In [51]:

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

In [2]:

```
with open('res/whitening_embeddings.pickle', 'rb') as f:
    sent_embdings = pickle.load(f)
```

In [3]:

```
with open("resources/新闻标题10000条.txt", 'r', encoding='utf-8') as f:
    lines = [line.replace("\u2022", "") for line in f.read().split("\n")]
```

In [86]:

```
# 暂时粗略将所有句向量分为10个簇
k = 10
K = KMeans(n_clusters=k)
model = K.fit(sent_embdings)
labels = model.predict(sent_embdings)
```

In [88]:

```
df = pd.DataFrame({"sents": lines, "labels": labels})
df[df['labels'] == 1]
```

Out[88]:

	sents	labels
11	2011年研究生入学考试今拉开帷幕 151万人报考	1
34	校园分裂的根源：物质差距首当其冲	1
54	四部门要求女生宿舍须实行封闭管理	1
57	中小学生每天下午在校都有课外活动 周四审批通过即实施	1
58	不吃早餐常熬夜 苏州一博士生开学第一天晕厥病危	1
...	...	...
9930	张异宾任南京大学党委书记	1
9934	国务院：加强党对高校的领导 建设世界一流大学	1
9938	悲剧！大一新生报到当天遭坠落天井盖砸中身亡[蜡烛]	1
9966	14岁少女因10元钱被7名同学殴打拍裸照	1
9970	杭州一位小学生，干了我们小时候都不敢干的事……网友全给跪了！	1

527 rows × 2 columns

可以发现这一类中似乎全是和“教育”这个主题相关的

In [61]:

```
# 列举离聚类中心最近的样本
centers = model.cluster_centers_

# centers[0]
```

In [53]:

```
def cosine(x, y):
    num = x.dot(y.T)
    denom = np.linalg.norm(x) * np.linalg.norm(y)
    return num / denom
```

In [91]:

```
cosines = []
for cluster in range(k):
    tmp_coss = []
    for i in range(len(lines)):
        tmp_coss.append(cosine(centers[cluster], sent_embdings[i]))
    cosines.append(tmp_coss)

# cosines[k, i]表示第i个样本和第k个聚类中心的余弦相似度
```

In [114]:

```
for cluster in range(k):
    idx = np.argpartition(np.array(cosines[cluster]), -5)
#     print(len(idx))
    print("第{}组".format(cluster))
    print("例句: ")
    for i, id in enumerate(idx[-5:]):
        print("{}.".format(i + 1) + lines[id])
    print("-----------------------------------------------")
```

```
第0组
例句:
1.多车碾压老人后逃逸 最后车辆停车报警被判赔
2.够任性！司机开报废车上高速 车灯坏了全靠乘客打手电[doge]
3.孕妇乘公交无人让座 司机发火停车指责乘客
4.货车撞死三轮车主逃逸 市民拦车追赶
5.北京一越野车失控后撞向公交车站，共致5名路人死亡
-----------------------------------------------
第1组
例句:
1.美国激辩“向中国学习”
2.特朗普发推谈朝核：中国赚我们钱还不帮忙 不会让朝鲜攻击美本土
3.李克强主持第四次中国—中东欧国家领导人会晤
4.外媒评习近平美国之行：展现中国大国形象
5.外交部：中方敦促韩国不使用武器对待中国渔民
-----------------------------------------------
第2组
例句:
1.春困？下面几招教你瞬间不困！[偷笑]
2.吃货们，你们吃全了吗？[馋嘴]
3.看到这画面，你能淡定吗？[偷笑]
4.今天，你表白了吗？[心]
5.它们是不是也感动了你？[心]
-----------------------------------------------
第3组
例句:
1.剑桥中国学生让家人汇500万炒房赚300多万
2.湖北彩民中奖1000万 给同事发千元红包[花心]
3.A股一个月蒸发4万亿，人均亏损8万元
4.网易云音乐获7.5亿元A轮融资：估值80亿元 用户破3亿
5.遗产税被曝80万起征 3000万遗产缴1034万
-----------------------------------------------
第4组
例句:
1.收评：沪指窄幅震荡微涨0.1% 次新股暴跌后大反弹
2.A股今日终结十连阳 沪指震荡下跌近0.8%
3.A股反弹又暴跌！沪指逼近4000点，创业板指跌8.05%
4.沪深股市今日普涨 沪指重上2500
5.收评：沪指震荡微升创业板跌近1% 次新股午后普涨
-----------------------------------------------
第5组
例句:
1.白岩松评医生手术台自拍 ：职业情商不够高
2.泰拳少年：境况堪比人妖
3.朱婷再获一大奖 当选国际体育新闻协会2016年杰出运动员
4.揭秘：应届大学毕业生为何爱“闪辞”
5.新浪微博发大招：手机游戏+电影点评+移动支付
-----------------------------------------------
第6组
例句:
```

1.裕丰牌花生油检出致癌物 塑化剂超标326倍
2.消费者网购化妆品汞超标17万倍 使用后中毒肾脏严重损害
3.乐百氏桶装水抽检不合格 致癌物质比规定超标40%
4.京津冀雾霾检出大量危险含氮有机颗粒物！
5.被忽略的污染：调查称室内含“塑化剂”等有毒物质

第7组
例句:
1.北京公积金调整：人均住房超29.4平米拒贷二套房
2.首个自住房项目昨起选房
3.购房一年买家被告知无该房 房产局回应：未搞清套数
4.自住型商品房，稳房价 OR推高房价？
5.广西首个限价房项目10月入市 共有近3000套房屋

第8组
例句:
1.女童想摸下男子怀中的小孩，男子暴怒一脚将她踢飞...[怒]
2.只因和女孩父母有矛盾 9岁女童遭虐打[怒]
3.女孩照顾残疾养父15年 亲生父母要接她 被女孩拒绝
4.一男一女让女童碰瓷撞车 称孩子是前妻的
5.男子勒死两子后服毒 称绝不让孩子活在单亲家庭

第9组
例句:
1.陕西富平贩婴案：副县长等6名责任人被免职
2.#湖北当阳爆炸事故# 已致21人死亡5人受伤
3.湖南衡阳破坏选举案50名嫌疑人被立案侦查
4.江西抚州烟花厂爆炸事故已致3人死亡 48人受伤
5.扬州仪征市发生一起刑事案件 现场5人死亡

通过以上的分组结果，可以发现 BERT-whitening + kmeans聚类算法不需样本训练，就能达到还不错的效果，后续可以通过聚类簇数的调整和使用少量样本进行bert微调，应该可以达到更好的效果。