generate_dummy_data

April 27, 2025

[ ]:
```
import random
import json
import os

# Set random seed for reproducibility
random.seed(42)

def generate_legal_corpus(num_documents=120, output_file="legal_dummy_corpus.
 ↪json"):
    """Generate a corpus of dummy legal documents across various legal domains.
 ↪"""

    # Document templates by legal domain
    templates = {
        "contract_law": [
            "The contract between {party1} and {party2} dated {date}␣
 ↪establishes a {duration} agreement for {service}. Consideration of ${amount}␣
 ↪shall be paid according to section {section}. Breach of contract will result␣
 ↪in {remedy}.",
            "AGREEMENT FOR {service}: This contract made on {date} by and␣
 ↪between {party1} ('Client') and {party2} ('Provider') establishes terms for␣
 ↪{service} to be performed over {duration}. Payment of ${amount} is due␣
 ↪{payment_terms}.",
            "EMPLOYMENT CONTRACT: {party1} agrees to employ {party2} as␣
 ↪{position} commencing on {date} for {duration}. Compensation shall be␣
 ↪${amount} per annum with benefits as outlined in Schedule A. Termination␣
 ↪requires {notice} notice.",
            "LEASE AGREEMENT: {party1} ('Landlord') hereby leases to {party2}␣
 ↪('Tenant') the property at {address} for {duration} beginning {date}.␣
 ↪Monthly rent is ${amount} due on the {day} of each month.",
            "NON-DISCLOSURE AGREEMENT: {party1} and {party2} agree to maintain␣
 ↪confidentiality regarding {subject} for a period of {duration} from {date}.␣
 ↪Breach will result in {remedy}."
        ],
        "tort_law": [
```

```
            "In the matter of {plaintiff} v. {defendant}, plaintiff alleges␣
↪{tort_type} resulting in {damages}. The incident occurred on {date} at␣
↪{location} when defendant allegedly {action}.",
            "COMPLAINT: {plaintiff} brings this action against {defendant} for␣
↪{tort_type} arising from events on {date}. Plaintiff suffered {damages}␣
↪estimated at ${amount} and seeks compensatory and punitive damages.",
            "The tort claim filed by {plaintiff} asserts that {defendant}'s␣
↪{action} constituted {tort_type} and breached the duty of care owed to␣
↪plaintiff, causing {damages} and economic losses of approximately ${amount}.
↪",
            "NEGLIGENCE CLAIM: {plaintiff} alleges that on {date}, {defendant}␣
↪failed to {duty} resulting in {damages}. The standard of care required␣
↪defendant to {standard}, which was breached by {action}.",
            "PRODUCT LIABILITY: {plaintiff} brings suit against {defendant}␣
↪alleging the {product} manufactured by defendant was defective and caused␣
↪{damages} on {date}. Plaintiff seeks ${amount} in damages."
        ],
        "criminal_law": [
            "STATE v. {defendant}: Defendant is charged with {crime} in␣
↪violation of Penal Code §{section}. The alleged offense occurred on {date}␣
↪at {location} when defendant allegedly {action}.",
            "INDICTMENT: The Grand Jury charges that on {date}, defendant␣
↪{defendant} did knowingly and intentionally {action}, constituting the␣
↪offense of {crime} under §{section}, punishable by up to {penalty}.",
            "CRIMINAL COMPLAINT: Officer {officer} attests that on {date},␣
↪{defendant} was observed {action} at {location}, constituting probable cause␣
↪for arrest on charges of {crime}.",
            "SEARCH WARRANT application states that probable cause exists to␣
↪believe evidence of {crime} will be found at {location} based on {evidence}␣
↪observed by Officer {officer} on {date}.",
            "PLEA AGREEMENT: Defendant {defendant}, charged with {crime},␣
↪agrees to plead guilty to {lesser_charge} in exchange for a recommended␣
↪sentence of {sentence}. Defendant waives right to trial."
        ],
        "property_law": [
            "DEED OF TRUST: {party1} ('Grantor') transfers to {party2}␣
↪('Trustee') the property at {address} to secure payment to {party3}␣
↪('Beneficiary') of ${amount} as evidenced by Promissory Note dated {date}.",
            "WARRANTY DEED: {party1} ('Grantor') conveys and warrants to␣
↪{party2} ('Grantee') the real property at {address}, together with all␣
↪improvements thereon, for consideration of ${amount} paid on {date}.",
            "EASEMENT: {party1} grants to {party2} a perpetual easement for␣
↪{purpose} over the property described as {description}. This easement shall␣
↪run with the land and bind all future owners.",
```

```
            "TITLE OPINION: Based on examination of records from {date1} to␣
↪{date2}, title to property at {address} is vested in {owner} subject to␣
↪{exceptions}.",
            "PROPERTY DISPUTE: {party1} claims adverse possession of␣
↪{description} against record owner {party2}, having maintained open,␣
↪notorious, and continuous possession for {duration} years."
        ],
        "constitutional_law": [
            "In {case}, the Court held that {right} protected under the␣
↪{amendment} Amendment was {ruling}. The majority opinion by Justice␣
↪{justice} established that {principle}.",
            "The constitutional challenge in {case} asserts that {law} violates␣
↪the {amendment} Amendment by {violation}. Petitioner seeks {remedy} based on␣
↪precedent established in {precedent}.",
            "AMICUS BRIEF: {organization} submits that interpretation of the␣
↪{amendment} Amendment in {case} should consider {principle}, as previously␣
↪recognized in {precedent}.",
            "DISSENTING OPINION: Justice {justice} argues that the majority's␣
↪interpretation of {amendment} Amendment in {case} fails to account for␣
↪{principle} and would lead to {consequence}.",
            "CONSTITUTIONAL ANALYSIS: The {law} must be subjected to {standard}␣
↪scrutiny under the {amendment} Amendment because it implicates {right}, a␣
↪fundamental right recognized in {precedent}."
        ],
        "administrative_law": [
            "REGULATORY FILING: {agency} proposes new rule §{section} regarding␣
↪{subject}. Public comments must be submitted by {date}. The proposed rule␣
↪would require {requirement}.",
            "ADMINISTRATIVE APPEAL: {party} contests {agency}'s determination␣
↪dated {date} regarding {subject}. Appellant argues that the agency {error}␣
↪and exceeded its statutory authority under {statute}.",
            "AGENCY DECISION: {agency} hereby approves/denies {party}'s␣
↪application for {permit} based on findings that {findings}. This decision is␣
↪appealable within {timeframe} days.",
            "NOTICE OF VIOLATION: {agency} finds that {party} violated␣
↪§{section} of {regulation} by {violation} on {date}. Proposed penalty is␣
↪${amount} unless remedied within {timeframe}.",
            "FREEDOM OF INFORMATION REQUEST: {party} requests all documents␣
↪held by {agency} relating to {subject} from {date1} to {date2} pursuant to 5␣
↪U.S.C. §552."
        ]
    }

    # Variables to fill in templates
    variables = {
```

```
      "party1": ["Smith Corp.", "Johnson LLC", "Acme Industries", "Blackstone␣
↪Properties", "FirstBank N.A.",
                 "TechSolutions Inc.", "Global Shipping Co.", "Jones␣
↪Manufacturing", "Metropolis City",
                 "State of Franklin", "United Insurance", "Green Energy LLC",␣
↪"MediCorp", "TransAmerica Logistics"],

      "party2": ["Doe Enterprises", "Miller Associates", "Omega Contractors",␣
↪"Redstone Investments", "SecondTrust",
                "InnovateTech Corp.", "Atlantic Freight Lines", "Brown␣
↪Fabricators", "Capital County",
                "Commonwealth of Jefferson", "National Assurance", "Blue␣
↪Power Inc.", "HealthSystems", "PacificRoute Services"],

      "party3": ["Secure Lending", "Guardian Trust", "Heritage Bank",␣
↪"Fidelity Investments", "Capital Finance",
                "Prosperity Funding", "Eagle Trust Company", "Cornerstone␣
↪Credit", "Municipal Bond Authority"],

      "date": ["January 15, 2022", "March 3, 2023", "November 12, 2021",␣
↪"July 8, 2022", "February 28, 2023",
               "April 17, 2022", "October 5, 2021", "June 23, 2023",␣
↪"September 9, 2022", "May 1, 2023",
               "December 11, 2021", "August 30, 2022"],

      "date1": ["January 1, 2020", "March 15, 2018", "November 30, 2019",␣
↪"July 1, 2017", "February 28, 2021"],

      "date2": ["December 31, 2022", "March 14, 2023", "October 31, 2022",␣
↪"June 30, 2022", "January 31, 2023"],

      "duration": ["one year", "three years", "five years", "ten years", "six␣
↪months", "two years", "indefinite"],

      "service": ["consulting services", "software development",␣
↪"construction", "financial advisory",
                "legal representation", "marketing services", "equipment␣
↪maintenance", "cloud hosting",
                "professional training", "security services", "product␣
↪distribution", "waste management"],

      "amount": ["10,000", "25,000", "50,000", "100,000", "250,000",␣
↪"500,000", "1,000,000", "75,000",
                "125,000", "300,000", "450,000", "850,000", "2,500,000"],
```

```
      "section": ["3.2", "7.1", "4.5", "10.3", "2.4", "5.7", "8.1", "6.3", "9.
↪2"],

      "remedy": ["liquidated damages", "specific performance", "injunctive␣
↪relief", "termination of contract",
                 "obligation to pay legal fees", "forfeiture of deposit",␣
↪"arbitration", "mediation"],

      "payment_terms": ["monthly", "quarterly", "upon completion", "net 30␣
↪days", "in advance",
                        "50% upon signing, 50% upon completion", "according to␣
↪project milestones"],

      "position": ["Chief Executive Officer", "General Counsel", "Sales␣
↪Director", "Chief Financial Officer",
                   "Operations Manager", "Marketing Specialist", "Software␣
↪Engineer", "Human Resources Director",
                   "Regional Manager", "Project Coordinator", "Research␣
↪Scientist", "Administrative Assistant"],

      "notice": ["30 days", "60 days", "90 days", "two weeks", "one month",␣
↪"immediate under certain circumstances"],

      "address": ["123 Main Street, Anytown, ST 12345", "456 Oak Avenue,␣
↪Metropolis, ST 67890",
                  "789 Pine Road, Capital City, ST 23456", "101 River Lane,␣
↪Westville, ST 34567",
                  "202 Mountain View Drive, Eastport, ST 45678", "303 Sunset␣
↪Boulevard, Northfield, ST 56789"],

      "day": ["1st", "5th", "10th", "15th", "20th", "25th", "last"],

      "subject": ["proprietary technology", "customer lists", "financial␣
↪projections", "merger negotiations",
                  "product formulations", "business strategies", "research␣
↪findings", "personnel information",
                  "acquisition targets", "regulatory compliance", "clinical␣
↪trial results"],

      "plaintiff": ["John Smith", "Jane Doe", "Robert Johnson", "Sarah␣
↪Williams", "Acme Corporation",
                    "Global Enterprises", "Mary Thompson", "James Wilson",␣
↪"Patricia Davis", "Michael Brown"],

      "defendant": ["XYZ Company", "David Miller", "Omega Corporation",␣
↪"Thomas Anderson", "City of Metropolis",
```

```
                    "Jennifer Lee", "United Industries", "Central Hospital",␣
↪"Richard Taylor", "Elizabeth Martin"],

      "tort_type": ["negligence", "defamation", "product liability", "medical␣
↪malpractice", "trespass",
                    "false imprisonment", "intentional infliction of emotional␣
↪distress", "battery",
                    "invasion of privacy", "conversion", "nuisance", "fraud"],

      "damages": ["severe physical injuries", "emotional distress", "property␣
↪damage", "loss of income",
                  "damage to reputation", "wrongful death", "permanent␣
↪disability", "medical expenses",
                  "pain and suffering", "loss of consortium", "diminished␣
↪quality of life"],

      "location": ["1234 Elm Street", "Central Park", "Westside Shopping␣
↪Mall", "Highway 101",
                   "Downtown Financial District", "Midtown Medical Center",␣
↪"Eastside Industrial Park",
                   "County Courthouse", "Northern University Campus",␣
↪"Southport Marina"],

      "action": ["failed to maintain safe premises", "published false␣
↪statements", "manufactured a defective product",
                 "ignored industry safety standards", "misrepresented material␣
↪facts", "operated a vehicle negligently",
                 "breached doctor-patient confidentiality", "trespassed on␣
↪private property",
                 "inappropriately accessed confidential information", "failed␣
↪to obtain informed consent"],

      "crime": ["aggravated assault", "wire fraud", "possession with intent␣
↪to distribute", "armed robbery",
                "embezzlement", "identity theft", "insider trading", "criminal␣
↪negligence", "bribery",
                 "tax evasion", "racketeering", "money laundering",␣
↪"cybercrime", "perjury"],

      "officer": ["J. Martinez", "S. Johnson", "D. Williams", "L. Thompson",␣
↪"M. Garcia", "K. Davis",
                  "R. Rodriguez", "C. Wilson", "B. Anderson", "T. Thomas", "H.␣
↪Jackson", "N. White"],

      "evidence": ["surveillance footage", "witness statements", "financial␣
↪records", "digital communications",
```

```
                    "physical evidence at the scene", "contraband in plain␣
↪view", "confidential informant information",
                    "bank transactions", "cell phone location data", "DNA␣
↪analysis"],

       "penalty": ["5 years imprisonment", "10 years imprisonment", "20 years␣
↪imprisonment",
                   "$250,000 fine", "$500,000 fine", "$1,000,000 fine",␣
↪"combination of imprisonment and fine",
                   "supervised release", "life imprisonment", "restitution to␣
↪victims"],

       "lesser_charge": ["misdemeanor assault", "simple possession",␣
↪"attempted fraud", "petty theft",
                         "obstruction", "disorderly conduct", "criminal␣
↪mischief", "reckless endangerment"],

       "sentence": ["2 years probation", "6 months imprisonment", "1 year␣
↪imprisonment followed by 3 years supervised release",
                    "time served plus community service", "deferred␣
↪adjudication", "weekend confinement",
                    "electronic monitoring for 90 days", "substance abuse␣
↪treatment program"],

       "description": ["the north 50 feet of Lot 7, Block 3", "Parcel 126-A as␣
↪recorded in Plat Book 7, Page 15",
                       "the southern boundary of Whispering Pines subdivision",␣
↪"a 20-foot wide strip along the western edge of the property",
                       "the access road leading from Highway 7 to the lake",␣
↪"all mineral rights beneath Section 23, Township 4N, Range 5W"],

       "purpose": ["ingress and egress", "utility maintenance", "drainage",␣
↪"construction access",
                   "conservation", "recreational use", "pipeline installation",␣
↪"agricultural access"],

       "owner": ["James and Mary Wilson", "Sunset Properties LLC", "Robert T.␣
↪Johnson Trust", "Green Acres Development Corp.",
                 "The Estate of Eleanor Smith", "Mountain View Homeowners␣
↪Association", "Riverfront Holdings Inc."],

       "exceptions": ["mortgage recorded in Book 456, Page 789", "utility␣
↪easement along the southern boundary",
                      "mineral rights reserved by previous owner",␣
↪"right-of-way for public road",
```

```
                    "tax lien in the amount of $5,240", "pending legal action␣
↪regarding boundary dispute",
                    "building code violations noted in city inspection␣
↪report"],

      "case": ["Smith v. Jones (2021)", "United States v. Thompson (2020)",␣
↪"In re Wilson Estate (2022)",
               "State v. Miller (2019)", "Citizens Group v. City of Metropolis␣
↪(2023)",
               "Department of Commerce v. Technology Corp. (2021)", "Johnson v.
↪ Board of Education (2022)"],

      "right": ["free speech", "equal protection", "due process", "privacy",␣
↪"religious freedom",
                "freedom from unreasonable search", "right to counsel", "right␣
↪to bear arms",
                "protection against self-incrimination", "right to jury␣
↪trial"],

      "amendment": ["First", "Fourth", "Fifth", "Sixth", "Eighth",␣
↪"Fourteenth", "Second", "Tenth"],

      "ruling": ["not violated by the challenged statute", "infringed by the␣
↪government action",
                 "subject to strict scrutiny", "protected even in the context␣
↪of", "limited in cases involving",
                 "balanced against compelling government interests", "broadly␣
↪interpreted to include"],

      "justice": ["Roberts", "Thomas", "Alito", "Sotomayor", "Kagan",␣
↪"Gorsuch", "Kavanaugh", "Barrett", "Jackson"],

      "principle": ["original understanding of the Constitution", "evolving␣
↪standards of decency",
                    "separation of powers", "federalism", "judicial␣
↪restraint", "stare decisis",
                    "proportionality", "equal dignity", "fundamental rights␣
↪analysis", "textualism"],

      "law": ["Senate Bill 247", "House Resolution 103", "Executive Order␣
↪14-32",
              "City Ordinance 2023-7", "State Tax Code §473.2", "Public Law␣
↪117-25",
              "Administrative Rule 42 CFR §438.6", "Local Zoning Regulation␣
↪§12-5-3"],
```

```
      "violation": ["imposing an undue burden", "creating a content-based␣
↪restriction",
                    "discriminating on the basis of protected status",␣
↪"exceeding congressional authority",
                    "infringing on state sovereignty", "constituting an␣
↪establishment of religion",
                    "denying procedural safeguards", "acting without rational␣
↪basis"],

      "remedy": ["declaratory relief", "permanent injunction", "class␣
↪certification",
                 "attorney's fees under 42 U.S.C. §1988", "compensatory␣
↪damages",
                 "prospective application only", "invalidation of the␣
↪statute", "remand with instructions"],

      "precedent": ["Marbury v. Madison", "Brown v. Board of Education", "Roe␣
↪v. Wade", "Miranda v. Arizona",
                    "District of Columbia v. Heller", "Obergefell v. Hodges",␣
↪"Citizens United v. FEC",
                    "Bostock v. Clayton County", "West Virginia v. EPA"],

      "organization": ["American Civil Liberties Union", "Chamber of␣
↪Commerce", "State Attorneys General",
                       "Constitutional Scholars", "Institute for Justice",␣
↪"Environmental Defense Fund",
                       "National Association of Manufacturers", "Legal Aid␣
↪Society", "Religious Freedom Institute"],

      "consequence": ["chilling protected speech", "expanding executive power␣
↪beyond constitutional limits",
                      "undermining precedent established for decades",␣
↪"leaving vulnerable populations unprotected",
                      "creating regulatory uncertainty", "blurring the line␣
↪between church and state",
                      "imposing unfair burdens on small businesses",␣
↪"violating principles of federalism"],

      "standard": ["strict", "intermediate", "rational basis", "exacting",␣
↪"heightened", "de novo"],

      "agency": ["Environmental Protection Agency", "Securities and Exchange␣
↪Commission", "Department of Labor",
                 "Federal Communications Commission", "Food and Drug␣
↪Administration", "Consumer Financial Protection Bureau",
```

```
                "Internal Revenue Service", "Department of Health and Human␣
↪Services", "Federal Trade Commission"],

      "subject": ["greenhouse gas emissions", "financial disclosures",␣
↪"workplace safety standards",
                  "broadband internet regulation", "pharmaceutical approval␣
↪process", "consumer lending practices",
                  "tax exemption requirements", "healthcare privacy rules",␣
↪"merger review procedures"],

      "requirement": ["quarterly reporting of compliance metrics",␣
↪"implementation of safety protocols",
                      "disclosure of financial interests", "obtaining prior␣
↪authorization",
                      "maintaining records for a period of five years",␣
↪"employee training on new procedures",
                      "installation of monitoring equipment", "submission of␣
↪annual certification"],

      "error": ["misinterpreted the statutory language", "failed to consider␣
↪relevant factors",
                "applied the wrong legal standard", "reached a conclusion␣
↪unsupported by substantial evidence",
                "denied procedural due process", "acted arbitrarily and␣
↪capriciously",
                "failed to provide adequate notice", "exceeded statutory␣
↪authority"],

      "statute": ["Administrative Procedure Act", "Clean Air Act",␣
↪"Securities Exchange Act",
                  "Communications Act", "Federal Food, Drug, and Cosmetic␣
↪Act", "Dodd-Frank Act",
                  "Internal Revenue Code", "Social Security Act", "Sherman␣
↪Antitrust Act"],

      "findings": ["the application did not meet regulatory requirements",␣
↪"the proposal would have significant environmental impacts",
                   "the party demonstrated financial responsibility", "public␣
↪interest considerations warranted approval",
                   "safety concerns could not be adequately addressed", "all␣
↪statutory prerequisites were satisfied"],

      "timeframe": ["30", "60", "90", "120", "15", "45", "180"],

      "permit": ["building expansion", "discharge permit", "operating␣
↪license", "broadcasting rights",
```

```
                "drug marketing approval", "financial services license",␣
↪"tax-exempt status",
                "transportation authority", "professional certification"],

      "violation": ["exceeding emission limits", "failing to file required␣
↪reports", "operating without proper license",
                    "misrepresenting material information", "failing to␣
↪implement required safeguards",
                    "marketing unapproved products", "improper handling of␣
↪sensitive data",
                    "non-compliance with accessibility requirements"]
  }

  # Generate documents
  documents = []
  doc_ids = []

  domains = list(templates.keys())
  doc_count = 0

  while doc_count < num_documents:
      # Select domain and template
      domain = random.choice(domains)
      template = random.choice(templates[domain])

      # Fill template with random variables
      doc_text = template
      for var in variables:
          if "{" + var + "}" in doc_text:
              value = random.choice(variables[var])
              doc_text = doc_text.replace("{" + var + "}", value)

      # Add document to corpus
      documents.append(doc_text)
      doc_ids.append(f"{domain}_{doc_count:03d}")
      doc_count += 1

  # Create corpus dictionary
  corpus = {
      "documents": documents,
      "doc_ids": doc_ids,
      "metadata": {
          "domains": domains,
          "count": len(documents),
          "generation_date": "2023-04-21"
      }
  }
```

```
    # Save to file
    with open(output_file, 'w') as f:
        json.dump(corpus, f, indent=2)

    print(f"Generated {len(documents)} legal documents across {len(domains)}␣
 ↪domains.")
    print(f"Saved to {output_file}")

    return corpus

# # Generate the corpus
# corpus = generate_legal_corpus(120, "legal_dummy_corpus.json")

# # Also create some sample queries
# sample_queries = [
#     "What are the essential elements of a valid contract?",
#     "How is negligence defined in tort law?",
#     "What constitutes probable cause for a search warrant?",
#     "What rights are protected under the First Amendment?",
#     "How does adverse possession work in property law?",
#     "What is the standard for proving defamation?",
#     "What are the remedies for breach of contract?",
#     "How does the Fourth Amendment limit police searches?",
#     "What is the process for appealing an administrative decision?",
#     "What constitutes insider trading under securities regulations?",
#     "How are easements created and terminated?",
#     "What is the difference between murder and manslaughter?",
#     "What are the requirements for a valid will?",
#     "How does eminent domain work?",
#     "What constitutes workplace discrimination?"
# ]

# # Save sample queries
# with open("legal_sample_queries.json", "w") as f:
#     json.dump(sample_queries, f, indent=2)

# print(f"Saved {len(sample_queries)} sample queries to legal_sample_queries.
 ↪json")
```

```
Generated 120 legal documents across 6 domains.
Saved to legal_dummy_corpus.json
Saved 15 sample queries to legal_sample_queries.json
```

[ ]: