HW2.1

张劲暾 PB16111485

1. 下载或者在线测试一到两个开源的中文分词工具，测试一下其对古汉语和古诗词的分词表现，给出测试结果和截图。然后讨论一下古汉语和古诗词分词与现代汉语分词相比的难点主要是什么？

- 分词代码：

```
import jieba
file = open('./poems')
file_cut = open('./poems_cut','w')
for line in file:
    file_cut.write(str(list(jieba.cut(line.replace('
','').replace(', ','').replace('。','').replace('; ','').replace('? ','').replace('! ','')
).replace('》','').replace('《','').replace('、','').replace('\n',''))))
    file_cut.write('\n')
```

- 原文：

> 《赴戍登程口占示家人》 力微任重久神疲，再竭衰庸定不支。 苟利国家生死以，岂因祸福避趋之？ 谪居正是君恩厚，养拙刚于戍卒宜。 戏与山妻谈故事，试吟断送老头皮。
>
> 秦孝公据崤函之固，拥雍州之地，君臣固守以窥周室，有席卷天下，包举宇内，囊括四海之意，并吞八荒之心。 当是时也，商君佐之，内立法度，务耕织，修守战之具，外连衡而斗诸侯。于是秦人拱手而取西河之外。 孝公既没，惠文、武、昭襄蒙故业，因遗策，南取汉中，西举巴、蜀，东割膏腴之地，北收要害之郡。诸侯恐惧，会盟而谋弱秦，不爱珍器重宝肥饶之地，以致天下之士，合从缔交，相与为一。 当此之时，齐有孟尝，赵有平原，楚有春申，魏有信陵。此四君者，皆明智而忠信，宽厚而爱人，尊贤而重士，约从离衡，兼韩、魏、燕、楚、齐、赵、宋、卫、中山之众。于是六国之士，有宁越、徐尚、苏秦、杜赫之属为之谋，齐明、周最、陈轸、 召滑、楼缓、翟景、苏厉、乐毅之徒通其意，吴起、孙膑、带佗、倪良、王廖、田忌、廉颇、赵奢之伦制其兵。尝以十倍之地，百万之众， 叩关而攻秦。秦人开关延敌，九国之师，逡巡而不敢进。秦无亡矢遗镞之费，而天下诸侯已困矣。于是从散约败，争割地而赂秦。 秦有余力而制其弊，追亡逐北，伏尸百万，流血漂橹；因利乘便，宰割天下，分裂山河。强国请服，弱国入朝。 延及孝文王、庄襄王，享国之日浅，国家无事。 及至始皇，奋六世之余烈，振长策而御宇内，吞二周而亡诸侯，履至尊而制六合，执敲扑而鞭笞天下，威振四海。 南取百越之地，以为桂林、象郡；百越之君，俯首系颈，委命下吏。乃使蒙恬北筑长城而守藩篱，却匈奴七百余里； 胡人不敢南下而牧马，士不敢弯弓而报怨。于是废先王之道，焚百家之言，以愚黔首；隳名城，杀豪杰；收天下之兵， 聚之咸阳，销锋镝，铸以为金人十二，以弱天下之民。然后践华为城，因河为池，据亿丈之城，临不测之渊，以为固。 良将劲弩守要害之处，信臣精卒陈利兵而谁何。天下已定，始皇之心，自以为关中之固，金城千里，子孙帝王万世之业也。 始皇既没，余威震于殊俗。然陈涉瓮牖绳枢之子，氓隶之人，而迁徙之徒也；才能不及中人，非有仲尼，墨翟之贤， 陶朱、猗顿之富；蹑足行伍之间，而倔起阡陌之中，率疲弊之卒，将数百之众，转而攻秦；斩木为兵，揭竿为旗， 天下云集响应，赢粮而景从。山东豪俊遂并起而亡秦族矣。 且夫天下非小弱也，雍州之地，崤函之固，自若也。陈涉之

> 位，非尊于齐、楚、燕、赵、韩、魏、宋、卫、中山之君也； 锄耰棘矜，非铦于钩戟长铩也；谪戍之众，非抗于九国之师也；深谋远虑，行军用兵之道，非及向时之士也。然而成败异变， 功业相反，何也？试使山东之国与陈涉度长絜大，比权量力，则不可同年而语矣。然 秦以区区之地，致万乘之势，序八州而朝同列，百有余年矣；然后以六合为家，崤函为宫；一夫作难而七庙隳，身死人手，为天下笑者， 何也？仁义不施而攻守之势异也。

- 分词后:

```
['赴', '戍', '登程', '口占示', '家人']
['力微任', '重久神', '疲再竭', '衰庸定', '不支']
['苟利国家生死以', '岂因祸福避趋之']
['谪居', '正是', '君恩厚', '养拙', '刚于', '戍卒', '宜']
['戏', '与', '山妻', '谈', '故事', '试吟', '断送', '老', '头皮']
[]
['秦孝公', '据', '崤函之固', '拥', '雍州', '之地', '君臣', '固守', '以', '窥', '周室', '有',
'席卷天下', '包举', '宇内', '囊括四海', '之意', '并吞', '八荒', '之心']
['当是', '时', '也', '商君佐', '之内', '立法', '度务', '耕织', '修守战', '之', '具外', '连衡',
'而斗', '诸侯', '于是', '秦人', '拱手', '而取', '西河', '之外']
['孝公', '既', '没惠', '文武', '昭', '襄蒙', '故业', '因', '遗策', '南取', '汉中', '西举', '巴
蜀', '东割', '膏腴之地', '北收', '要害', '之', '郡']
['诸侯', '恐惧', '会盟', '而谋弱', '秦不爱珍', '器重', '宝', '肥饶', '之地', '以致', '天下', '之
士合', '从', '缔交', '相与为一']
['当此', '之时', '齐有', '孟尝', '赵有', '平原', '楚有', '春申', '魏有', '信陵此', '四君者',
'皆', '明智', '而', '忠信', '宽厚', '而', '爱人', '尊贤', '而', '重士', '约', '从', '离衡']
['兼', '韩', '魏燕楚', '齐', '赵', '宋卫', '中山', '之众', '于是', '六', '国之士', '有宁越',
'徐尚', '苏秦', '杜赫', '之', '属', '为', '之谋齐', '明周', '最', '陈轸']
['召滑楼', '缓', '翟景苏', '厉乐毅', '之徒通', '其意', '吴起', '孙膑', '带佗', '倪良王', '廖田
忌', '廉颇', '赵奢', '之伦制', '其兵', '尝以', '十倍', '之地', '百万', '之众']
['叩关', '而攻', '秦', '秦人', '开关', '延敌', '九国', '之师', '逡巡', '而', '不敢', '进秦',
'无亡', '矢', '遗镞', '之费', '而', '天下', '诸侯', '已困', '矣', '于是', '从散', '约', '败争',
'割地', '而赂', '秦']
['秦有', '余力', '而制', '其弊', '追亡逐北', '伏尸', '百万', '流血', '漂', '橹', '因利乘便', '宰
割', '天下', '分裂', '山河', '强国', '请服', '弱国', '入朝']
['延及', '孝', '文王', '庄襄王', '享国', '之日', '浅', '国家', '无', '事']
['及至', '始皇', '奋', '六世', '之余烈', '振', '长策', '而', '御宇', '内吞', '二周', '而亡',
'诸侯', '履', '至尊', '而制', '六合', '执敲', '扑', '而', '鞭笞', '天下', '威振', '四海']
['南取', '百越', '之地', '以为', '桂林', '象', '郡', '百', '越之君', '俯首', '系颈委命', '下
吏', '乃', '使', '蒙恬', '北筑', '长城', '而守', '藩篱', '却', '匈奴', '七百余里']
['胡人', '不敢', '南下', '而', '牧马', '士', '不敢', '弯弓', '而', '报怨', '于是', '废先', '王
之道', '焚', '百家', '之言以', '愚', '黔首', '隳', '名城', '杀', '豪杰', '收', '天下', '之兵']
['聚', '之', '咸阳', '销', '锋镝', '铸', '以为', '金', '人', '十二', '以弱', '天下', '之民',
'然后', '践', '华为', '城因', '河为', '池据', '亿丈', '之', '城临', '不测', '之渊', '以为',
'固']
['良将', '劲弩守', '要害', '之处', '信臣', '精卒', '陈利兵', '而', '谁', '何', '天下', '已',
'定', '始皇', '之心', '自', '以为', '关中', '之固', '金城千里', '子孙', '帝王', '万世之', '业',
'也']
['始皇', '既', '没余', '威震', '于殊', '俗然', '陈涉', '瓮牖绳枢', '之子氓', '隶之人', '而', '迁
徙', '之徒', '也', '才能', '不及', '中人非', '有', '仲尼', '墨', '翟之贤']
['陶朱', '猗顿', '之富', '蹑足', '行伍', '之间', '而', '倔起', '阡陌', '之中', '率', '疲弊',
'之', '卒', '将', '数百', '之众', '转而', '攻秦', '斩木为', '兵', '揭竿', '为旗']
['天下', '云集响应', '赢', '粮而景', '从', '山东', '豪俊', '遂', '并', '起而亡', '秦族', '矣']
```

```
['且夫', '天下', '非小弱', '也', '雍州', '之地', '崤函之固', '自若', '也', '陈涉', '之', '位
非', '尊于', '齐楚', '燕赵', '韩魏', '宋卫', '中山', '之君', '也']
['锄', '耰', '棘', '矜', '非', '铦', '于', '钩', '戟', '长', '铩', '也', '谪戍', '之众',
'非', '抗于', '九国', '之师', '也', '深谋远虑', '行军', '用兵之道', '非及', '向', '时之士',
'也', '然而', '成败', '异变']
['功业', '相反', '何', '也', '试使', '山东', '之国', '与', '陈涉', '度长絜大', '比权量力',
'则', '不可', '同年而语', '矣', '然']
['秦以', '区区', '之', '地致', '万乘', '之势', '序八州', '而', '朝', '同列', '百有', '余年',
'矣', '然后', '以', '六合', '为家', '崤', '函为', '宫', '一夫', '作难', '而', '七庙', '隳',
'身死', '人手', '为', '天下', '笑者']
['何', '也', '仁义', '不施', '而', '攻守', '之势', '异', '也']
```

- 难点：

1. 古诗文中有一些特定的词语，名称等未登录词，给分词带来困难。
2. 古代的一些语法规则，助词和语气词使用，通假多意等情况难以处理。
3. 一些修辞手法如互文等手段，需要根据上下文的特定格式和内容判断，难以处理。

2.假设词典中包括词｛王公，公子，研究，研究生，生命，起源｝以及所有单字集合，请分别给出句子“王公子在研究生命的起源”的FMM和BMM分词结果。

FMM：

> 王公/子/在/研究生/命/的/起源

BMM：

> 王/公子/在/研究/生命/的/起源

3.下面的句子存在哪种类型的分词歧义？为什么？

吉林省长春药店： `交叉歧义`

> 吉林省/长春/药店
>
> 吉林/省长/春药店

东北大学生联合会: `交叉歧义`

> 东/北大/学生/联合会
>
> 东北/大学生/联合会

人大代表群体性事件: `组合歧义`

> 人大代表/群体性/事件
>
> 人大/代表/群体性/事件