Lab 6: Word2Vec

1 Intro

skip-gram Skip-gram算法是指在给出目标单词（中心单词）的情况下，预测它的上下文单词（除中心单词外窗口内的其他单词，本实验中窗口大小是2，也就是左右各两个单词）。

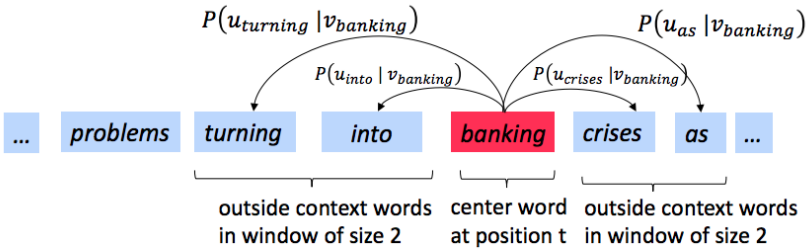

Figure 1: skip-gram

求两个词向量的相似度：设u_o和v_c分别是外部的词向量(outside vector)和中心单词的词向量(center vector)，矩阵U的每一列代表每一个外部词向量u_o，矩阵V的每一列代表所有的中心词向量v_c，U和V向量都包含了所有的单词。

使用softmax函数计算u_o, v_c的相似度为：

$$P(O = o|C = c) = \frac{exp(u_o^T v_c)}{\sum_{w \in Vocab} exp(u_w^T v_c)}$$

损失函数为：

$$J(v_c, o, U) = -logP(O = c|C = c)$$

对损失函数求导：

$$\frac{\partial J(v_c, w_{c+j}, U)}{\partial v_c} = U^T(\hat{y} - y)$$
$$\frac{\partial J(v_c, w_{c+j}, U)}{\partial U} = (\hat{y} - y)^T v_c$$

2 Environment

创建环境：

```
conda env create -f env.yml
conda activate a2
```

退出环境：

```
conda deactivate
```

3 TODO

在word2vec.py文件中完成下列任务

1. 实现sigmoid 函数
2. 实现损失函数
3. 实现损失函数求导得到gradCenterVec, gradOutsideVecs

Note: How to run

1. 终端输入‘sh get datasets.sh’，完成数据准备
2. 输入‘python run.py’即可（训练时间可能较长，会有若干小时，可以通过运行‘python word2vec.py’来检验是否有bug）

4 Submit

最终提交实验代码以及实验报告，报告中包括但不限于词向量的认识，实验生成图片的理解等。

- 2021xxxxxx_xiaoming_lab6.zip (./code ./report.pdf)
- Email xihuaw@ruc.edu.cn, DDL 2022.11.04 20:00