HW2-report

数据预处理和可视化

新闻数据读入与建立数据框对象

通过 `preprocess` 函数，读取xml类型文件中的标题，内容，类别，发布时间信息，其中发布时间统计了其年月日信息（publication_day_of_week和publication_day_of_month一个表示的是星期几一个表示的是日期，这里我都先保存下来了，尽管之后没太用上），多个类别用逗号分割，结果保存在 `nyt_corpus_processed.csv` 中

```
Preprocess successfully
                                                 title  ...                categories
0                  The Man on the Spot Over Rail Safety  ...              U.S., Travel
1    BEIRUT REFUGEES'S PLIGHT TERMED GRIM DESPITE A...  ...              Travel, World
2                JOHNSON'S 39 LEAD LAKERS PAST CELTICS  ...                    Sports
3                 Moore Sculpture Approved to Be Altar  ...       Travel, World, Arts
4                 POSTINGS: Circa 1821; The Vinyl Goes  ...                    Travel
..                                                 ...  ...                       ...
495                                  ROOM TO IMPROVE  ...    Home and Garden, Style
496  Criminal Charges Are Expected Against Marines,...  ...  Travel, World, Washington
497                            Notes From an Ex-Senator  ...   Books, Arts, Washington
498                                  METROPOLITAN DIARY  ...       New York and Region
499  Arts, Briefly; Changing of the Guard At the Vi...  ...                      Arts

[500 rows x 7 columns]
```

对新闻全文进行预处理

这一步也实现在函数 `preprocess` 中，我对正文部分进行了去除标点符号、停用词、数字、空白字符，将大写字母都转化为小写，以及词干化处理

```
content = ''
for block in content_blocks:
    for p in block.findall(".//p"):
        if p.text is not None:
            content += p.text.strip() + ' '

content = content.strip()

content = re.sub(r'[^\\w\\s]', '', content)
content = re.sub(r'\\d+', '', content)
content = content.lower()

words = word_tokenize(content)

words = [word for word in words if word not in stop_words]

words = [porter.stem(word) for word in words]

processed_content = ' '.join(words)
```

将每一篇新闻的全文表示成 BagOfWords 向量

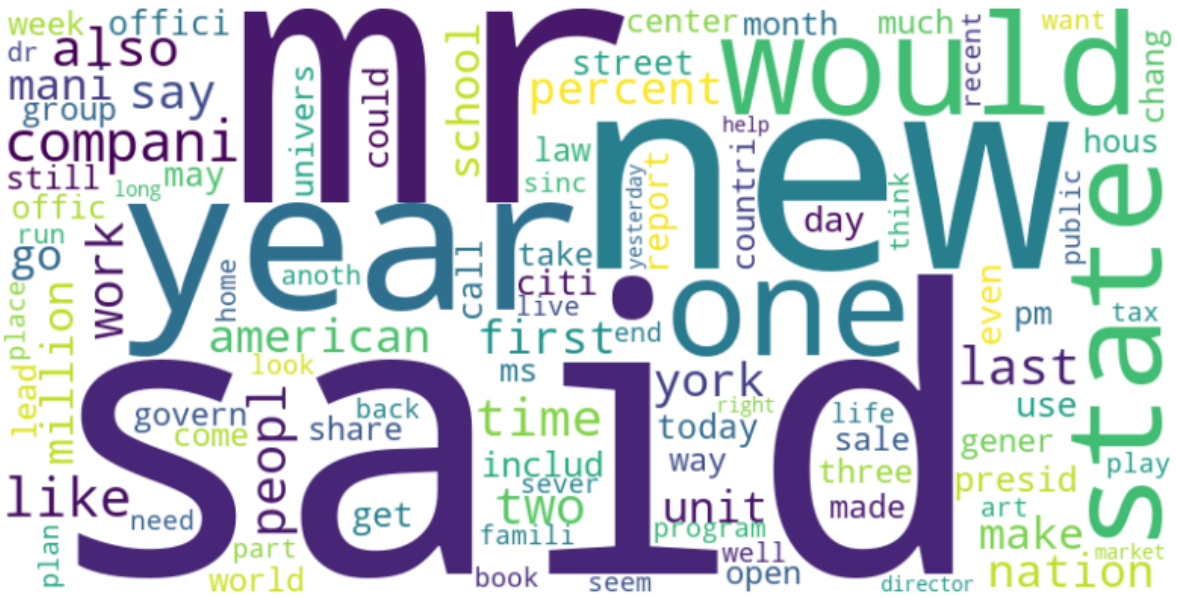
Bag of Words（BoW）是一种用于文本处理和自然语言处理的基本技术。它将文本表示为一个无序的词汇集合，忽略了文本中单词的顺序和语法。每个文档或句子被看作是一个袋子，里面装着各种单词，而且每个单词的出现都是独立的，没有考虑它们的上下文关系，基于上述定义，我将新闻中的正文内容处理为词袋向量

```
Wordbag build successfully
                                                 title  publication_day_of_month  publication_month  ...  zurich zygmont éragni
0                  The Man on the Spot Over Rail Safety                        20                  1  ...       0       0      0
1    BEIRUT REFUGEES'S PLIGHT TERMED GRIM DESPITE A...                        15                  2  ...       0       0      0
2                JOHNSON'S 39 LEAD LAKERS PAST CELTICS                        16                  2  ...       0       0      0
3                  Moore Sculpture Approved to Be Altar                        19                  2  ...       0       0      0
4                  POSTINGS: Circa 1821; The Vinyl Goes                         1                  3  ...       0       0      0
..                                                 ...                       ...                ...  ...     ...     ...    ...
495                                ROOM TO IMPROVE                        19                  4  ...       0       0      0
496  Criminal Charges Are Expected Against Marines,...                        27                  4  ...       0       0      0
497                          Notes From an Ex-Senator                        29                  4  ...       0       0      0
498                                METROPOLITAN DIARY                         4                  6  ...       0       0      0
499  Arts, Briefly; Changing of the Guard At the Vi...                         7                  6  ...       1       0      0

[500 rows x 16680 columns]
```

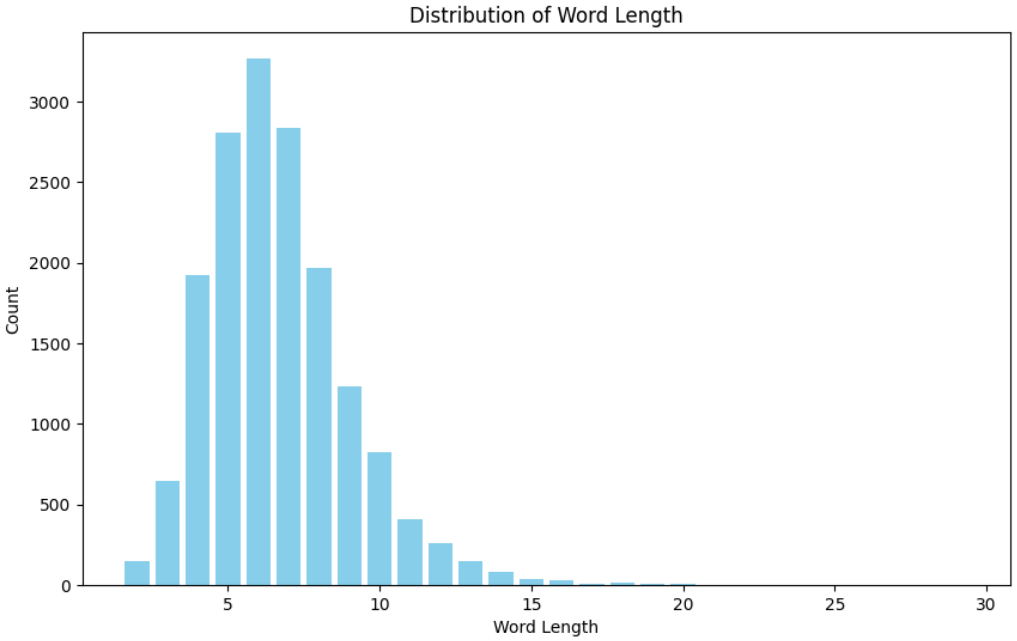
“云图”绘制

统计词频信息并绘制云图如下，可以看到 said 是出现频率最高的词，compani这样的词出现是由于词频的统计是基于之前的词干化结果，因此有一些单词的形式和常见形式有所差异



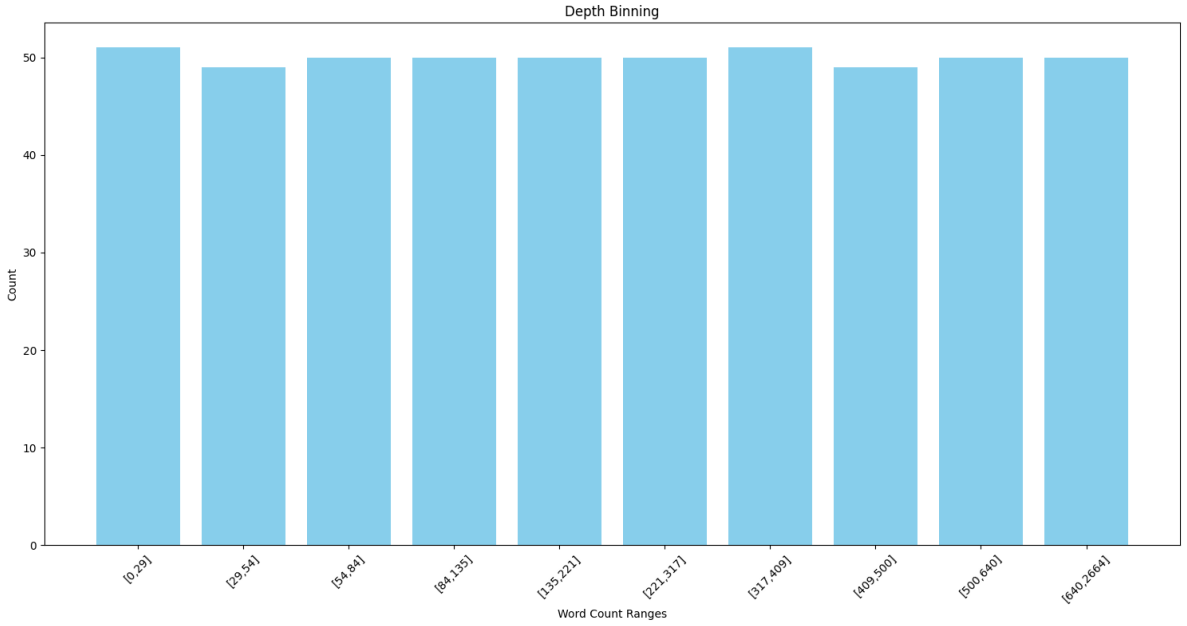
单词长度柱状图

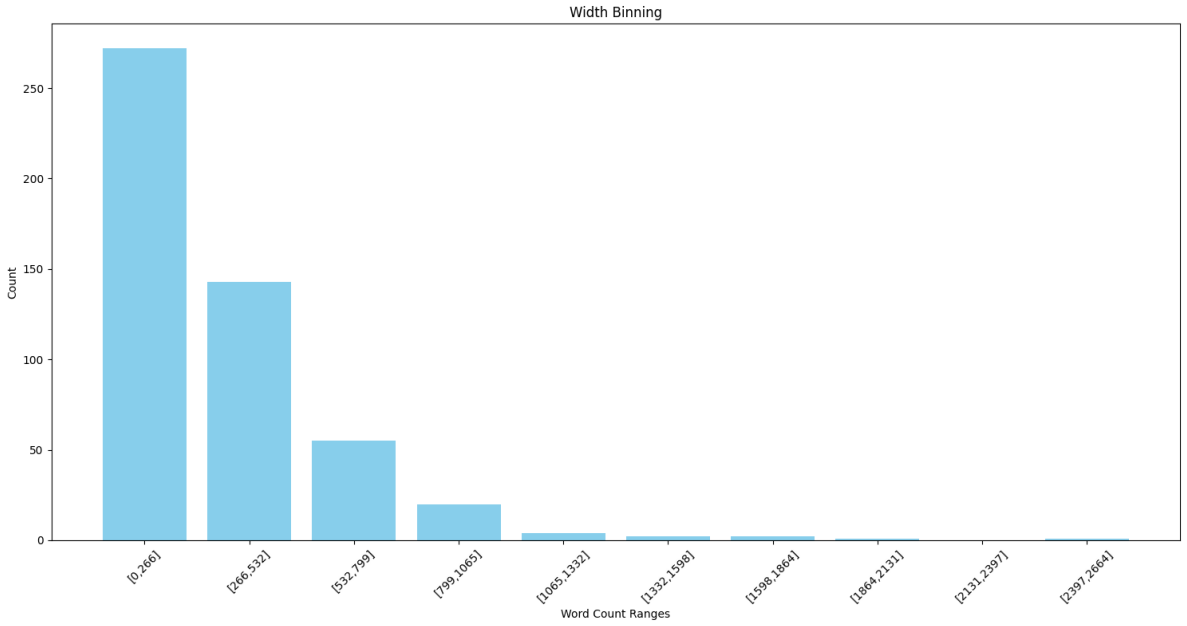
可以看到，大部分的词长度都在5~10之间



全文单词数等深分箱和等宽分箱

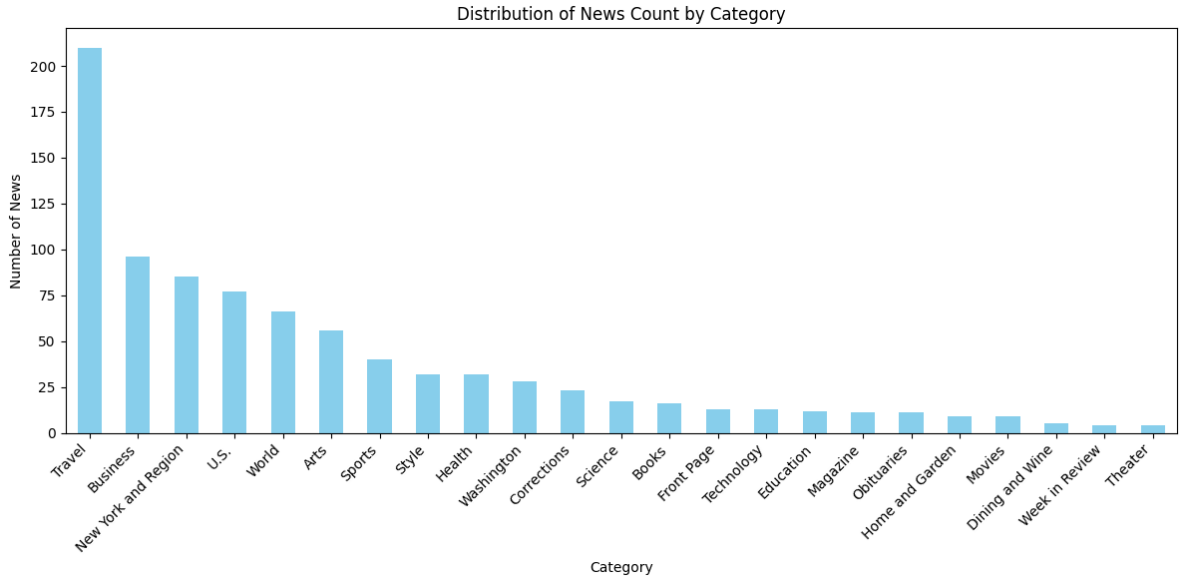
等深和等宽分箱绘制如下，可以看出，大部分新闻词量都在300以下





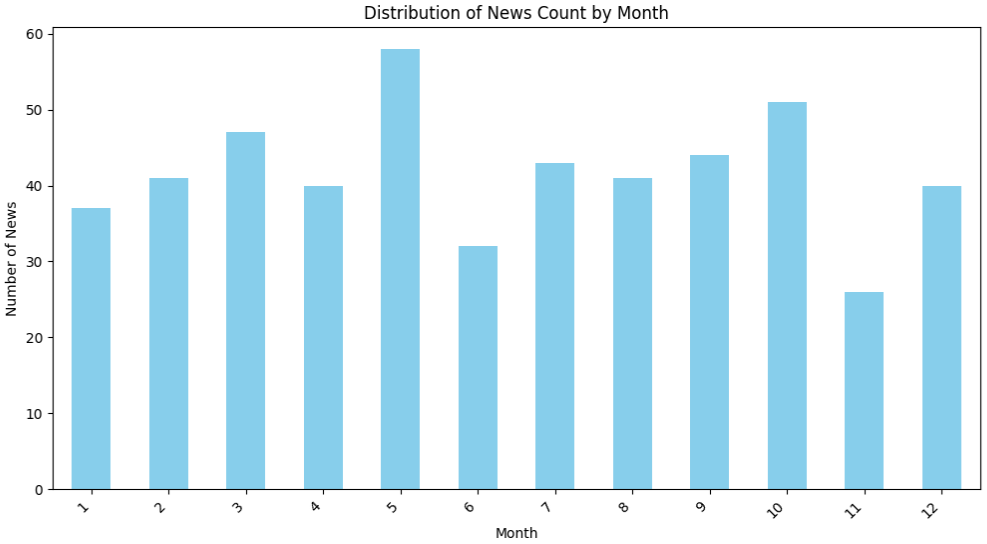
基于类别的新闻数量的分布情况

可以看到，Travel类别的新闻数量最多，其余常见类别还包括business，国家地区名等，类别分布具有比较明显的长尾性



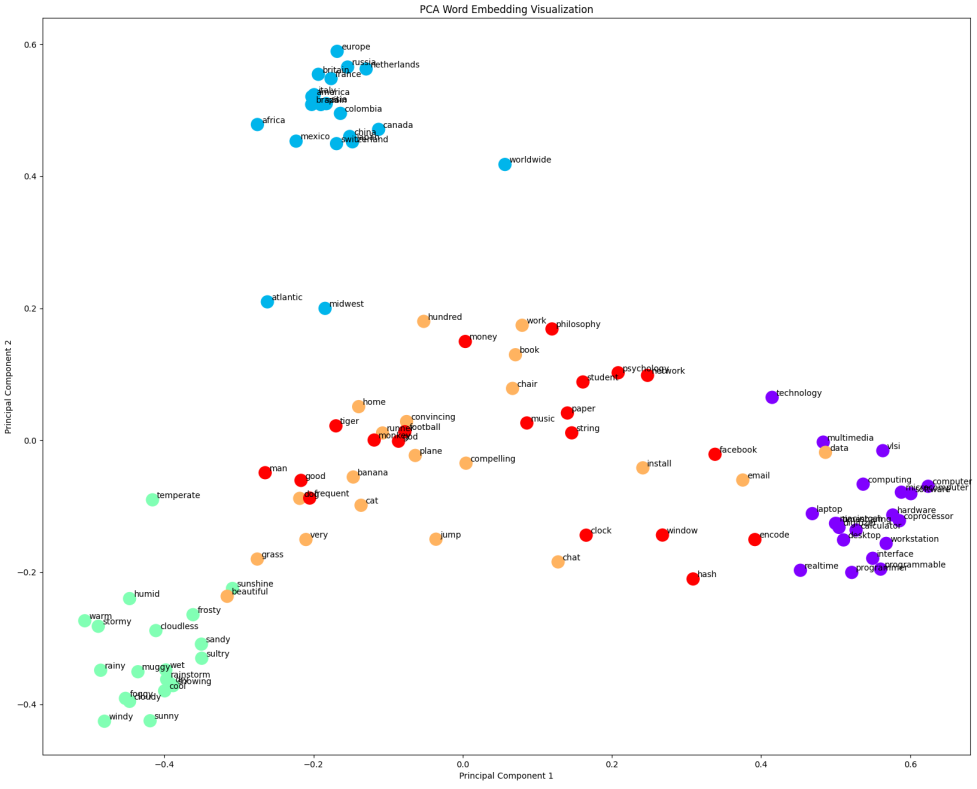
基于月份的新闻数量的分布情况

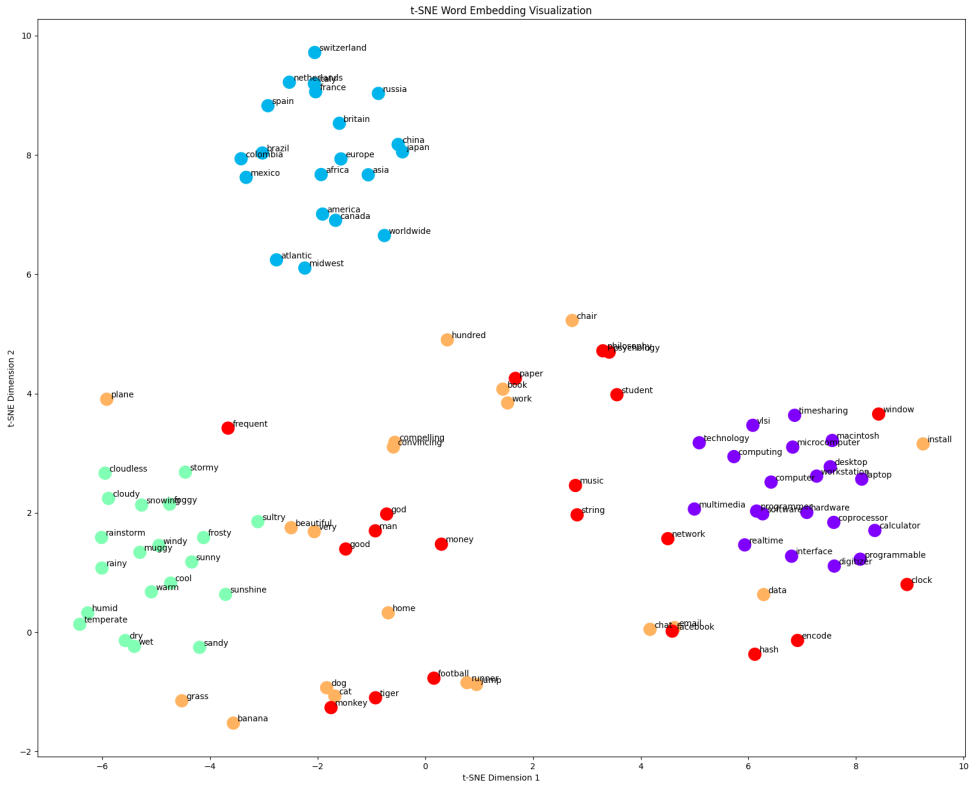
基于月份的新闻数量分布情况如下，可以看到分布相对均匀，其中5月份和10月份最多，1-5月和6-10月两段的分布有比较明显的相似规律，可能和新闻运营的规律有关



高维向量可视化

本节中我采用tsne和pca两种方法对高维词向量进行了降维压缩处理并以散点的形式绘制在了平面图上，通过对词的观察，我发现这100个词大致成每20个一组的分类，比如0-20为cs相关术语，20-40为地理词汇，40-60为天气相关的词汇，而60-100的类别相对模糊不清，为了便于观察发现规律我将他们按照每二十个一组进行了染色





分析：

- 可以看到，两种方法都可以对词向量进行有效的降维处理，特别是在分类特征上，前三组类别明显的词经过两种方法压缩得到的散点也呈现出比较好的聚类效果（蓝色紫色和青色），而后两组特征不是很明显的词则混在了一起，这和我们的词义的观察是一致的
- 尽管 t-SNE 是对 PCA 方法的一种改进，但仅从分类出发似乎二者都能在对词向量降维的同时保留较多的语义特征，经过多次实验，我发现 t-SNE 相比 PCA 对于同类的词降维后得到的点不会过于聚集，相比之下PCA在处理地理相关词汇时容易把所有词都压缩到一个很小的范围内，这样做可能会导致同类词的差异在低维空间中被抹去，导致一定程度上的信息丢失。
- t-SNE 相比 PCA 在捕捉词的细节关联上更有优势，例如，在后40个类别不明显的词中，词对（“hash”，“encode”），（“god”，“man”）等在t-SNE中距离要比PCA更近，说明 t-SNE 在保留词的细节特征上更有优势
- 通过实验我发现 t-SNE 有更多可以调的参数，这些参数会很大程度上影响最后的降维效果，而且算法的随机性更大，这一方面可能使其不适用于需要确定性结果的场景，但另一方面也提升了输出的多样性，在机器学习或深度学习等场景中可能相比PCA更适用。