# Chinese Keyword Extraction

Caroline Chen
*Columbia University*
yc3756@columbia.edu

January 3, 2021

**Abstract**

The keyword for Chinese video transcriptions on the UI is missing. It is necessary to provide tags for user to better understand the uniqueness of each individual Chinese video transcription. In this report, we use natural language processing techniques to find tags from the eight Chinese video transcriptions. This includes two parts: keyword extraction and common word removal. The final result will then be incorporated into the UI.

## 1 Introduction

Thanks to the advancement of technology, there are a variety of online videos available to us. These videos are in different languages and come from different cultural background. The goal of this project is to discover the cultural difference based on the analysis of news videos. We specifically focus on the Go match between Kejie and AlphaGo. Kejie is the No. 1 Go player in China while AlphaGo is a AI developed by google. This event is covered by large amount of both Chinese and US media. We hence focus on this specific news topic to analyze the cultural difference between China and US.

My research this semester focus on generate unique tags for individual Chinese news video transcription among 7 others. We explore Chinese NLP library to pre-process data. With the help of NLP technique such as TF-IDF algorithm and LDA topic modelling, we will be able to extract keyword from documents and further find out the uniqueness of individual documents. There are also many Chinese news training corpus available. These corpus enable us to train our model and algorithm with more specific data set than the general corpus.

The main contribution of my research can be summarized as follow:

**(1)** Keyword extraction using TF-IDF algorithm.

**(2)** Generate IDF corpus from two Chinese news corpus that can be used as training data for Jieba.

**(3)** Common word removal with IDF corpus.

**(4)** LDA topic modeling based on the eight Chinese news transcripts.

# 2 Related Work

## 2.1 Data Set

The data for the analysis is the video transcript from the audio corresponding to a particular video frame. Each text corresponds to one of the eight Chinese news bubbles on UI.

## 2.2 Jieba

In order to do keyword extraction on our video transcript, we need to tokenize the Chinese text. Jieba*, a Chinese word segmentation module, is used to achieve this task. Jieba is "based on a prefix dictionary structure to achieve efficient word graph scanning." It "builds a directed acyclic graph (DAG) for all possible word combinations" and "use dynamic programming to find the most probable combination based on the word frequency". "For unknown words, Jieba use a HMM-based model with the Viterbi algorithm"[2]. In our project, we use Jieba to remove stopwords, tokenize sentences and compute TF-IDF.

## 2.3 TF-IDF Algorithm

TF-IDF is used to measure how important a word is to a document in a corpus. TF is term frequency. IDF is inverse document frequency and it is first introduced in 1972 by *Karen Spärck Jones*.

**TF: term frequency**

$$\text{tf}(t) = \frac{\text{number of times term t appears in a document}}{\text{Total number of terms in the document}}$$

**IDF: inverse document frequency**[3]

$$\text{idf}(t) = \log(\frac{\text{Total number of documents}}{\text{Number of documents with term t in it}})$$

$$\text{tf-idf} = \text{tf}(t) \cdot \text{idf}(t)$$

*https://github.com/fxsjy/jieba

### 2.4 LDA topic modeling

Latent Dirichlet Allocation is used to classify tokens in a document into specific abstract topics. The algorithm is described in *Thomas Hofmann*'s paper[1]. The Python library gensim[†] is used to compute the LDA topic modeling. The class gensim.models.ldamulticore[‡] is used.

## 3 Experiment

### 3.1 Chinese News Corpus

A Chinese news corpus is required to serve as training corpus of tf-idf algorithm. We want a more specific data set than a general Chinese corpus, since we want our model to learn vocabulary and idf that have distribution more similar our text. Hence, a Chinese news corpus is required rather than a general Chinese corpus. We explore multiple Chinese news corpus, including some from the Linguistic Data Consortium. Since the match between Kejie and AlphaGo occurs in 2017, we want a news corpus latter than that. We experiment with two Chinese corpus: toutiao news description [§] (in 2018) and 250 million pieces of Chinese news (from 2014 to 2016)[¶].

### 3.2 Stopword Removal

We need to remove stopwords from our data set since we only want to focus on the words that define the meaning of the text. We first experiment with Jieba's stopword corpus, which only includes 51 Chinese stopwords. This is way too few for us. We then find a stop word corpus[‖] which contains 794 Chinese stopwords. We use this stopword corpus to remove stopwords for our project.

### 3.3 IDF Corpus

To calculate the tf-idf of tokenized words. We calculate the idf score of each individual words using the selected Chinese news corpus. We only use the news content. The news content is tokenized into words with Jieba. The idf score is then calculated with stopwords removed. The tokenized word and its idf score in the corpus is stored as the following format. This news idf corpus contains 688,100 words with its idf value in the corpus (all stopwords removed).

[†] https://pypi.org/project/gensim/

[‡] https://radimrehurek.com/gensim/models/ldamulticore.html#module-gensim.models.ldamulticore

[§] https://github.com/skdjfla/toutiao-text-classfication-dataset/blob

[¶] https://github.com/brightmart/nlp$_c hinese_c orpus$

[‖] https://github.com/stopwords-iso/stopwords-zh

```
积极 2.6187561857974053
对方 3.227008076887095
顺境 7.5853592097431575
抱怨 4.3441700859109655
帮助 2.240819074530935
才 1.6622695932188083
善人 8.35854909797664
以 1.115671699720891
信心 3.586453040672567
坚信 5.240107670430209
深深地 5.897062722396738
践行 4.989339391807881
```

### 3.4 TF-IDF

With the idf corpus described above, we use Jieba's tf-idf function to extract key words from our 8 pieces of Kejie and AlphaGo news. We first experiment with selecting words with top 5 tf-idf score as the key words for the 8 pieces of news.

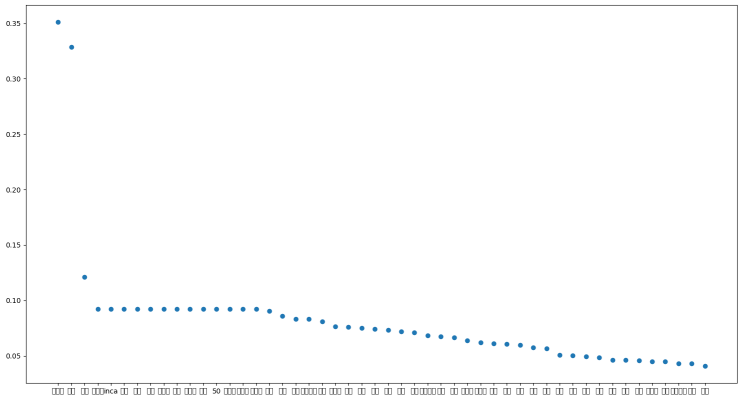
```
1:
阿尔法,围棋,认输,手柯洁,inca
2:
albergo,13,中令,李三,言面
3:
围棋,阿尔法,觉得,下载,谢谢
Common 1:
立足于,看好,真正,认为,情况
Common 2:
柯洁,阿尔法,围棋,赛后,缺陷
Common 3:
李世石,胜利,对尔法,长舒,人机
Common 4:
Alphago,柯洁,巨大进步,落败,谷歌
Common 5:
人工智能,令人生畏,围棋,进步,深刻
```

In order to select keyword that is a good summary of text, we then need to decide the cutoff point for tf-idf score. We want to decide the number of keyword we need for each document based on the tf-idf score. The tf-idf score of tokens in each document is shown as follow

```
1:
[('阿尔法', 0.35105149139505704), ('围棋', 0.328749475298605), ('认输', 0.12094622696663791), ('手柯洁', 0.09220426931043282), ('inca', 0.09220426931043282), ('可杰', 0.09220426931043282), ('投子',
0.09220426931043282), ('绝黑', 0.09220426931043282), ('徐柯洁', 0.09220426931043282), ('真本', 0.09220426931043282), ('从四得', 0.09220426931043282), ('尔法', 0.09220426931043282), ('50',
0.09220426931043282), ('般尔法', 0.09220426931043282), ('围棋子', 0.09220426931043282), ('依曼体', 0.09220426931043282), ('对决', 0.09026094537808725), ('赛后', 0.08578301141269591), ('哈萨',
0.08319925055085815), ('围棋比赛', 0.08319925055085815), ('法为', 0.08084120077666322), ('围棋赛', 0.07625418612145483), ('缺陷', 0.07614716444070874), ('此曲', 0.07515966650977643), ('症是',
0.07419423179128348), ('柯洁', 0.07333062100720295), ('三番', 0.07183618201708855), ('白起', 0.07118009424927269), ('地球科学', 0.06851269752439869), ('中盘', 0.06764908674031814), ('比斯',
0.06650349704871848), ('一个劲', 0.06405683326809556), ('他妈的', 0.06217507548969801 5), ('棋局', 0.06100211775493841), ('拼杀', 0.06064568241117004), ('棋手', 0.05997619035367605), ('乌镇',
0.05738053781822496), ('人机', 0.05649354122281321 4), ('完美', 0.05095118680762216), ('比赛', 0.05019051779865015), ('一书', 0.04928261387740045 4), ('奋力', 0.04865340062976243), ('歌声',
0.04631556472847893), ('团队', 0.04612084677864 06), ('句子', 0.04590120178452 13), ('执行官', 0.04479586617658251), ('获胜', 0.04477068421726309), ('人工智能', 0.04291152212975412), ('终极',
0.04280915057063479 56), ('巅峰', 0.04102257340335302)]
2:
[('albergo', 0.3515287767460251), ('13', 0.17576438837301256), ('中午', 0.17576438837301256), ('李三', 0.17576438837301256), ('意面', 0.17576438837301256), ('15', 0.17576438837301256), ('12',
0.17576438837301256), ('福斯', 0.16493396367676343), ('钟燕', 0.15859857136257335), ('比赛', 0.14351351183095152), ('十只', 0.13693772191907506), ('五局', 0.13452911759777414), ('李世石', 0.1314954298721342),
('对局', 0.1314954298721342), ('三局', 0.12610729727738259), ('两局', 0.12113145772509881), ('人类', 0.11384471421311508), ('照常', 0.11179025458829224), ('围棋', 0.10444644788132762), ('一局',
0.10246207771318219), ('爱奇艺', 0.10070155954905544), ('双方', 0.09866542013349 17), ('实例', 0.09666134098196622), ('理事', 0.09560972095233955), ('首尔', 0.08949156153141496), ('如期', 0.08930517761164507),
('赛前', 0.08510628875774595), ('宣告', 0.08401254791245 92), ('三月', 0.08388106046050162), ('谷歌', 0.08258346122897434), ('四季', 0.08116377358755365), ('大战', 0.07298530465862121), ('战胜',
0.07109425753680554), ('场地', 0.06977493211357569), ('约定', 0.06790903310490955), ('落后', 0.06603207452797129), ('同一', 0.06515213220723651), ('胜利', 0.06296892419146476), ('赢得', 0.05952127416780445),
('保留', 0.05801703739178 8075), ('体育', 0.05746811260971147), ('酒店', 0.05733988888244 8076), ('韩国', 0.05703156873439068), ('更为', 0.05373396514681 2716), ('一句', 0.04964183916036947), ('提前',
0.04844492483796206), ('实在', 0.04798192800036246), ('终于', 0.04674273750222 45), ('举行', 0.04591415408882 0505), ('报告', 0.04462038819181684)]
3:
[('围棋', 0.15915649200964208), ('阿尔法', 0.10197209988142134), ('赞得', 0.09295846771202944), ('下载', 0.09055960236121707), ('谢谢', 0.06093044129449528), ('其实', 0.055281465660975554), ('普下',
0.05356628978987049), ('要注', 0.05356628978987049), ('单关', 0.05356628978987049), ('算太准', 0.05356628978987049), ('那个游戏', 0.05356628978987049), ('33', 0.05356628978987049), ('先拉后',
0.05356628978987049), ('300', 0.05356628978987049), ('Whisky', 0.05356628978987049), ('阿法', 0.04833480270097474), ('人字', 0.04833480270097474), ('回时', 0.04696488807025197), ('开拓', 0.04393308857320 15),
('比赛', 0.04373745112245379 54), ('赔不起', 0.04366418721044270 5), ('这题', 0.04310331561207898), ('难下', 0.04214774087178302), ('大图', 0.04195400863682 4566), ('边有', 0.04099935012503593), ('全力',
0.04012164154402160 5), ('下得', 0.03930089801104197), ('太狠', 0.03906856389594943), ('对齐', 0.03769864926522667), ('弟子', 0.03769864926522667), ('狠瞧', 0.03706278549082418), ('放心大胆',
0.03650191389246045), ('一盘棋', 0.03600197151232705), ('取下', 0.03588261328175474 5), ('贪恋', 0.03588261328175474 5), ('没想到', 0.03540350928062 3954), ('那种', 0.03505769762294211), ('后面',
0.03479555787636983), ('匀速', 0.03457112766337395), ('把戏', 0.03406941092214621), ('张家', 0.03275939525431325 5), ('国女', 0.03230060065342 61), ('转手', 0.03163690795183 1966), ('一想', 0.03149615175068175),
('近视', 0.03065112619285898 3), ('莫属', 0.03038755431151058), ('非常感谢', 0.03038755431151058), ('妙票', 0.03013780919068895), ('解禁', 0.02983482968649 7463), ('一次', 0.02847873600755 7764)]
Common 1:
[('立足于', 1.0258617758443402), ('看好', 0.6614854338314361), ('真正', 0.39677472447916573), ('认为', 0.29967810640355 33), ('情况', 0.25339701763353 63), ('可能', 0.23809053500994712)]
Common 2:
[('柯洁', 0.7455279802398965), ('阿尔法', 0.5948372493082911), ('围棋', 0.55704772203374 73), ('赛后', 0.43606364134787 09), ('缺陷', 0.38708141922 76943), ('发布会', 0.30589794062111147), ('完美',
0.25900186627207 93), ('举行', 0.24487548847370935), ('新闻', 0.23388889592675502), ('表现', 0.19335670705590458), ('表示', 0.13487960203809396), ('没有', 0.08898346389486639)]
Common 3:
[('李世石', 0.9350786124240654), ('胜利', 0.6716685247089574), ('对尔法', 0.6249400475484891), ('长舒', 0.48244286102284817), ('人机', 0.38290066828795562), ('连败', 0.37194688690401 46), ('较量',
0.30517659123810514), ('一口气', 0.28578486796987884), ('大战', 0.25950330545287 54), ('人类', 0.20239960304553 79), ('首次', 0.16452180182005 79), ('一场', 0.15831170583962687), ('相信', 0.14649794933408 8), ('获
得', 0.11226336476490502), ('之后', 0.09825296370464229)]
Common 4:
[('Alphago', 0.7539838339509185), ('柯洁', 0.6390239830626 7685), ('巨大进步', 0.6202836764579476), ('落败', 0.5075651521058851), ('谷歌', 0.37752439373245 41), ('大脑', 0.31080113662458835), ('半年',
0.28859432270804325), ('训练', 0.25780561200105 65), ('自我', 0.24651563002543672), ('负责人', 0.20717021646333 62), ('取得', 0.19808391816933 86), ('过去', 0.17712318460441456), ('方面', 0.12291296518150215),
('表示', 0.11561108746122 34)]
Common 5:
[('人工智能', 0.8054162615123 08), ('令人生畏', 0.6741549389296003), ('围棋', 0.51419789726192 06), ('进步', 0.2928021640829510 7), ('深刻', 0.28423923106650556), ('印象', 0.28073227095822995), ('令人',
0.25865104499193 19), ('几年', 0.25240723973376833), ('挑战', 0.24668188044723466), ('取得', 0.21332114264390312), ('过去', 0.19074804495860 03), ('一直', 0.14126586921588363)]
```

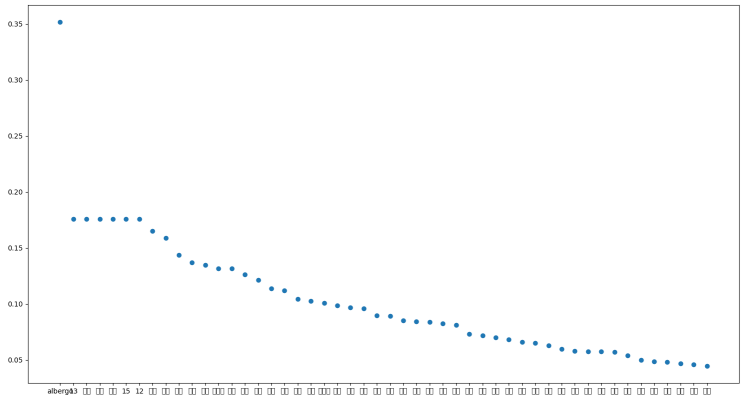

The tf-idf score for each document is plotted to find the number of tags we want for each document.

**1**



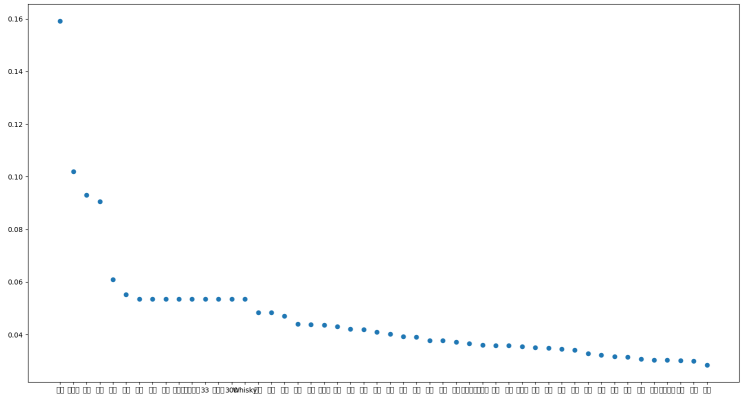

Looking at the plot, we select the top 3 tokens as tags for this document. The top 3 word in this document with its tf-idf value are (' 阿尔法', 0.35105149139505704), (' 围棋', 0.328749475298605), (' 认输', 0.12094622696663791)

**2**



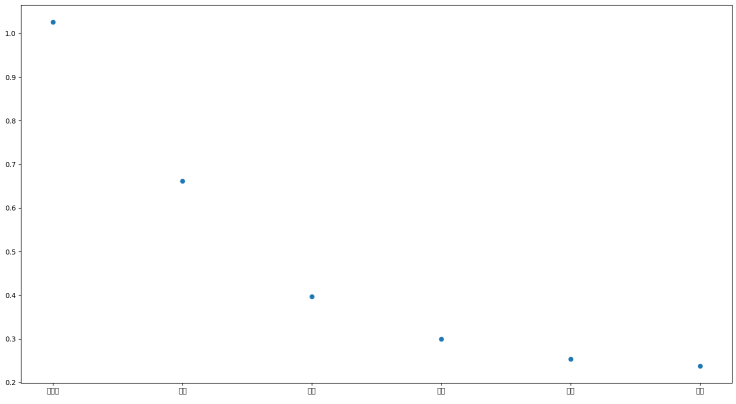
Looking at the plot, we select the top 7 tokens as tags for this document. The top 7 tokens are ('albergo', 0.3515287767460251), ('13', 0.17576438837301256), (' 中令', 0.17576438837301256), (' 李三', 0.17576438837301256), (' 言面', 0.17576438837301256), ('15', 0.17576438837301256), ('12', 0.17576438837301256).

**3**



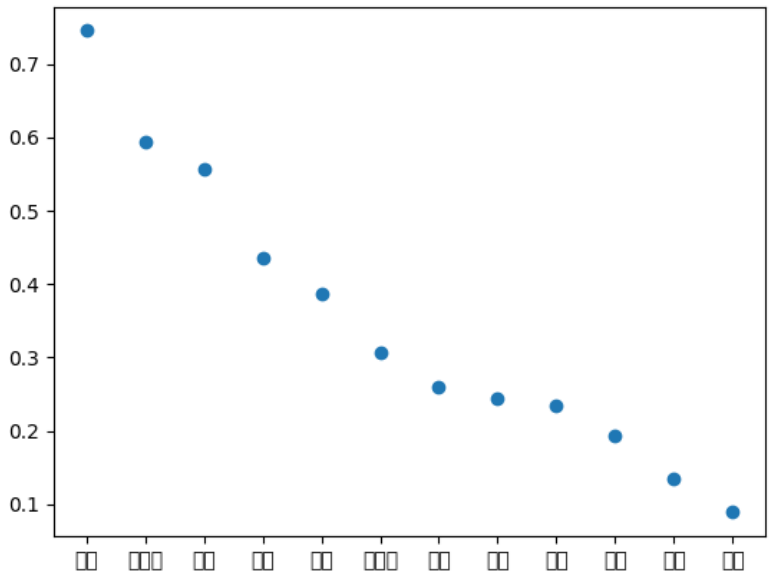
We select the top 5 tokens as tags: (' 围棋', 0.15915649200964208), (' 阿尔法', 0.10197209988142134), (' 觉得', 0.09295846771202944), (' 下载', 0.09055960236121707), (' 谢谢', 0.06093044129449528).

**Common 1**



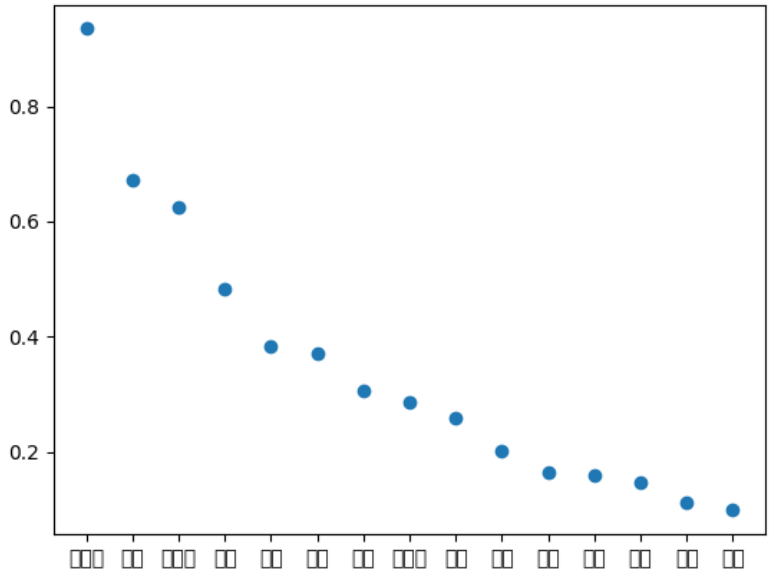
There are only 6 tokens in this documents with stopwords removed. We select the top 2 tokens as tags: (’ 立足于’, 1.0258617758443402), (’ 看好’, 0.6614854338314361).

**Common 2**



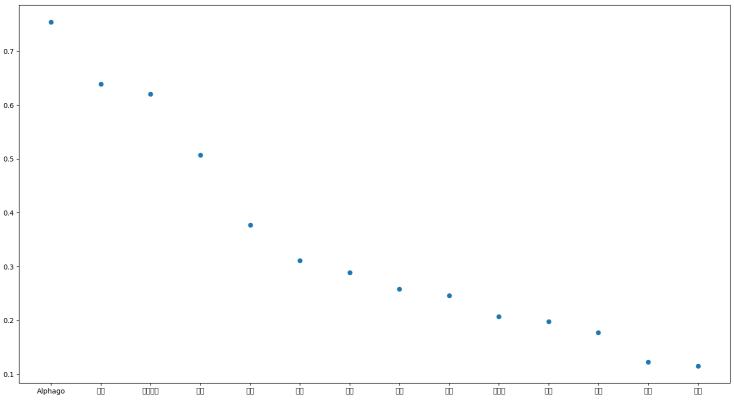
We select the top 4 tokens as tags: (' 柯洁', 0.7455279802398965), (' 阿尔法', 0.5948372493082911), (' 围棋', 0.5570477220337473), (' 赛后', 0.4360636413478709).

**Common 3**



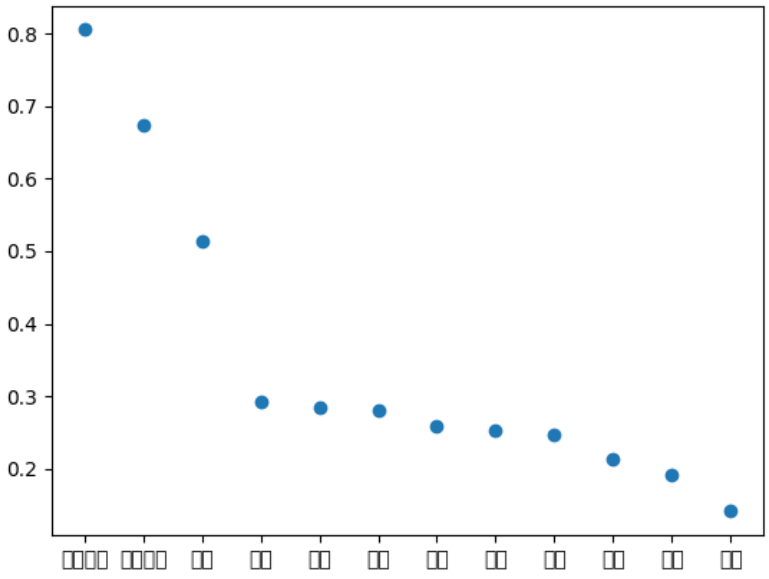

We select the top 4 tokens as tags: (' 李世石', 0.9350786124240654), (' 胜利', 0.6716685247089574), (' 对尔法', 0.6249400475484891), (' 长舒', 0.48244286102284817).

**Common 4**



We select the top 4 tokens as tags: ('Alphago', 0.7539838339509185), (' 柯洁', 0.6390239830627685), (' 巨大进步', 0.6202836784579476), (' 落败', 0.5075561521058051).

**Common 5**



We select the top 3 tokens as tags: (' 人工智能', 0.805416261512308), (' 令人生畏', 0.6741549389296003), (' 围棋', 0.5141978972619206).

For the 5 common news text (the last 5 documents), the keyword we selected based on the graph have tf-idf value greater than 0.4. However, the selection based on the graph is relatively arbitrary. Future can potentially focus on refining this result with the support of some algorithm and extending the data set to get a more accurate result.

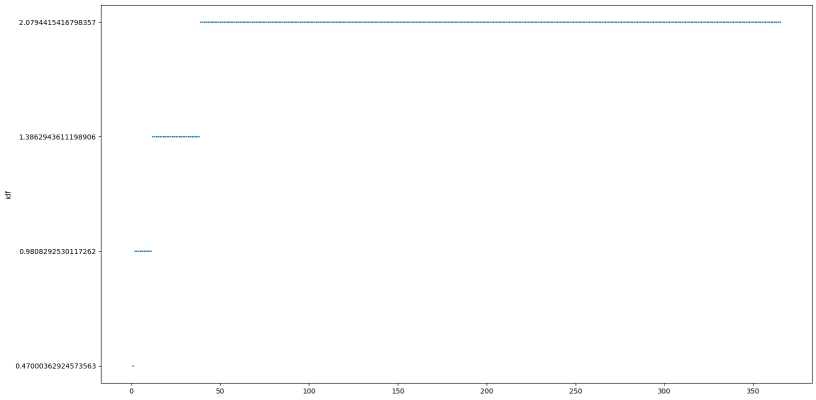
## 3.5 Common word removal

In order to show the uniqueness of each video in the tags, the common word in the 8 texts need to be removed.

**idf value** To remove common words, we first experiment with the idf weight of words in our 8 pieces of text. After stopwords, punctuation, and spaces are removed, we construct a small idf corpus in the following format.

```
一点 1.3862943611198906
albergo 2.0794415416798357
可杰 2.0794415416798357
思维 2.0794415416798357
巨大进步 2.0794415416798357
匀速 2.0794415416798357
这题 2.0794415416798357
今天 1.3862943611198906
巅峰 2.0794415416798357
实际上 2.0794415416798357
轻 2.0794415416798357
厉害 2.0794415416798357
贪恋 2.0794415416798357
全力 2.0794415416798357
出 2.0794415416798357
实在 2.0794415416798357
类 2.0794415416798357
定价 2.0794415416798357
乐队 2.0794415416798357
普下 2.0794415416798357
懂 2.0794415416798357
查 2.0794415416798357
想 2.0794415416798357
连败 2.0794415416798357
Whisky 2.0794415416798357
笑 2.0794415416798357
人工智能 1.3862943611198906
形式 2.0794415416798357
```

The idf weight of words in the 8 pieces of text is calculated and plotted. Words with smaller idf weight are more common.




This corpus contains 365 words and its idf weight and is constructed from 8 documents. The words and its idf weight correspond to the above graph:

```
2.0794415416798357': ['albergo', '可杰', '思维', '巨大进步', '匀速', '这题', '巅峰', '实际上', '轻', '厉害', '贪恋',
'全力', '出', '实在', '类', '定价', '乐队', '普下', '懂', '查', '想', '连败', 'Whisky', '笑', '形式', '久', '人才',
'思想','两个', '终极', '钱', '照常', '不会', '级', '最近', '狗', '历经', '旗下', '地球科学', '早就', '之前', '盘', '时间',
'机会', '半小时', '酒店', '蝶', '其实', '意义', '先拉后', '边有', '段', '台湾', '相信', '喜欢', '这场', '变化', '拉', '四季',
 '最初', '把戏', '提', '世界', '应该', '近期', '快乐', '棋', '法为', '首席', '变成', '一场', '合适', '参与', '获胜', '奋力',
 '改变', '围棋子', '一次', '理解', '言面', '昨天', '基本', '董', '一口气', '包括', '将会', '同学', '希望', '地方',
 '转手', '闺女', '更为', '临期', '解禁', '乌镇', '好玩', '场地', '白', '令人', '落败', '尔法', '落后', '认输', '一系列',
 '觉得', '负责人', '进入', '成员', '数字', '谢谢', '报告', '公司', '团队', '返还', '三分', '中国', '三局', '李三',
 '围棋比赛', '获得', '请求', '为期', '带来', '字', '中盘', '真本', '徐柯洁', '手上', '角落', '以后', '好像', '战胜',
 '百度', '真的', '双方', '完成', '越来越', '单关', '知识', '爱奇艺', '计算', '期', '开拓', '从四得', '一句', '职业',
 '一个劲', '拼', '挑战', '后面', '完全', '口', '首次', '棋局', '韩国', '难下', '非常感谢', '肯定', '阿法', '白起',
 '四川', '有意思', '机', '即时', '影响', '样子', '奇', '继续', '围棋赛', '回时', '体现', '行', '科技', '接受',
 '体育', '近视', 'Alphago', '值得', '不行', '接下来', '踩', '谷', '相爱', '宣告', '日记', '开局', '之后', '对齐',
 '这时候', '三月', '提前', '对手', '观看', '此曲', '打破', '这种', '看法', '深刻', '一想', '开展', '句子',
 '依曼体', '集', '手续', '程序', '没', '国际', '哈萨', '办法', '不再', '带', '今后', '联系', '五局', '三番', '代表',
 '放心大胆', '自由', '拼杀', '约', '子', '那个游戏', '有太多', '两局', '展开', '首尔', '自我', '一局', '较量', '终于',
 '最后', '约定', '种', '实例', '歌声', '症是', '绝黑', '学习', '贯彻落实', '大师', '人字', '要注', '局', '进步', '对局',
 '三个', '顶尖', '点评', '最终', '比较', '气候', '苦', '他妈的', '对不起', '十只', '赛前', '那种', '贯彻', '钟路', '一书',
 '赔不起', '对尔法', '加载', '公布', '开心', '知道', '冲击', '项目', '般尔法', '加', 'R', '大胆', '投子', '具体', '非常',
 '很瘦', '感谢', '名人', '几年', '理事', '探讨', '手柯洁', '长舒', '爱', '比斯', '没想到', '下得', '大脑', '棋手', '同一',
 '当时', '一盘棋', '浙江', '我会', '半年', '更好', '一处', '问题', '令人生畏', '方面', '如期', '执行官', '保留', '慢',
 '中令', '下载', '天', '北京', '理念', '张家', '先', '取下', '活动', '算太准', '莫属', '调子', '爱好者', '钞票',
 '太狠', '开', 'inca', '赢得', '一下', '题', '训练', '去年'],

'1.3862943611198906': ['一点', '今天', '人工智能', '结束', '新闻', '人类', '找', '赛后', '印象', '完美', '一直', '对决',
 '量', '缺陷', '李世石', '享受', '心态', '过去', '发布会', '实际', '表现', '已经', '人机', '很多', '取得', '谷歌', '胜利']

'0.9808292530117262': ['没有', '柯洁', '大战', '前', '表示', '比赛', '进行', '举行', '阿尔法', '太'],

'0.4700036292457356': ['围棋']
```

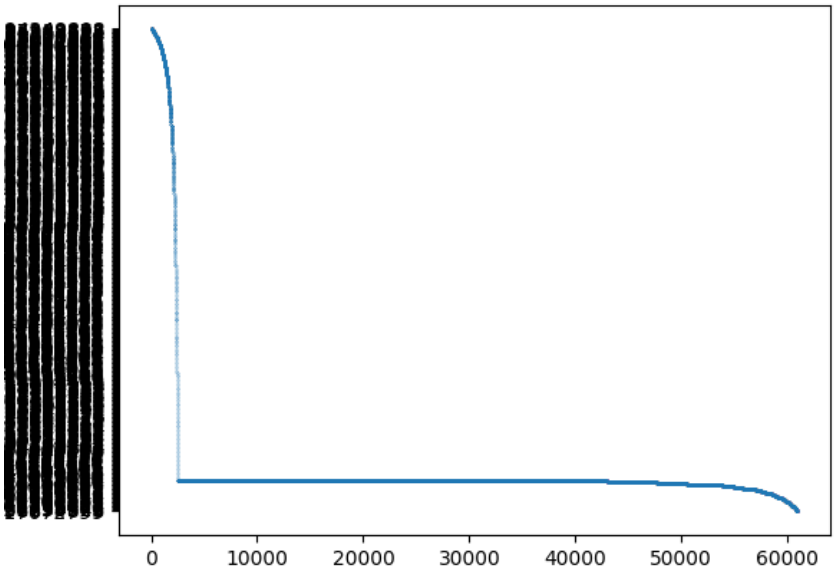
More than 300 words out of 365 words have idf weight of 2.07, which indicates these words only appears once. The one most frequent word appears in 5

documents.

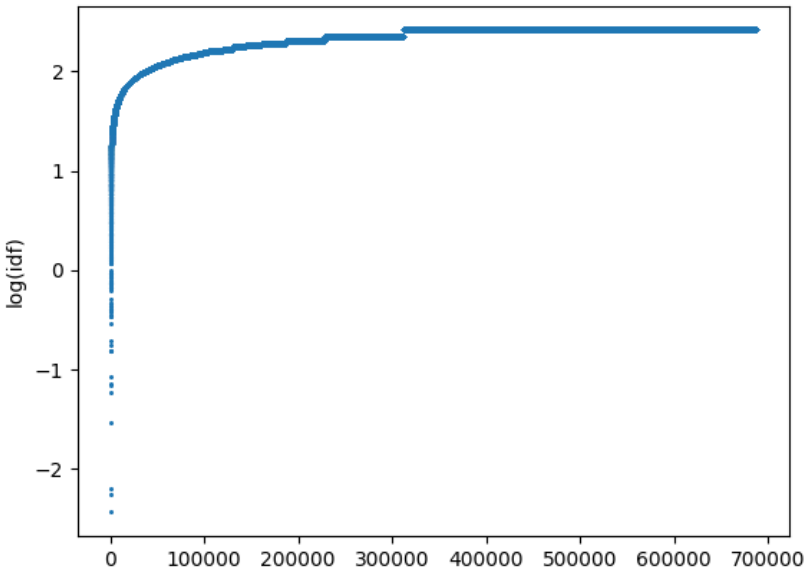
$$\ln(\frac{8}{1}) = 2.079$$
$$\ln(\frac{8}{2}) = 1.386$$
$$\ln(\frac{8}{3}) = 0.981$$
$$\ln(\frac{8}{5}) = 0.470$$

However, the idf graph also indicates that the data set is too small for idf to display a curve. Based on this result, we decide the token ' 围棋' with idf weight 0.470 and ' 没有', ' 柯洁', ' 大战', ' 前', ' 表示', ' 比赛', ' 进行', ' 举行', ' 阿尔法', ' 太' with idf weight 0.981 should be removed from our key word set.

For reference, we also plot the idf value of the large news corpus which contains 688,100 words with inverse y axis.




We can see there is a clear cutoff point in this large idf corpus. The tokens with idf weight smaller than this cutoff point can be seen as common words in the corpus. After taking the log of idf, we get the following graph




The result of a larger idf corpus indicates that extending our idf corpus (currently only contains 8 documents) might be helpful in finding a cutoff point of common words.

**Latent Dirichlet Allocation (LDA)** LDA is used to model topics from our 8 texts. Since our data set our small, we experiment with $k = 2$ topics. The 8 Chinese news text are used as the training corpora. Gensim python library is used to do the computation. The code is in the Appendix.

```
lda_model_tfidf = models.LdaMulticore(corpus_tfidf, num_topics=5, id2word=dictionary, passes=2, workers=4)
```

The result of selecting 2 topics is:

```
Topic: 0 Word: 0.004*"人工智能" + 0.003*"过去" + 0.003*"取得" + 0.003*"几年" + 0.003*"令人生畏" + 0.003*"进步" + 0.003*"挑战" + 0.003*"负责人" + 0.003*"半年" + 0.003*"深刻"
Topic: 1 Word: 0.004*"胜利" + 0.003*"李世石" + 0.003*"盘" + 0.003*"人类" + 0.003*"长舒" + 0.003*"对尔法" + 0.003*"狗" + 0.003*"之后" + 0.003*"连败" + 0.003*"相信"
```

This result can potentially used to sort documents into topics and select common words from there.

# 4 Final Result

After the experiments above, we decide to construct the final keyword tags based on the TF-IDF score and removing common word from there. The final keyword for each documents is listed as follow:

| | |
|---|---|
| 1 | 认输 |
| 2 | albergo, 1 月 13 日, 中令, 李三, 言面, 三月 15 日 12 点 |
| 3 | 觉得, 下载, 谢谢 |
| Common 1 | 立足于, 看好 |
| Common 2 | 赛后 |
| Common 3 | 李世石, 胜利, 长舒 |
| Common 4 | Alphago, 巨大进步, 落败 |
| Common 5 | 人工智能, 令人生畏 |

**Note** In text 2, the tokens 13, 12, 15 are manually turned into 1 月 13 日 and 三月 15 日 12 点. Jieba fails to tokenize these numbers as date and time as a whole but tokenize them into individual numbers instead.

# 5 Future Work

**1.** Refine the process of deciding cutoff in keyword extraction based on TF-IDF value. The current cutoff point is made relatively arbitrary from the TF-IDF value plot. Research to see if there any algorithm to decide the cutoff point.

**2.** Define cutoff for common word removal in keyword extraction based on the idf weight. Extend the data set to see if there is a clearer trend. Explore algorithms to define the middle frequency.

**3.** Remove common words based on LDA modeling result.

# 6 References


[1] Thomas Hofmann. “Probabilistic Latent Semantic Analysis”. In: (Jan. 2013).

[2] *Jieba*. `https://github.com/fxsjy/jieba`. Accessed: 2020-12-20.

[3] Karen Jones. “IDF term weighting and IR research lessons”. In: *Journal of Documentation - J DOC* 60 (Oct. 2004), pp. 521–523. DOI: `10.1108/00220410410560591`.

# 7 Appendix

## 7.1 TF-IDF keyword extraction with Jieba

```
import sys

import jieba
import math
import os
import jieba.analyse
from optparse import OptionParser

def small_corpus_idf(directory):
    documents = []
    word_set = set()
    stop_words = []
    with open("stopword.txt","r") as f:
        lines = f.readlines()
        for line in lines:
            stop_words.append(line.rstrip('\n'))
    for filename in os.listdir(directory):
        l = []
        dic_file = os.path.join(directory, filename)
        with open(dic_file, "r") as file:
            lines = file.readlines()
            for line in lines:
                jiecut = jieba.lcut(line)
                l += jiecut
        word_set.update(l)
        documents.append(l)
    total_document = len(documents)
    diction = {}
    for word in word_set:
        if word.isspace() or word.isnumeric():
            continue
        if word in stop_words:
            continue
        count = 0
        for d in documents:
            if word in d:
                count += 1
        idf = math.log(total_document / count)
        print(word + " " + str(idf))
        diction[word] = idf
    a = sorted(diction.items(), key=lambda x: x[1], reverse=False)
    return [each[0] for each in a[0:30]]
    # print(a[0:10])


def jieba_analyze(filename):
    USAGE = "usage:    python extract_tags.py [file name] -k [top k]"
    parser = OptionParser(USAGE) #
    parser.add_option("-k", dest="topK")
    # opt, args = parser.parse_args()

    jieba.analyse.set_stop_words("stopword.txt")
```

```
    content = open(filename, 'rb').read()
    jieba.analyse.set_idf_path("/static/id/news_idf.txt")
    tags1 = jieba.analyse.extract_tags(content, topK=5)

    print(",".join(tags1))
```

## 7.2 LDA topic modeling with gensim

select_topics.py

```
import numpy as np
from collections import defaultdict
import os
import jieba
from gensim import models

matrix = np.zeros((365,8))

def extract_idf(filename):
    idf_dict = defaultdict(float)
    index_dict = {}
    i = 0
    with open(filename,"r") as f:
        lines = f.readlines()
        for line in lines:
            word = line.split()[0]
            idf_dict[word] = float(line.split()[1])
            if not word in index_dict:
                index_dict[word] = i
                i += 1
    return idf_dict, index_dict

idf_dict, index_dict = extract_idf("chi_small_idf.txt")

def small_corpus_idf(directory):
    documents = []
    stop_words = []
    with open("stopword.txt","r") as f:
        lines = f.readlines()
        for line in lines:
            stop_words.append(line.rstrip('\n'))
    for filename in os.listdir(directory):
        l = []
        dic_file = os.path.join(directory, filename)
        with open(dic_file, "r") as file:
            lines = file.readlines()
            for line in lines:
                tmp = defaultdict(int)
                jiecut = jieba.lcut(line)
                for word in jiecut:
                    if word.isspace() or word.isnumeric() or word in
                        stop_words:
                        continue
                    else:
                        tmp[word] += 1
                documents.append(tmp)
```

```
    return documents

documents = small_corpus_idf("static/small_corpus")
corpus_tfidf = []
for i in range(len(documents)):
    document = documents[i]
    total_word = 0
    tmp = []
    for key, value in document.items():
        total_word += value

    for key, value in document.items():
        tf = value / total_word
        tfidf = idf_dict[key] * tf
        matrix[index_dict[key]][i] = tfidf
        tmp.append((index_dict[key], tfidf))
    corpus_tfidf.append(tmp)
dictionary = {y:x for x,y in index_dict.items()}

lda_model_tfidf = models.LdaMulticore(corpus_tfidf, num_topics=5,
    id2word=dictionary, passes=2, workers=4)
for idx, topic in lda_model_tfidf.print_topics(-1):
    print('Topic: {} Word: {}'.format(idx, topic))
```

## 7.3 First 100 tokens in the news idf corpus

华夏 5.399596075925945
全然 6.47823623140714
一部分 3.2740439553139282
一刻起 6.986240978831489
秋天 4.605131122725132
贫病 9.30301070681749
是因为 2.8769846771137053
积极 2.6187561857974053
对方 3.227008076887095
顺境 7.5853592097431575
抱怨 4.3441700859109655
帮助 2.240819074530935
才 1.6622695932188083
善人 8.35854909797664
以 1.115671699720891
信心 3.586453040672567
坚信 5.240107670430209
深深地 5.897062722396738
践行 4.989339391807881
校训 7.814933651387658
它 1.8409598898134976
自己 1.1302009632938024
勇气 4.525088415051596
您 2.38813796006649

应有 4.6531403419114925
训 7.278628942320682
高高地 8.476332133633022
周边 3.667201215747496
温暖 3.821182015339911
希望 1.94965427470222
资本 3.214938121189584
家里 3.20911851213632
和 0.3423962295915959
人世间 6.918187515586473
千年 5.020409852281621
主要 1.6506006901294765
美丽 3.239890170803074
力量 2.983785225935419
重要 1.6276635971102118
小学 3.8325423766798767
与 0.6961596434707181
母校 6.461429113090758
漂泊 6.9051154340191205
我们 1.069887073229418
她 2.0846244223980284
阳光 3.5622995209281845
由此 3.8875454268954557
微 2.479413735842577
自愿 4.9499716090168615
乞丐 6.512722407478309
家庭 2.73774573678213
然后 2.3298690519425143
两个 2.0833686666867557
明天 3.9009770327241173
种子 4.804789599172364
其实 1.9348512981341446
而是 2.4828383967239414
才能 2.27721845616948
改变 2.5852060117937996
另 2.713887746328236
校长 4.741643143487792
完善 2.985330423255485
教师 3.9593103344216365
从 0.9253423430331922
选择 1.732935711070668
伟大 4.115624900976736
俞正强 11.248920855872804
愿意 2.9912758976645764
新闻线索 7.138046991699493
是 0.317082547709233

一位 2.4884678095575326
按 2.0775292340351963
下 1.1606148663441342
地 1.4057673423763757
无限 4.006838496615843
事例 6.336265970136752
而 0.886996310462406
开始 1.4359599425362344
孩子 2.4122562695153786
人生 2.9397363281865063
宇宙 5.054515464768132
辛苦 4.168894355950214
小学校 7.752413294406324
想 1.6613776855064286
关于 2.2339872802392318
最 1.2566443405226169
因此 1.9732610177759788
这些 1.5116649989816828
不是 1.5381688984795034
中 0.6301088661653788
ID 3.667201215747496
规矩 5.194481509603434
这篇 5.654209476270965
一样 1.9367461780153534
努力 2.6455499682155144
强大 3.273700017219394
特别 1.9724182816429527
从今天起 6.7056260736028
国家 1.965422963278895
知识 2.8144573120555627

## 7.4 Small idf corpus constructed from our 8 Chinese news text

一点 1.3862943611198906 albergo 2.0794415416798357 可杰 2.0794415416798357 思维 2.0794415416798357 巨大进步 2.0794415416798357 匀速 2.0794415416798357 这题 2.0794415416798357 今天 1.3862943611198906 巅峰 2.0794415416798357 实际上 2.0794415416798357 轻 2.0794415416798357 厉害 2.0794415416798357 贪恋 2.0794415416798357 全力 2.0794415416798357 出 2.0794415416798357 实在 2.0794415416798357 类 2.0794415416798357 定价 2.0794415416798357 乐队 2.0794415416798357 普下 2.0794415416798357 懂 2.0794415416798357 查 2.0794415416798357 想 2.0794415416798357 连败 2.0794415416798357 Whisky 2.0794415416798357 笑 2.0794415416798357 人工智能 1.3862943611198906 形式 2.0794415416798357 久 2.0794415416798357 人才 2.0794415416798357 没有 0.9808292530117262 结束 1.3862943611198906 思想 2.0794415416798357 两

个 2.0794415416798357 终极 2.0794415416798357 钱 2.0794415416798357 照
常 2.0794415416798357 不会 2.0794415416798357 级 2.0794415416798357 最
近 2.0794415416798357 狗 2.0794415416798357 新闻 1.3862943611198906 历经
2.0794415416798357 旗下 2.0794415416798357 地球科学 2.0794415416798357
早就 2.0794415416798357 之前 2.0794415416798357 盘 2.0794415416798357 时
间 2.0794415416798357 机会 2.0794415416798357 人类 1.3862943611198906 半
小时 2.0794415416798357 酒店 2.0794415416798357 蝶 2.0794415416798357 其
实 2.0794415416798357 意义 2.0794415416798357 先拉后 2.0794415416798357
柯洁 0.9808292530117262 边有 2.0794415416798357 段 2.0794415416798357 台
湾 2.0794415416798357 相信 2.0794415416798357 喜欢 2.0794415416798357 这
场 2.0794415416798357 变化 2.0794415416798357 拉 2.0794415416798357 四
季 2.0794415416798357 最初 2.0794415416798357 把戏 2.0794415416798357 提
2.0794415416798357 世界 2.0794415416798357 应该 2.0794415416798357 近期
2.0794415416798357 快乐 2.0794415416798357 棋 2.0794415416798357 法为 2.0794415416798357
首席 2.0794415416798357 变成 2.0794415416798357 一场 2.0794415416798357
合适 2.0794415416798357 参与 2.0794415416798357 获胜 2.0794415416798357
奋力 2.0794415416798357 改变 2.0794415416798357 围棋子 2.0794415416798357
一次 2.0794415416798357 理解 2.0794415416798357 言面 2.0794415416798357
昨天 2.0794415416798357 基本 2.0794415416798357 董 2.0794415416798357 一
口气 2.0794415416798357 包括 2.0794415416798357 将会 2.0794415416798357
同学 2.0794415416798357 希望 2.0794415416798357 地方 2.0794415416798357
转手 2.0794415416798357 闺女 2.0794415416798357 找 1.3862943611198906 更
为 2.0794415416798357 赛后 1.3862943611198906 临期 2.0794415416798357 解
禁 2.0794415416798357 乌镇 2.0794415416798357 好玩 2.0794415416798357 场
地 2.0794415416798357 白 2.0794415416798357 令人 2.0794415416798357 落败
2.0794415416798357 尔法 2.0794415416798357 落后 2.0794415416798357 认输
2.0794415416798357 一系列 2.0794415416798357 印象 1.3862943611198906 觉得
2.0794415416798357 负责人 2.0794415416798357 进入 2.0794415416798357 成
员 2.0794415416798357 大战 0.9808292530117262 数字 2.0794415416798357 谢
谢 2.0794415416798357 报告 2.0794415416798357 公司 2.0794415416798357 团
队 2.0794415416798357 返还 2.0794415416798357 完美 1.3862943611198906 三
分 2.0794415416798357 中国 2.0794415416798357 三局 2.0794415416798357 李
三 2.0794415416798357 围棋比赛 2.0794415416798357 获得 2.0794415416798357
请求 2.0794415416798357 为期 2.0794415416798357 一直 1.3862943611198906
带来 2.0794415416798357 字 2.0794415416798357 中盘 2.0794415416798357 真
本 2.0794415416798357 对决 1.3862943611198906 前 0.9808292530117262 徐柯
洁 2.0794415416798357 手上 2.0794415416798357 角落 2.0794415416798357 以
后 2.0794415416798357 好像 2.0794415416798357 战胜 2.0794415416798357 百
度 2.0794415416798357 真的 2.0794415416798357 双方 2.0794415416798357 完
成 2.0794415416798357 越来越 2.0794415416798357 单关 2.0794415416798357
知识 2.0794415416798357 量 1.3862943611198906 爱奇艺 2.0794415416798357
计算 2.0794415416798357 期 2.0794415416798357 开拓 2.0794415416798357 从
四得 2.0794415416798357 一句 2.0794415416798357 职业 2.0794415416798357
一个劲 2.0794415416798357 拼 2.0794415416798357 挑战 2.0794415416798357
后面 2.0794415416798357 完全 2.0794415416798357 口 2.0794415416798357 首
次 2.0794415416798357 棋局 2.0794415416798357 韩国 2.0794415416798357 难

下 2.0794415416798357 非常感谢 2.0794415416798357 表示 0.9808292530117262 肯定 2.0794415416798357 阿法 2.0794415416798357 白起 2.0794415416798357 四川 2.0794415416798357 有意思 2.0794415416798357 机 2.0794415416798357 即时 2.0794415416798357 影响 2.0794415416798357 样子 2.0794415416798357 奇 2.0794415416798357 继续 2.0794415416798357 围棋赛 2.0794415416798357 回时 2.0794415416798357 体现 2.0794415416798357 行 2.0794415416798357 比赛 0.9808292530117262 科技 2.0794415416798357 围棋 0.47000362924573563 接受 2.0794415416798357 体育 2.0794415416798357 近视 2.0794415416798357 Alphago 2.0794415416798357 值得 2.0794415416798357 不行 2.0794415416798357 接下来 2.0794415416798357 踩 2.0794415416798357 谷 2.0794415416798357 相爱 2.0794415416798357 宣告 2.0794415416798357 日记 2.0794415416798357 开局 2.0794415416798357 之后 2.0794415416798357 对齐 2.0794415416798357 这时候 2.0794415416798357 三月 2.0794415416798357 / 2.0794415416798357 提前 2.0794415416798357 对手 2.0794415416798357 观看 2.0794415416798357 此曲 2.0794415416798357 打破 2.0794415416798357 缺陷 1.3862943611198906 李世石 1.3862943611198906 这种 2.0794415416798357 看法 2.0794415416798357 深刻 2.0794415416798357 一想 2.0794415416798357 开展 2.0794415416798357 句子 2.0794415416798357 依曼体 2.0794415416798357 集 2.0794415416798357 手续 2.0794415416798357 程序 2.0794415416798357 没 2.0794415416798357 国际 2.0794415416798357 哈萨 2.0794415416798357 享受 1.3862943611198906 办法 2.0794415416798357 不再 2.0794415416798357 带 2.0794415416798357 今后 2.0794415416798357 联系 2.0794415416798357 五局 2.0794415416798357 心态 1.3862943611198906 三番 2.0794415416798357 过去 1.3862943611198906 代表 2.0794415416798357 放心大胆 2.0794415416798357 自由 2.0794415416798357 拼杀 2.0794415416798357 约 2.0794415416798357 子 2.0794415416798357 那个游戏 2.0794415416798357 有太多 2.0794415416798357 两局 2.0794415416798357 展开 2.0794415416798357 发布会 1.3862943611198906 首尔 2.0794415416798357 自我 2.0794415416798357 一局 2.0794415416798357 较量 2.0794415416798357 终于 2.0794415416798357 最后 2.0794415416798357 进行 0.9808292530117262 约定 2.0794415416798357 种 2.0794415416798357 实例 2.0794415416798357 歌声 2.0794415416798357 症是 2.0794415416798357 绝黑 2.0794415416798357 学习 2.0794415416798357 举行 0.9808292530117262 贯彻落实 2.0794415416798357 大师 2.0794415416798357 人字 2.0794415416798357 要注 2.0794415416798357 局 2.0794415416798357 进步 2.0794415416798357 对局 2.0794415416798357 三个 2.0794415416798357 实际 1.3862943611198906 顶尖 2.0794415416798357 点评 2.0794415416798357 最终 2.0794415416798357 比较 2.0794415416798357 气候 2.0794415416798357 苦 2.0794415416798357 他妈的 2.0794415416798357 对不起 2.0794415416798357 十只 2.0794415416798357 赛前 2.0794415416798357 那种 2.0794415416798357 贯彻 2.0794415416798357 钟路 2.0794415416798357 一书 2.0794415416798357 赔不起 2.0794415416798357 对尔法 2.0794415416798357 加载 2.0794415416798357 公布 2.0794415416798357 表现 1.3862943611198906 开心 2.0794415416798357 知道 2.0794415416798357 已经 1.3862943611198906 冲击 2.0794415416798357 项目 2.0794415416798357 般尔法 2.0794415416798357 加 2.0794415416798357 R 2.0794415416798357 大胆 2.0794415416798357 投子 2.0794415416798357 人机 1.3862943611198906 具体 2.0794415416798357 非常 2.0794415416798357 很多 1.3862943611198906 很瘦 2.0794415416798357 感谢

2.0794415416798357 名人 2.0794415416798357 几年 2.0794415416798357 理事 2.0794415416798357 探讨 2.0794415416798357 手柯洁 2.0794415416798357 长舒 2.0794415416798357 爱 2.0794415416798357 阿尔法 0.9808292530117262 比斯 2.0794415416798357 没想到 2.0794415416798357 下得 2.0794415416798357 大脑 2.0794415416798357 取得 1.3862943611198906 棋手 2.0794415416798357 同一 2.0794415416798357 当时 2.0794415416798357 一盘棋 2.0794415416798357 浙江 2.0794415416798357 我会 2.0794415416798357 半年 2.0794415416798357 更好 2.0794415416798357 一处 2.0794415416798357 问题 2.0794415416798357 令人生畏 2.0794415416798357 方面 2.0794415416798357 如期 2.0794415416798357 执行官 2.0794415416798357 保留 2.0794415416798357 慢 2.0794415416798357 中令 2.0794415416798357 下载 2.0794415416798357 天 2.0794415416798357 北京 2.0794415416798357 理念 2.0794415416798357 张家 2.0794415416798357 先 2.0794415416798357 取下 2.0794415416798357 活动 2.0794415416798357 算太准 2.0794415416798357 莫属 2.0794415416798357 调子 2.0794415416798357 爱好者 2.0794415416798357 钞票 2.0794415416798357 太狠 2.0794415416798357 开 2.0794415416798357 太 0.9808292530117262 inca 2.0794415416798357 谷歌 1.3862943611198906 赢得 2.0794415416798357 一下 2.0794415416798357 题 2.0794415416798357 胜利 1.3862943611198906 训练 2.0794415416798357 去年 2.0794415416798357