实验1

词向量

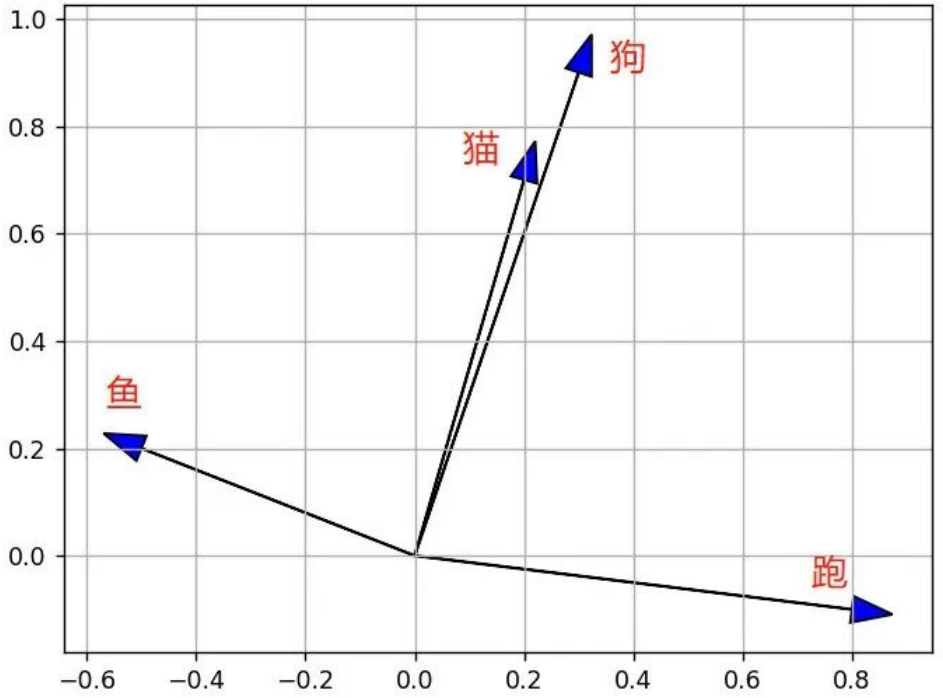
文本向量化

- 前言

文本向量化是将文本表示成一系列能够表达文本语义的向量。

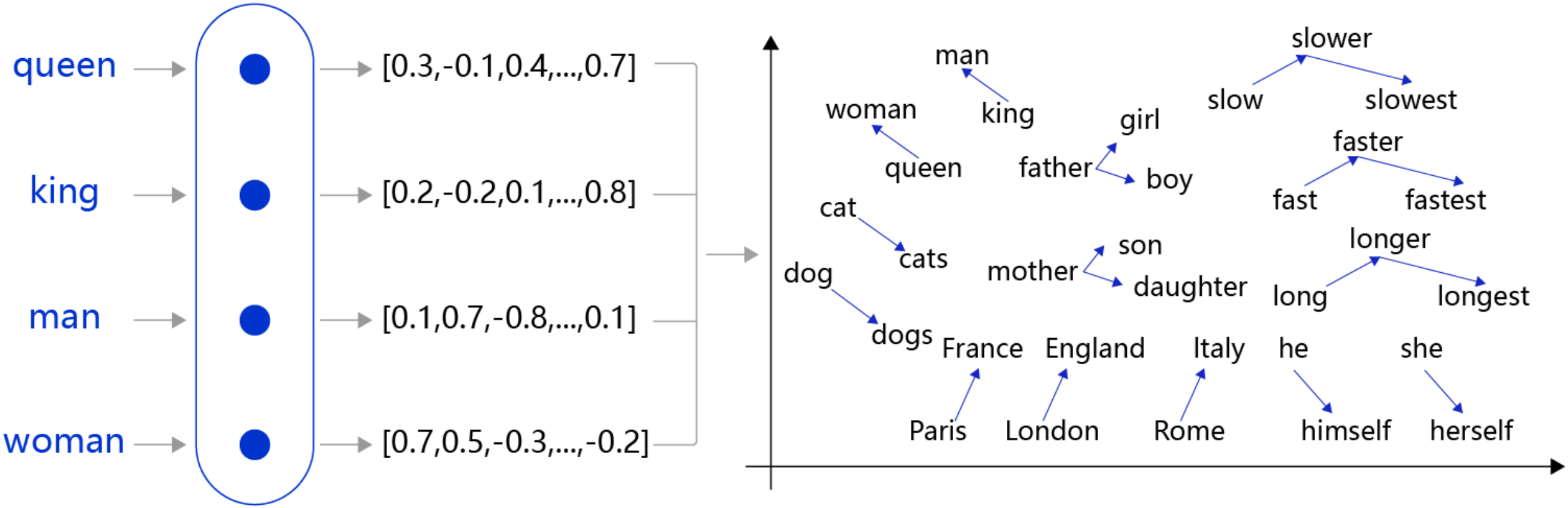
比如有四个单词：“猫”、“狗”、“鱼”、“跑”，通过向量转换可以得到如下的向量：

- 猫：[0.2, 0.7]
- 狗：[0.3, 0.9]
- 鱼：[-0.5, 0.2]
- 跑：[0.8, -0.1]



将四个向量画在坐标图上

• 词向量模型

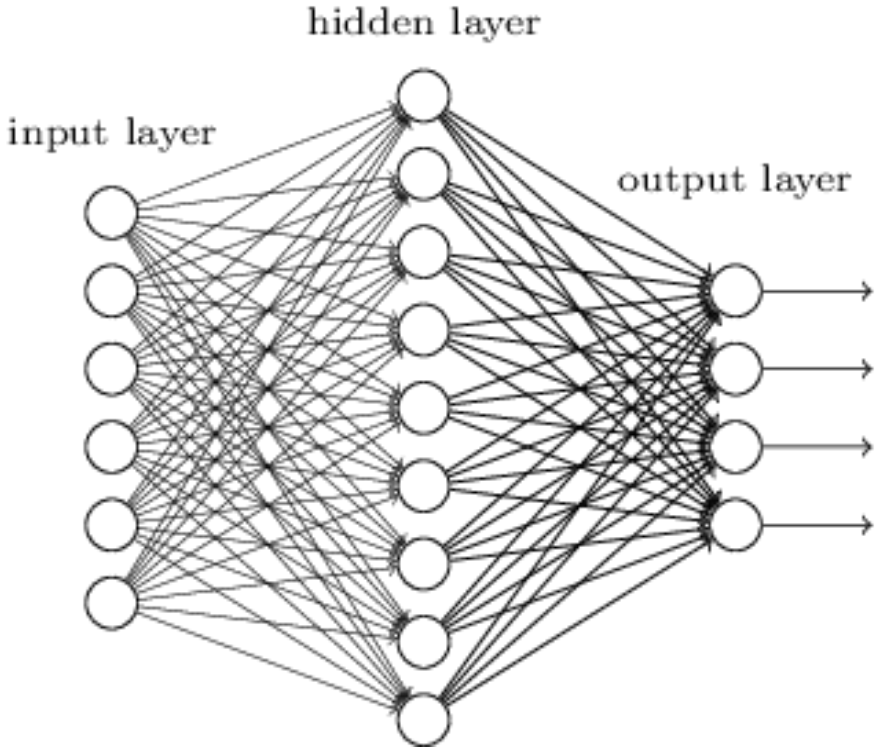


在自然语言处理任务中，词向量（Word Embedding）是表示自然语言里单词的一种方法，即把每个词都表示为一个N维空间内的点，即一个高维空间内的向量。通过这种方法，实现把自然语言计算转换为向量计算。

- one-hot表示
  - one-hot表示用一个长的向量表示一个词，向量长度为词典的大小，每个向量只有一个维度为1，其余维度全部为0，为1的位置表示该词语在词典的位置。
  - 例如，有两句话“小张喜欢看电影，小王也喜欢。”和“小张也喜欢看足球比赛。”。
  - 首先对这两句话分词后构造一个词典，词典的键是词语，值是ID。
  - {"小张": 1, "喜欢": 2, "也": 3, "看": 4, "电影": 5, "足球": 6, "比赛": 7, "小王": 8 }。
  - 然后根据ID值对每个词语进行向量化，用0和1代表这个词是否出现。
  - 如“小张” 的one-hot表示为[1, 0, 0, 0, 0, 0, 0, 0]
  - “小张喜欢看电影，小王也喜欢。”是什么？

- DNN训练词向量

输入是某个词，一般用one-hot表示该词（长度为词汇表长度）



输出层的神经元个数和输入相同，隐藏层再到输出层时最后需要计算每个位置的概率，使用softmax计算，每个位置代表不同的单词。

隐藏层有N个神经元，代表我们想要的词向量的维度，输入层与隐藏层全连接

- **实验数据**

中华人民共和国海商法：

第一章　　总　　则

第一条　　为了调整海上运输关系、船舶关系，维护当事人各方的合法权益，促进海上运输和经济贸易的发展，制定本法。

第二条　　本法所称海上运输，是指海上货物运输和海上旅客运输，包括海江之间、江海之间的直达运输。

本法第四章海上货物运输合同的规定，不适用于中华人民共和国港口之间的海上货物运输。

- **示例代码下载**

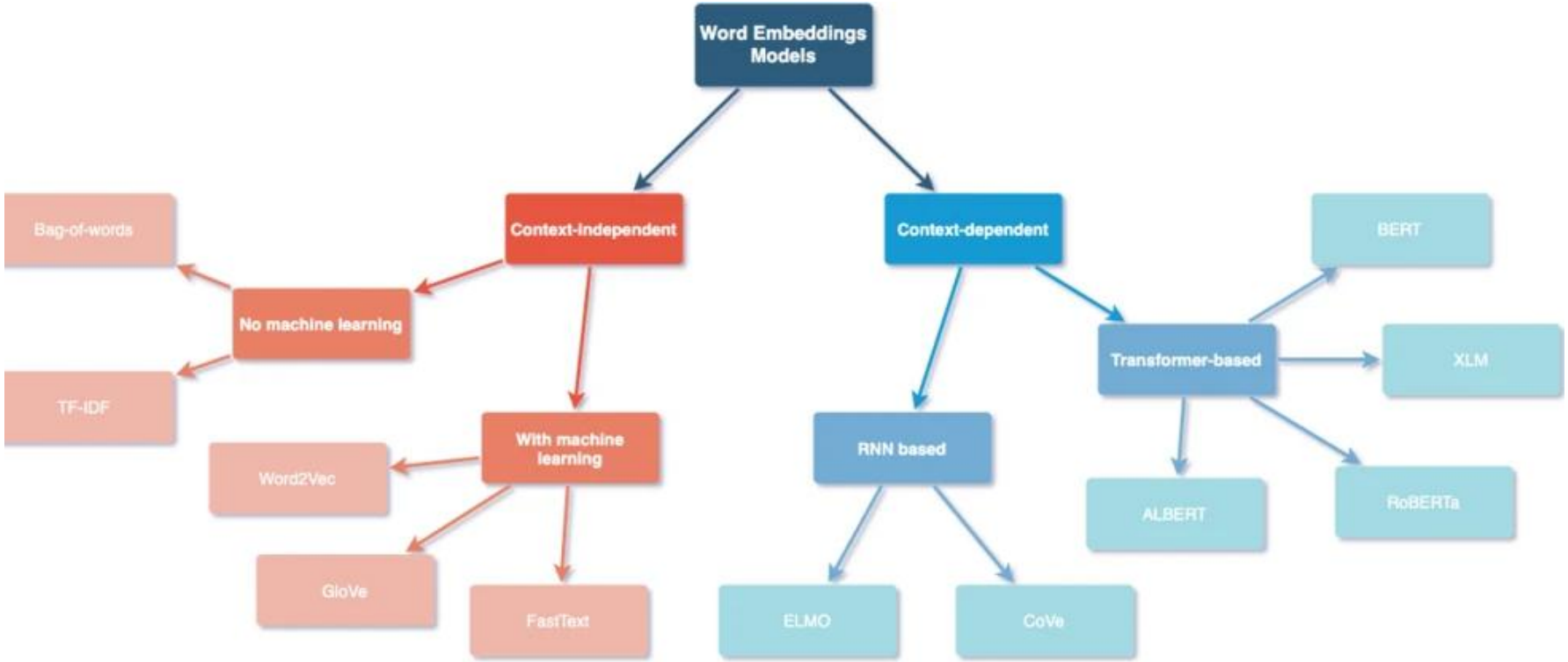
  下载代码，成功复现，可以解释代码（6-7分）

- 完善代码

增加文本预处理（1分）

- 完善代码

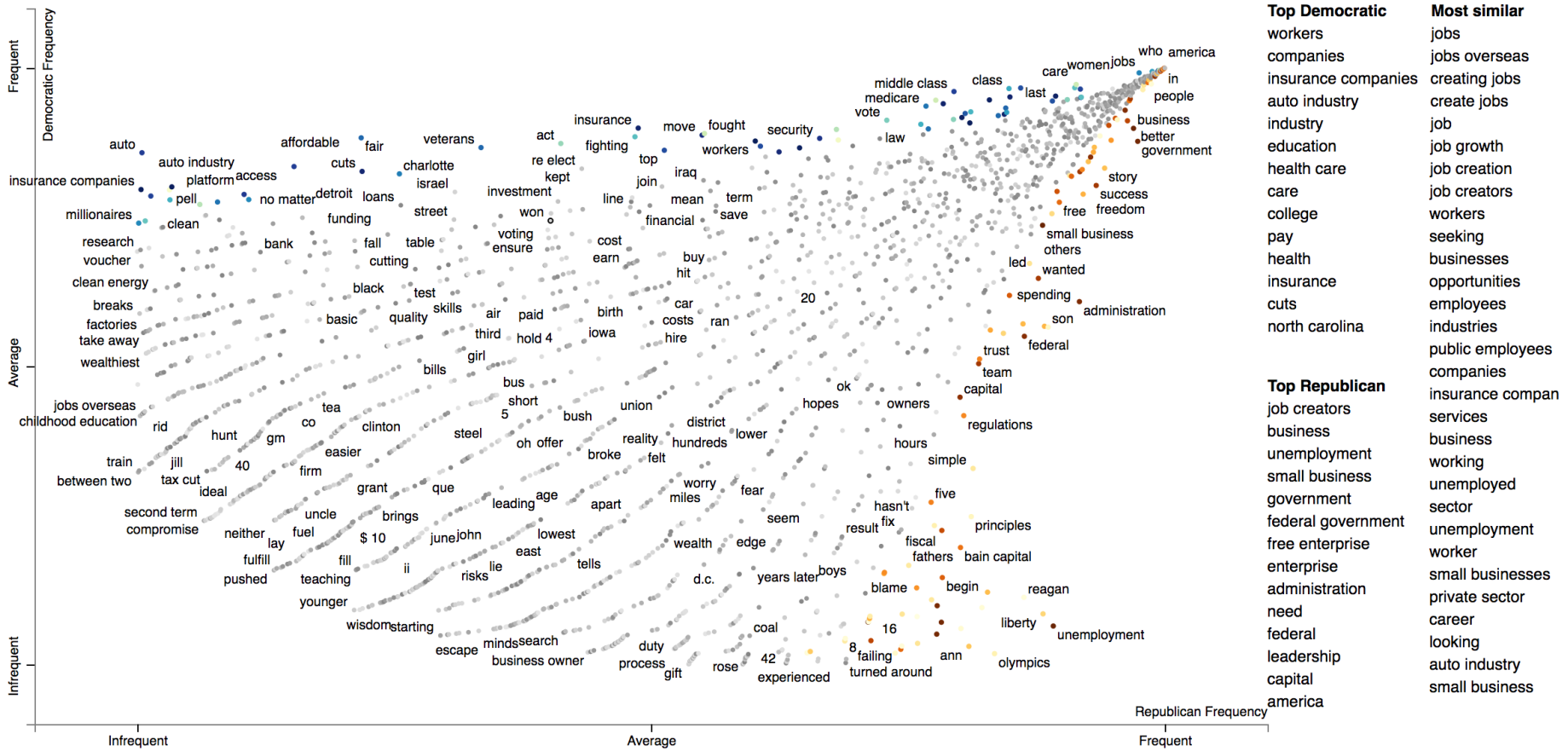
使用其他词向量模型，推荐Word2Vec，Glove，Bert
（每个模型1分）



Word2vec：https://github.com/dav/word2vec
GloVe：https://github.com/stanfordnlp/GloVe
中文bert模型：https://github.com/ymcui/Chinese-BERT-wwm

- 完善代码

词向量结果可视化，映射至低纬空间（2分）



Top Democratic
workers
companies
insurance companies
auto industry
industry
education
health care
care
college
pay
health
insurance
cuts
north carolina

Top Republican
job creators
business
unemployment
small business
government
federal government
free enterprise
enterprise
administration
need
federal
leadership
capital
america

Most similar
jobs
jobs overseas
creating jobs
create jobs
job
job growth
job creation
job creators
workers
seeking
businesses
opportunities
employees
industries
public employees
companies
insurance compan
services
business
working
unemployed
sector
unemployment
worker
small businesses
private sector
career
looking
auto industry
small business

Democratic document count: 123; word count: 76,864
Republican document count: 66; word count: 58,138

- **完善代码**

  其他优化（每个1分）

  - 改变训练模型，换成卷积神经网络，循环神经网络和长短时记忆网络等。
  - 优化训练过程，更换损失函数、优化算法，增加批量归一化等。
  - 自行设计。

谢谢

FOR YOUR LISTENING

陈雅妮.