# 凯撒密码的统计法破解分析实验报告

## 1. 问题描述

凯撒密码（Caesar Cipher）是一种经典的替换加密算法。其加密方法是将明文中的所有字母按字母表中的顺序向后或向前移动一定的偏移量，从而生成密文。尽管这种加密方法简单，但其安全性较低，因为通过字母频率分析法，攻击者可以通过分析密文中每个字母的频率并与字母表中的标准频率分布进行比较，从而破解密码。

本次实验的任务是通过编写一个Python程序，使用字母频率分析法破解凯撒密码，并对以下几个因素进行分析：

1. 错误容忍度（tolerance 超参数）对破解精度的影响；
2. 加密文本长度对破解精度的影响；
3. 不同类型文本（如学术、新闻、小说等）对破解精度的影响；
4. 同一类型中不同主题的文本对破解精度的影响。

## 2. 主要算法或模型

### 2.1 凯撒密码加密算法

凯撒密码的加密方法通过将每个字母移动固定的偏移量（本实验中固定为偏移量为3）实现。通过遍历文本中的每个字符，程序对字母进行移动，并保留非字母字符（如标点符号或空格）。

**加密公式**：

$$C_i = (P_i + k) \mod 26$$

其中，$C_i$ 为密文字母，$P_i$ 为明文字母，$k$ 为偏移量，本实验中为3。

### 2.2 字母频率统计与分析法

为了破解凯撒密码，我们首先统计密文中每个字母出现的频率，然后与已知的标准字母频率分布进行比较。标准字母频率表基于英语语言的字母使用频率排序（如：e, t, a, o, i, n 等）。通过计算密文字母的频率分布与标准分布的相似性，可以推断出密文字母与明文字母的对应关系，进而进行破解。

### 2.3 精度计算

实验中的精度计算方法是通过比较解密后的文本与原始明文的相似度来衡量。我们采用了一个错误容忍度参数（tolerance），用以控制解密结果与标准字母频率的匹配程度。例如，当 tolerance = 1 时，解密的字母允许在标准频率表中前后相差 1 位仍然视为正确匹配。

**精度公式**：

$$Accuracy = \frac{\text{正确解密的字母数}}{\text{文本中的总字母数}}$$

## 3. 实验数据

实验数据由不同类型和主题的文本组成，包括：

- **学术**（academic）类文本，如计算机科学、文学、地理等；
- **新闻**（news）类文本，包含经济、军事、体育等主题；
- **小说**（novel）类文本，涵盖文学、历史、科幻等类型。

### 实验数据样本

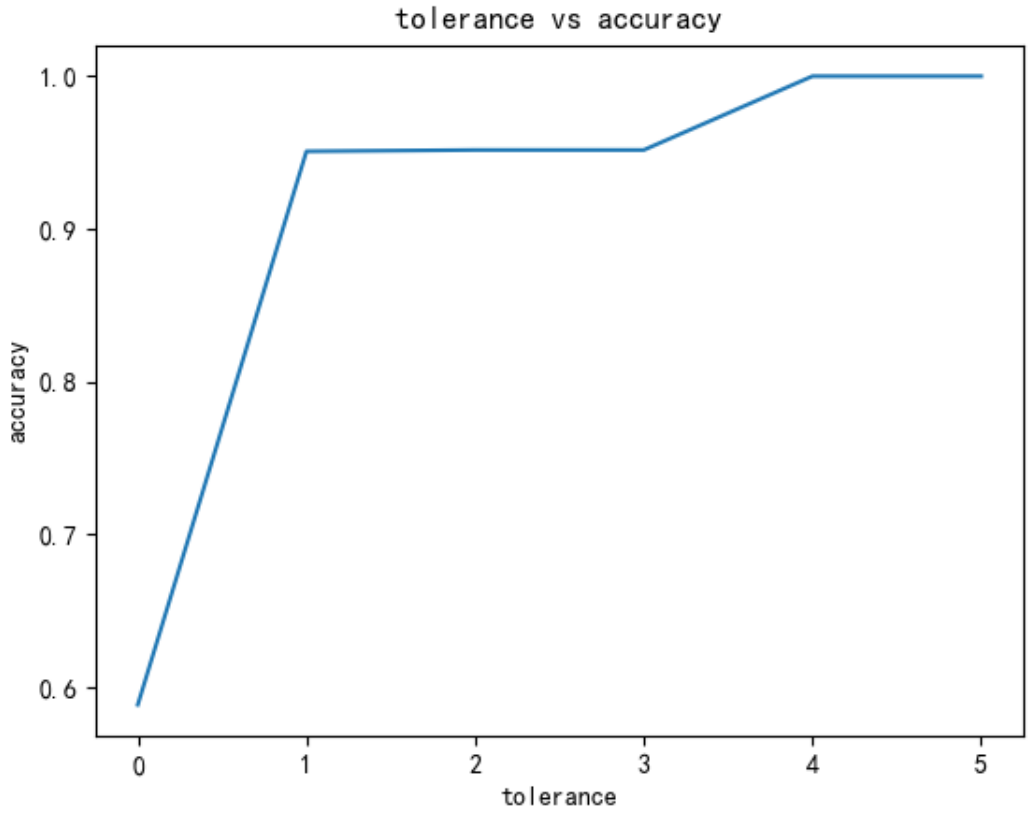
- origin（原始文本）：553,772 字符
- academic：198,000 字符
  - academic_cs：149,634 字符
  - academic_geo：57,008 字符
  - academic_literature：128,438 字符
  - academic_philo：1,261,451 字符
  - academic_phy：12,313 字符
  - academic_social：198,000 字符
- news：12,442 字符
  - news_economic：3,768 字符
  - news_literary：7,430 字符
  - news_military：12,442 字符
  - news_politics：7,881 字符
  - news_sports：9,108 字符
  - news_tech：5,341 字符
- novel：754,635 字符
  - novel_classical：1,018,279 字符
  - novel_history：3,198,748 字符
  - novel_literature：754,635 字符
  - novel_scific：267,782 字符
  - novel_tales：200,608 字符

# 4. 实验结果展示与分析

## 4.1 错误容忍度对破解精度的影响

在这部分实验中，我们设置了不同的错误容忍度（tolerance），分别为 0、1、2、3、4、5。结果表明，当错误容忍度从 0 增大到 1 时，破解精度迅速提高，随后精度变化趋于平缓。

| 容忍度 (tolerance) | 破解精度 (accuracy) |
| --- | --- |
| 0 | 0.588 |
| 1 | 0.951 |
| 2 | 0.952 |
| 3 | 0.952 |
| 4 | 1.000 |
| 5 | 1.000 |

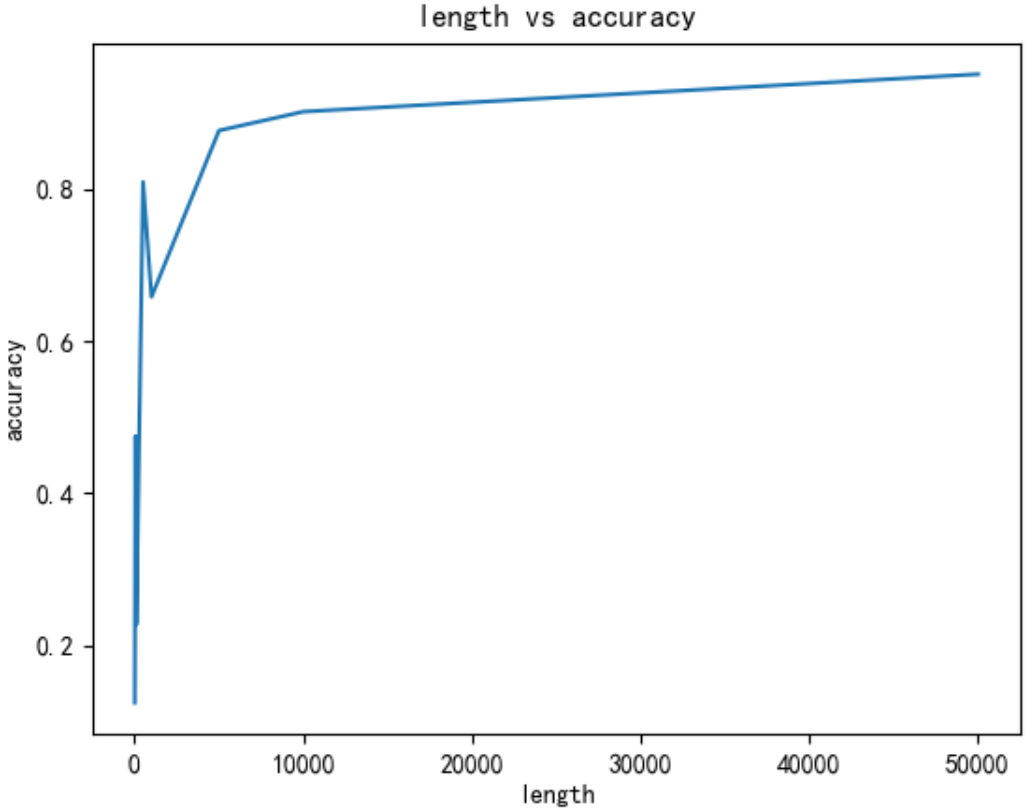



**分析**：随着容忍度增加，算法对密文字母的解密允许更多偏差，因而精度显著提升，特别是在容忍度为 1 时，精度提升尤为明显。

## 4.2 加密文本长度对破解精度的影响

文本长度对字母频率分析的效果也有显著影响。我们选择了不同长度的文本（10、50、100、500、1000、5000、10000、50000）进行实验，结果如下：

| 文本长度 (length) | 破解精度 (accuracy) |
| --- | --- |
| 10 | 0.125 |
| 50 | 0.475 |
| 100 | 0.228 |
| 500 | 0.809 |
| 1000 | 0.658 |
| 5000 | 0.876 |
| 10000 | 0.901 |
| 50000 | 0.950 |

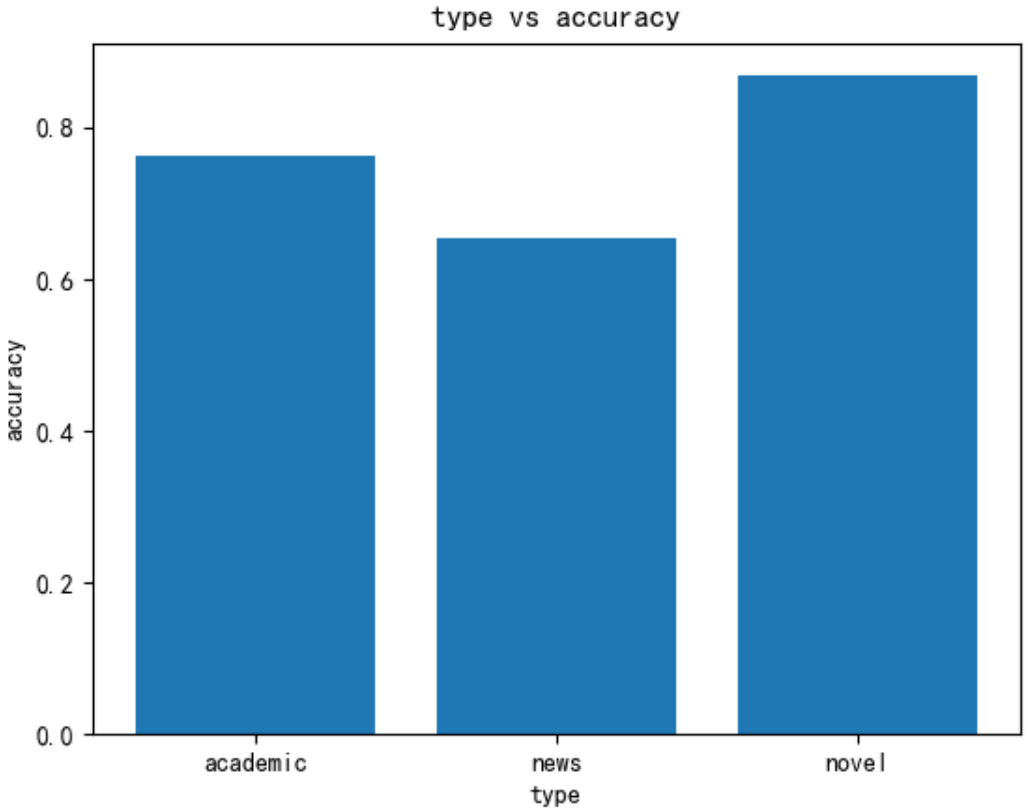



**分析**：随着文本长度的增加，破解精度显著提高。特别是当文本长度达到 5000 个字母时，破解精度已经接近 90%，这说明文本越长，字母频率统计越可靠，进而提高了破解的准确性。

## 4.3 不同类型文本对破解精度的影响

我们在三类文本（学术、新闻、小说）上分别进行了实验，并统一选取了长度为 12,442 字符的文本进行比较。结果如下：

| **文本类型 (type)** | **破解精度 (accuracy)** |
| --- | --- |
| 学术 (academic) | 0.763 |
| 新闻 (news) | 0.654 |
| 小说 (novel) | 0.868 |




**分析**：小说文本的破解精度最高，新闻文本的精度最低。这可能与不同文本类型的字母使用频率分布差异有关。学术类文本和新闻类文本可能存在更多的术语和专业词汇，使字母分布更均匀，从而降低了字母频率分析法的效果。

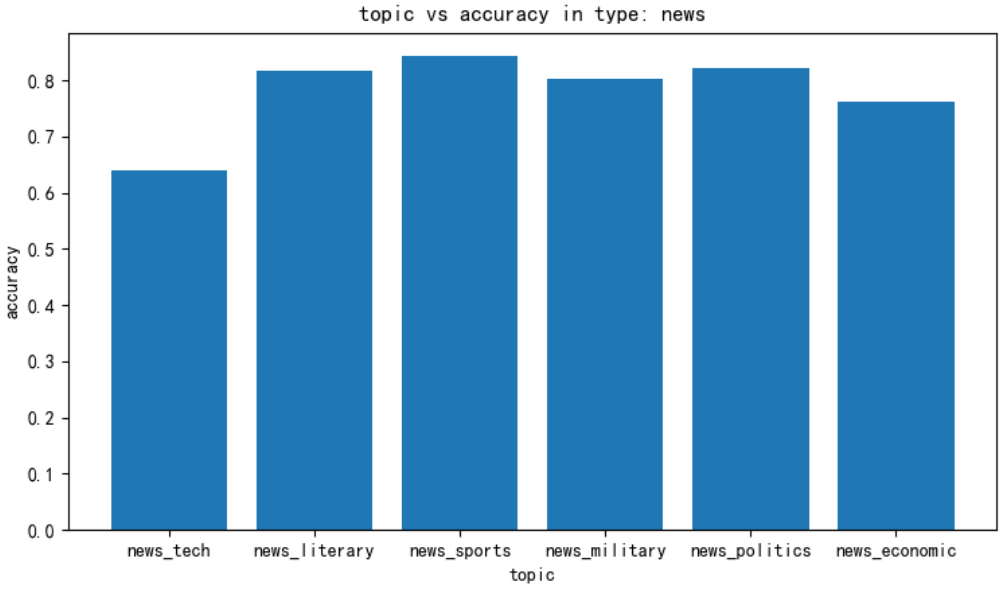
## 4.4 不同主题的文本对破解精度的影响

在同一类型的文本中，我们选择了不同的主题进行分析。破解精度结果如下：

### 新闻类文本（news）

在新闻类文本中，我们选择了六个不同主题（经济、军事、体育、政治、娱乐、科技）进行分析，并统一选取了长度为 3,768 字符的文本进行比较。

| 主题 (topic) | 破解精度 (accuracy) |
| --- | --- |
| 经济 (news_economic) | 0.761 |
| 军事 (news_military) | 0.802 |
| 体育 (news_sports) | 0.843 |
| 政治 (news_politics) | 0.822 |
| 娱乐 (news_literary) | 0.817 |
| 科技 (news_tech) | 0.641 |

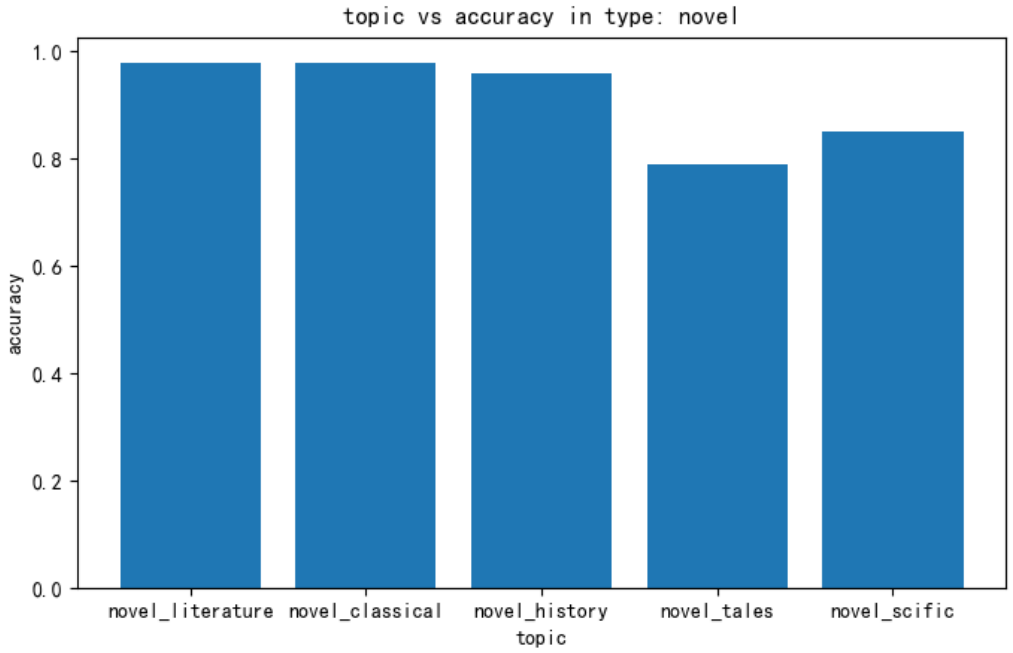



**分析**：不同主题的文本在破解精度上有明显差异。体育类和军事类文本的破解精度相对较高，可能是因为这些文本中使用了更多常见的词汇，使得字母频率更接近标准分布。而科技类文本的破解精度较低，可能是因为科技类文本中的术语更加专业和集中，词汇使用的多样性较低，导致字母频率分布与标准表的差异较大。此外，科技文本中的特定术语通常含有一些不常用的字母，这也进一步增加了破解的难度。

## 小说类文本（novel）

在小说类文本中，我们选择了五个不同主题（童话、经典文学、历史、科幻小说、文学小说）进行分析，并统一选取了长度为 200,608 字符的文本进行比较，结果如下：

| **主题 (topic)** | **破解精度 (accuracy)** |
| --- | --- |
| 童话 (novel_tales) | 0.789 |
| 经典文学 (novel_classical) | 0.977 |
| 历史小说 (novel_history) | 0.958 |
| 科幻小说 (novel_scific) | 0.851 |
| 文学小说 (novel_literature) | 0.978 |

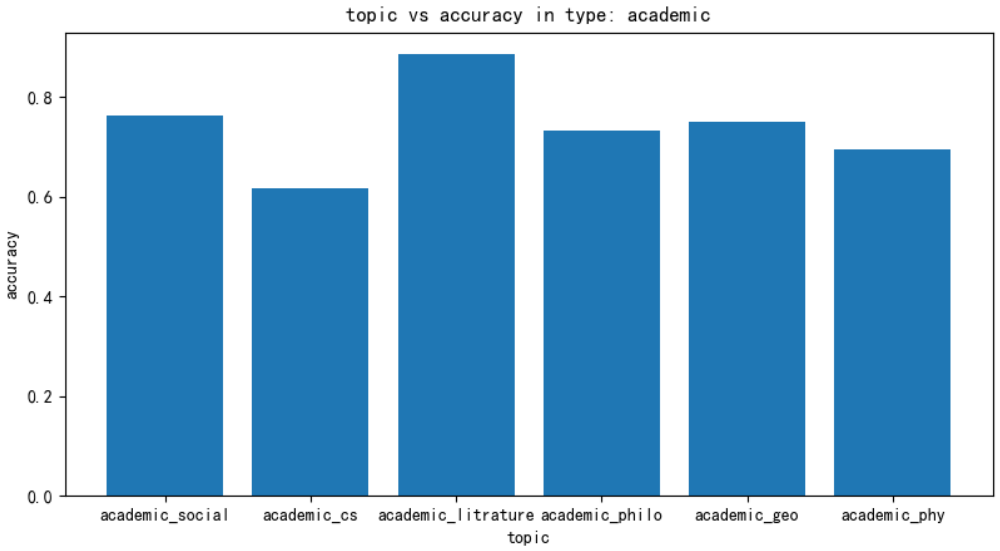



**分析**：经典文学与文学小说的破解精度非常高，分别为 0.977 和 0.978。这类文本包含较多常用词汇和标准语句，字母频率分布接近标准字母表，因此更容易被破解。科幻小说的破解精度为 0.851，虽然仍然较高，但由于科幻小说中可能包含一些专业术语和未来设定的专有名词，导致字母频率分布与标准表稍有差异。童话类文本的破解精度较低，为 0.789。童话文本的语言风格较为独特，包含一些少见的词汇和表达方式，导致字母频率与标准分布不完全吻合，从而降低了解密精度。

## 学术类文本（academic）

在学术类文本中，我们选择了六个不同的主题（物理学、地球科学、计算机科学、哲学、文学、社会科学）进行分析，并统一选取了长度为 12,313 字符的文本进行比较，结果如下：

| **主题 (topic)** | **破解精度 (accuracy)** |
| --- | --- |
| 物理 (academic_phy) | 0.695 |
| 地球科学 (academic_geo) | 0.749 |
| 计算机科学 (academic_cs) | 0.616 |
| 文学 (academic_literature) | 0.885 |
| 哲学 (academic_philo) | 0.732 |
| 社会科学 (academic_social) | 0.762 |




**分析**：文学类学术文本的破解精度最高，为 0.885。这是因为学术文学文本中的词汇更接近日常用语，字母频率分布与标准频率表的偏差较小。计算机科学类学术文本的破解精度最低，仅为 0.616。这类文本中常用大量专业术语和技术符号，字母分布显著偏离标准表，增加了破解难度。物理类学术文本的破解精度为 0.695，处于中等水平，说明物理文本中的术语也影响了字母频率分布，但影响较为适中。

# 5. 总结

本实验通过对凯撒密码的破解过程进行了详细分析，探讨了错误容忍度、加密文本长度、文本类型和文本主题对破解精度的影响。实验结果表明：

- 错误容忍度对破解精度有显著影响，适当增加容忍度可以显著提升破解效果；
- 文本长度越长，破解精度越高；
- 不同类型文本和不同主题的文本由于字母频率分布的差异，导致破解精度存在差异。

通过字母频率分析法，我们可以在一定条件下有效破解凯撒密码，这也表明该加密方法在现代环境下的安全性较低，实际应用中需谨慎使用。