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。

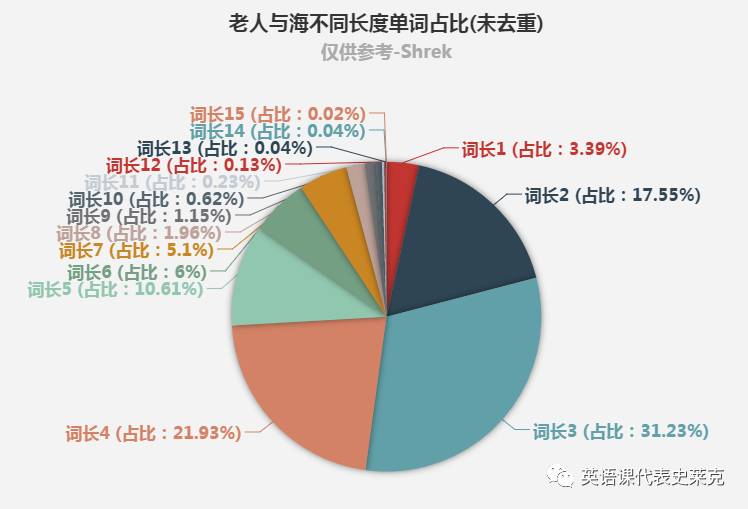
正文

一晃十几年，回想小时看的书，故事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。

说做就做。小时候，老人与海是我最爱读的小说，所以就先从这本入手。本来买来英文书读就可以了，但我迫切地想知道小说到底讲了些什么，而且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。

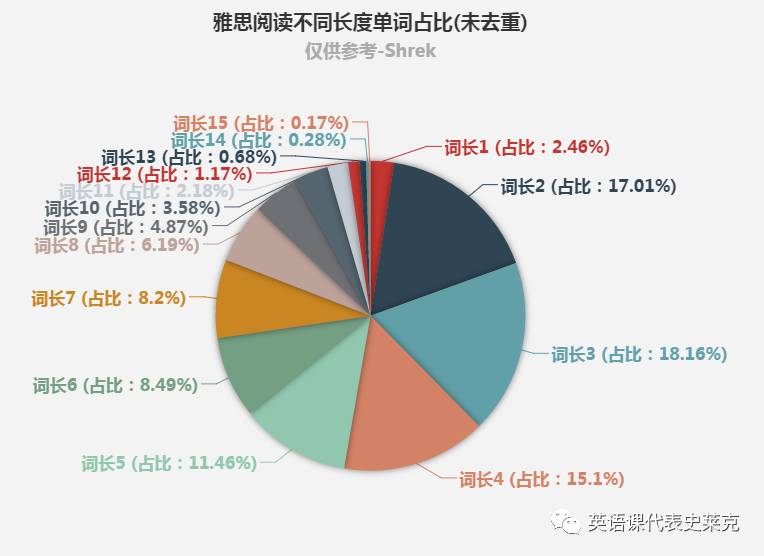
这是一部由2516个单词组成的小说。小说总字数为26780。看起来平均每个单词出现了10次左右。

1. 首先，我统计了不同长度单词数量的分布。



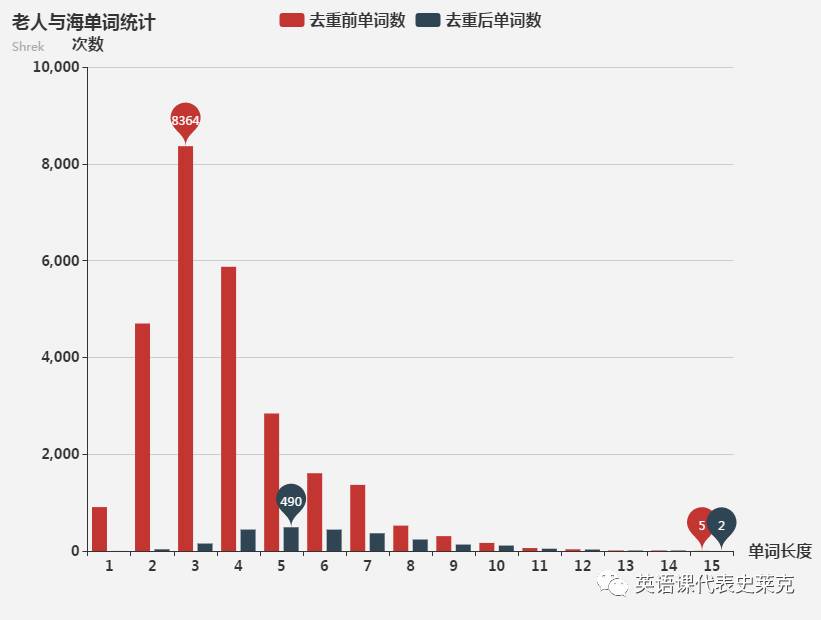
占比最多的是词长为3的单词。打开单词列表，我发现这其中大部分是初中词汇，例如：all, sky, sea，当然也有生僻一些的单词，比如:rig(用具), fin(鱼鳍)等。

词长大于4的单词占1/4。相比之下，雅思阅读中，词长大于4的单词占了1/2之多。



说明老人与海的单词要比雅思阅读的容易一些。这也好解释，因为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，去重前大部分的单词分布在词长2-4的部分，而去重后，则大部分分布在词长4-7部分。可见词长1-4的单词虽然少，但却被频繁使用，可谓语言中的基石。而词长越长的单词，使用次数越少。

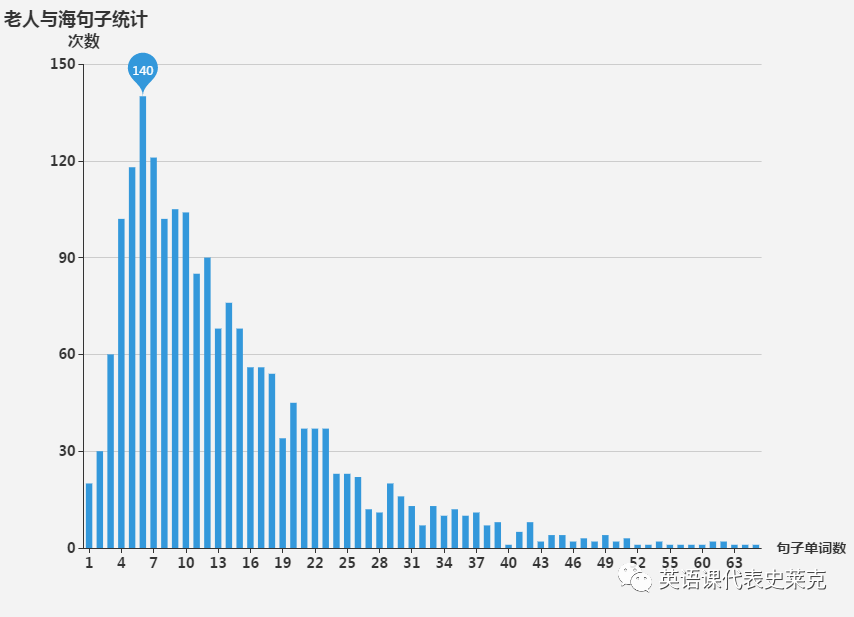
词长1的单词只有3个，却在小说中出现了907次。这三个单词是：a, i, j。a和I是当之无愧，但j是什么单词？我打开小说，开始搜索，发现了原文：“"Tell me about the great John J. McGraw." He said Jota forJ.”原来J是middle name Jota的缩写。

词长2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va，找到了原文："Que Va," the boy said. 这是英语吗？抱着疑问我查了词典，原来是西班牙语！意为“哪里，哪里”，是一种语气词，表示没什么。

接着我又搜索了la，原文是：“He always thought of the sea as la mar which is what people call herin Spanish when they love her.”la mar（乍一看以为是某护肤品牌la mer），也是西班牙语。意为the sea。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，不只可以帮助英语学习，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。

3. 接着，我又统计了不同长度句子在数量上的分布。



整部小说由1918句话组成。包含相同单词个数的句子被分在一组。这样，我得到了63个组。由6个单词组成的句子是最多的，共140句。

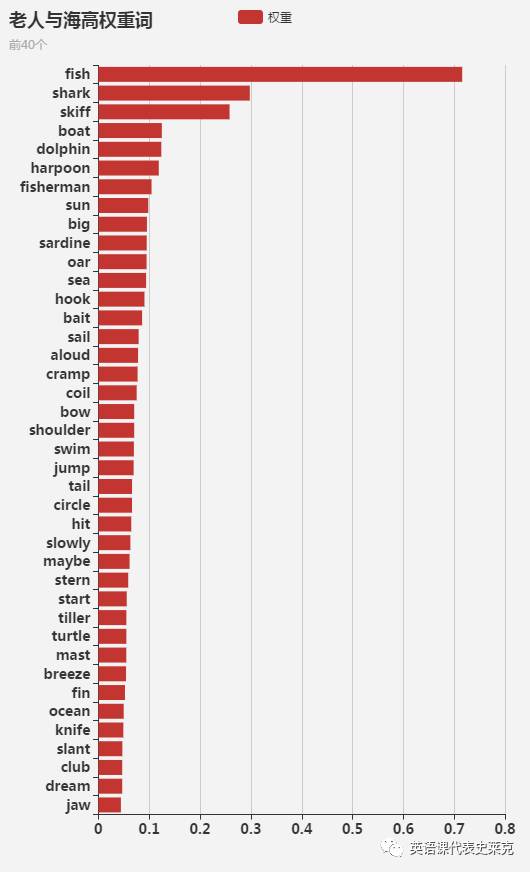
作为一个被中国应试培养出来的学生，我最关心的就是长难句，毕竟那才是考试的拦路虎。于是我被好奇心驱使，找到了最长的那句话。这句话由69个单词组成，这句话是：“After that he began to dream of the long yellow beach and he saw thefirst of the lions come down onto it in the early dark and then the other lionscame and he rested his chin on the wood of the bows where the ship lay anchoredwith the evening off-shore breeze and he waited to see if there would be morelions 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，那我要给你看看雅思剑桥8阅读真题中的一句话，这篇阅读的名称是The Nature of Genius。这句话有118个单词，几乎是上一句话的一倍。这句话是：“In the mythology of giftedness, it is popularly believed that ifpeople are talented in one area, they must be defective in another, thatintellectuals are impractical, that prodigies burn too brightly too soon andburn 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, thatgiftedness is the same as having a high IQ, that some races are moreintelligent or musical or mathematical than others, that genius goesunrecognised and unrewarded, that adversity makes men wise or that people withgifts have a responsibility to use them”。看完这句话，我觉得海明威友好多了。

4. 下一步，为了弄清楚这本小说在说什么，我尝试找出这篇文章最重要的单词。

在这里，我使用了TF-IDF统计方法。这种算法会给一篇文档中每个单词的评分，评分越高，这个单词对这篇文档越重要。（举个栗子：现有10篇文章，10篇文章中world都出现了20多次，而只有第一篇文章中出现了 food，出现次数为5次，那么在第一篇文章中，food的评分比world要高，因为world在每篇文章都出现了很多次，所以world并不能有效的区别每篇文章,但是food可以。因此对第一篇文章来说，food要比world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。明白了统计方法，现在我们来看看统计结果。

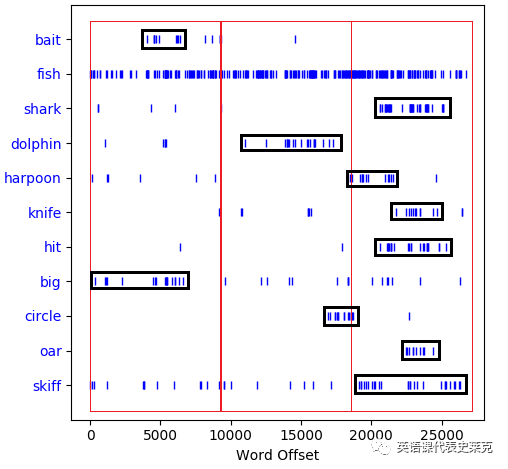
​

上图是老人与海权重最高的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。单词9 sun和单词32 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。

5. 本着小说故事情节会不断推进的原则，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，引起我注意的词是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。

第二部分里频繁出现了dolphin，难道是渔民想抓海豚？小说的前两部分好像风平浪静，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？

用统计的视觉去理解小说纵然新颖有趣，但如果想真正体会故事的精彩，还是要打开书卷，逐句品读。

本书篇幅短小，难度适中，非常适合作为你的第一本英文原版小说。同时这本小说也很适合备考四、六级的学生提高阅读能力。

如果了你对这本书感兴趣，欢迎加入我们的英文读书群。在这里，有一群志同道合的人在等你。

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。微信：shrek4j。