浙江工业大学

文本分析与挖掘

上机实验

计算机科学与技术学院

文本数据无监督分析

一、 实验目的

1. 对文本数据进行聚类，对比不同预处理和表示对聚类结果有效性的影响。
2. 熟悉无标签数据的聚类分析过程。
3. 熟悉文本主题检测方法和结果，对比不同方法的有效性。

二、实验内容

1.用 k-means 算法对新闻数据进行聚类和评估

a. 从 20newsgroups 中抽取两个子集 multi3 和 multi5，其中 multi3 包含 comp.graphics、rec.autos、talk.politics.guns 三个类别，multi5 包含 comp.graphics、comp.windows.x、rec.sport.hockey、rec.autos、talk.politics.guns 五个类别。每个类别随机采样 200 个样本，所以 multi3 共有 600 个文档，multi5 共有 1000 个文档。

b. 进行适当预处理后基于 TF-IDF 表示，对 multi3 和 multi5 数据集用 k-means 聚类（k 分别取 3 和 5），得到 10 次结果的平均 NMI（归一化互信息）。注意：聚类时用全部数据集，不用分训练和测试。观察讨论数据集分布和类别个数是否会增加聚类难度。

c. 改变实验设置（可以是预处理、表示、聚类算法等），对比结果并讨论。

2.主题检测和方法对比

a. 对 multi5 数据集进行预处理和词袋表示后，用 LDA 进行主题检测：设置主题数目 topic_num=5,打印每个主题 top-10 关键词,根据得到的结果适当调整预处理和表示模型。观察并定性讨论检测到的主题、计算主题的 coherence 分数、以及每个主题在文档中的分布情况（每个文档关联到最相关的主题，再打印主题包含的文档百分比。参考课件：对主题模型的解释和分析）。

b. 用矩阵分解方法 LSI 重复以上实验，对比每个主题的 top-10 关

键词以及 coherence 分数。

2.主题级别特征对分类的影响（选做）

a. 基于 TF-IDF 用支持向量机对 multi5 进行分类；把基于非负矩阵分解 NMF 得到的 doc-topic 矩阵与 TD-IDF 矩阵合并，即对每个文档增加主题级别特征，用支持向量机进行分类，对比只用 TF-IDF 的结果并讨论。

注意：

可以用 gensim 实现 LSI、LDA 主题检测，以及 coherence 计算。非负矩阵分解可以用 sklearn.decomposition 里面的 NMF，具体用法参照官方文档或课件。