数据挖掘第三次作业

李晨昊 2017011466

2020 年 3 月 17 日

目录

1 数据预处理和可视化 **1**

1.1 数据读入 1

1.2 预处理 1

1.3 词云图 2

1.4 单词长度直方图 2

1.5 新闻词数直方图 3

1.6 新闻类别直方图 4

1.7 新闻月份直方图 5

2 高维向量可视化 **5**

1 数据预处理和可视化

1.1 数据读入

我使用 Python 的`xml.etree.ElementTree`库来读入新闻数据。对于缺失的数据，我使用`None`来表示，作业指导中提到的`NA`其实没有什么意义。

观察数据可以发现一篇新闻可能有多个重复的类别，所以解析新闻类别的时候不应该使用列表来存储，而是应该使用集合。

1.2 预处理

我使用了 Python 的`nltk`库来进行预处理，需要下载`nltk`的`punkt`包和`stopwords`包。

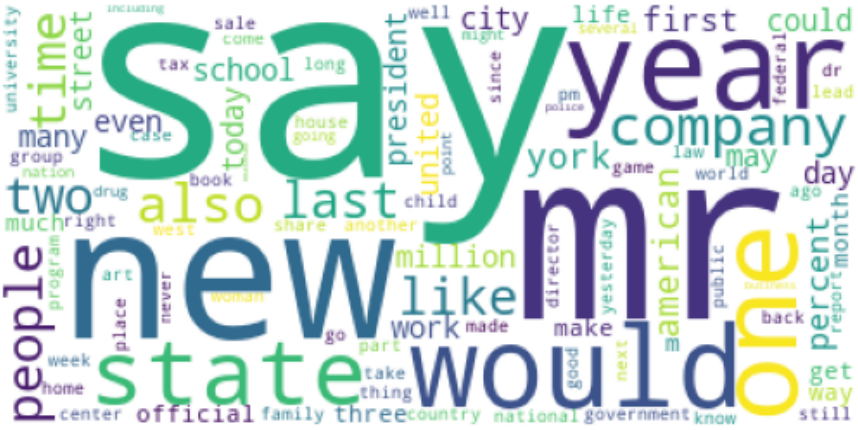
在读入数据的同时，为每个新闻的文本进行预处理，预处理按照顺序包括大小写转换，去除标点符号和数字，分词，去除停用词，以及词干化处理。最后用得到的词序列构造出一

个Counter对象，这就相当于 BagOfWords 向量。同时也维护一个全局的Counter对象，记录所遇到的所有词。

进行词干化处理的时候，nltk库无法处理一些非常规的情形，最明显的例子是无法正确识别 said 的词干 say，而这又是最常见的词之一，所有我对这个词做了一个特判的处理。

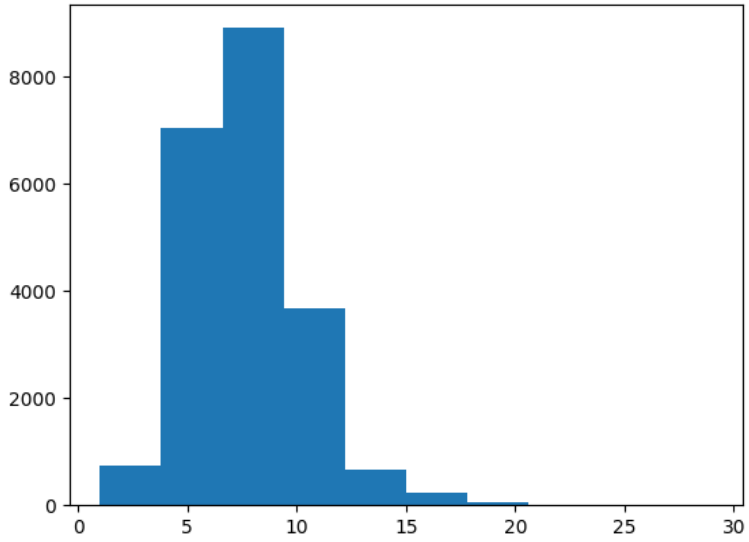
1.3 词云图

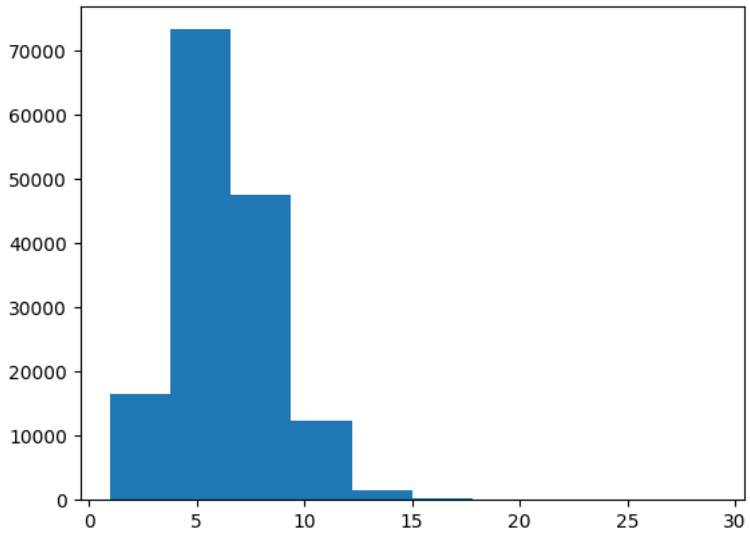
我使用了 Python 的wordcloud库来绘制云图。直接利用预处理中构造的全局的Counter就可以绘制云图了，结果如下：



1.4 单词长度直方图

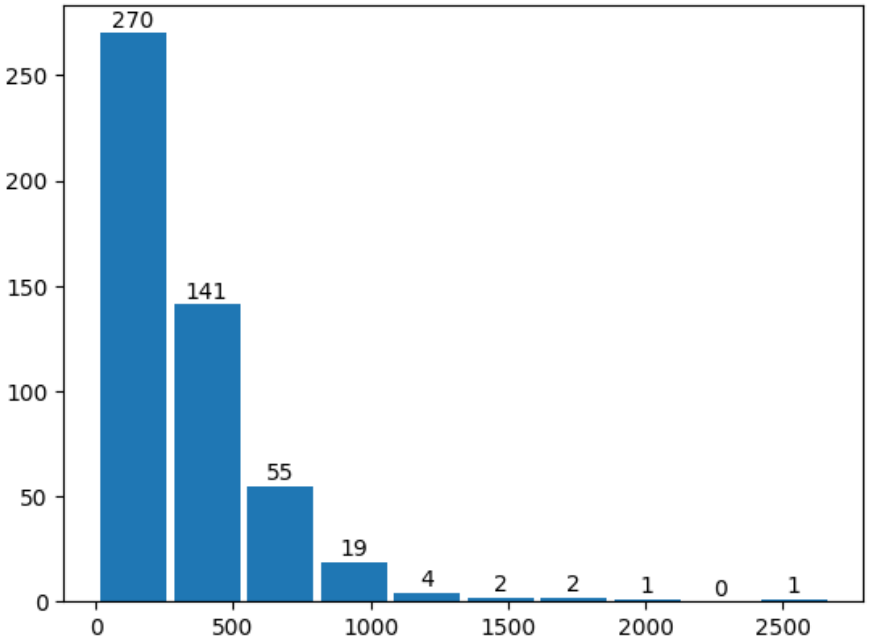
其实这个问题有一定的歧义：在统计某一长度的单词数量的时候是否考虑某个单词的出现次数。使用matplotlib的hist函数来绘制直方图，如果不考虑出现次数，则使用全局的Counter的所有的键的长度的序列；如果考虑出现次数，还需要在键的长度的序列中，键的长度都重复这个键对应的值那么多次，即这个单词出现的次数那么多次。不考虑和考虑的结果分别如下：

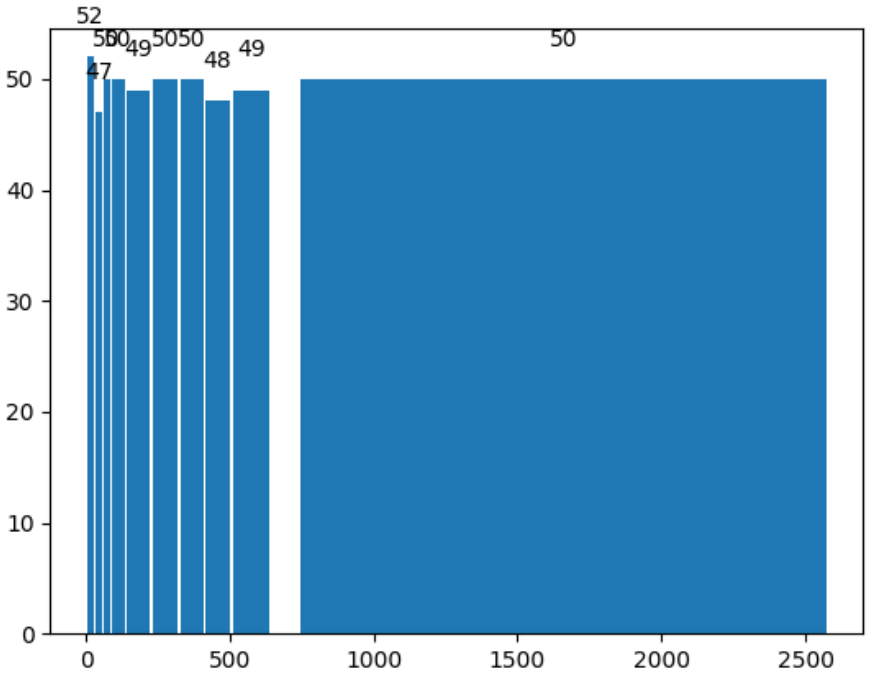




1.5 新闻词数直方图

绘制等宽直方图和等深直方图的数据统计分别使用pandas的cut和qcut函数，得到数据后使用matplotlib的bar函数绘制，等宽直方图和等深直方图的结果分别如下：

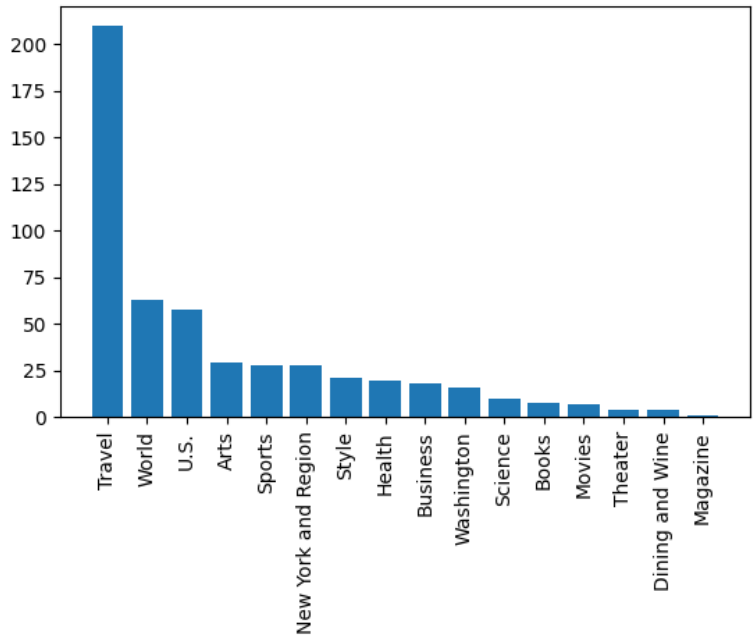




等深直方图并不是很好看，不过这是数据的分布的不均匀导致的，没有什么很好的解决方案。

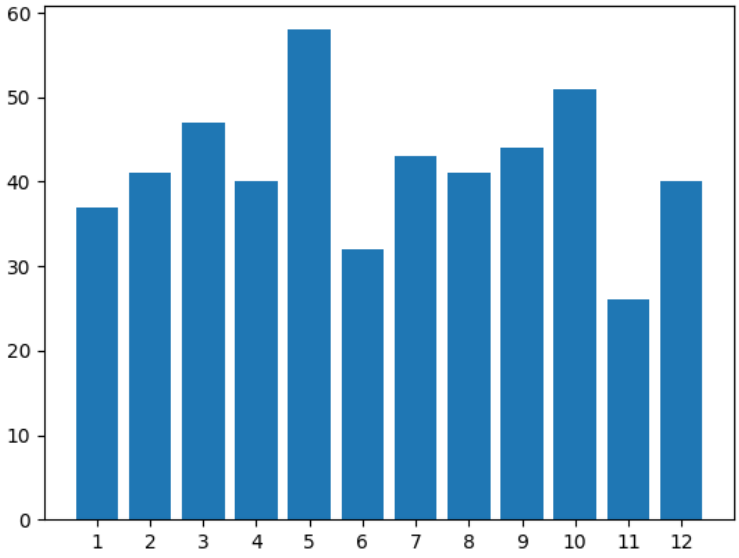
1.6 新闻类别直方图

方法与上面类似，不再赘述，结果如下：



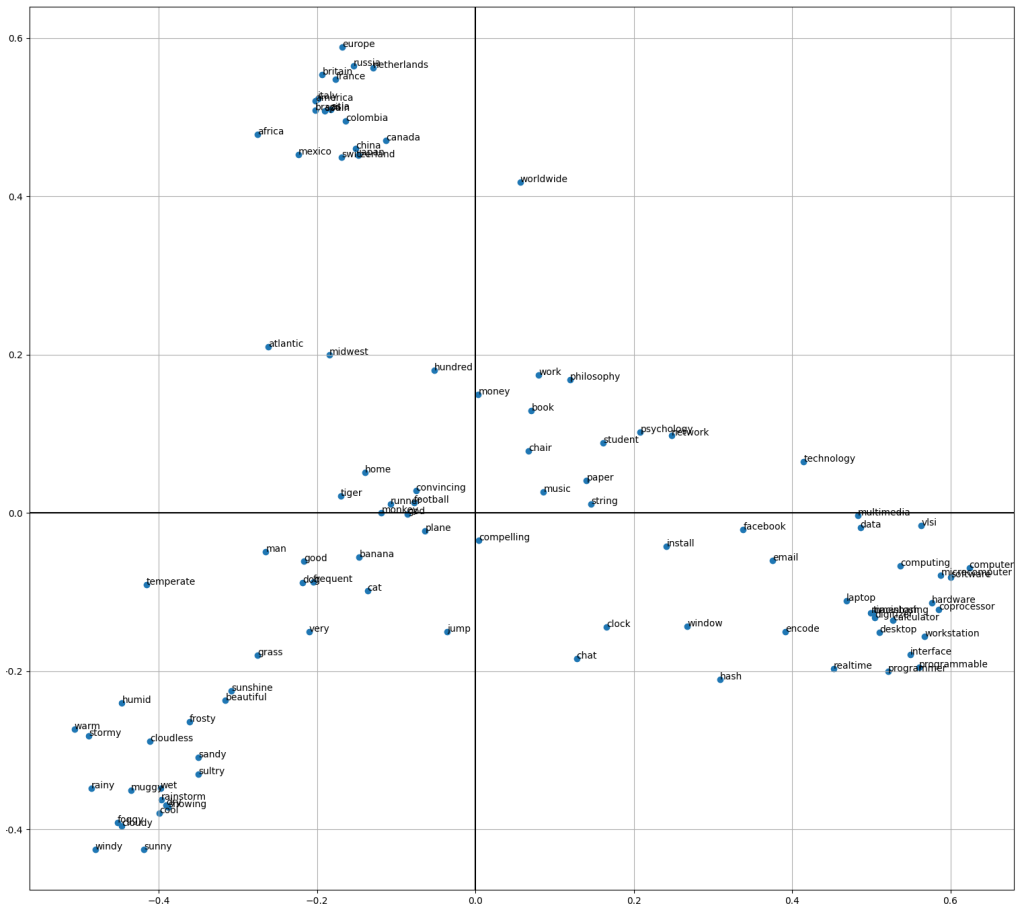
1.7 新闻月份直方图

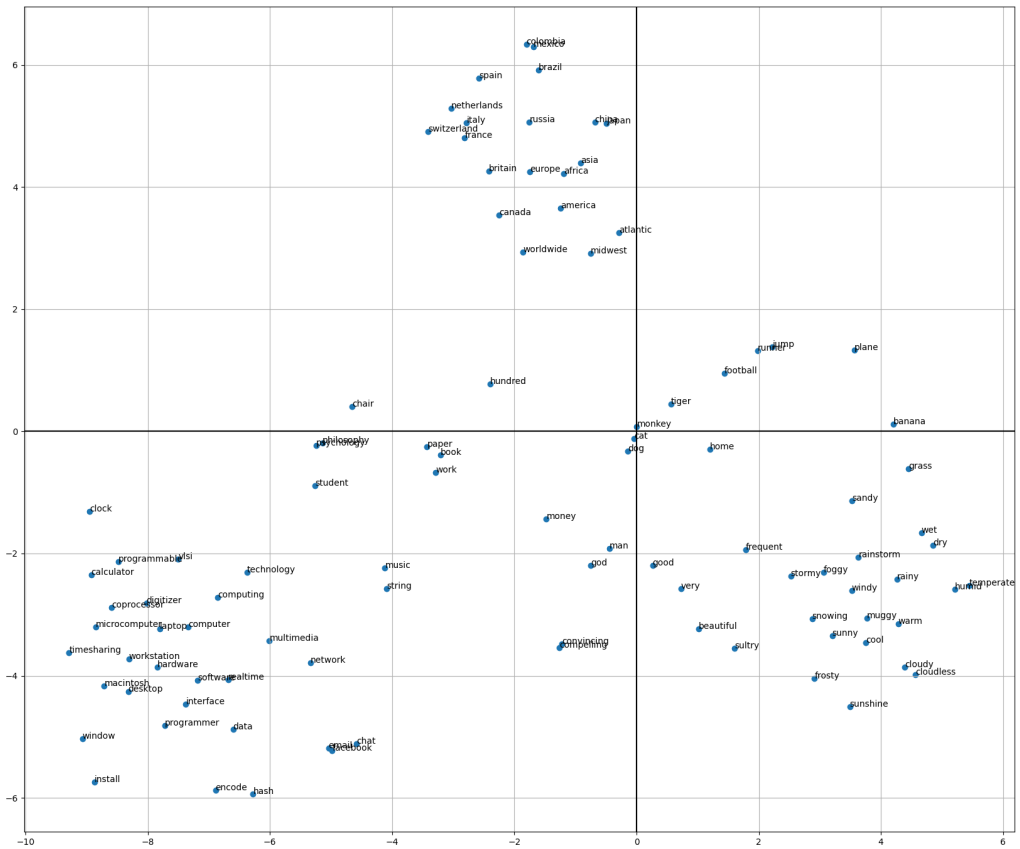
方法与上面类似，不再赘述，结果如下：



2 高维向量可视化

我使用 Python 的`sklearn`库来进行向量的降维，读入数据后直接调用对应的函数即可，使用 PCA 和 t-SNE 两种方法的结果分别如下：


europe
russia
netherlands
britain
france
italy
africa
colombia
canada
china
mexico
switzerland
worldwide
atlantic
midwest
hundred
work
philosophy
money
book
psychology
network
student
chair
technology
home
paper
tiger
convincing
music
string
football
monkey
plane
compelling
multimedia
data
vlsi
facebook
install
man
good
banana
email
computing
temperate
frequent
cat
laptop
hardware
coprocessor
calculator
very
jump
clock
window
encode
desktop
workstation
grass
chat
interface
realtime
programmable
programmer
hash
sunshine
beautiful
humid
frosty
warm
stormy
cloudless
sandy
sultry
rainy
muggy
wet
rainstorm
snowing
cool
foggy
cloudy
windy
sunny


colombia
mexico
brazil
spain
netherlands
italy
switzerland
france
russia
china
japan
asia
britain
europe
africa
canada
america
atlantic
worldwide
midwest
jump
plane
football
hundred
chair
tiger
monkey
banana
cat
philosophy
technology
paper
book
dog
home
work
grass
student
sandy
clock
money
wet
dry
programmable
vlsi
man
frequent
rainstorm
calculator
technology
music
god
good
stormy
foggy
rainy
digitizer
string
very
windy
temperate
computing
coprocessor
microcomputer
laptop
computer
multimedia
beautiful
snowing
muggy
warm
sunny
cool
timesharing
workstation
hardware
convincing
compelling
sultry
network
cloudy
cloudless
macintosh
desktop
software
realtime
frosty
interface
sunshine
programmer
data
window
email
chat
facebook
install
encode
hash

简单分析可以发现，两种方法的结果都算不错，例如 america 和 europe 这些词在两张图中都相当接近，体现了它们在文本中相似的地位。

不过也可以看出对于一些词，t-SNE 方法的表现比 PCA 方法更好。例如 convincing 和 compelling 这两个词都可以表示令人信服的意思，在 PCA 方法的结果中这两个词相距的比较远，而在 t-SNE 方法的结果中这两个词几乎贴到一起去了。从这可以看出 t-SNE 方法的降维结果更优一些。