# 의생명 문헌 기반

# 약물 유사도 계산 방법 소개 및 실습

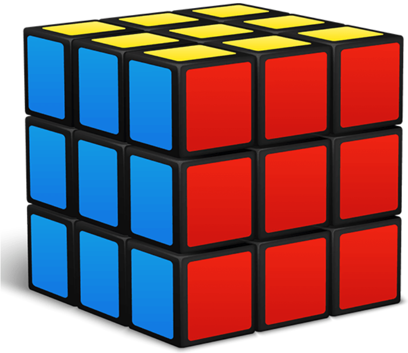
서울대학교 의생명지식공학연구실
심용선
yongsun0926@snu.ac.kr

서울대학교
SEOUL NATIONAL UNIVERSITY

BIKE
Biomedical Knowledge
Engineering Laboratory
Seoul National University

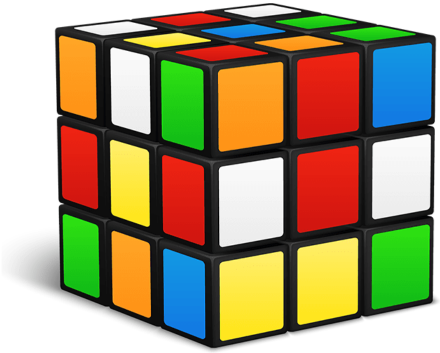
# Data Type

## Structured Data



- 고정된 필드에 저장된 데이터
- 관계형 데이터 베이스, 스프레드시트 등
- 구조 변경에 있어 제한됨
- 별도의 분석 처리 기술 없이 간단한 쿼리를 통하여 원하는 데이터 추출 가능

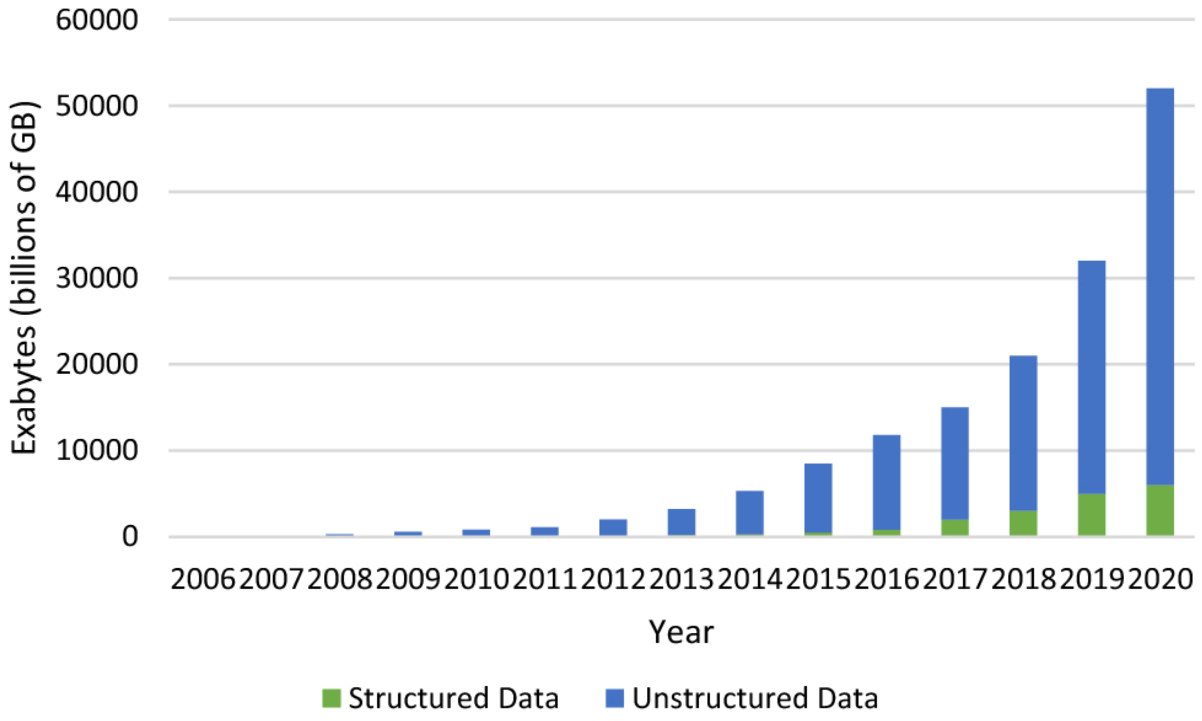
## Unstructured Data



- 고정된 필드에 저장되어 있지 않은 데이터
- 텍스트 문서, 이미지/동영상/음성 데이터 등
- 구조 변경에 있어 자유로움
- 다양하고 방대한 양의 데이터를 처리할 수 있는 별도의 분석 처리 기술 필요

# Structured/Unstructured Data Growth Rate Comparison

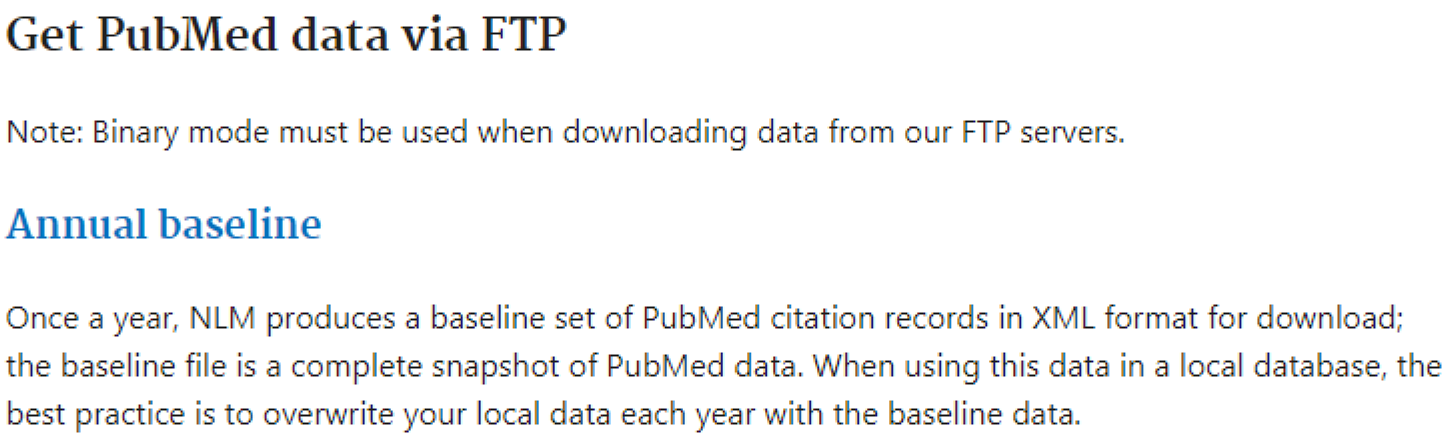
## 정형 · 비정형데이터 증가 추세



Azad, P., Navimipour, N. J., Rahmani, A. M., & Sharifi, A. (2020). The role of structured and unstructured data managing mechanisms in the Internet of things. Cluster computing, 23(2), 1185-1198.

# Text Data Collection

## Download



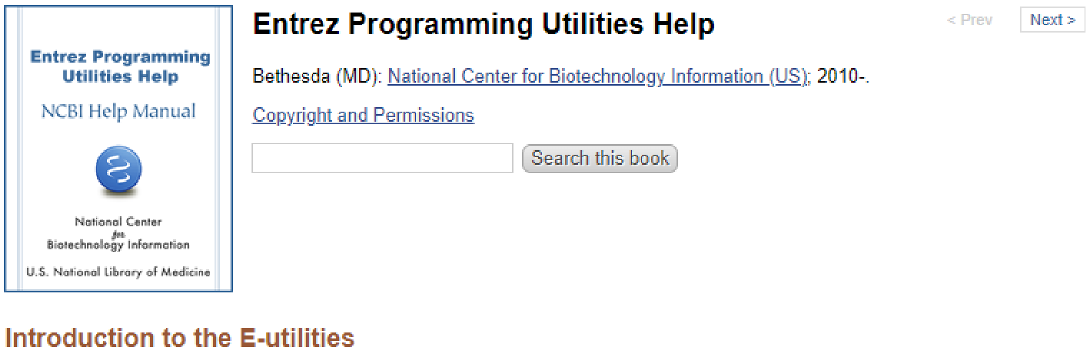
### Get PubMed data via FTP

Note: Binary mode must be used when downloading data from our FTP servers.

#### Annual baseline

Once a year, NLM produces a baseline set of PubMed citation records in XML format for download; the baseline file is a complete snapshot of PubMed data. When using this data in a local database, the best practice is to overwrite your local data each year with the baseline data.

## Use API



Entrez Programming Utilities Help
NCBI Help Manual
National Center for Biotechnology Information
U.S. National Library of Medicine
< Prev
Next >
Search this book

**Entrez Programming Utilities Help**

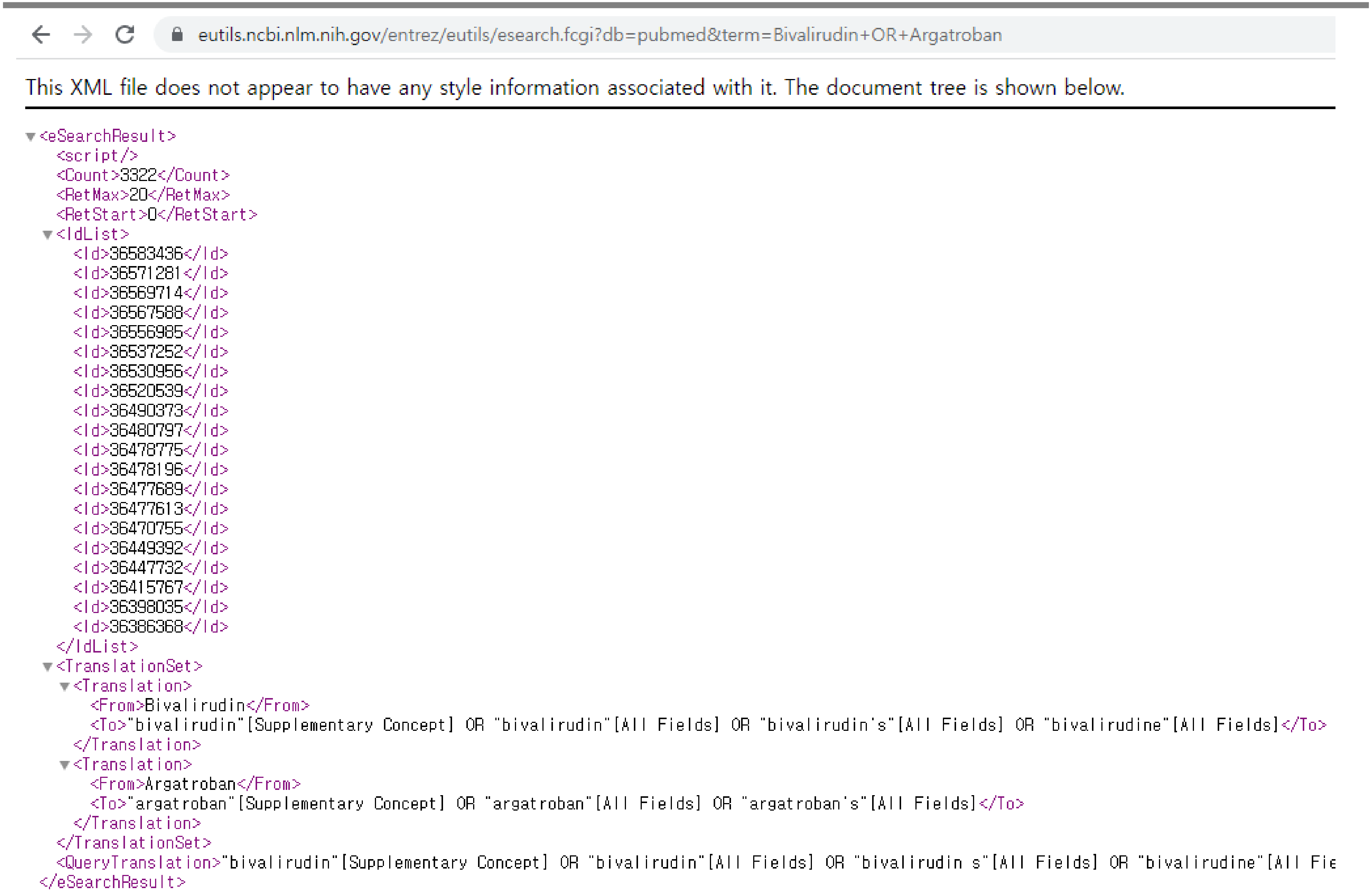
Bethesda (MD): National Center for Biotechnology Information (US); 2010-.

Copyright and Permissions

**Introduction to the E-utilities**

**https://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=Bivalirudin+OR+Argatroban**

# Text Data Collection

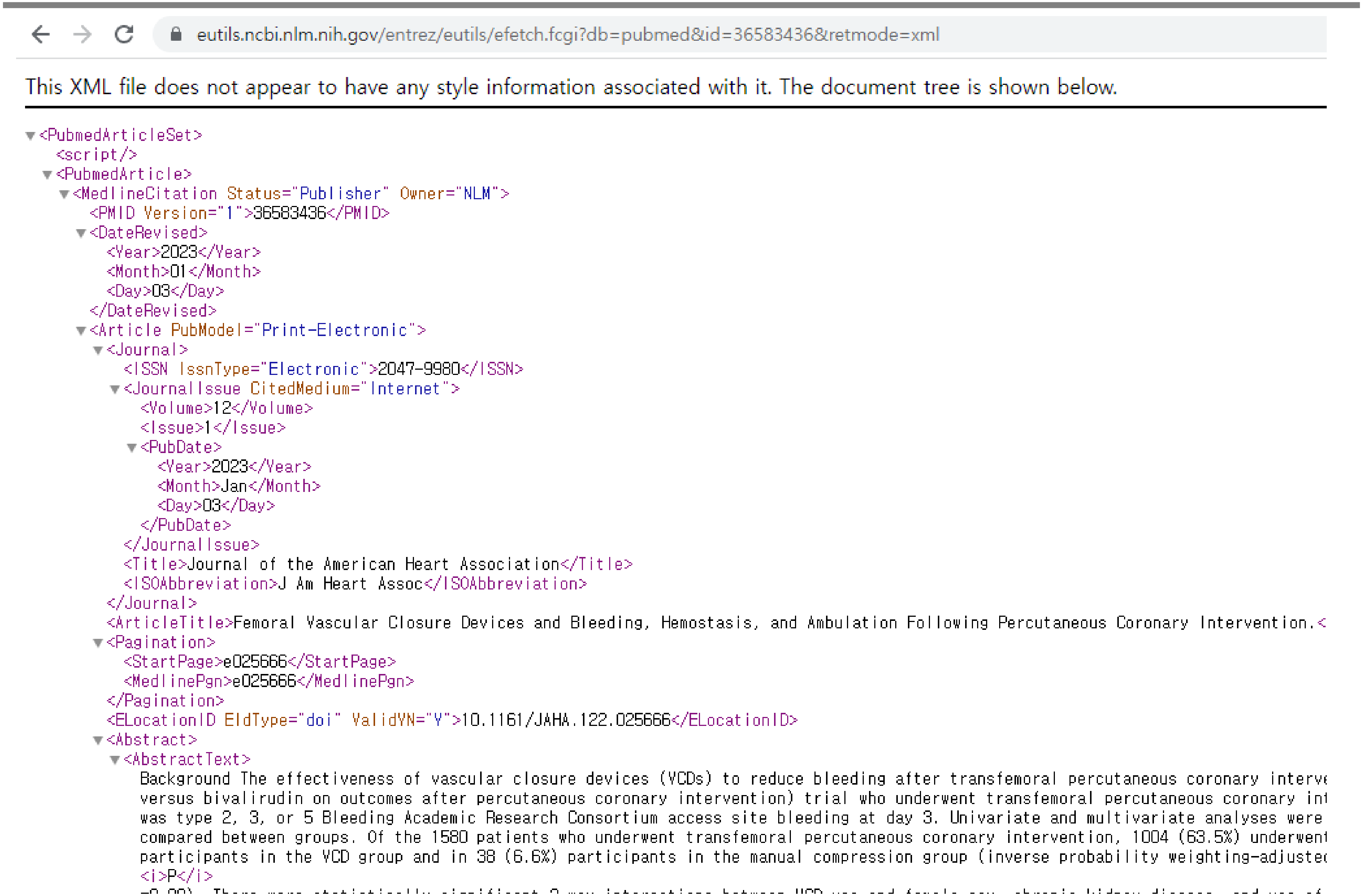


eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?db=pubmed&term=Bivalirudin+OR+Argatroban

This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
▼<eSearchResult>
   <script/>
   <Count>3322</Count>
   <RetMax>20</RetMax>
   <RetStart>0</RetStart>
 ▼<IdList>
     <Id>36583436</Id>
     <Id>36571281</Id>
     <Id>36569714</Id>
     <Id>36567588</Id>
     <Id>36556985</Id>
     <Id>36537252</Id>
     <Id>36530956</Id>
     <Id>36520539</Id>
     <Id>36490373</Id>
     <Id>36480797</Id>
     <Id>36478775</Id>
     <Id>36478196</Id>
     <Id>36477689</Id>
     <Id>36477613</Id>
     <Id>36470755</Id>
     <Id>36449392</Id>
     <Id>36447732</Id>
     <Id>36415767</Id>
     <Id>36398035</Id>
     <Id>36386368</Id>
   </IdList>
 ▼<TranslationSet>
   ▼<Translation>
       <From>Bivalirudin</From>
       <To>"bivalirudin"[Supplementary Concept] OR "bivalirudin"[All Fields] OR "bivalirudin's"[All Fields] OR "bivalirudine"[All Fields]</To>
     </Translation>
   ▼<Translation>
       <From>Argatroban</From>
       <To>"argatroban"[Supplementary Concept] OR "argatroban"[All Fields] OR "argatroban's"[All Fields]</To>
     </Translation>
   </TranslationSet>
   <QueryTranslation>"bivalirudin"[Supplementary Concept] OR "bivalirudin"[All Fields] OR "bivalirudin s"[All Fields] OR "bivalirudine"[All Fie
</eSearchResult>
```

# Text Data Collection



eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=pubmed&id=36583436&retmode=xml

This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
▼<PubmedArticleSet>
   <script/>
 ▼<PubmedArticle>
  ▼<MedlineCitation Status="Publisher" Owner="NLM">
     <PMID Version="1">36583436</PMID>
   ▼<DateRevised>
      <Year>2023</Year>
      <Month>01</Month>
      <Day>03</Day>
     </DateRevised>
   ▼<Article PubModel="Print-Electronic">
    ▼<Journal>
       <ISSN IssnType="Electronic">2047-9980</ISSN>
     ▼<JournalIssue CitedMedium="Internet">
        <Volume>12</Volume>
        <Issue>1</Issue>
      ▼<PubDate>
         <Year>2023</Year>
         <Month>Jan</Month>
         <Day>03</Day>
        </PubDate>
       </JournalIssue>
       <Title>Journal of the American Heart Association</Title>
       <ISOAbbreviation>J Am Heart Assoc</ISOAbbreviation>
      </Journal>
      <ArticleTitle>Femoral Vascular Closure Devices and Bleeding, Hemostasis, and Ambulation Following Percutaneous Coronary Intervention.<
    ▼<Pagination>
       <StartPage>e025666</StartPage>
       <MedlinePgn>e025666</MedlinePgn>
      </Pagination>
      <ELocationID EIdType="doi" ValidYN="Y">10.1161/JAHA.122.025666</ELocationID>
    ▼<Abstract>
     ▼<AbstractText>
        Background The effectiveness of vascular closure devices (VCDs) to reduce bleeding after transfemoral percutaneous coronary interve
        versus bivalirudin on outcomes after percutaneous coronary intervention) trial who underwent transfemoral percutaneous coronary int
        was type 2, 3, or 5 Bleeding Academic Research Consortium access site bleeding at day 3. Univariate and multivariate analyses were
        compared between groups. Of the 1580 patients who underwent transfemoral percutaneous coronary intervention, 1004 (63.5%) underwent
        participants in the VCD group and in 38 (6.6%) participants in the manual compression group (inverse probability weighting-adjuste
        <i>P</i>
```

# Text Preprocessing

## Tokenization

- 문서를 토큰 단위로 분리하는 기법
- 토큰의 단위가 상황에 따라 다르지만, 보통 의미 있는 단위로 토큰을 정의

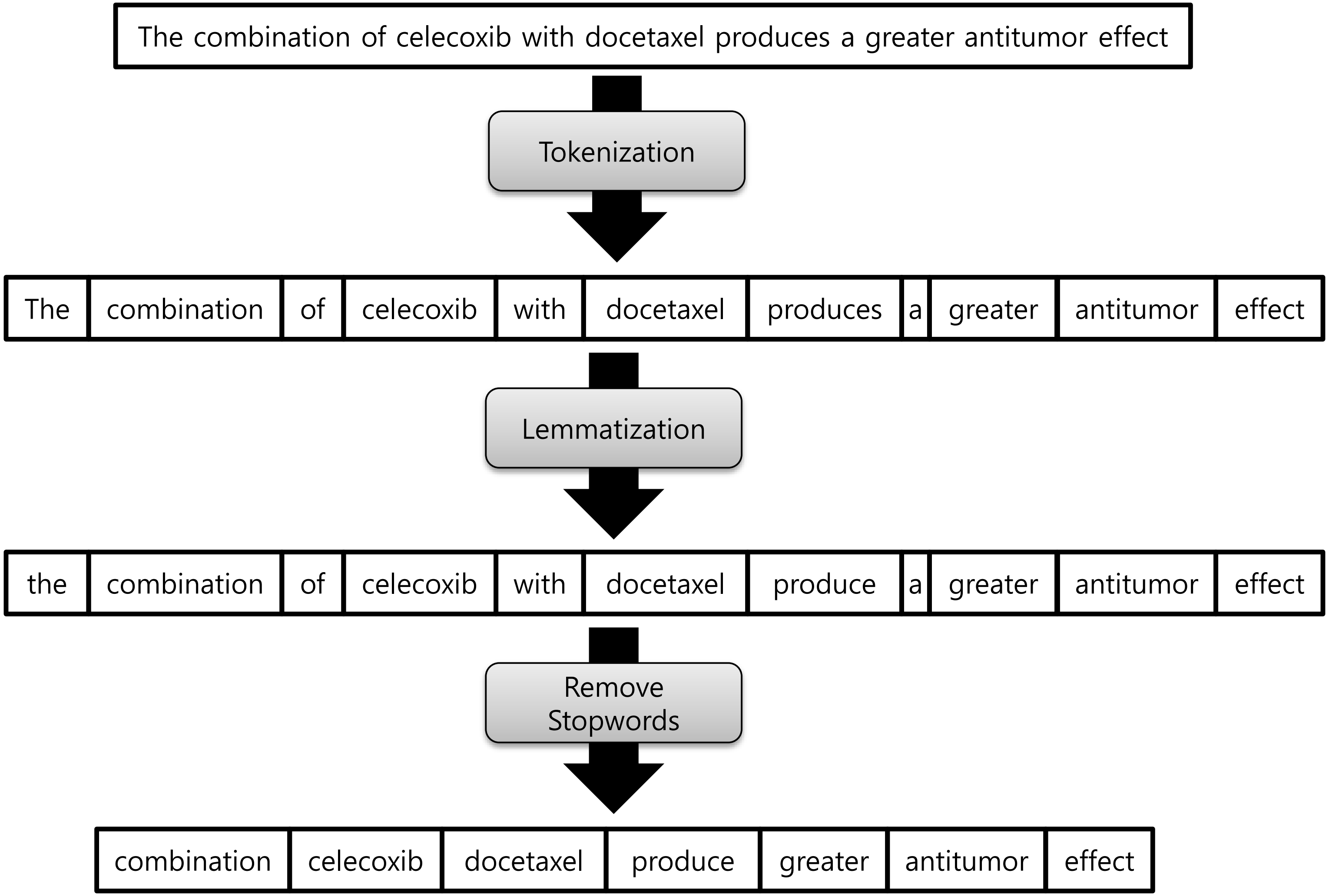
## Lemmatization

- 다양한 형태로 표현되어 있는 단어를 일반형태로 변형하는 기법
- 단어의 의미적인 단위를 고려하고, 형태소 분석을 통해 수행
- Lemmatization을 수행할 경우, 품사 정보가 남아있기 때문에 의미론적 관점에서 더 효과적

## Remove Stopwords

- 자주 등장하지만 분석에 있어 큰 의미가 없는 단어들을 제거하는 과정
- 예를 들면, I, a, the, 조사, 접미사 같은 단어들은 문장에서는 자주 등장하지만 실제 분석에 있어 의미가 거의 없음

# Text Preprocessing



The combination of celecoxib with docetaxel produces a greater antitumor effect

Tokenization

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| The | combination | of | celecoxib | with | docetaxel | produces | a | greater | antitumor | effect |

Lemmatization

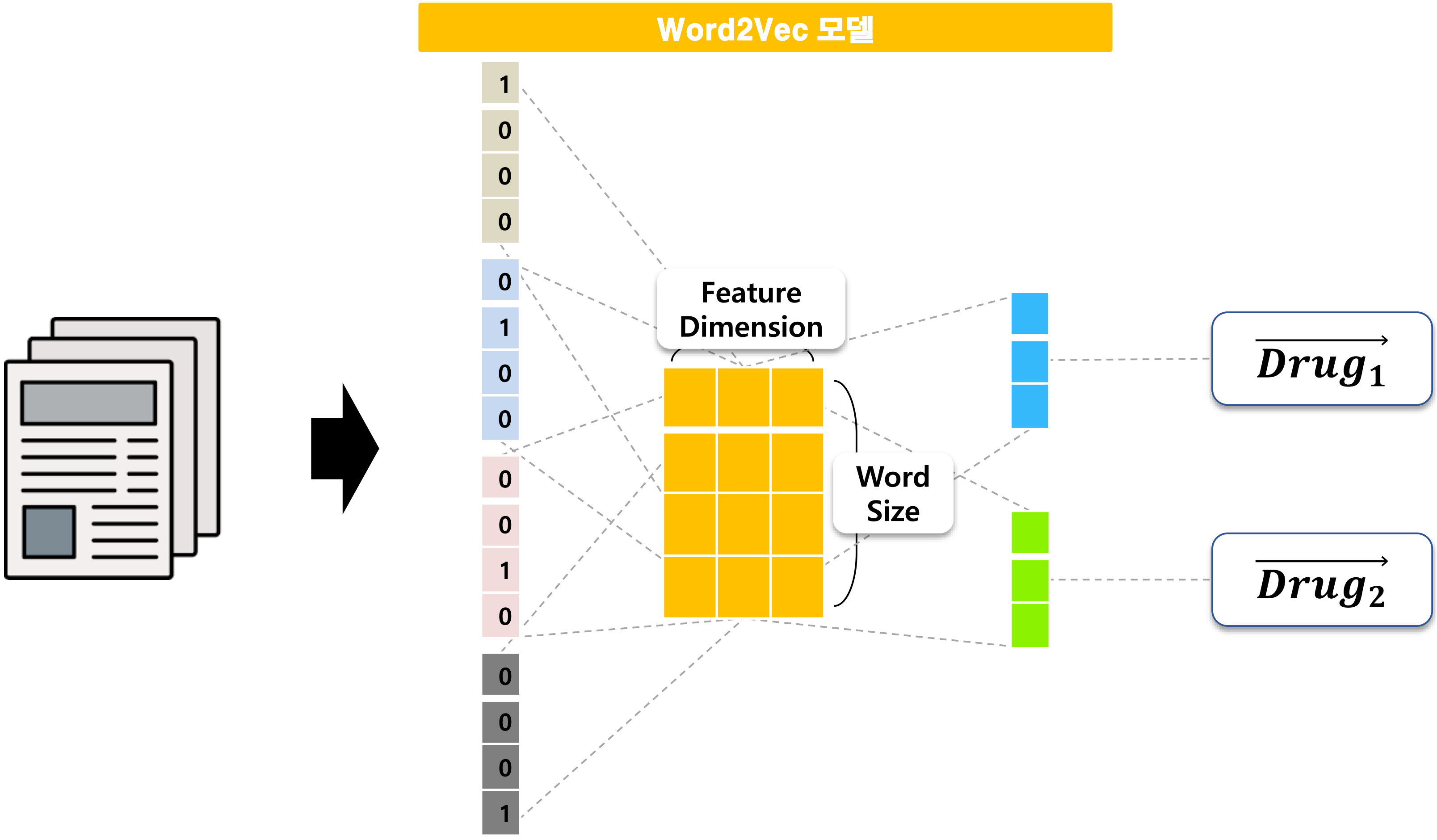
| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| the | combination | of | celecoxib | with | docetaxel | produce | a | greater | antitumor | effect |

Remove Stopwords

| | | | | | | |
|---|---|---|---|---|---|---|
| combination | celecoxib | docetaxel | produce | greater | antitumor | effect |

# Word2Vec 모델 기반 약물 단어 임베딩

## 의생명 문헌 기반 약물 벡터 추출



전처리된 의생명 문헌

One-hot Word Vector

Lookup Table

Word Vector

# Drug Similarity

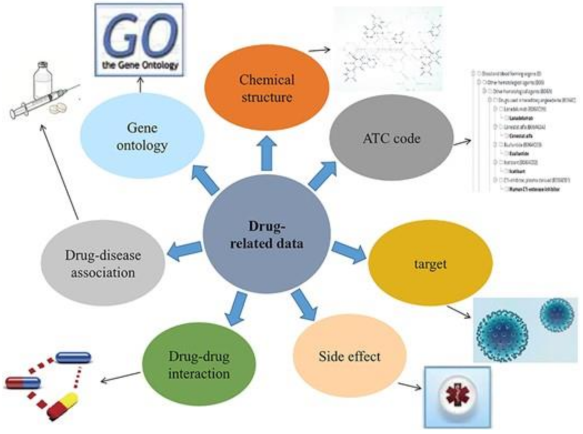
## 약물 유사성 연구의 배경

- 약물 유사성은 서로 다른 약물 간의 유사성을 비교하는 연구
- 약물이 유사한 특성을 가질 수 있다는 가정에 기초하여 많은 연구에서 약물-약물 유사성을 활용하여 잠재적인 약물 관련 정보를 발견

## 약물 조합 연구의 방법

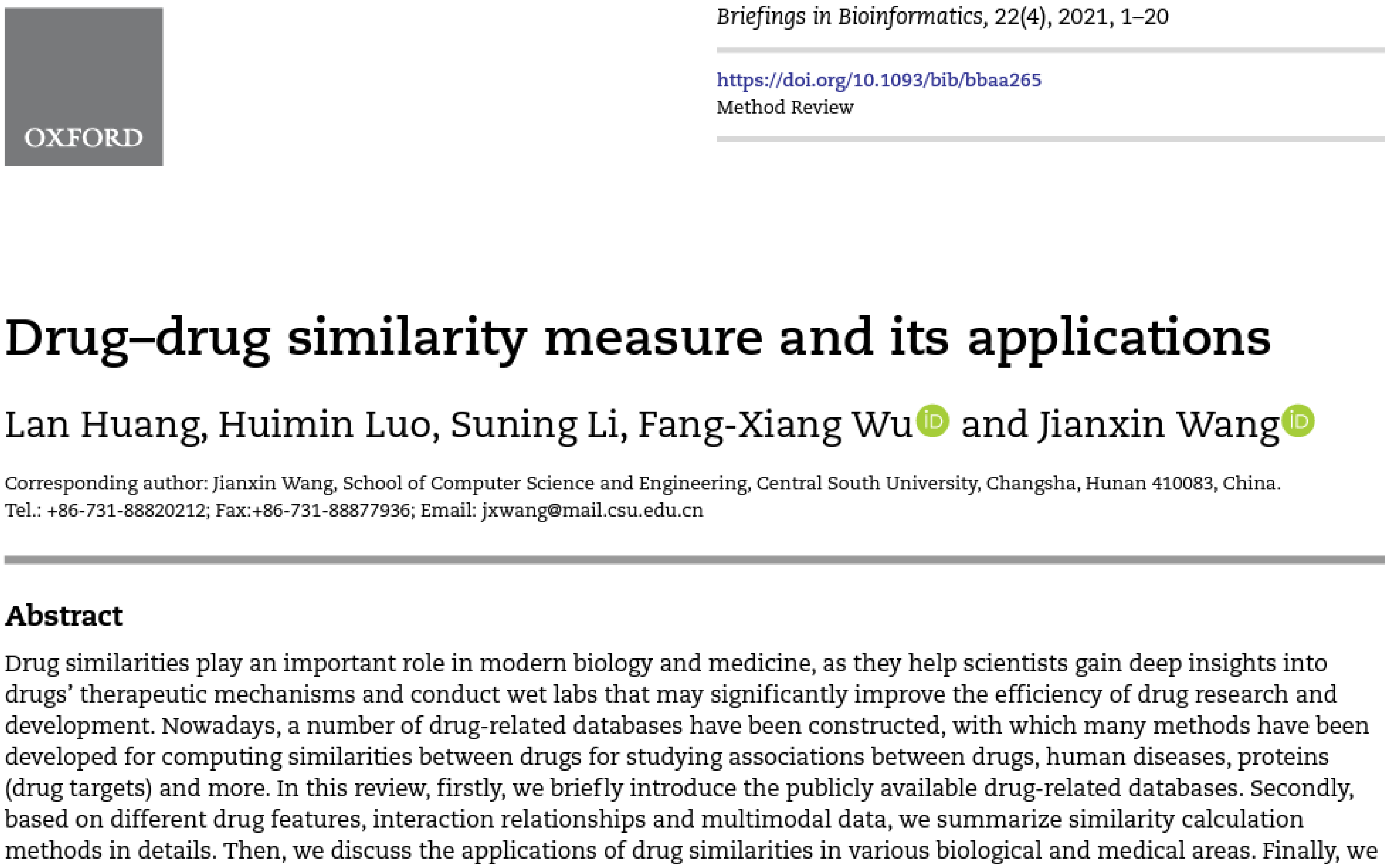
- 다양한 약물 관련 데이터를 기반으로 약물 유사성을 측정
- 서로 다른 유형의 약물 관련 정보를 통합하여 다양한 약물-약물 유사성 측정을 설계할 수 있으며 개발된 약물 유사성은 생물 의학 연구를 개선하는 데 추가로 사용

**약물 관련 데이터 종류**



Huang, Lan, et al. "Drug–drug similarity measure and its applications." Briefings in Bioinformatics 22.4 (2021): bbaa265.

# Drug Similarity review paper



OXFORD

*Briefings in Bioinformatics*, 22(4), 2021, 1–20

https://doi.org/10.1093/bib/bbaa265
Method Review

## Drug–drug similarity measure and its applications

Lan Huang, Huimin Luo, Suning Li, Fang-Xiang Wu and Jianxin Wang

Corresponding author: Jianxin Wang, School of Computer Science and Engineering, Central South University, Changsha, Hunan 410083, China.
Tel.: +86-731-88820212; Fax:+86-731-88877936; Email: jxwang@mail.csu.edu.cn

### Abstract

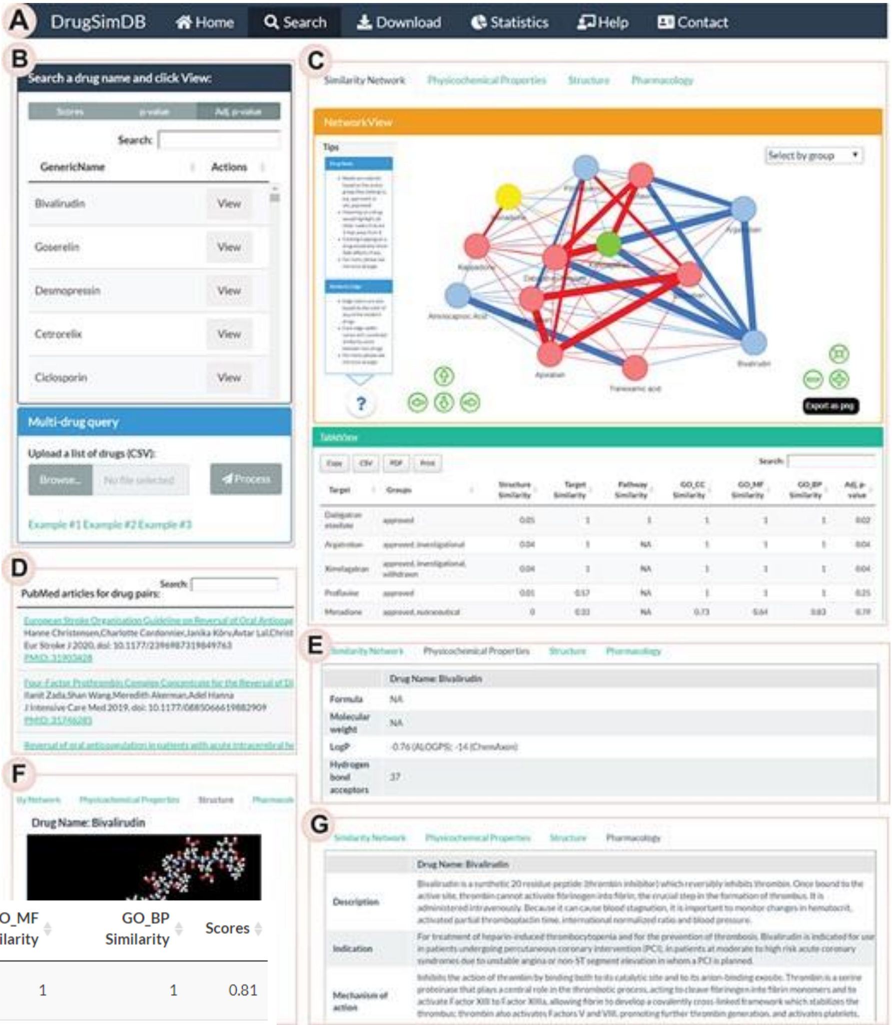
Drug similarities play an important role in modern biology and medicine, as they help scientists gain deep insights into drugs' therapeutic mechanisms and conduct wet labs that may significantly improve the efficiency of drug research and development. Nowadays, a number of drug-related databases have been constructed, with which many methods have been developed for computing similarities between drugs for studying associations between drugs, human diseases, proteins (drug targets) and more. In this review, firstly, we briefly introduce the publicly available drug-related databases. Secondly, based on different drug features, interaction relationships and multimodal data, we summarize similarity calculation methods in details. Then, we discuss the applications of drug similarities in various biological and medical areas. Finally, we

# DrugSimDB

## http://vafaeelab.com/drugSimDB.html

## 스코어 계산 방법

- Structure similarity
  - 화학구조 기반 약물 유사성 계산
- Target Similarity
  - 타겟 단백질 기반 약물 유사성 계산
- Pathway Similarity
  - 패스웨이 기반 약물 유사성 계산
- GO-CC/MF/BP Similarity
  - Gene Ontology Cellular Component
  - Gene Ontology Molecular Function
  - Gene Ontology Biological Process



| GenericName | Actions |
|---|---|
| Bivalirudin | View |

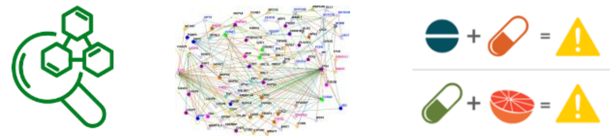
| Target | Groups | Structure Similarity | Target Similarity | Pathway Similarity | GO_CC Similarity | GO_MF Similarity | GO_BP Similarity | Scores |
|---|---|---|---|---|---|---|---|---|
| Argatroban | approved, investigational | 0.04 | 1 | NA | 1 | 1 | 1 | 0.81 |

# 실습 링크

**https://colab.research.google.com/drive/1xwNZgg78ACBiIcYdbykNmIpI6yPlyQdW?usp=share_link**

# *In-silico* 기반 의생명 연구 공통 프로세스

**Problem**



**Process**

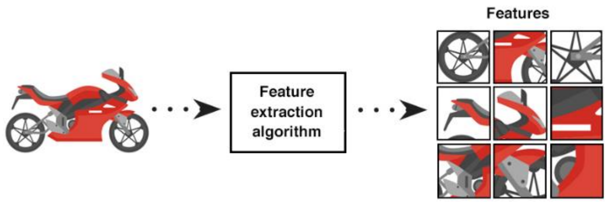
## 1 Data Collection

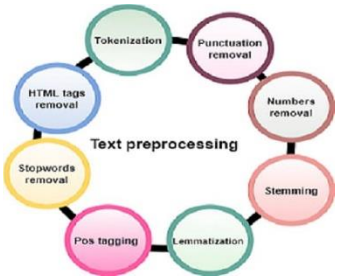
DRUGBANK

KEGG Kyoto Encyclopedia of Genes and Genomes

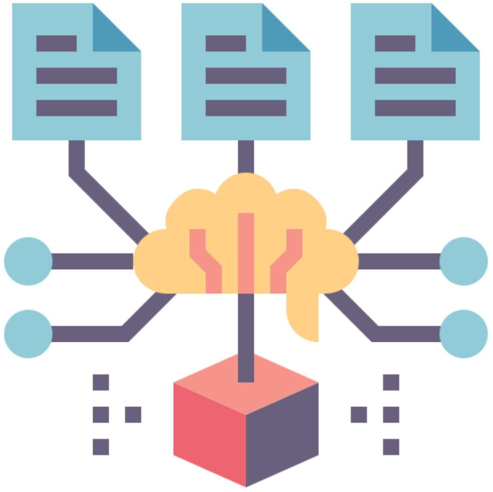
PubMed

## 2 Data Preprocessing





## 3 Applying algorithms



Problem resolve

**Output**

# Q&A