# Word2vec

Word2vec，是一群用来产生词向量的相关模型。这些模型为浅而双层的神经网络，用来训练以重新建构语言学之词文本。网络以词表现，并且需猜测相邻位置的输入词，在word2vec 中词袋模型假设下，词的顺序是不重要的。训练完成之后，word2vec 模型可用来映射每个词到一个向量，可用来表示词对词之间的关系，该向量为神经网络之隐藏层。

随着计算机应用领域的不断扩大，自然语言处理受到了人们的高度重视。机器翻译、语音识别以及信息检索等应用需求对计算机的自然语言处理能力提出了越来越高的要求。为了使计算机能够处理自然语言，首先需要对自然语言进行建模。自然语言建模方法经历了从基于规则的方法到基于统计方法的转变。从基于统计的建模方法得到的自然语言模型称为统计语言模型。有许多统计语言建模技术，包括 n－gram、神经网络以及 log_linear 模型等。在对自然语言进行建模的过程中，会出现维数灾难、词语相似性、模型泛化能力以及模型性能等问题。寻找上述问题的解决方案是推动统计语言模型不断发展的内在动力。在对统计语言模型进行研究的背景下，Google 公司在 2013 年开放了 Word2vec 这一款用于训练词向量的软件工具。Word2vec 可以根据给定的语料库，通过优化后的训练模型快速有效地将一个词语表达成向量形式，为自然语言处理领域的应用研究提供了新的工具。Word2vec 依赖 skip-grams 或连续词袋（CBOW）来建立神经词嵌入。Word2vec 为托马斯·米科洛夫（Tomas Mikolov）在 Google 带领的研究团队创造。该算法渐渐被其他人所分析和解释。

# 依赖

## 词袋模型

词袋模型（Bag-of-words model）是个在自然语言处理和信息检索(IR)下被简化的表达模型。此模型下，像是句子或是文件这样的文字可以用一个袋子装着这些词的方式表现，这种表现方式不考虑文法以及词的顺序。最近词袋模型也被应用在计算机视觉领域。词袋模型被广泛应用在文件分类，词出现的频率可以用来当作训练分类器的特征。关于"词袋"这个用字的由来可追溯到泽里格·哈里斯于 1954 年在 Distributional Structure 的文章。

## Skip-gram 模型

Skip-gram 模型是一个简单但却非常实用的模型。在自然语言处理中，语料的选取是一个相当重要的问题：第一，语料必须充分。一方面词典的词量要足够大，另一方面要尽可能多地包含反映词语之间关系的句子，例如，只有“鱼在水中游”这种句式在语料中尽可能地多，模型才能够学习到该句中的语义和语法关系，这和人类学习自然语言一个道理，重复的次数多了，也就会模仿了； 第二，语料必须准确。 也就是说所选取的语料能够正确反映该语言的

语义和语法关系，这一点似乎不难做到，例如中文里，《人民日报》的语料比较准确。 但是，更多的时候，并不是语料的选取引发了对准确性问题的担忧，而是处理的方法。 n 元模型中，因为窗口大小的限制，导致超出窗口范围的词语与当前词之间的关系不能被正确地反映到模型之中，如果单纯扩大窗口大小又会增加训练的复杂度。Skip-gram 模型的提出很好地解决了这些问题。顾名思义，Skip-gram 就是“跳过某些符号”，例如，句子“中国足球踢得真是太烂了”有 4 个 3 元词组，分别是“中国足球踢得”、“足球踢得真是”、“踢得真是太烂”、“真是太烂了”，可是我们发现，这个句子的本意就是“中国足球太烂”可是上述 4 个 3 元词组并不能反映出这个信息。Skip-gram 模型却允许某些词被跳过，因此可以组成“中国足球太烂”这个 3 元词组。 如果允许跳过 2 个词，即 2-Skip-gram 。

# 实验结果

## 神雕侠侣（杨过）:

[('小龙女', 0.9683679342269897), ('陆无双', 0.9629747271537781), ('黄蓉', 0.9519241452217102), ('周伯通', 0.9390044808387756), ('武三通', 0.9183982610702515), ('郭靖', 0.9177126884460449), ('武修文', 0.9151492118835449), ('裘千尺', 0.9151337146759033), ('一眼', 0.9079431891441345), ('一灯', 0.9058586955070496)]

## 鹿鼎记（韦小宝）:

[('忙', 0.9307552576065063), ('康熙', 0.9267476201057434), ('海老公', 0.9074211120605469), ('陈近南', 0.887366771697998), ('多隆', 0.8853891491889954), ('点头', 0.8850342631340027), ('低声', 0.883967936038971), ('大声', 0.8771711587905884), ('双儿', 0.8762736916542053), ('微笑', 0.8744578957557678)]

## 倚天屠龙记（张无忌）:

[('张翠山', 0.970447838306427), ('周芷若', 0.9606700539588928), ('赵敏', 0.948734700679779), ('殷素素', 0.92494797706604), ('谢逊', 0.9176401495933533), ('杨不悔', 0.8957116603851318), ('怔', 0.8930292129516602), ('一眼', 0.8905831575393677), ('忙', 0.875346302986145), ('转头', 0.8731189370155334)]

## 射雕英雄传（郭靖）:

[('欧阳锋', 0.9491536021232605), ('黄蓉', 0.944141149520874), ('陆冠英', 0.9204340577125549), ('洪七公', 0.9178668260574341), ('黄药师', 0.9069315195083618), ('裘千仞', 0.8955918550491333), ('丘处机', 0.8955711722373962), ('忙', 0.8825729489326477), ('完颜洪烈', 0.8770402669906616), ('拖雷', 0.8679450750350952)]

## 天龙八部（乔峰）:

[(' 王 夫 人 ', 0.9792421460151672), (' 阿 紫 又 ', 0.9753959774971008), (' 钟 夫 人 ', 0.9737507104873657), ('伯父', 0.9722434282302856), ('小人', 0.971569299697876), ('包不同', 0.9714356064796448), ('段誉忙', 0.9713180065155029), ('段誉笑', 0.9703710079193115), ('甥儿', 0.9696488380432129), ('段誉叹', 0.969330370426178)]

# 代码

https://github.com/Libra-ChosenOne/NLP_homework_4