Report of Verification of Zipf's Law

Weidong Xu
2697782204@qq.com

Abstract

This is a report of Zipf Law's verification by way of jieba tokenization(Chinese characters). One of four great classic novels——Romance of Three Kingdoms(三国演义) is chosen as the tested datasets for this report.(https://pan.baidu.com/s/193xE7xcd_tMLrVZ-VpcspQ?pwd=Ed67 提取码：Ed67) This report contains two parts. The first is the experiment of verification, while the other shows the calculation process of words' average information entropy.

Introduction

Zipf's law, named after linguist George Zipf, is a statistical principle that describes the relationship between the frequency of words in a language and their rank or position in a frequency table. It states that the frequency of any word is inversely proportional to its rank. The most common word appears approximately twice as often as the second most common word, three times as often as the third most common word, etc.

Experimental Studies

Figure 1 shows the relevant result between two factors: frequency and rank. A dense map is drawn based on the result, which is similar to an inverse proportional function. The words that appear in a descending order of frequency are as follows:

['曰', '之', '也', '吾', '与', '将', '而', '了', '有', '在', '为', '我', '不', '去', '又', '皆', '来', '乃', '于', '曹操', '见', '矣', '遂', '是', '孔明', '人', '将军', '等', '今', '欲', '此', '却说', '便', '到', '至', '若', '兵', '从', '已', '汝', '可', '中', '操', '得', '被', '杀', '蜀', '上', '以', '玄德', '关公', '亦', '使', '丞相', '问', ' ', '二人', '其', '走', '不可', '只', '荆州', '却', '令', '寨', '玄德曰', '下', '孔明曰', '出', '说', '不能', '如此', '无', '请', '死', '张飞', '一', '后', '斩', '并', '商议', '如何', '他', '主公', '即', '听', '军士', '言', '张', '就', '军', '吕布', '左右', '军马', '时', '且', '者', '回', '赵云', '刘备', '引兵', '次日', '二', '看', '大喜', '诸', '大', '所', '你', '孙权', '引', '更', '天下', '郤', '东吴', '于是', '某', '今日', '正', '不敢', '往', '前', '如', '但', '则', '尽', '忽', '魏兵', '陛下', '瑜', '臣', '都督', '司马懿', '人马', '不知', '日', '帐', '周瑜', '既', '都', '自', '一人', '汉中', '只见', '众将', '虽', '懿', '退', '还', '起', '待', '望', '云长', '谁', '未', '众', '后主', '取', '投', '袁绍', '当', '大叫', '擒', '上马', '朕', '马超', '各', '非', '事', '急', '太守', '再', '此人', '夫人', '马', '背后', '天子', '先主', '贼', '城中', '一面', '必', '何不', '把', '乎', '后人', '大军', '地', '先生', '百姓', '唤', '何故', '然后', '先锋', '黄忠', '不如', '魏延', '内', '诸葛亮', '守', '云', '的', '用', '愿', '叱', '着', '岂', '赶来', '勿', '入', '原来', '忽报', '令人', '之后', '要', '骑', '教', '救', '喊声', '江东', '徐州', '忽然', '下马', '正是', '故', '城', '奏', '笑', '先', '因此', '僮', '成都', '不见', '未知', '才', '辽', '破', '大事', '诗', '起兵', '诏', '军中', '马岱', '甚', '接应', '大败', '讫', '书', '耳', '敢', '一军', '引军', '闻', '进兵', '姜维', '可以', '庞德', '以为', '叹', '多', '毕', '郡', '不得', '心中', '彼', '拜', '刘表', '和', ' ', '孟获', '赍', '下文', '道', '一声', '蜀兵', '追赶', '命', '粮草', '遣', '一齐', '分解', '回报', '分付', '董卓', '只得', '出马', '因', '夏侯惇', '曹兵', '报', '三千', '随后', '过', '受', '报知', '大将', '许都', '前面', '且说', '坐', '送', '李', '恐', '奉', '孙策', '三', '洛阳', '哭', '鲁肃', '之兵',

...]

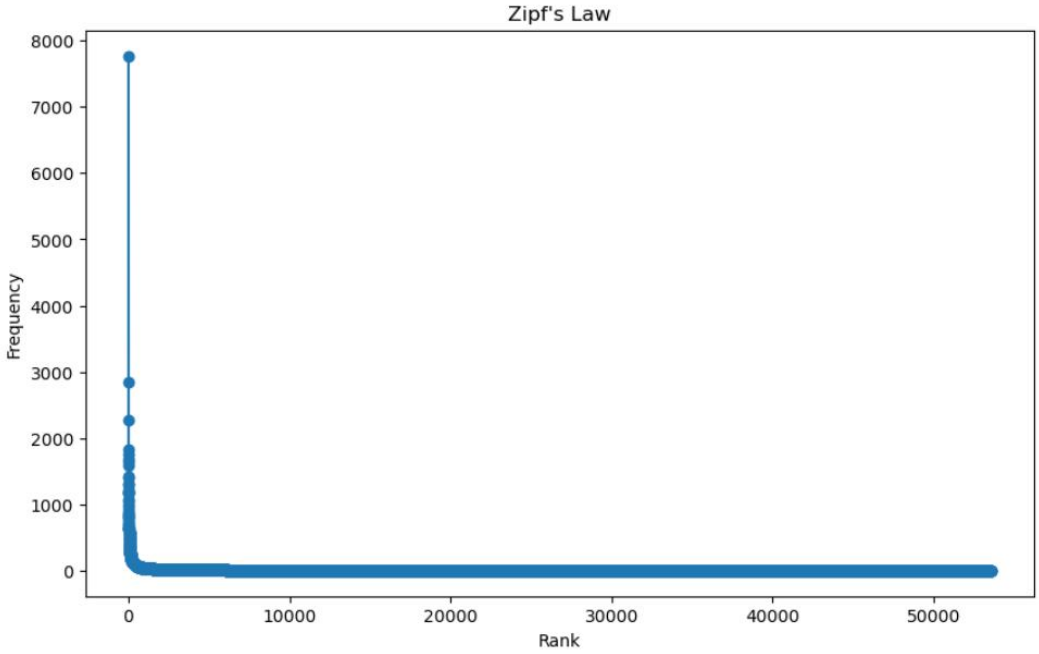

Figure 1 Frequency-Rank relation

In information theory, the entropy of a random variable is the average level of "information", "surprise", or "uncertainty" inherent to the variable's possible outcomes. Given a discrete random variable X, which is distributed according to $p : X \to [0,1]$, the entropy is

$$H(X) = -\sum_{x \in \chi} p(x) \log p(x).$$

The calculation result is shown as follows:

```
Average Entropy: 12.680905521821787
```

Conclusions

This experiment shows that Zipf's Law can be equivalently applicable to Chinese contexts. The essential two factors, frequency and rank are also in an inverse proportional distribution. The distribution pattern of word frequency has rich connotations, and the academic community believes that normal distribution is a typical distribution for describing natural sciences, while Zipf distribution will become a typical distribution for revealing social science laws.