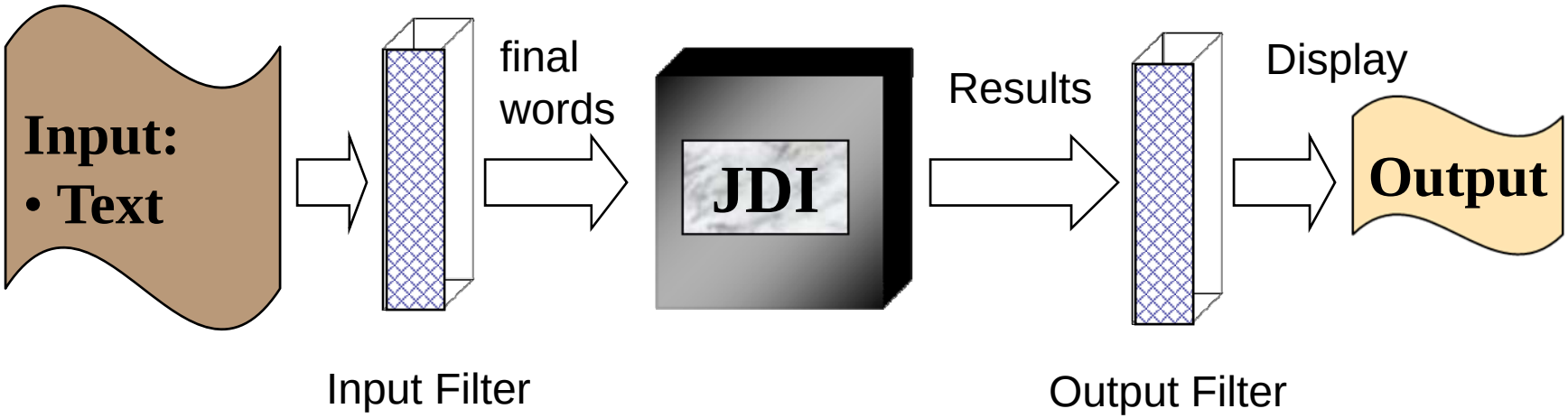
- Surgery|JD115

Word	Count
group	128
surgery	5043
…	…
heart	605
…	…

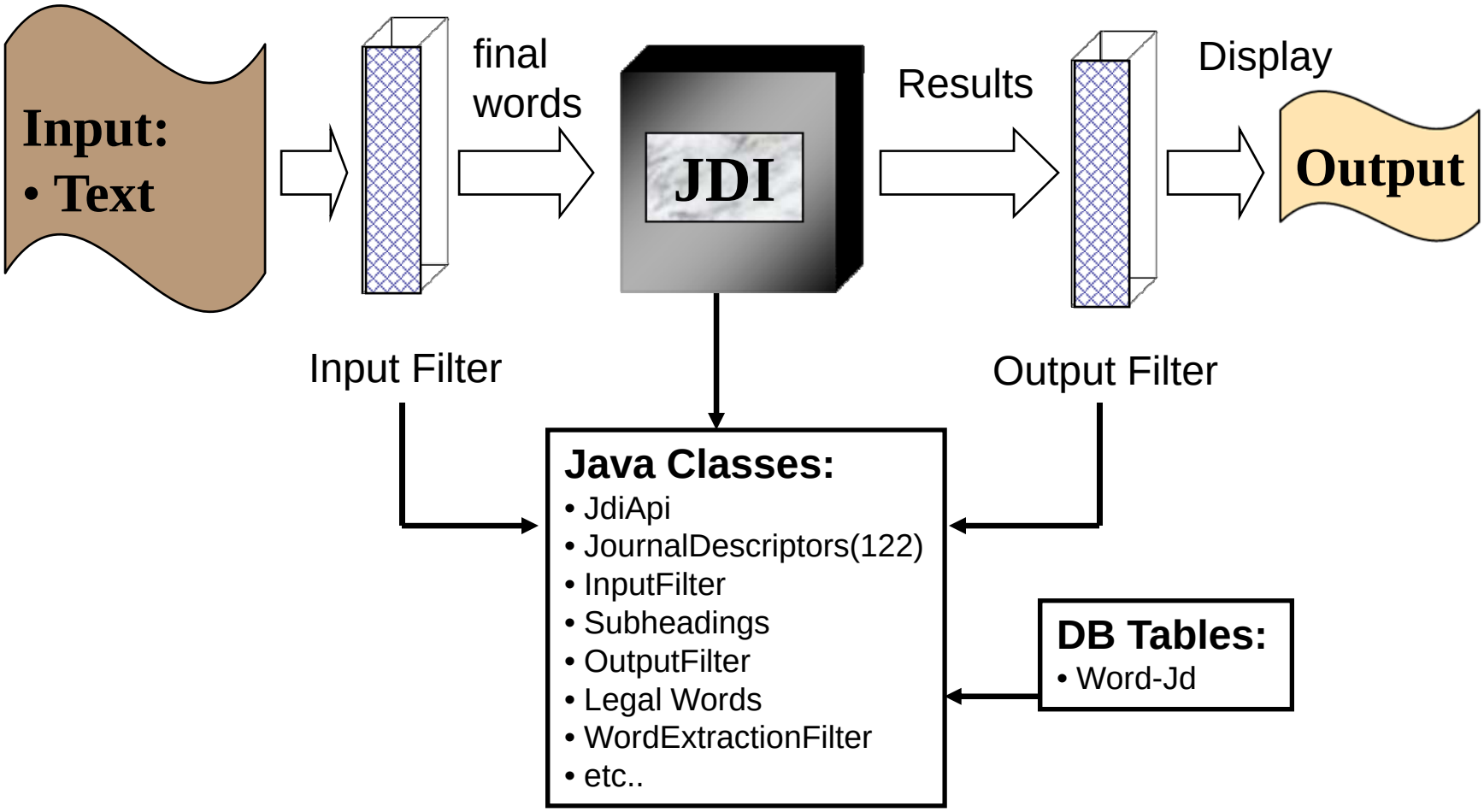
WC/DC Scores

Word	JDID	WC score	DC score
a00	JD036	0.5608018	0.3280480
…	…	…	…
heart	JD017	0.00180937	0.00524491
heart	JD018	0.04665726	0.09366356
heart	…	…	…
heart	JD114	0.00075555	0.00223503
heart	JD115	0.00665273	0.01562358
…	…	…	…

TC Tools - JDI


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

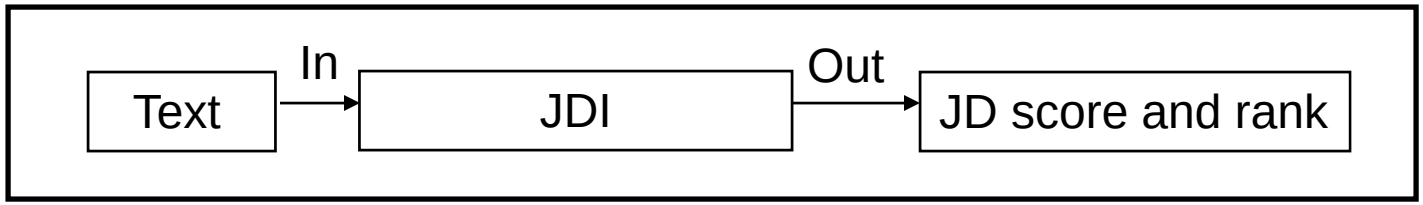
TC Tools - JDI


Input:
• Text
final
words
JDI
Results
Display
Output
Input Filter
Output Filter
Java Classes:
• JdiApi
• JournalDescriptors(122)
• InputFilter
• Subheadings
• OutputFilter
• Legal Words
• WordExtractionFilter
• etc..
DB Tables:
• Word-Jd

TC Tools - JDI

- Input filter
  - Tokenize and filter out words for processing
  - Word extraction filter (filter out irrelevant words)
  - Unique word filter (filter out duplicated words)
  - Legal word filter: stopwords filter, restrictwords filter, word length, word count, document count, etc.
- Process:
  - Get JD scores from DB for each final word
  - Calculate average JD scores for the input
- Output filter
  - Ranked JD list with scores (0 ~ 1)
  - Cluster display
  - Display number
  - Candidate only display
  - etc..

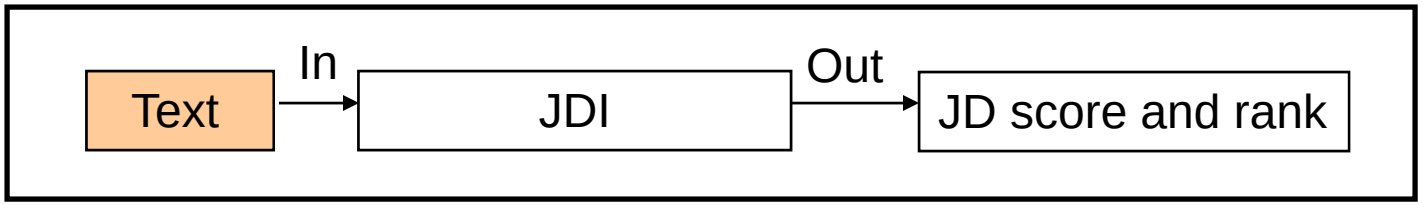
JDI - Example



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

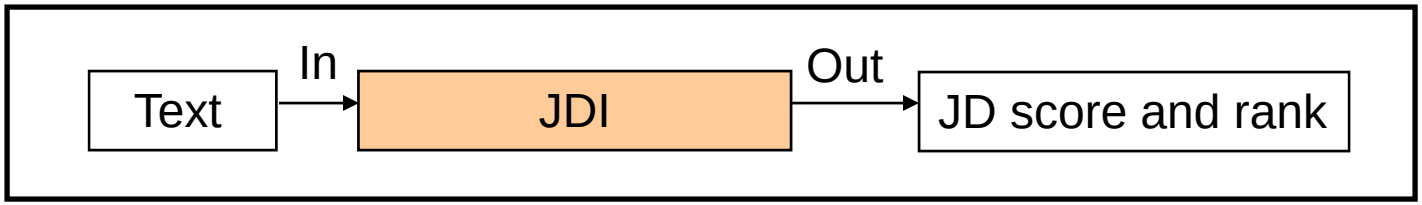
--- JD scores and rank based on document count for word ---
JD018|Cardiology
1|0.123606|JD018|Cardiology
2|0.086522|JD099|Pulmonary Disease (Specialty)
3|0.062557|JD124|Vascular Diseases
4|0.045034|JD115|Surgery
5|0.024740|JD120|Transplantation
6|0.024412|JD005|Anesthesiology
7|0.023319|JD030|Diagnostic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

JDI - Example



Input: **`The Heart Valve`**

JDI - Example

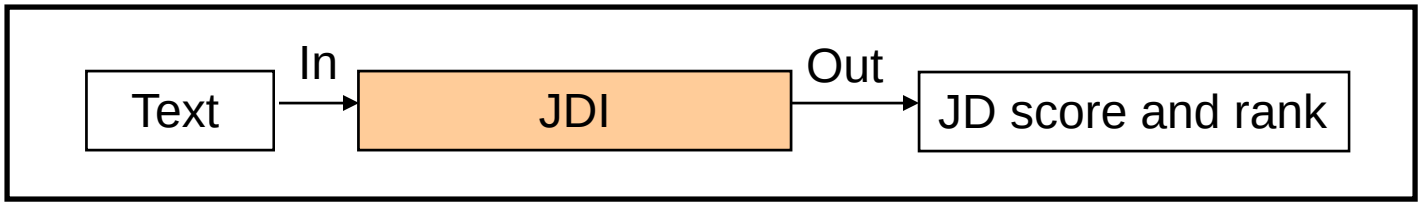


Input: ~~The~~ **Heart Valve**

Algorithm: Java Program

Filtered out (stop word)

JDI - Example



Input: **The Heart Valve**

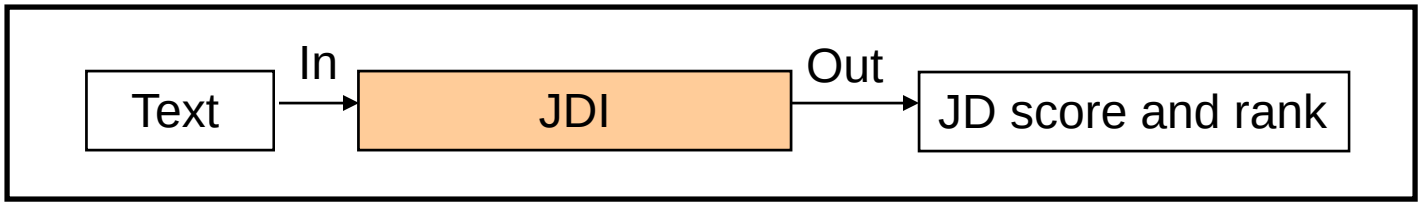
DB: Word-JD table

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

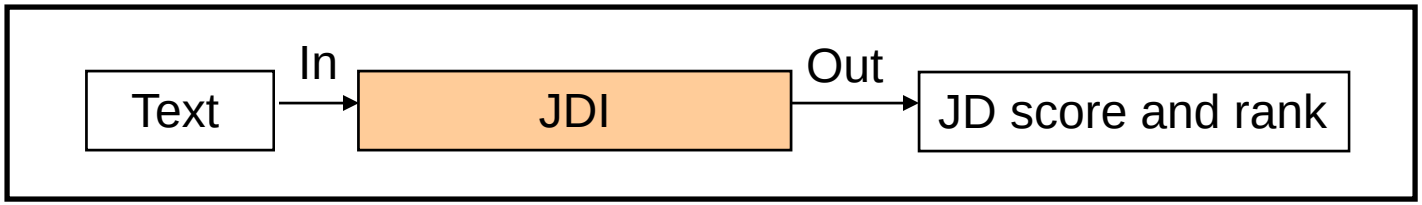
Heart	
JD001	0.0001787949
JD002	0.0015644556
…..	…..
JD018	0.09365937
…	…

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**
…	…

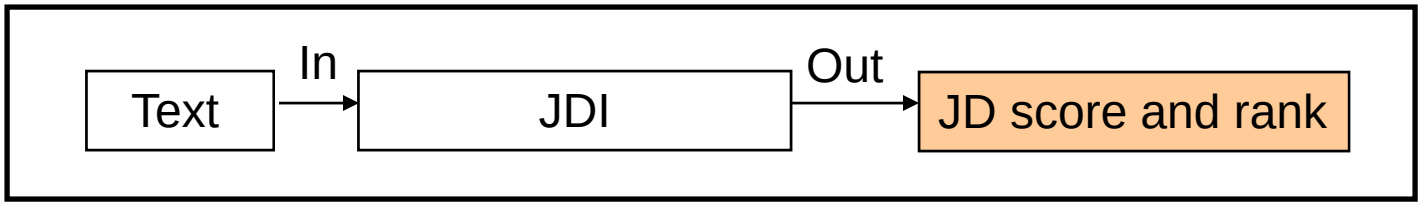
Cardiology

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

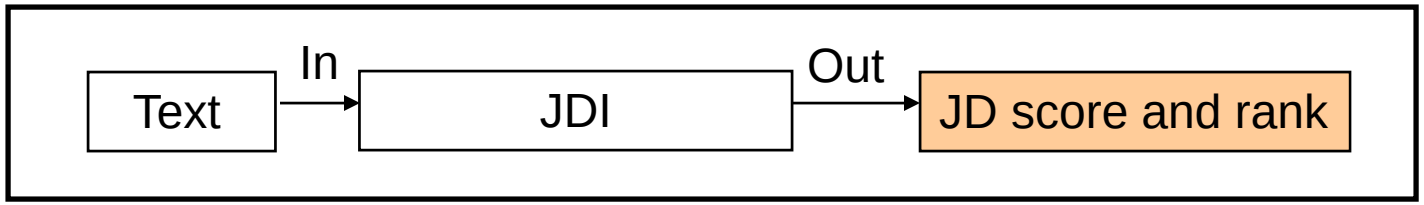
Filtered out (stop word)

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 - Summary

- JDI tool
- JDI methodology
  - Training set (citations & descriptors)
  - Word-Jd table

- What if training set is not available?

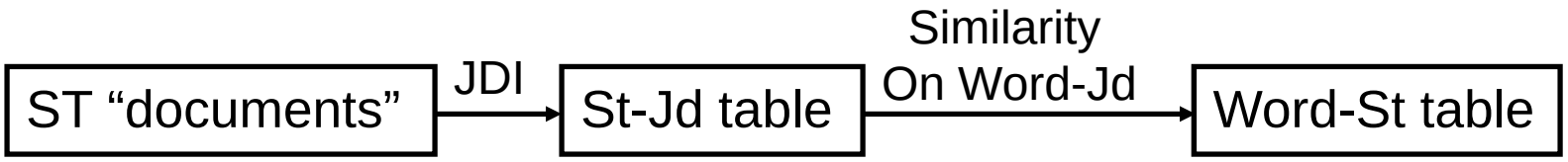
Tools – Semantic Type Indexing

- **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].

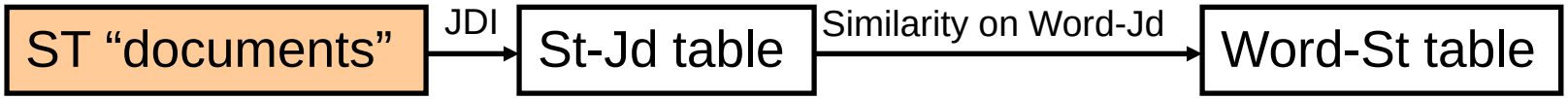
- **STI Tool:**
  - Use JDI methodology as basis
  - Calculate the average ST scores for input text from Word-St table
  - Print out ranked ST list with scores

STI: Word-St Table

- Generate ST “documents” (all words associated with ST).
- Apply JDI on ST “documents” to generate St-Jd table
- Calculate similarity (cosine coefficient) on JDI of ST “documents” (St-Jd) and JDI on individual training set words (Word-Jd) to generate Word-St table



STI: Word-St Table



- **MRCON**

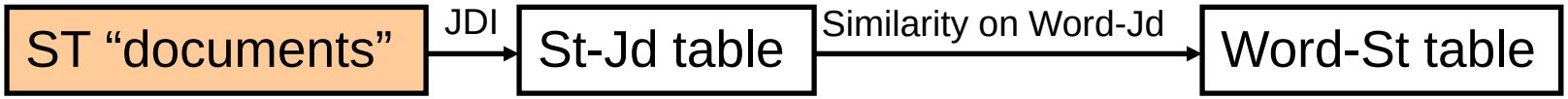
```
…
C0018786|SPA|P|L0352821|PF|S0461126|TESTS AUDITIVOS|3|
C0018786|SPA|S|L2710424|PF|S3184484|Prueba de audicion|3|
C0018787|DUT|P|L2060782|PF|S2399004|Hart|3|
C0018787|ENG|P|L0018787|PF|S0047194|Heart|0|
C0018787|ENG|P|L0018787|VC|S0375948|HEART|0|
C0018787|ENG|P|L0018787|VC|S0419735|heart|0|
…
```

- **MRSTY**

```
…
C0018786|T060|Diagnostic Procedure|
C0018787|T023|Body Part, Organ, or Organ Component|
C0018789|T047|Disease or Syndrome|
…
```

BPOC

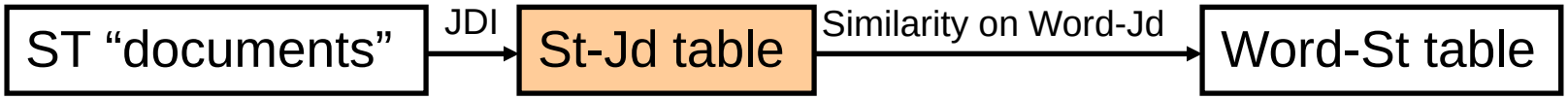
STI: Word-St Table



- **ST “documents”**

ST	Words
aapp	fetuins, ache, actin, maltase, aminoacid, diaminoacid, ...
acab	stump, ankylosis, scar, corn, hernia, keloid, kyphosis, …
acty	activities, bioterrorism, burial, interment, civilization, …
…	…
antb	calcimycin, topicycline, albicdin, ansamycins, menomycin, …
…	…
bpoc	liver, jaw, skin, kidney, liver, heart, hand, finger, stomach, …
…	…

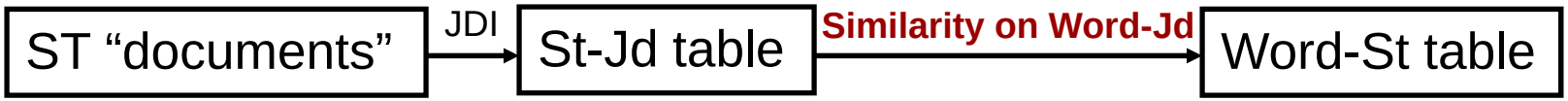
STI: Word-St Table



- **ST-Jd table**

ST	JDID	JD	WC	DC
…	…	…	…	…
bpoc	**JD001**	**Acquired Immunodeficiency Syndrome**	**0.002833920**	**0.003035284**
bpoc	**JD002**	**Aerospace Medicine**	**0.007579821**	**0.005971246**
bpoc	**JD003**	**Allergy and Immunology**	**0.012128671**	**0.012012365**
…	…	…	…	…
bpoc	**JD134**	**Women’s Health**	**0.03777063**	**0.048345257**
…	…	…	…	…

STI: Word-St Table



- **ST-Jd table**

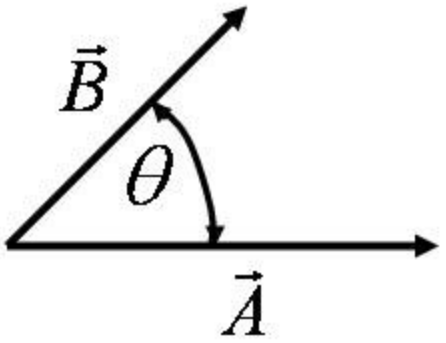
ST	JDID	WC	DC
…	…	…	…
bpoc	JD001	0.002833920	0.003035284
bpoc	JD002	0.002833920	0.003035284
…	…	…	…
bpoc	JD134	0.03777063	0.048345257
…	…	…	…

- **Word-Jd table (JDI)**

Word	JDID	WC	DC
…	…	…	…
heart	JD001	0.00004663	0.00017879
heart	JD002	0.00004664	0.00015644
…	…	…	…
heart	JD134	0.00020986	0.00040228
…	…	…	…

STI: Cosine Coefficient

$$\vec{A} = (a_1, a_2, \ldots, a_n), \vec{B} = (b_1, b_2, \ldots, b_n)$$



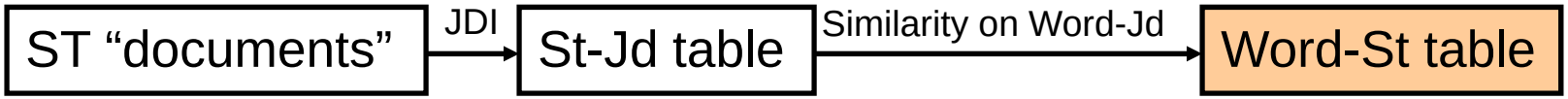
$$\vec{A} \cdot \vec{B} = |\vec{A}| \cdot |\vec{B}| \cdot \cos(\theta)$$

$$\Rightarrow \cos(\theta) = \frac{\vec{A} \cdot \vec{B}}{|\vec{A}| \cdot |\vec{B}|}$$

$$\Rightarrow \cos(\theta) = \frac{(a_1 \cdot b_1 + a_2 \cdot b_2 + \ldots + a_n \cdot b_n)}{\sqrt{a_1^2 + a_2^2 + \ldots + a_n^2} \cdot \sqrt{b_1^2 + b_2^2 + \ldots + b_n^2}}$$

$$\Rightarrow \cos(\theta) = \frac{\sum_{i=0}^{n} a_i \cdot b_i}{\sqrt{\sum_{i=0}^{n} a_i^2} \cdot \sqrt{\sum_{i=0}^{n} b_i^2}}$$

STI: Word-St Table



- **ST-Jd table**

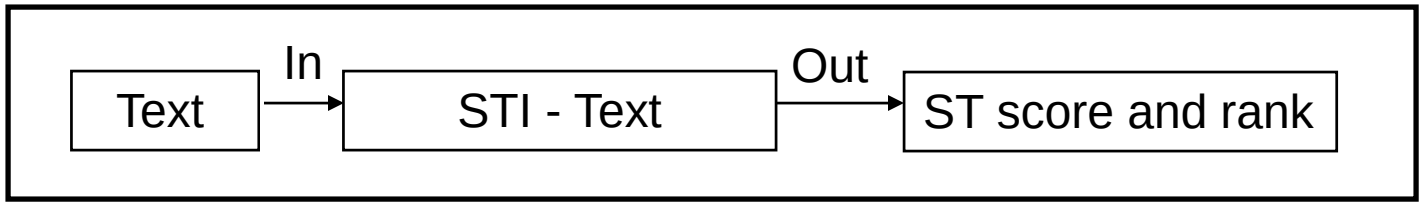
ST	JDID	WC	DC
…	…	…	…
bpoc	**JD001**	0.002833920	0.003035284
bpoc	**JD002**	0.002833920	0.003035284
…	…	…	…
bpoc	**JD134**	0.03777063	0.048345257
…	…	…	…

- **Word-Jd table (JDI)**

Word	JDID	WC	DC
…	…	…	…
heart	**JD001**	0.00004663	0.00017879
heart	**JD002**	0.00004664	0.00015644
…	…	…	…
heart	**JD134**	0.00020986	0.00040228
…	…	…	…

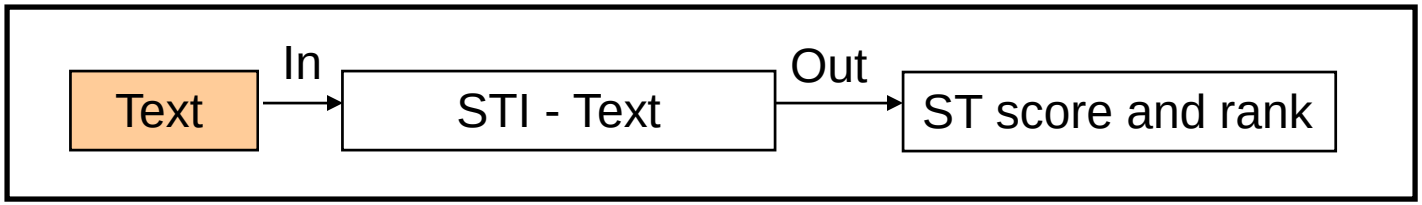
Word	St	WC	DC
…	…	…	…
heart	**blor**	0.3981	0.4469
heart	**bmod**	0.3299	0.3868
heart	**bodm**	0.1714	0.2251
heart	**bpoc**	0.3400	0.3903
…	…	…	…

Tools - STI



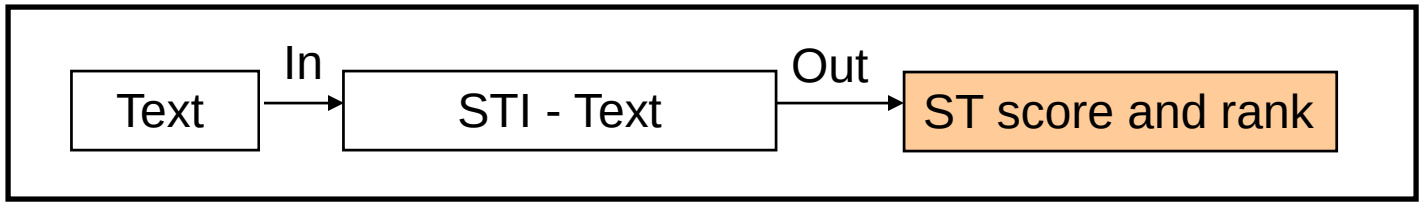
- **Input filter:**
  - Tokenize and filter out words for processing
  - Word extraction filter (filter out irrelevant words)
  - Unique word filter (filter out duplicated words)
  - Legal word filter: stopwords filter, restrictwords filter, word length, word count, document count, etc.
- **Process:**
  - Get ST scores from DB for each final word
  - Calculate average ST scores for the input
- **Output filter**
  - Ranked ST list with scores (0 ~ 1)
  - Cluster display
  - Display number
  - Candidate only display
  - etc..

Tools – STI Inputs



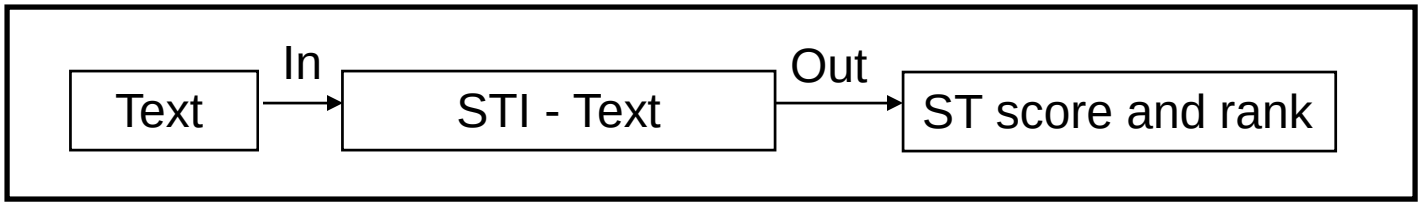
- Text
  - word
  - phrase
  - sentence
  - paragraph
  - page
  - document
  - etc..

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

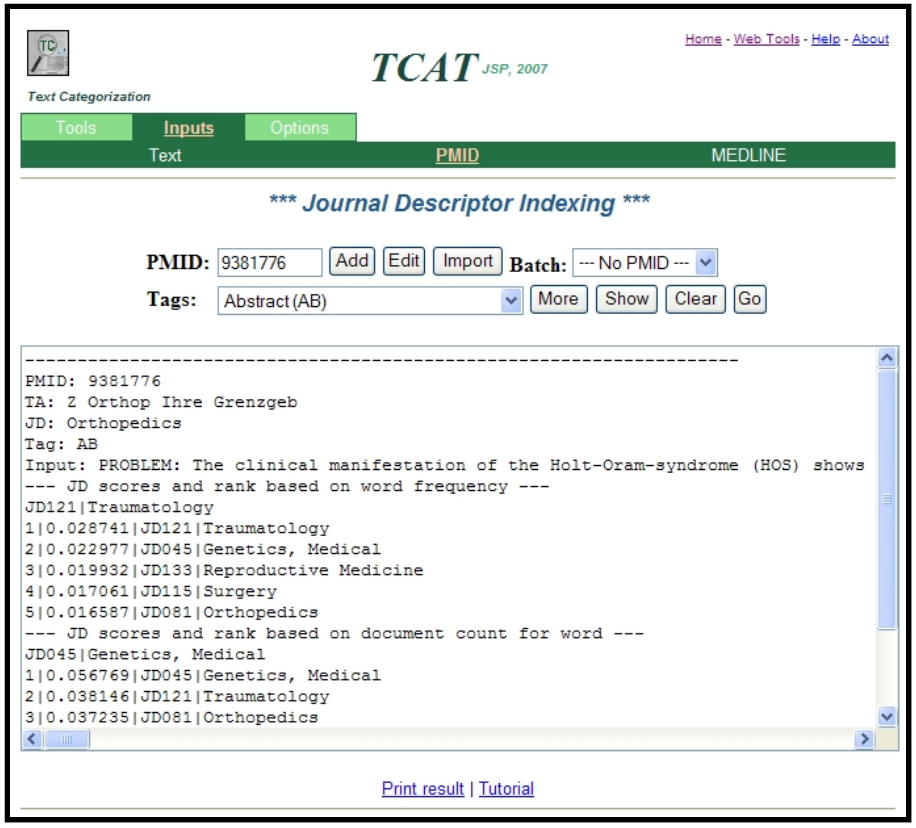
STI - Summary

- STI tool
  - For different disciplines
  - Word-St table
  - Use JDI methodology as basis

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

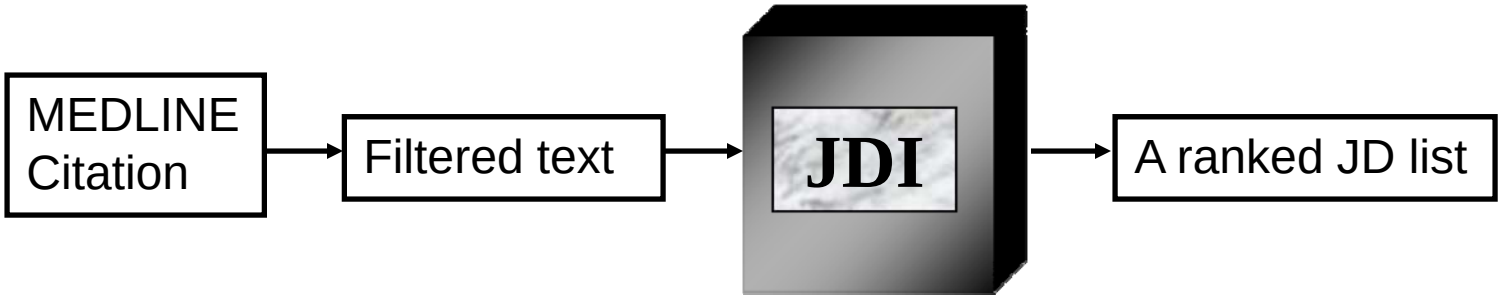


Application:
JDI for Text Categorization

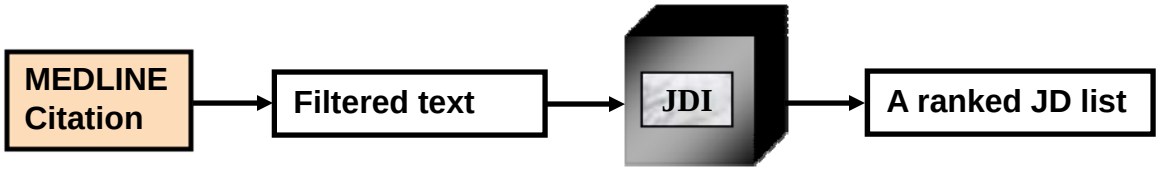
- JDI to index and categorize MEDLINE
- Inputs:
  - Title
  - Abstract
  - Title & abstract
- Outputs:
  - A ranked JD list with scores

Application: JDI for TC

- Procedures:
  - Select the MEDLINE citation
  - Filtered text
    - Identify fields of interest (TI, AB, TI & AB)
    - Filter out irrelevant characters and words
  - Apply JDI
  - Get results



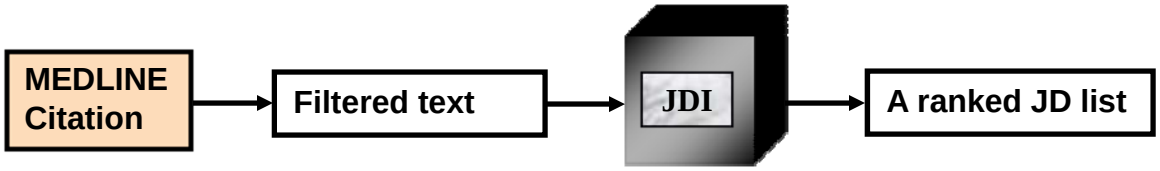
Application: JDI for TC



- 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
…
SO - J Pediatr Surg. 2004 Nov;39(11):1744-5; author reply 1745.

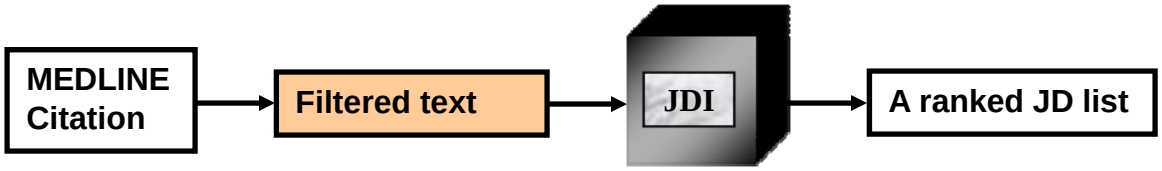
Application: JDI for TC



- MEDLINE Citation:

PMID- 15547873
OWN - NLM
STAT- MEDLINE
…
TI - Outcome and complications after resection of hepatoblastoma.
…
SO - J Pediatr Surg. 2004 Nov;39(11):1744-5; author reply 1745.

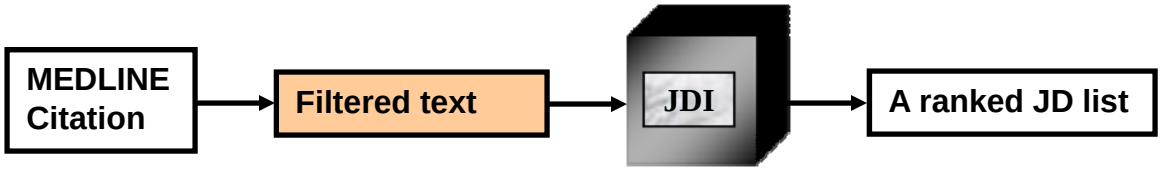
Application: JDI for TC



- MEDLINE Citation:

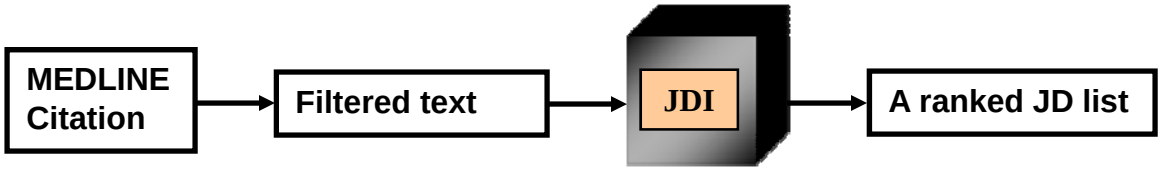
PMID- 15547873
OWN - NLM
STAT- MEDLINE
…
TI - Outcome and complications after resection of hepatoblastoma.
…
SO - J Pediatr Surg. 2004 Nov;39(11):1744-5; author reply 1745.

Application: JDI for TC



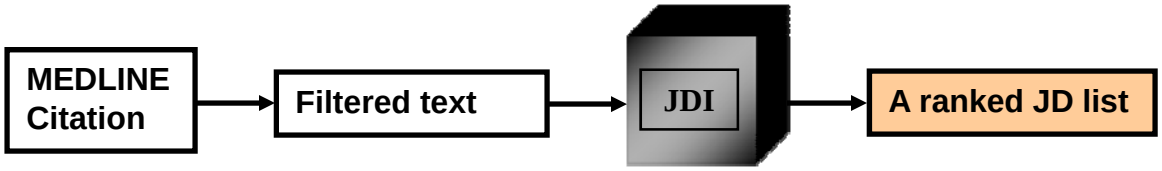
- MEDLINE Tokenizer (MLT) with tag TI:
  - Outcome and complications after resection of hepatoblastoma.

Application: JDI for TC



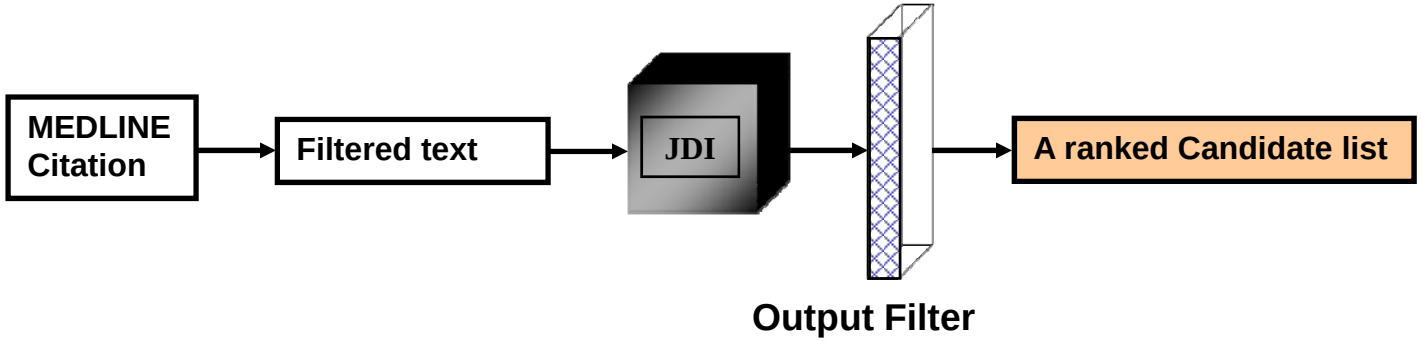
- Input Text (title):
  - Outcome and complications after resection of hepatoblastoma.
- Input Filter:
  - Word Extraction Filter:
    - outcome and complications after resection hepatoblastoma
  - Legal words Filter:
    - resection hepatoblastoma
  - Unique words Filter
    - resection hepatoblastoma
  - 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

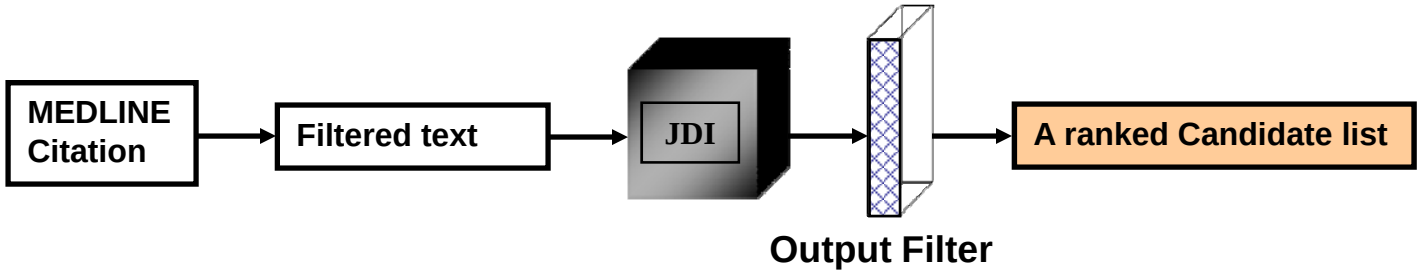
Application:
JDI to identify text
in a domain of interest



Example: to identify text in molecular biology domain

- Run text through JDI
- Apply candidates output filter option
- Check if either [Biochemistry] or [Molecular Biology] is in the top 15 rank JDs

Application: JDI to IT



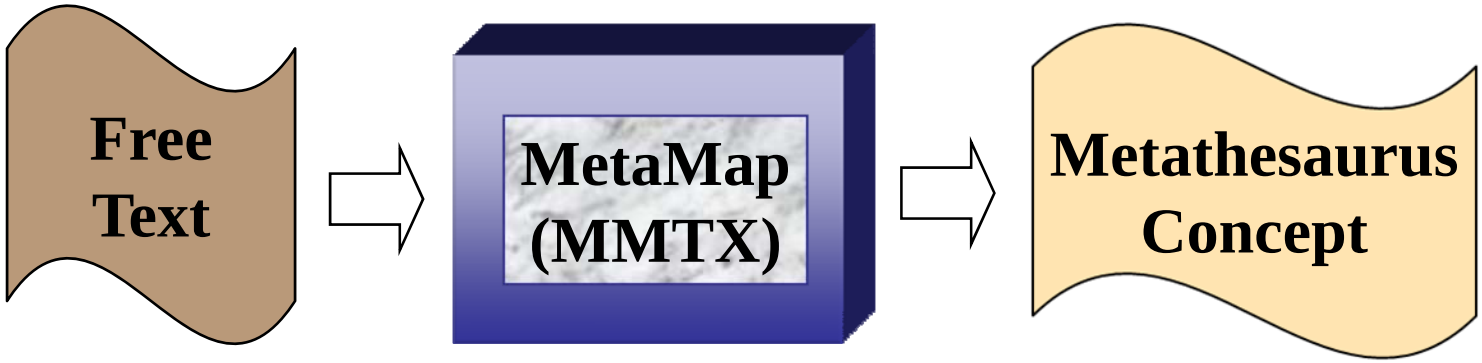
PMID: 12928053
TI: Methylenetetrahydrofolate reductase and angiotensin converting enzyme gene polymorphisms in two genetically and diagnostically distinct cohort of Alzheimer patients.
AB: The role of methylenetetrahydrofolate reductase (MTHFR) and angiotensin converting enzyme (ACE) gene polymorphisms as risk factors for the occurrence of Alzheimer's disease (AD) is still controversial. In this ….

JDI Outputs (with candidate filter option):

--- JD scores and rank based on word frequency ---
10|0.008049|JD067|Molecular Biology
17|0.006159|JD012|Biochemistry
--- JD scores and rank based on document count for word ---
8|0.018241|JD067|Molecular Biology
16|0.012244|JD012|Biochemistry

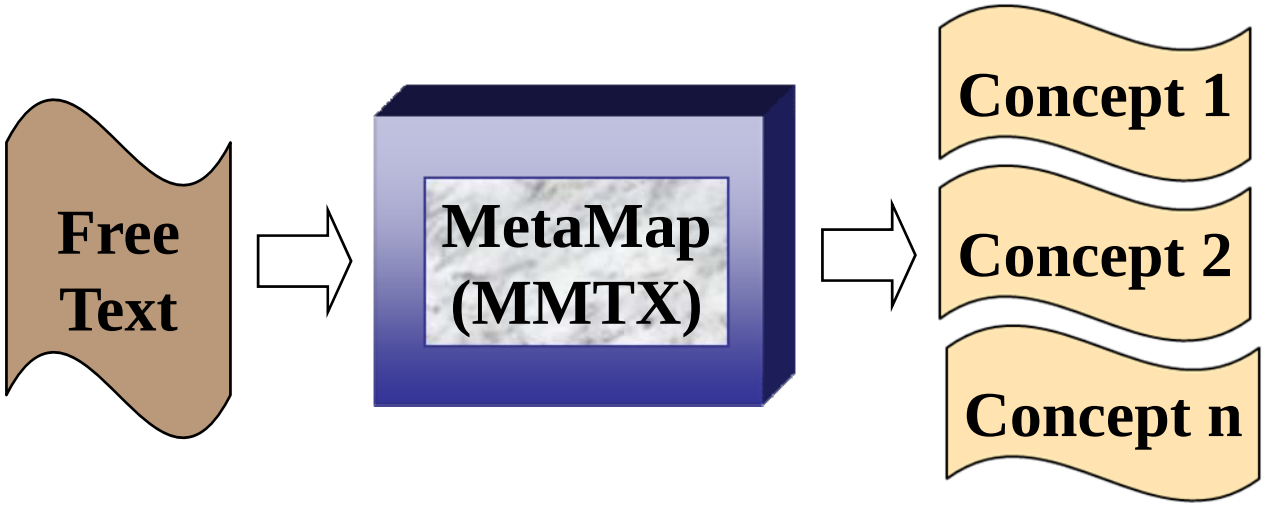
Application:
STI for Word Sense Disambiguation

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



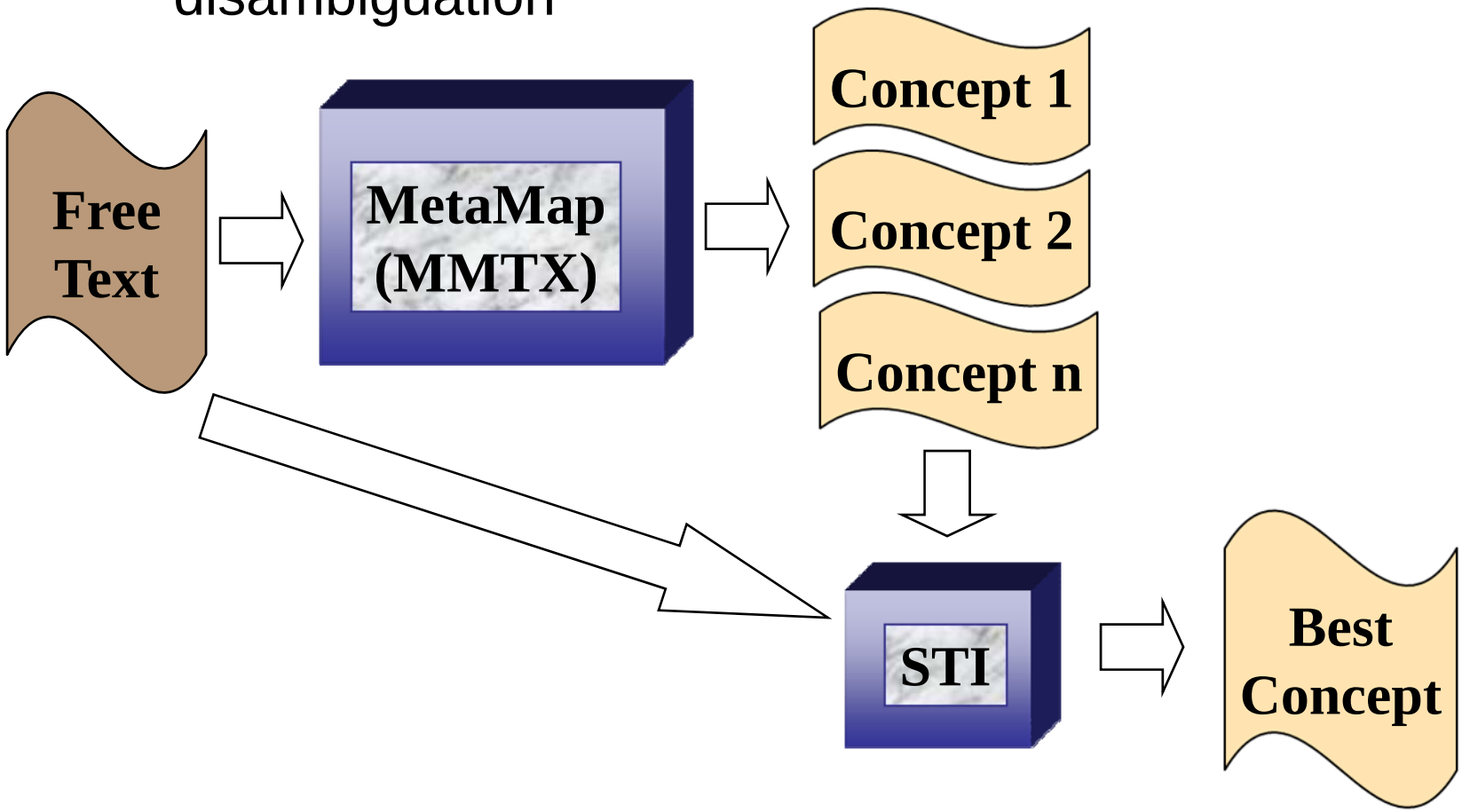
Application: STI for WSD

- Multiple mapped concepts with same confidence score generate ambiguity



Application: STI for WSD

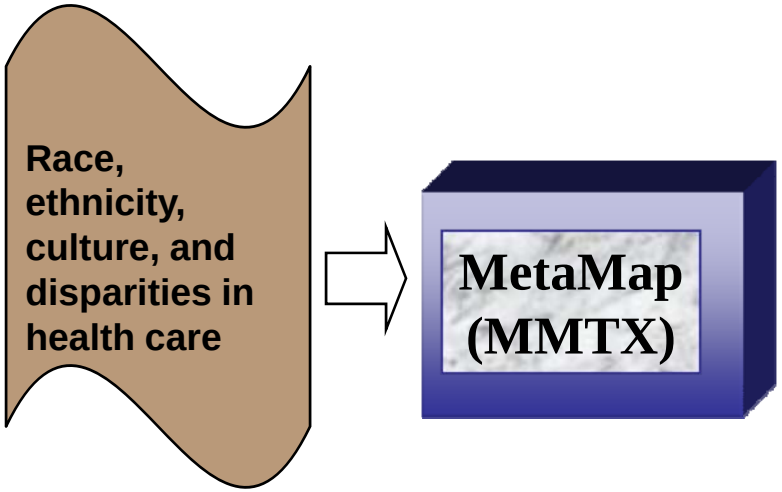
- Apply STI with candidate only option for disambiguation



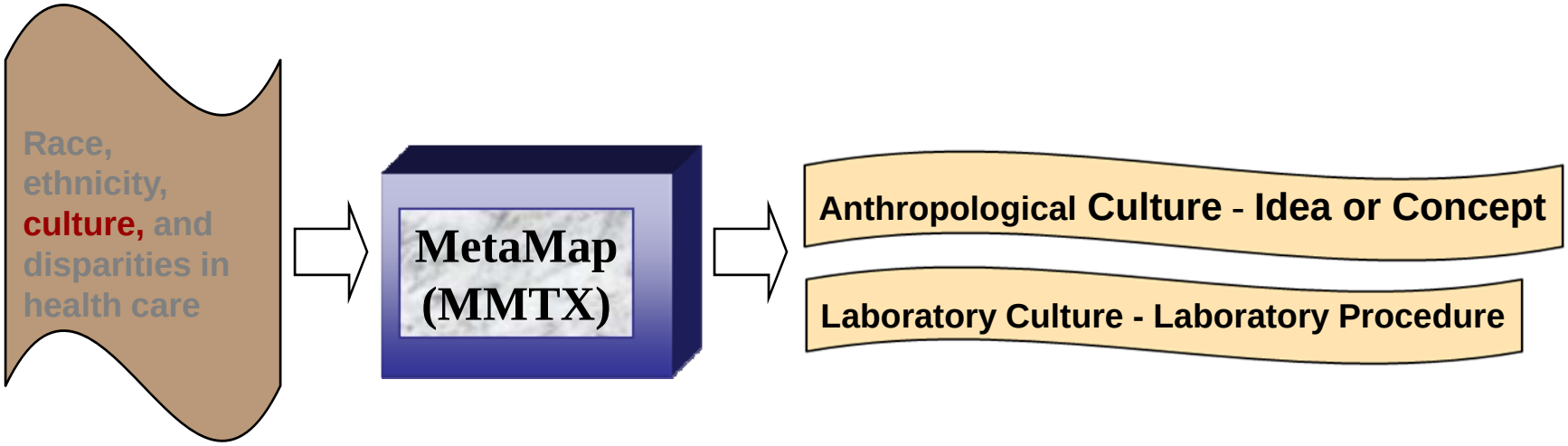
Application: STI for WSD

- **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 concept mapping can cause ambiguity

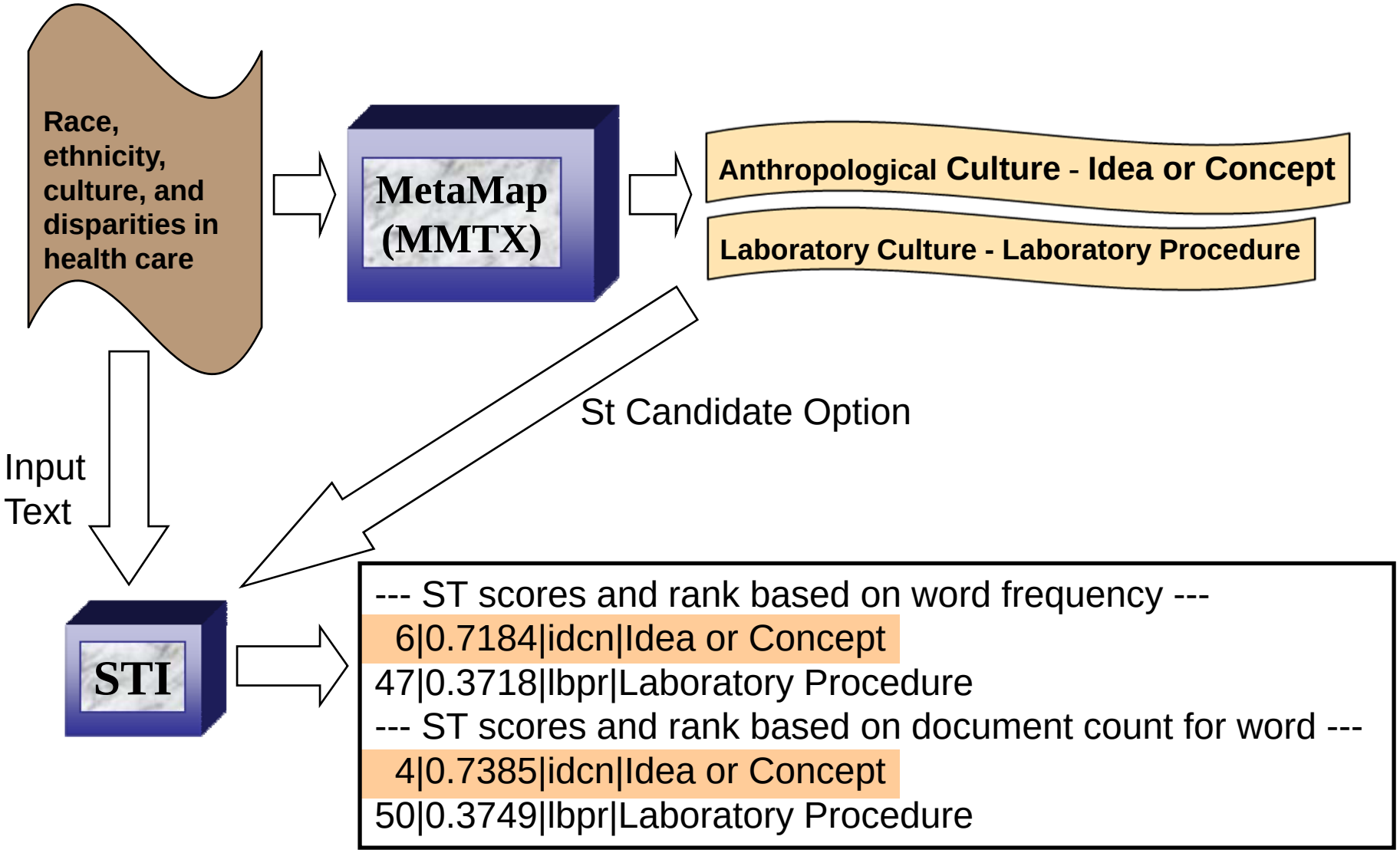
Application: STI for WSD



Application: STI for WSD



Application: STI for WSD



TCAT

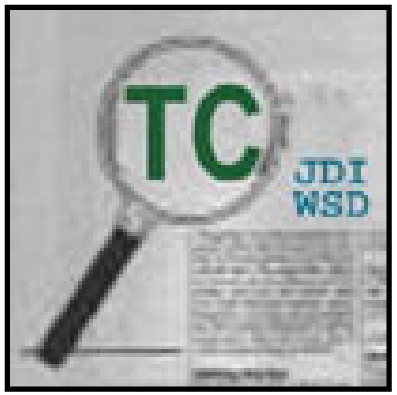
- Text Categorization Application Tools
- A showcase of applying TC package on our research projects
  - Use HTML as front end GUI
  - Use TC Java APIs as back end algorithm
  - Design to ease the processes of our existing research

- Demo
  - Apply JDI for TC on MEDLINE
    - PMID: 15547873
  - Apply JDI to identify text in a domain of interest
    -PMID: 12928053
    -JD candidate: Molecular Biology, Biochemistry
  - Apply STI for WSD
    -Input: Race, ethnicity, culture, and disparities in health care
    - ST candidate: idcn, lbpr

Future Plan

- Tool package
  - Annual release with updated training set
  - Automated training set generation
- Research:
  - Use JD to index and retrieve MEDLINE
  - Apply TC tools on more medical databases
  - Apply TC tools on more WSD applications
  - Use WSD for automatic indexing
  - Automatic stopwords determination
  - Enhance training set
  - Quality control
  - General categorization of specialized medical collection
  - Apply JDI methodology to different set of descriptor (Library of Congress, class number)

Thank You!



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