```
{
 "cells": [
  {
   "cell_type": "code",
   "execution_count": 5,
   "metadata": {},
   "outputs": [],
   "source": [
    "import requests;\n",
    "from bs4 import BeautifulSoup;\n",
    "\n",
    "#定义获取豆瓣影评数据的方法\n",
    "def getComment(url,commentList):\n",
    "     #1.1---添加请求头(为了伪装的更像)\n",
    "     header={\n",
    "                         \"User-Agent\":\"Mozilla/5.0  (Windows  NT  10.0;  Win64;  x64)
AppleWebKit/537.36 (KHTML, like Gecko) Chrome/80.0.3987.132 Safari/537.36\"\n",
    "     }\n",
    "     #1.2---发请求，获取响应\n",
    "     response = requests.get(url=url,headers = header);\n",
    "     #1.3---使用 bs4 跟 html5lib 解析网页内容\n",
    "     if(response.status_code==200):\n",
    "          soup = BeautifulSoup(response.content,\"html5lib\");\n",
    "          #1.4---获取所有存放评论区域的 div,class=\"mod-hd\"\n",
    "          commentItemList = soup.find_all(\"div\",attrs={\"class\":\"comment-item\"});\n",
    "          #1.5---遍历列表，获取每一个评论的作者、打分和正文\n",
    "          for commentItem in commentItemList:\n",
    "               #1.6---获取存放作者、打分、评论正文的 div,class=\"comment\"\n",
    "               comment = commentItem.find(\"div\",attrs={\"class\":\"comment\"});\n",
    "               #1.7---获取存放作者、打分的 div,class=\"comment-info\";\n",
    "                                                                  commentInfo      =
comment.find(\"span\",attrs={\"class\":\"comment-info\"});\n",
    "               #1.8---获取作者名字   None\n",
    "               author = commentInfo.find(\"a\").text;\n",
    "               #1.9---获取打分   因为豆瓣机制中是可以不打分的\n",
    "               star = commentInfo.find_all(\"span\")[1].get(\"title\");\n",
    "               #1.10---获取评论正文  span class=\"short\"\n",
    "                                                                  commentText      =
comment.find(\"span\",attrs={\"class\":\"short\"}).text.replace(\"\\n\",\"\");\n",
    "               #1.11---将作者、打分、评论拼接成列表\n",
    "               yingping = [author,star,commentText]\n",
    "               #1.12---将每条评论到添加到 commentList 中\n",
    "               commentList.append(yingping);\n",
    "          return commentList;\n",
    "\n",
    "import csv;\n",
    "#将评论写入到 csv 文件中\n",
    "def writeComment(commentList):\n",
    "    with open(\"流浪地球.csv\",\"w\",newline=\"\",encoding=\"utf-8\") as file:\n",
    "        csvWriter = csv.writer(file);\n",
    "        csvWriter.writerows(commentList);\n",
    "\n",
    "\n",
    "if __name__ == '__main__':\n",
    "    #  定义存储评论的列表\n",
    "    commentList = []\n",
    "    for i in range(10):\n",
    "                                                                        baseUrl        =
\"https://movie.douban.com/subject/26266893/comments?start=%d&limit=20&sort=new_score
&status=P\" % (\n",
    "                      i * 20)\n",
    "        #调用获取影评的方法\n",
    "        commentList = getComment(baseUrl,commentList);\n",
    "    #调用写入 csv 的方法\n",
    "    writeComment(commentList);"
   ]
  },
  {
   "cell_type": "code",
   "execution_count": 36,
   "metadata": {},
   "outputs": [],
   "source": [
    "import csv;\n",
    "import jieba;\n",
    "from wordcloud import WordCloud;\n",
    "from PIL import Image;\n",
    "import numpy;\n",
    "def readData():\n",
    "    #  豆瓣的 1 星到 5 星分别代表什么：很差、较差、还行、推荐、力荐\n",
    "    stars = (\"很差\",\"较差\",\"还行\",\"推荐\",\"力荐\");\n",
    "    #定义列表，存储最终的评论结果\n",
    "    commentList = [];\n",
    "    #  读取 csv 文件内容\n",
    "    with open(\"流浪地球.csv\",\"r\",encoding=\"utf-8\") as file:\n",
    "        #  获取 csv 的读编辑对象\n",
    "        csvReader = csv.reader(file);\n",
    "        #  遍历所有评论\n",
    "        for item in csvReader:\n",
    "            #不要没打分的评论\n",
    "            if(item[1] in stars):\n",
    "                commentList.append(item[2]);\n",
    "        return commentList\n",
    "  #定义生成词云图的方法\n",
    "def generateWordCloud():\n",
    "    #定义最终评论变量\n",
    "    finalComment = \"\";\n",
    "    #1.1---读取所有评论\n",
    "    comments = readData();\n",
    "    for comment in comments:\n",
    "        finalComment+=comment;\n",
    "    #1.2---将所有的评论拼接成一个完整的字符串，再做分词\n",
    "    finalComment = \" \".join(jieba.cut(finalComment));\n",
    "    #1.3---读取并设置词云轮廓\n",
    "    image = numpy.array(Image.open(\"tangguo.jpg\"));\n",
    "    #1.4---生成词云对象\n",
    "    word = WordCloud(\n",
    "        font_path= \"C:/Windows/Fonts/simhei.ttf\",\n",
    "        background_color= \"white\",\n",
    "        mask= image\n",
    "    ).generate(finalComment)\n",
    "    #1.5---生成本地词云文件\n",
    "    word.to_file(\"流浪地球.jpg\");\n",
    "\n",
    "if __name__ == '__main__':\n",
    "    generateWordCloud();\n",
    "    pass;"
   ]
  },
  {
   "cell_type": "code",
   "execution_count": null,
   "metadata": {},
   "outputs": [],
   "source": []
  }
 ],
 "metadata": {
  "kernelspec": {
   "display_name": "Python 3",
   "language": "python",
   "name": "python3"
  },
  "language_info": {
   "codemirror_mode": {
    "name": "ipython",
    "version": 3
   },
   "file_extension": ".py",
   "mimetype": "text/x-python",
   "name": "python",
   "nbconvert_exporter": "python",
   "pygments_lexer": "ipython3",
   "version": "3.6.4"
  }
 },
 "nbformat": 4,
 "nbformat_minor": 2
}
```