自然语言处理第一次作业（A）

作者：张宗桁
学号：2023K8009991013

2025 年 11 月 3 日

目录

1 样本的爬取过程

本次实验利用 Python 语言的 Scrapy 框架进行网络数据的爬取。Scrapy 是一个为了爬取网站数据、提取结构性数据而编写的应用框架。

我们主要选取了几个大型新闻门户网站以及维基百科的中文和英文版面作为主要数据源。爬取过程中，我们设置了合理的 'DOWNLOAD_DELAY'（注意这里的下划线需要转义）以避免对目标服务器造成过大压力。所有爬取到的原始网页均以 HTML 格式存储在本地文件系统中。

2 样本的清洗

数据清洗是本实验至关重要的一步，其目的是从原始 HTML 文本中提取出干净、可用的纯文本内容。我们主要使用了 Python 的 BeautifulSoup 库和正则表达式 (re) 库。

清洗步骤如下：

1. **去除 HTML 标签：** 使用 BeautifulSoup 的 `get_text()` 方法提取所有文本。
2. **去除脚本和样式：** 剔除 `<script>` 和 `<style>` 标签内的内容。
3. **去除无效字符：** 使用正则表达式去除多余的空白符、换行符和特殊控制字符。
4. **统一编码：** 所有文本统一保存为 UTF-8 编码。

最终，我们得到了三个不同规模的中文和英文语料库：

- **小规模样本：** 约 1MB 文本。
- **中规模样本：** 约 10MB 文本。
- **大规模样本：** 约 100MB 文本。

3 英语字母和单词或汉字的概率和熵

在此部分，我们分别对中文和英文样本进行信息熵的计算。信息熵的计算公式为：

$$H(X) = -\sum_{i=1}^{n} P(x_i) \log_2 P(x_i)$$

其中 $P(x_i)$ 是一个符号（字母、单词或汉字）出现的概率。

3.1 中文

对于中文，我们以“汉字”为单位进行统计。

3.1.1 小规模样本

（此处分析小规模中文样本的汉字频率和信息熵...）

3.1.2 中规模样本

（此处分析中规模中文样本的汉字频率和信息熵...）

3.1.3 大规模样本

（此处分析大规模中文样本的汉字频率和信息熵...）

3.2 英文

对于英文，我们分别以“字母”（Letter）和“单词”（Word）为单位进行统计。

3.2.1 小规模样本

字母熵： （分析...）

单词熵： （分析...）

3.2.2 中规模样本

字母熵： （分析...）

单词熵： （分析...）

3.2.3 大规模样本

字母熵： （分析...）

单词熵： （分析...）

4 验证齐夫定律 (Zipf's Law)

齐夫定律指出，在一个大型语料库中，任意单词的出现频率 f 与其在频率表中的排名 r 成反比，即：

$$f \propto \frac{1}{r}$$

取对数后，$\log(f)$ 和 $\log(r)$ 应呈线性关系。我们仅对英文样本进行单词级别的验证。

4.1 小规模样本

（分析...）

4.2 中规模样本

（分析...）

4.3 大规模样本

（分析...）我们在大规模样本上的对数-对数坐标图（见图 1）清晰地展示了这一线性关系，从而验证了齐夫定律。

图 1: Zipf 定律对数-对数图