注：因设备原因，过大数据集难以计算，故缩减数据集进行运算。

```python
import numpy as np
import pandas as pd
import jieba
from sklearn.cluster import KMeans


def counts(values):  # 计数并排序
    d = {}
    for i in values:
        if i in d:
            d[i] += 1
        else:
            d[i] = 1
    l = sorted(d.items(), key = lambda x:x[1], reverse = True)
    return l


def main():
    # 读取弹幕集
    D = pd.read_csv(r'week2_danmuku.csv', encoding = 'utf-8').iloc[:, 0]
    # 停用词表stops
    with open(r'stopwords_list.txt', encoding='utf-8') as f:
        stop = f.read().split('\n')
    # 动态修改jieba分词词典
    for i in ['怡饭', '大佬', '盛月社', '户部巷', '周黑鸭', '云南白药']:
        jieba.add_word(i)
    # 分词
    tokens = D[:1000000].apply(lambda x : [i for i in jieba.lcut(x) if i not in stop])
    all_words = []  # 全部词
    for i in tokens:
        all_words.extend(i)
    word_count = pd.DataFrame(counts(all_words))  # 词频

    # 输出10个最高频和最低频词
    print("10 high-frequency words:\n", word_count[0:9])
    print("10 high-frequency words:\n", word_count[-10:-1])

    # 筛选词频数大于5的特征词
    characters = word_count[word_count.iloc[:, 1] > 5]
    characters.tail()  # 查看特征集的最后5行
    vocabulary = characters.iloc[:, 0]

    tokens = tokens[pd.Series(len(x) for x in tokens) > 3]  # 筛选分词数大于5的弹幕
     # 弹幕分词向量化(one-hot编码)
    vectors = np.array(list(tokens.apply(lambda x : [int(i in x) for i in vocabulary])))

    # K-Means聚类
    mC = KMeans(n_clusters = 5, init = 'k-means++')  # 欧式距离K-Means聚类
    mC.fit(vectors)
    labels = mC.labels_
    for i in range(6):  # 输出每类的5条弹幕
        print(tokens[labels == i][:5])


if __name__ == "__main__":
    main()
```

输出高频词：

```
10 high-frequency words:  10 low-frequency words:
      0      1                    0  1
0  哈哈哈哈  11247         5324   好几年  1
1    武汉   6500          5325    写个  1
2     吃   5841          5326   过生日  1
3     蒜   4482          5327  电视广告  1
4    好吃   3085          5328    硬推  1
5     藕   2899          5329   第二颗  1
6    真的   2295          5330    好美  1
7    热情   1934          5331    炸鸡  1
8    萝卜   1715          5332    蛮多  1
```

特征集的最后 5 个词及其频数：

```
      0   1
7    户部巷  28
8    啊啊啊  26
9     大蒜  24
10    热情  24
11    好吃  16
```

欧式距离 K-Means 聚类（k=5）结果：

```
4                  [辽, 两, 小时, 前]
58               [每次, 他俩, 视频, 饿]
61        [肯定, 大佬, 配, 音, 省钱, 盗月社]
83      [啊啊啊, 真的, 有种, 人生, 一串, 感觉]
111           [两位, 视频, 舒适, ❤, ]
Name: content, dtype: object
215      [武汉, 捂汗, 傻傻, 分不清楚]
257        [⊙, ∀, ⊙, 武汉, 诶]
267        [八月, 下旬, 武汉, 耍耍]
304         [武汉, 长沙, 还会, 远]
367     [天气, 武汉, 玩, 我敬, 汉子]
Name: content, dtype: object
104                            [小时, 前, 热热, 吃]
222                         [小时, 前, 抢救, 抢救, 吃]
405                      [刚到, 武汉, 明天, 吃, 吃, 吃]
414    [万松园, 当地人, 吃, 位置, 户部巷, 吉庆街, 户部巷, 稍微, 一点]
504                          [减肥, 计划, 中, 帮, 吃]
Name: content, dtype: object
1419            [广西, 嗦, 粉, 10086]
1426               [南昌, 嗦, 拌, 粉]
1431               [江西, 嗦, 拌, 粉]
1433     [万人, 血书, 江西, 嗦, 粉, 10000]
1441               [强, 推广, 西, 粉]
Name: content, dtype: object
354            [武汉, 小吃, 真的, 好吃, 肯定, 吃, 全乎]
392         [户部巷, 感觉, 好吃, 排队, 买, 长沙, 臭豆腐]
397           [户部巷, 感觉, 不太, 好吃, 游客, 感受]
425     [吉庆街, 骏骏, 牛肉, 粉, 旁边, 一家, 豆皮, 巨, 好吃]
430              [学校, 门口, 户部巷, 烤, 面筋, 好吃]
```

可以看到，距离小的样本（此处体现为同一类簇）语义更相似，距离大的样本（此处体现为不同类簇）语义差别较大。

可以通过距离的聚类方法，是具有相同语义特征的弹幕聚在一起，进而找到代表性弹幕。