作者简介：一个喜欢英语的程序员，“让英语学习变得更有趣“

作者微信：shrek4j

**前言**

作为一个从小生长在海边的孩子，我却一直没有见过海的样子。但这并没有妨碍年幼的我对大海的幻想和憧憬。

童年的大海，是海洋馆里的蝙蝠（后来才知道叫蝠鲼），是千奇百怪的水母，是色彩斑斓的珊瑚……

后来长大些，大海变成了家里书架上的一本本书名：海底两万里、金银岛、鲁滨逊漂流记……

现在还知道，大海见证了哥伦布的大航海，促成了达尔文的进化论，也承载着路飞的海贼梦。

**正文**

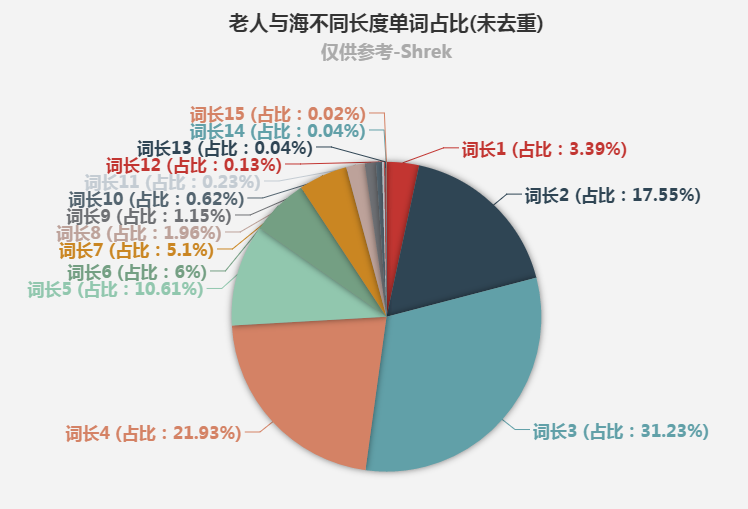
一晃十几年，回想小时看的关于大海的书，情节早已忘得一干二净。感慨时光飞逝的同时，也勾起了我重温老书的兴趣。

小时候，**老人与海**是我最爱读的小说，所以就从这本入手。由于我迫切地想知道小说到底讲了些什么，也想发掘一些之前忽视的东西。于是这次，**我尝试通过统计分析来了解这部小说**。

这是一部由2516个单词组成的小说。小说总字数为26780。看起来平均每个单词出现了10次左右。

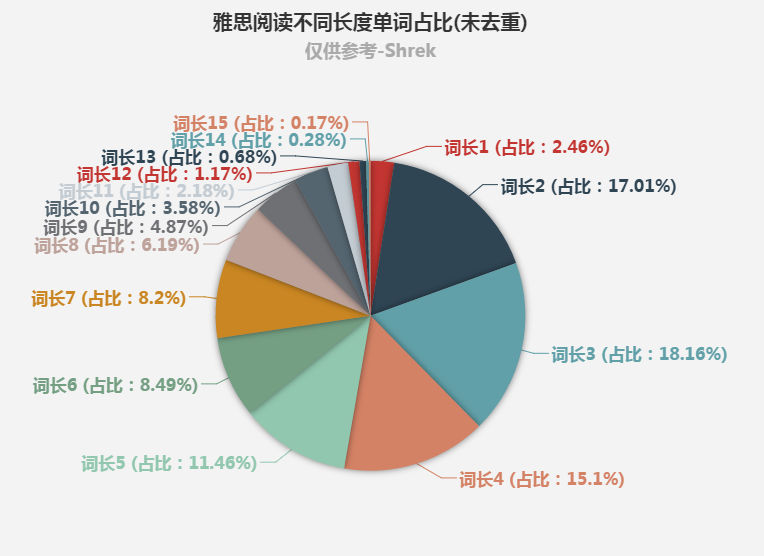
**1. 首先，我想知道这本书单词的难易程度。**

通常，**越短的单词越简单**，于是我统计了不同长度单词数量的分布。



**占比最多的是3个字母的单词**。打开单词列表，我发现这其中大部分是初中词汇，例如：all, sky, sea，当然也有生僻一些的单词，比如:rig(用具), fin(鱼鳍)等。

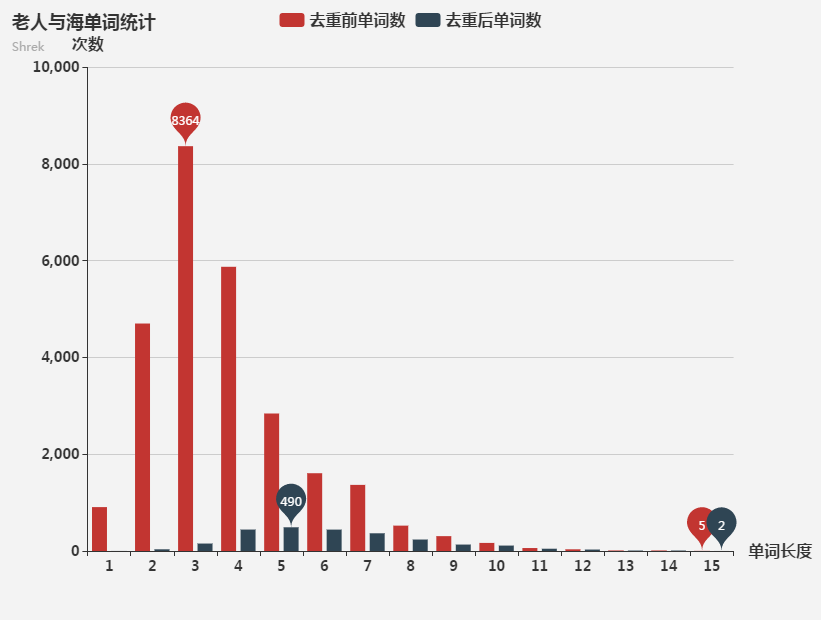
**老人与海中，4个字母的单词只占1/4**。为了对比，我又统计了60篇雅思阅读真题的单词分布。如下图，**雅思阅读中，词长大于4的单词占了1/2之多**。



这说明**老人与海的单词要比雅思阅读的容易很多**。细想下也合理，老人与海的内容更生活化一些，而雅思阅读的取材则是科学美国人(Scientific American)和经济学人(The Economist)这类专业性较强的文章。

**2.** **然后，我想知道不同单词在小说中的出现频率。**

于是，我又对比了去重前后的单词数量的分布。



**（去重：相同单词重复出现只计算一次）**

红色为去重前单词数量，蓝色为去重后单词数量。很明显，去重前大部分是2-4个字母的单词，而去重后，则大部分是4-7个字母的单词。可见**2-4个字母的单词虽然少，但却被频繁使用**，可谓语言中的基石。而另一方面，**超过4个字母的单词，单词越长，使用次数越少**。而这些也是学习英语时，最令我们头疼的单词。

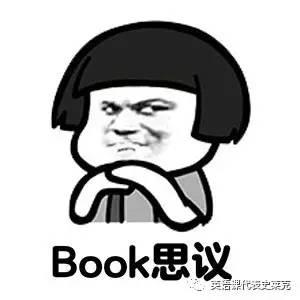
**3.** **那最短的和最长的单词都是什么呢？**

经统计发现，**1个字母的单词只有3个，却在小说中出现了907次**。这三个单词是：**a, i, j**。a和I是当之无愧，但j是什么单词？我打开小说，开始搜索，发现了原文：“"Tell me about the great John **J.** McGraw." He said Jota for **J**.”原来J是middle name Jota的缩写。

**2个字母的单词有35个，在小说中出现了4700次**。一番观察，我发现了几个奇怪的单词，如va, la等。在小说中搜索la，找到了原文。原文为：“He always thought of the sea as **la mar** which is what people call herin Spanish when they love her.”**la mar**（乍一看以为是那个死贵的护肤品牌la mer），原来是西班牙语，意为**the sea(大海)**。



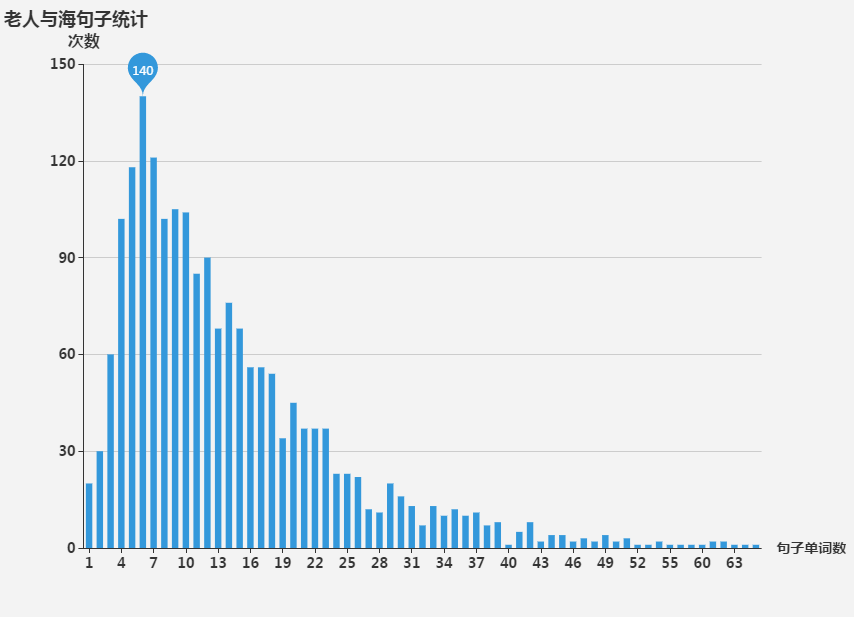
看到这个词，我马上想到了英语中的marine（海洋的，海军）。词源上，marine从拉丁单词mare（sea）演化而来。而西班牙语是罗曼语，也是latin的分支，所以**西班牙语中的mar和英语中的marine都是由拉丁单词mare变化来的**。看来学好词源，不只可以帮助学习英语，还能融会贯通其他罗曼语。



除了这些，我还很好奇这本书最长的单词是哪些。虽然他们只有2个，一共才出现了5次。经过统计，发现他们是conscientiously（良心上地）和phosphorescence（磷光）。第一个单词还算正常，至于第二个单词，如果你不考gre，还是忽略它吧。

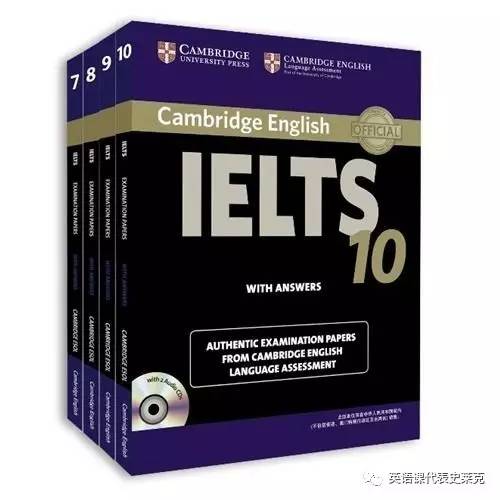
**4. 接着，我想看看这本小说的长难句是哪些。**

所以我又统计了不同长度语句数量的分布。



经统计，整部小说由**1918句话**组成。相同长度的句子被分在一组，一共有63组。**由6个单词组成的句子是最多的，共140句。**

作为一个被中国应试培养出来的学生，我最关心的就是长难句，毕竟那才是考试的拦路虎。于是我统计出了最长的那句话。这句话由**69个单词**组成，它是：“After that he began to dream of the long yellow beach and he saw the first of the lions come down onto it in the early dark and then the other lions came and he rested his chin on the wood of the bows where the ship lay anchored with the evening off-shore breeze and he waited to see if there would be more lions and he was happy.”不难看出，这句话虽然长，但是由多个简单句组成的复合句。只要把每个简单句攻克就可以了。



如果你觉得这句话已经挺长了，那我们来比较下雅思阅读中的长句。在统计了60篇雅思阅读真题后，我找到了雅思阅读中最长的句子。

这句话来自**雅思剑桥8阅读真题**，这篇阅读的名称是“The Nature of Genius”。这句话由**118个单词**组成，长度几乎是上一句话的一倍。它是：“In the mythology of giftedness, it is popularly believed that if people are talented in one area, they must be defective in another, that intellectuals are impractical, that prodigies burn too brightly too soon and burn out, that gifted people are eccentric, that they are physical weaklings, that there's a thin line between genius and madness, that genius runs infamilies, that the gifted are so clever they don't need special help, that giftedness is the same as having a high IQ, that some races are more intelligent or musical or mathematical than others, that genius goes unrecognised and unrewarded, that adversity makes men wise or that people with gifts have a responsibility to use them”。

看完这句话，我觉得海明威友好多了。

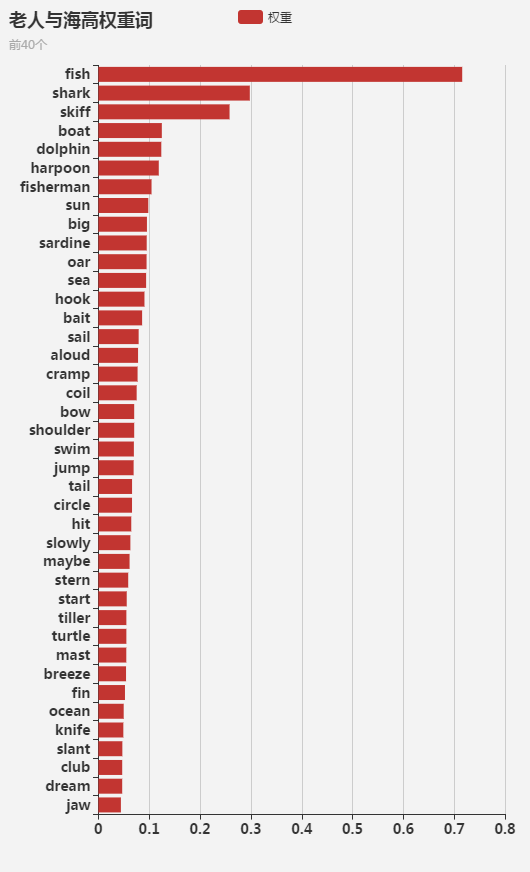
**这个对比，也在语句层面上证明了老人与海要比雅思阅读简单许多。**

**5. 下一步，为了弄清楚这本小说在说什么，我尝试找出最重要的单词。**

在这里，我使用了TF-IDF统计方法。这种算法会**给一系列文章中每篇文章的每个单词的打分，打分越高，这个单词对这篇文档越重要**。

为了尽量得到准确的结果，我的语料库包含了几十部小说和著作，涵盖范围从《血字的研究》，到《哈姆雷特》，再到《国富论》。

现在我们来看看统计结果。



上图是**对老人与海最重要的40个单词**。首先吸引我的单词有：**fish ,dolphin(海豚),sardine(沙丁鱼),turtle**等。看着这些词，我仿佛已经潜到了海里，看着一群群鱼儿从身边游过。

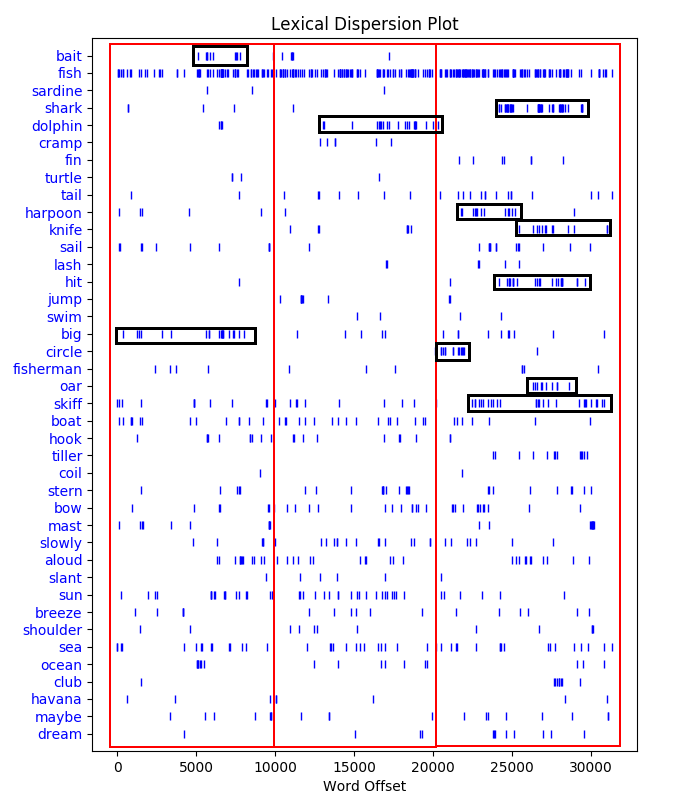


此外，还出现了**boat,oar(桨),mast(桅杆),stern(船尾),tiller(舵柄),harpoon(鱼叉)**。一个渔夫划着小船，拿着鱼叉盯着海里的画面已经跃然纸上了。**sun和breeze(微风)**，也许意味着有时天气还不错。

现在我已经知道故事发生的大背景了。但具体发生了什么，我依然云里雾里。毕竟一部获得诺贝尔文学奖的小说不大可能只是讲述渔民轻松惬意的捕鱼生活。

**6. 为了进一步了解故事情节，我想在时间轴上重新审视这些单词。**

于是我使用了另一个工具：**单词位置分布图**。



上图中，蓝色竖条标记单词在文章中出现的位置。我用红框将小说分成了三部分，每部分约9000字，然后用黑框选出了一些位置。

一眼就能发现**fish**是全篇的重点。无疑这是一个关于捕鱼的故事。

先来看第一部分，密集出现的词是**big和bait**，也许说明渔民发现了很大的鱼，然后开始准备放鱼饵了。在第二部分里，频繁出现了**dolphin**，难道是渔民想抓海豚？

小说的前两部分风平浪静，好像没有什么可圈可点的地方。



开始看第三部分，这部分信息量最大。

首先出现的词是**circle**,是鱼在绕着船游吗？接着看，发现第三部分反复提到了**shark**。也许小船遭遇了鲨鱼的袭击！那渔民的反应是怎么样的呢？

往下找，看到先频繁出现**harpoon(鱼叉)**，然后是**knife(刀子)**，难道渔民在向鲨鱼反击？接着，我发现了单词**hit**，这个词和**shark**的出现位置高度重合，也许是鲨鱼在撞击小船，也许是渔民在反击。

接着往下看，**oar(船桨)和skiff(轻舟)**也被反复提及，也许是深处险境的渔夫在用船桨击打鲨鱼，也许是他在拼命划着船想逃离危险。

一场激烈的搏斗正在展开，渔夫的命运将会如何？



毋庸置疑，用统计的方法可以更加量化地理解小说。但如果想体会情节的精彩，还是要打开书卷，逐句品读。

​

作者保留文章所有权。未经授权，不得转载。如需转载，请联系作者。作者微信：shrek4j。