# 实验一

Word2vec是Google在2013年开源的一款用于词向量计算的工具，一经发布就引起了工业界和学术界的关注。首先，Word2vec可以在百万数量级的词典和上亿的数据集上进行高效地训练；其次，该工具训练得到的词向量（word embedding），可以很好地度量词与词之间的相似性。Word2vec不是一种深度学习算法，其后面只是一个浅层神经网络，包含两种模型：CBOW模型和Skip-gram模型。

本章实验主要是基于Python和gensim框架实现Word2vec在Wikipedia语料集上面的应用，并且获取词的词向量以及寻找相近词。

## 实验预备知识

* 熟悉gensim框架中Word2vec子模块
* 相应Python语言实战基础，对Python基础的数据格式，数据类型以及相关的字符处理函数有所了解，并熟悉对文件的读写操作
* 对Word2vec技术的相关理论有所了解，该方法常用于文本数据向量化，且往往在后续任务中效果更佳。

## 数据集

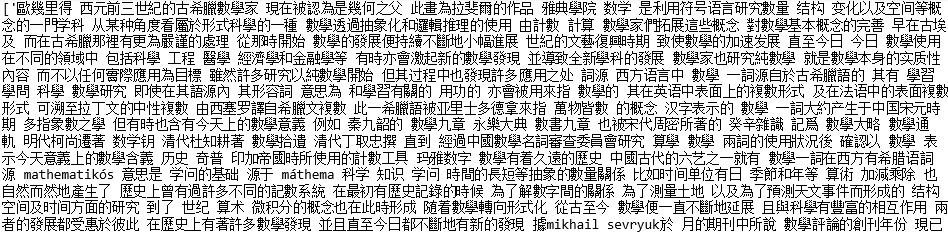
Wikipedia语料集:[zhwiki-latest-pages-articles.xml.bz2](https://www.cnblogs.com/chenbjin/p/5635853.html)

## 参考与提示

1. [中文维基百科语料上的Word2Vec实验](https://blog.csdn.net/yangyangrenren/article/details/56286394)
2. 过程的文件自己进行命名
3. 步骤提示：

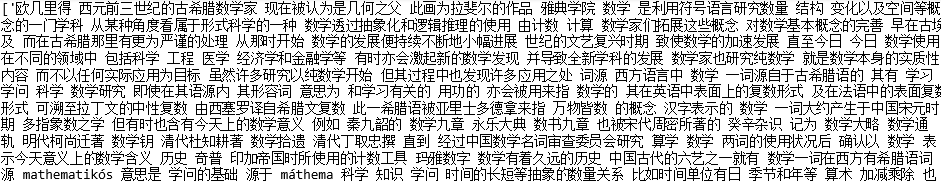
* 安装所需的包
* 导入包和模块
* 加载数据并提取内容

提示：使用WikiCorpus类中的get\_texts()方法读取文件，每篇文章转换为一行文本，并去掉标签符号等内容，并查看结果



* 预处理：

1. 繁体字转换简体字，zhconv.convert(line, 'zh-hans')



1. 分词

**参考操作**（部分）：

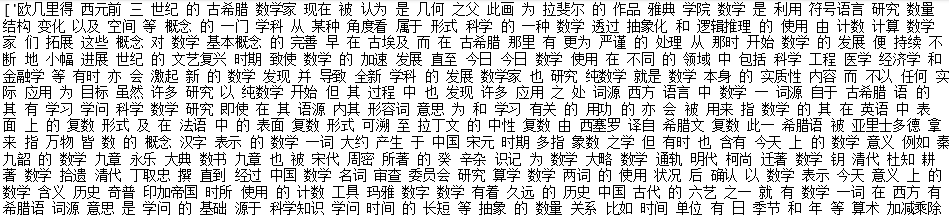
' '.join(jieba.cut(line.split('\n')[0].replace(' ', ''))) + '\n'

1. 去除非中文词

参考操作（部分）：

cn\_reg = '^[\u4e00-\u9fa5]+$' #符合中文的正则表达式，^：字符串开始位置； $：字符串结尾位置

if re.search(cn\_reg, word) #查看是否是中文



1. 训练词向量并保存模型

参考操作

先加载之前的处理后的txt文件

训练后保存模型

word2vec.LineSentence(input\_file\_name)

word2vec.Word2Vec(……#参数设置)

请自行查阅操作方法

1. 加载模型测试

word2vec.Word2Vec.load(model\_path)

wordvec.wv.get\_vector("华为")

#获得与"华为"最相近的5个词

wordvec.wv.most\_similar("华为",topn=5)

#请自行查阅函数意义