作业

- 根据之前提供的《人民日报》语料，自行构造一个一定规模或者具有一定特点的中文文档集，对之进行LSI分析（SVD分解）。要求任选至少两个不同的k1和k2作为语义空间维度大小，如k1=100, k2=200，在这两种k截断的情况下，做如下各任务：
- (1) 考察不同k的选取对term-doc矩阵近似程度的变化情况；
- (2) 任选4个词（其中**一对词为近义词，其余两个词任意**），并任选**不含**上述任何词的文档2篇，**仅含其中1个词**的文档2篇，**仅含其中2个词**的文档2篇，**至少包含其中3个词**的文档2篇，共8篇。画出将这4个词和8篇文档投射到同一个k维潜在语义空间上的情形，并进行适当分析和比较；
- (3) 分别根据不同k截断下的term-term、doc-doc矩阵，对上述4个词和8篇文档分别计算词和词之间、文档和文档之间的两两相似度，并进行适当分析和比较。

作业

- 文档选取要求：
  - 文档规模不少于10000篇，“具有一定特点”指可以选取特定栏目的文档，选择更大规模的文档或能根据栏目文本特点进行特定新颖的分析可酌情加分(不超过10%)
- 提示：
  - 使用语言以及工具包：Python，sklearn或nltk
  - 可视化工具：tSNE可视化工具
  - 可考虑去掉停用词，并适当过滤低频词，防止矩阵过大
- 提交内容
  - PDF以及源代码
- 停用词参考： https://github.com/stopwords-iso/stopwords-zh

工具包

- SVD分解python工具包：
  - Numpy linalg.svd() / SciPy linalg.svd()　精确求解
  - sklearn.decomposition.TruncatedSVD　可制定求解的奇异值个数
- 高维向量可视化tsne
  - t-SNE: T-Distribution Stochastic Neighbour Embedding
  - 将高维向量降维到二维或三维向量实现可视化，高维空间建模为高斯分布，低维空间建模为t分布，寻找分布之间的转化映射
  - sklearn.manifold.TSNE
- 参考链接
  - https://docs.scipy.org/doc/scipy/reference/reference/generated/scipy.linalg.svd.html
  - https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.TruncatedSVD.html
  - https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html