

数值线性代数大作业：潜在语义分析与截断 SVD

1. 学习目标与阅读材料

完成本作业后，你能：

- 1. 将一组具有交叉术语的科学短摘要转化为文档-词项 TF-IDF 矩阵；
- 2. 用截断 SVD 分析重构误差、累计能量和主题检索之间的差异；
- 3. 从右奇异向量解释潜在方向，并识别过度平滑和跨主题桥接文档；
- 4. 将新查询一致地折叠到潜在空间，并用检索结果解释秩选择。

先复习第 2 章 SVD（幻灯片 2.42—2.49）与最佳低秩近似（幻灯片 2.50）；潜在语义分析示例见幻灯片 2.64—2.73。

2. 数学问题与算法接口

本题用 12 篇科学短摘要构造 TF-IDF 矩阵，通过截断 SVD 进行潜在语义分析、查询折叠和秩选择。数据和 token 序列均已给定。

2.1 固定语料与 token

原始短摘要

以下文本用于说明语境。各小节的主题标签只用于后续定量评价，不参与矩阵构造。

多孔介质流（D1—D4）

D1

In a heterogeneous porous reservoir, Darcy flow links permeability to pressure and mass conservation. A finite-volume discretization on the mesh produces a sparse system, for which multigrid preconditioning supports a parallel solve.

D2

Two-phase porous flow couples pressure, saturation, and transport in a reservoir. An implicit discretization on the mesh uses Newton-Krylov iteration and a preconditioner to solve the mass-conservative sparse system.

D3

Reactive transport in a porous reservoir combines flow, diffusion, chemical reaction, and mass conservation. Domain decomposition partitions the mesh, while a physics-based preconditioner accelerates the parallel linear solver for the coupled system.

D4

A poroelastic reservoir couples Darcy pressure to solid displacement and stress. The monolithic block system uses an approximate Schur-complement preconditioner to obtain robust Krylov convergence in parallel.

流固耦合（D5--D8）

D5

Fluid-structure coupling transfers incompressible pressure and velocity to an elastic wall. A partitioned interface iteration preserves stability against added-mass effects while enforcing displacement and traction continuity in the coupled system.

D6

A monolithic fluid-structure formulation solves pressure, velocity, and solid displacement in one coupled system across the interface. Block factorization with a Schur-complement preconditioner accelerates parallel GMRES.

D7

An immersed flexible membrane represents a structure on a fixed fluid mesh. In this coupled system, the fluid-structure coupling uses conservative interface-force transfer to limit artificial energy growth and preserve numerical stability during large displacement.

D8

Vortex-induced motion couples fluid flow to an elastic structure. The resulting system uses a stable interface treatment, adaptive mesh refinement, and a parallel linear solver to predict the displacement over long time intervals.

高性能迭代求解（D9--D12）

D9

A GMRES solver treats the nonsymmetric sparse system arising from an implicit multiphysics discretization. Domain-decomposition preconditioning improves Krylov convergence and residual reduction, but global orthogonalization and communication limit strong parallel scaling.

D10

A conjugate-gradient Krylov solver with multigrid preconditioning treats pressure and diffusion systems on a distributed mesh. Sparse matrix operations, residual convergence, and communication determine parallel scaling.

D11

A field-split Krylov method uses block preconditioners for coupled pressure, velocity, and displacement. Approximate Schur complements reuse porous-flow and fluid-structure solvers, making the monolithic system scalable in parallel.

A communication-avoiding Krylov solver with a preconditioner reduces synchronization in a parallel sparse system. Fewer reductions improve scaling per iteration, but loss of orthogonality can damage residual convergence and numerical stability.

规范化 token 序列

矩阵统一由下面的 token 序列构造，避免分词、词形还原和复合词处理的差异。原始短摘要只用于语义解释。

`added_mass` 和 `conjugate_gradient` 各视为一个词项。

```
documents_tokens = [
    # D1
    ["porous", "flow", "darcy", "pressure", "permeability",
     "heterogeneous", "reservoir", "mass", "discretization", "mesh",
     "multigrid", "preconditioner", "parallel", "sparse", "system"],

    # D2
    ["porous", "flow", "pressure", "saturation", "transport",
     "reservoir", "mass", "discretization", "mesh", "newton",
     "krylov", "preconditioner", "sparse", "system"],

    # D3
    ["porous", "flow", "transport", "diffusion", "reaction",
     "reservoir", "mass", "mesh", "domain", "decomposition",
     "preconditioner", "parallel", "solver", "system"],

    # D4
    ["porous", "poroelastic", "pressure", "displacement", "stress",
     "coupling", "monolithic", "block", "schur", "preconditioner",
     "krylov", "parallel", "reservoir", "system"],

    # D5
    ["fluid", "structure", "coupling", "pressure", "elastic", "wall",
     "displacement", "interface", "velocity", "traction", "partitioned",
     "added_mass", "stability", "iteration", "system"],

    # D6
    ["fluid", "structure", "coupling", "pressure", "velocity",
     "displacement", "interface", "monolithic", "block", "schur",
     "preconditioner", "gmres", "parallel", "system"],

    # D7
    ["fluid", "structure", "coupling", "flexible", "membrane",
     "displacement", "interface", "force", "mesh", "immersed",
     "conservative", "stability", "energy", "system"],

    # D8
```

```

["fluid", "structure", "coupling", "flow", "elastic",
 "displacement", "interface", "mesh", "adaptive", "stability",
 "parallel", "solver", "time", "system"],

# D9
["krylov", "gmres", "sparse", "system", "preconditioner", "domain",
 "decomposition", "convergence", "parallel", "scaling",
 "orthogonalization", "communication", "residual", "solver"],

# D10
["krylov", "conjugate_gradient", "multigrid", "sparse", "system",
 "pressure", "diffusion", "mesh", "preconditioner", "parallel",
 "scaling", "communication", "solver", "residual"],

# D11
["krylov", "block", "schur", "preconditioner", "coupling",
 "pressure", "velocity", "displacement", "porous", "fluid",
 "structure", "monolithic", "parallel", "scalable", "solver",
 "system"],

# D12
["krylov", "sparse", "solver", "preconditioner", "parallel",
 "scaling", "communication", "synchronization", "orthogonalization",
 "residual", "convergence", "stability", "iteration", "system"],
]

```

2.2 统一数学约定

令 $m = 12$ ，词表为全部 token 的字典序并集，词表大小记为 n 。设 c_{ij} 是词项 t_j 在文档 D_i 中的计数，定义

$$\text{tf}_{ij} = \frac{c_{ij}}{\sum_{\ell=1}^n c_{i\ell}}, \quad \text{df}_j = \#\{i : c_{ij} > 0\}, \quad (2)$$

$$\text{idf}_j = \ln\left(\frac{m}{\text{df}_j}\right) + 1, \quad A_{ij} = \text{tf}_{ij} \text{idf}_j. \quad (3)$$

矩阵方向固定为

$$A \in \mathbb{R}^{m \times n}, \quad (5)$$

行为文档、列为词项。使用自然对数；不采用软件包默认的平滑 IDF，也不在 TF-IDF 后额外进行行归一化。

令 $r = \min(m, n)$ 。约化 SVD 为

$$A = U \Sigma V^T, \quad U \in \mathbb{R}^{m \times r}, \quad \Sigma \in \mathbb{R}^{r \times r}, \quad V \in \mathbb{R}^{n \times r}, \quad A_k = U_k \Sigma_k V_k^T. \quad (7)$$

本题将矩阵的 **Frobenius 范数平方** 称为其“能量”：

$$E(A) := \|A\|_F^2 = \sum_{i=1}^m \sum_{j=1}^n |A_{ij}|^2 = \sum_{j=1}^n \sigma_j^2. \quad (8)$$

这里的“能量”是衡量矩阵元素整体平方大小的数学术语，不具有本题语料中流体或结构的物理能量含义。第 j 个奇异分量的能量为 σ_j^2 ；保留前 k 个奇异分量得到的累计能量及其占总能量的比例分别为

$$E(A_k) = \sum_{j=1}^k \sigma_j^2, \quad \eta_k := \frac{E(A_k)}{E(A)}. \quad (9)$$

因此，“保留至少 70% 的能量”指 $\eta_k \geq 0.70$ 。该比例描述平方范数的保留程度，不能直接解释为保留了相同比例的语义信息或达到了相同的检索命中率。它与重构误差的关系将在任务 2 中推导。

定义文档和词项潜在坐标

$$X_k = U_k \Sigma_k = AV_k, \quad T_k = V_k \Sigma_k. \quad (10)$$

奇异向量可逐列成对变号；只要奇异值、子空间、重构量和余弦相似度一致，列符号差异不影响结果。

本文中的余弦相似度统一定义为

$$\cos(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\|_2 \|\mathbf{y}\|_2}. \quad (11)$$

2.3 近邻和检索指标

对每篇文档， $\mathcal{N}_2^{(k)}(i)$ 表示在 X_k 的行余弦中排除自身后最相近的两篇文档； $\mathcal{N}_2^{(A)}(i)$ 表示原始 TF-IDF 空间中的对应集合。如遇数值完全相同的相似度，以文档编号较小者优先。

对给定的 k ，汇总每篇文档的两个最近邻中与其主题标签相同的关系数，再除以全部 $2m$ 个近邻关系，得到同标签近邻比例：

$$\frac{1}{2m} \sum_{i=1}^m \#\{j \in \mathcal{N}_2^{(k)}(i) : \ell_j = \ell_i\}, \quad (13)$$

其中 ℓ_i 是 2.1 节给出的主题标签。原始空间的同标签近邻比例相应由 $\mathcal{N}_2^{(A)}(i)$ 计算。

2.4 新查询

将三个查询依次记为 q_1, q_2, q_3 。原始问题和唯一计算用 token 如下：

1. How can multigrid preconditioning accelerate pressure transport through a heterogeneous permeability field in a porous reservoir?

```
["porous", "pressure", "permeability", "transport",
 "reservoir", "multigrid", "preconditioner"]
```

2. Which interface treatment preserves stability against added-mass effects for an elastic wall in fluid-structure coupling?

```
["fluid", "structure", "interface", "elastic", "wall",
 "added_mass", "stability", "coupling"]
```

3. Find scalable block Krylov methods with a Schur-complement preconditioner for coupled pressure and displacement.

```
["coupling", "pressure", "displacement", "schur", "block",  
"krylov", "preconditioner", "scalable"]
```

教师预设相关集仅用于评价：

$$\mathcal{R}_1 = \{D_1, D_2, D_3, D_4\}, \quad \mathcal{R}_2 = \{D_5, D_6, D_7, D_8\}, \quad \mathcal{R}_3 = \{D_{11}, D_4, D_6\}. \quad (14)$$

对查询编号 $a \in \{1, 2, 3\}$ ，分别在原始空间和秩 k 潜在空间中按余弦相似度排序。前三篇检索结果中属于 $\mathcal{R}_a$ 的文档数除以 3，称为该查询在相应空间中的前三命中比例。

所有查询排序如遇数值完全相同的相似度，同样以文档编号较小者优先。

3. 作业要求

配套学生程序由同目录的 `lsa_data.py`（固定输入）、`lsa_student.py`（待实现接口）和 `test_lsa.py`（公开测试）组成，运行说明见 `README.md`。学生应完成计算接口和实验入口；公开测试调用学生实现，不提供本题完整数值答案。作业包括下列 4 个任务，不另设扩展实验。比较原始空间与秩 3、5 的结果时，复用同一次 SVD，并在报告中整理题目指定的结果。

任务 1：从语料到 TF-IDF 矩阵（25 分）

- 由 token 列表生成字典序词表、计数矩阵 C 和 TF-IDF 矩阵 A 。
- 报告 `system`、`pressure`、`added_mass` 的 df 和 IDF，并解释常见词与稀有词权重的差别。
- 报告 A 的形状、非零元个数、数值秩和 Frobenius 范数。
- 在原始 TF-IDF 空间中，计算同标签近邻比例，只列出 D4、D6、D11 的两个最近邻及相似度。结合文档内容解释跨标签近邻为何可能合理。

任务 2：SVD、误差与能量选阶（25 分）

- 计算约化 SVD，列出全部奇异值，并报告相对重构残差以及 U, V 的正交性残差。
- 按 2.2 节的能量定义，对全部 $k = 1, \dots, 12$ 计算累计能量比例 η_k 与相对 Frobenius 重构误差 ρ_k ：

$$\eta_k = \frac{\sum_{j=1}^k \sigma_j^2}{\sum_{j=1}^{12} \sigma_j^2}, \quad \rho_k = \frac{\|A - A_k\|_F}{\|A\|_F}, \quad (15)$$

并用 Eckart--Young--Mirsky 定理推导 $\rho_k = \sqrt{1 - \eta_k}$ 。

- 分别给出累计能量比例至少为 70%、80% 的最小秩，并解释能量准则不能直接保证检索质量。
- 对 $k = 5$ 给出 $\|A - A_5\|_2$ 、 $\|A - A_5\|_F$ 和相对 Frobenius 误差。

任务 3：潜在方向与桥接文档（25 分）

- 检查维数并证明

$$A_k = X_k V_k^\top = U_k T_k^\top, \quad X_k X_k^\top = A_k A_k^\top. \quad (16)$$

解释为什么 $X_k T_k^\top \neq A_k$ 。

2. 计算 X_5 及其行余弦矩阵 S_5 ，仅报告 D4、D6、D11 的两个最近邻及相似度，与原空间比较。无需抄写完整坐标或相似度矩阵。
3. 第 ℓ 个方向上词项 t_j 的载荷为 $V_{j\ell}$ 。只解释第 2 个方向：列出载荷最大和最小的各 4 个词项，说明其两侧的语义对比及整体变号的不影响性；载荷相同时按词项字典序排列。

任务 4：查询折叠与检索比较（25 分）

1. 对查询 q_a ，以 $c_a(t)$ 表示词项 t 的计数，令 $L_a = \sum_t c_a(t)$ 为查询 token 总数，采用 $\text{tf}_a(t) = c_a(t)/L_a$ ，并结合训练语料的 IDF 构造行向量 $q_a \in \mathbb{R}^{1 \times n}$ 。证明当文档坐标为 $X_k = A V_k$ 时，与之尺度一致的查询坐标为

$$z_a^{(k)} = q_a V_k. \quad (17)$$

解释为什么必须使用训练语料的词表和 IDF。

2. 对每个查询，在原始 TF-IDF 空间以及 $k = 3, 5$ 的潜在空间中，报告前三篇文档、相似度和前三命中比例。
3. 分析查询 3 在 $k = 3$ 和 $k = 5$ 时的差异，结合累计能量推荐一个用于本题检索的秩，并说明推荐结论的适用范围。

提交要求

- 可使用 Python/NumPy 或 MATLAB。
- 提交一份 Markdown 报告 `学号_姓名_LSA.md`（建议 4—6 页），包含推导、指定数值表和必要解释。代码应能从 2.1 节 token 列表直接复现结果，不要求绘图。
- 报告软件环境。

随报告提交完整代码、结果表和运行说明，注明 Python/NumPy 或 MATLAB 版本、token 输入来源、排序规则、随机种子（若使用），并指出各图表对应的生成脚本。

4. 扩展阅读材料

- Golub and Van Loan, *Matrix Computations*: 只读 SVD 与最佳低秩近似部分。
- Deerwester et al., “Indexing by latent semantic analysis,” 1990: 只读 LSA 的低秩表示及查询思想，不要求复现论文实验。