Word2vec给的二进制bin文件如何转成文本

发表于 **2014 年 12 月 20 日**

用word2vec工具跑词向量时有参数可选，保存为文本格式或二进制格式。而Mikilov公开的在Google News 上跑的词向量为了节省存储空间，保存为二进制了，解压前占1.5G，解压后占3.4G左右（注：超过了2G，则程序一定要在64位机子上调用GCC进行编译）。为了便于在其他地方使用，我们将其转为文本格式会很方便一些。转二进制为文本格式，需要知道二进制存储格式及内容。还好，Thomas Mensink 帮我们做了此工作：https://groups.google.com/forum/#!topic/word2vec-toolkit/5Qh-x2O1lV4，感谢 Glenn Murray 的释疑。

```
1.  //  Copyright 2013 Google Inc. All Rights Reserved.
2.  //
3.  //  Licensed under the Apache License, Version 2.0 (the ”License”);
4.  //  you may not use this file except in compliance with the License.
5.  //  You may obtain a copy of the License at
6.  //
7.  //      http://www.apache.org/licenses/LICENSE-2.0
8.  //
9.  //  Unless required by applicable law or agreed to in writing, software
10. //  distributed under the License is distributed on an ”AS IS” BASIS,
11. //  WITHOUT WARRANTIES OR CONDITIONS OF ANY KIND, either express or implied.
12. //  See the License for the specific language governing permissions and
13. //  limitations under the License.
14.
15. #include <stdio.h>
16. #include <string.h>
17. #include <math.h>
18. #include <malloc.h>
19.
20. const long long max_size = 2000;         // max length of strings
21. const long long N = 40;                  // number of closest words that will be shown
22. const long long max_w = 50;              // max length of vocabulary entries
23.
24. int main(int argc, char **argv) {
25.   FILE *f;
26.   char file_name[max_size];
27.   float len;
28.   long long words, size, a, b;
29.   char ch;
30.   float *M;
31.   char *vocab;
32.   if (argc < 2) {
33.     printf(“Usage: ./distance <FILE>\nwhere FILE contains word projections in the BINARY FORMAT\n”);
```

```
    return 0;
  }
  strcpy(file_name, argv[1]);
  f = fopen(file_name, "rb");
  if (f == NULL) {
    printf("Input file not found\n");
    return -1;
  }
  fscanf(f, "%lld", &words);
  fscanf(f, "%lld", &size);
  vocab = (char *)malloc((long long)words * max_w * sizeof(char));
  M = (float *)malloc((long long)words * (long long)size * sizeof(float));
  if (M == NULL) {
    printf("Cannot allocate memory: %lld MB    %lld  %lld\n", (long long)words * size * sizeof(float) / 1048576, words, size);
    return -1;
  }
  for (b = 0; b < words; b++) {
    fscanf(f, "%s%c", &vocab[b * max_w], &ch);
    for (a = 0; a < size; a++) fread(&M[a + b * size], sizeof(float), 1, f);
    len = 0;
    for (a = 0; a < size; a++) len += M[a + b * size] * M[a + b * size];
    len = sqrt(len);
    for (a = 0; a < size; a++) M[a + b * size] /= len;
  }
  fclose(f);
  //Code added by Thomas Mensink
  //output the vectors of the binary format in text
  printf("%lld %lld #File: %s\n",words,size,file_name);
  for (a = 0; a < words; a++){
    printf("%s ",&vocab[a * max_w]);
    for (b = 0; b< size; b++){ printf("%f ",M[a*size + b]); }
    printf("\n");
  }

  return 0;
}
```

替换掉word2vec中的distance.c文件，进行make。

比较直接的方式 就是执行如下指令

./distance ori.bin tar.txt

即可~~~~

此条目由 **jacoxu** 发表在 **Deep Learning** 分类目录，并贴了 **Word2vec** 标签。将**固定链接**
[http://jacoxu.com/word2vec%e7%bb%99%e7%9a%84%e4%ba%8c%e8%bf%9b%e5%88%b6bin%e6%96%87%e4%bb%b6%e5%a6%82%e4%bd%95%e8%bd%ac%e6%88%90%e6%96%87
加入收藏夹。

《WORD2VEC给的二进制BIN文件如何转成文本》上有 2 条评论

陈缘
在 **2017 年 5 月 12 日上午 11:13** 说道：

你好，请问我执行./distance ori.bin tar.txt命令 报错：bash: ./distance: 没有那个文件或目录
请问这是为什么？

jacoxu
在 **2017 年 5 月 22 日上午 9:25** 说道：

你好，请问是否成功make了distance.c文件并得到distance文件？

?>