统计方法

1. 高频词，观察，正常的喜欢抽奖
2. 不同语料词库 集合 A（自杀top1000）-B(正常top1000)，做个词云
3.

4. 根据word embedding结果，分析近邻词

> Word: 吃药
> ('吃药', 0.0)
> ('严重', 0.03874201789284121)
> ('旁边', 0.03907254000253807)
> ('演唱会', 0.03959063171829224)
> ('随便', 0.039676641110651154)
> ('无论', 0.04009705245593089)
> ('呼吸', 0.040441373196000295)
> ('照顾', 0.040781383683768185)
> ('记录', 0.04136005146988998)
>
> Word: 自杀
> ('勇敢', 0.03753376097938794)
> ('奇怪', 0.03838413283664744)
> ('呵呵', 0.03903476579059143)
> ('随便', 0.03905248312863383)
> ('分手', 0.03915504773076543)
> ('世上', 0.039538301042447474)
>
> Word: 作业
> ('复习', 5.268356063861754e-09)
> ('抱怨', 0.03767612014478099)
> ('无力', 0.038336747718780394)
> ('睡眠', 0.03880065739207785)
> ('只不过', 0.03940206762564953)
> ('位置', 0.039474194957598345)
> ('一周', 0.03997180765601731)
> ('懂得', 0.04037634909745646)
> ('网友', 0.040421693515838714)
>
> Word: 绝望
> ('意义', 0.0)
> ('是因为', 0.051160863150700425)
> ('值得', 0.05145259938307053)
> ('哪里', 0.05219584265932755)
> ('就算', 0.05273962402550551)
> ('一生', 0.05284762076858523)
> ('吃饭', 0.05386954072205553)
> ('本来', 0.05526760768254513)
> ('眼睛', 0.05568184526429394)
>
> Word: 抑郁
> ('每次', 0.05123636099870571)
> ('害怕', 0.07930934977188313)
> ('这些', 0.0908300682430974)
> ('情绪', 0.09087768943356675)
> ('不了', 0.09474769709513162)

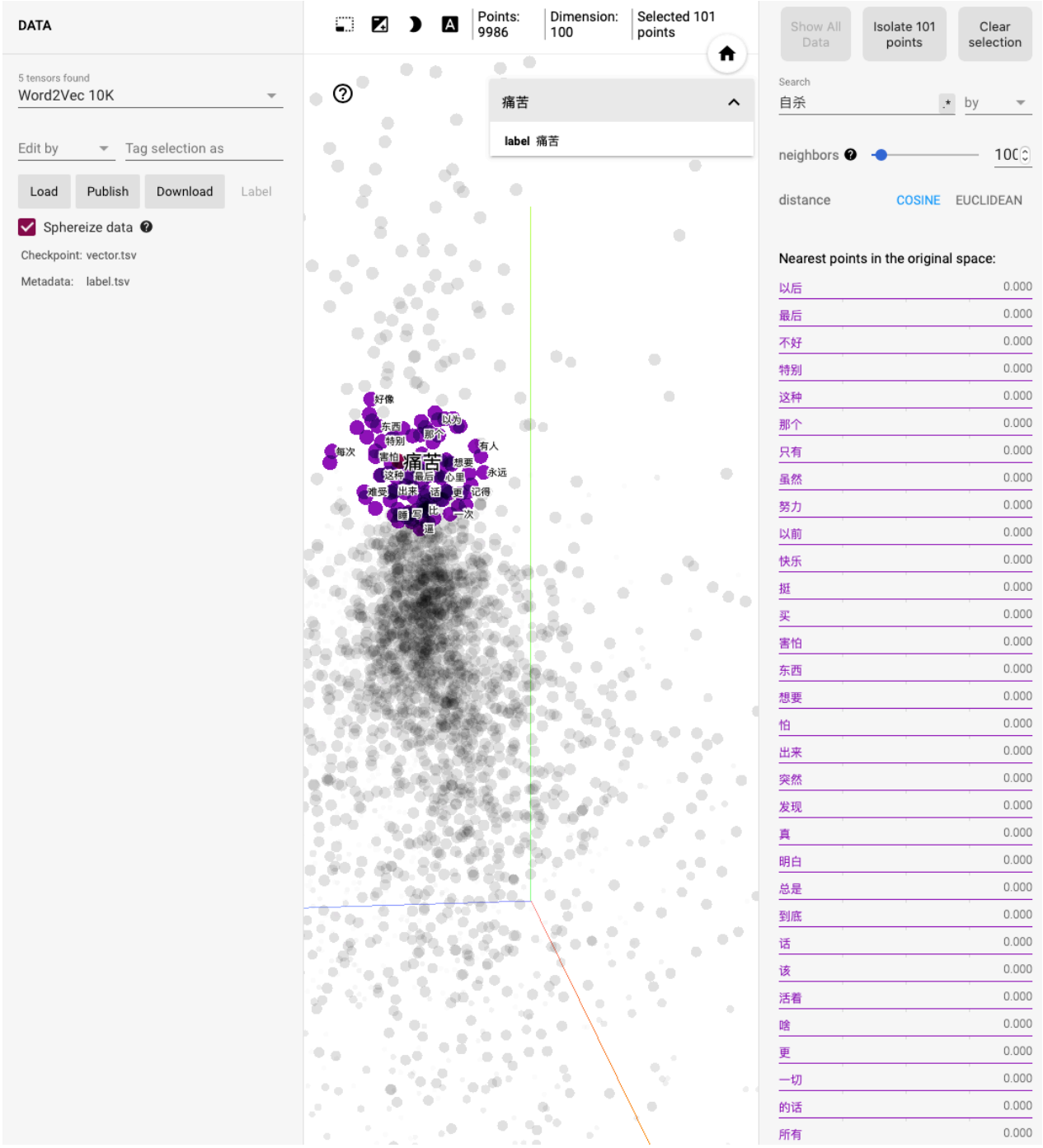
('那种', 0.09478192080182556)

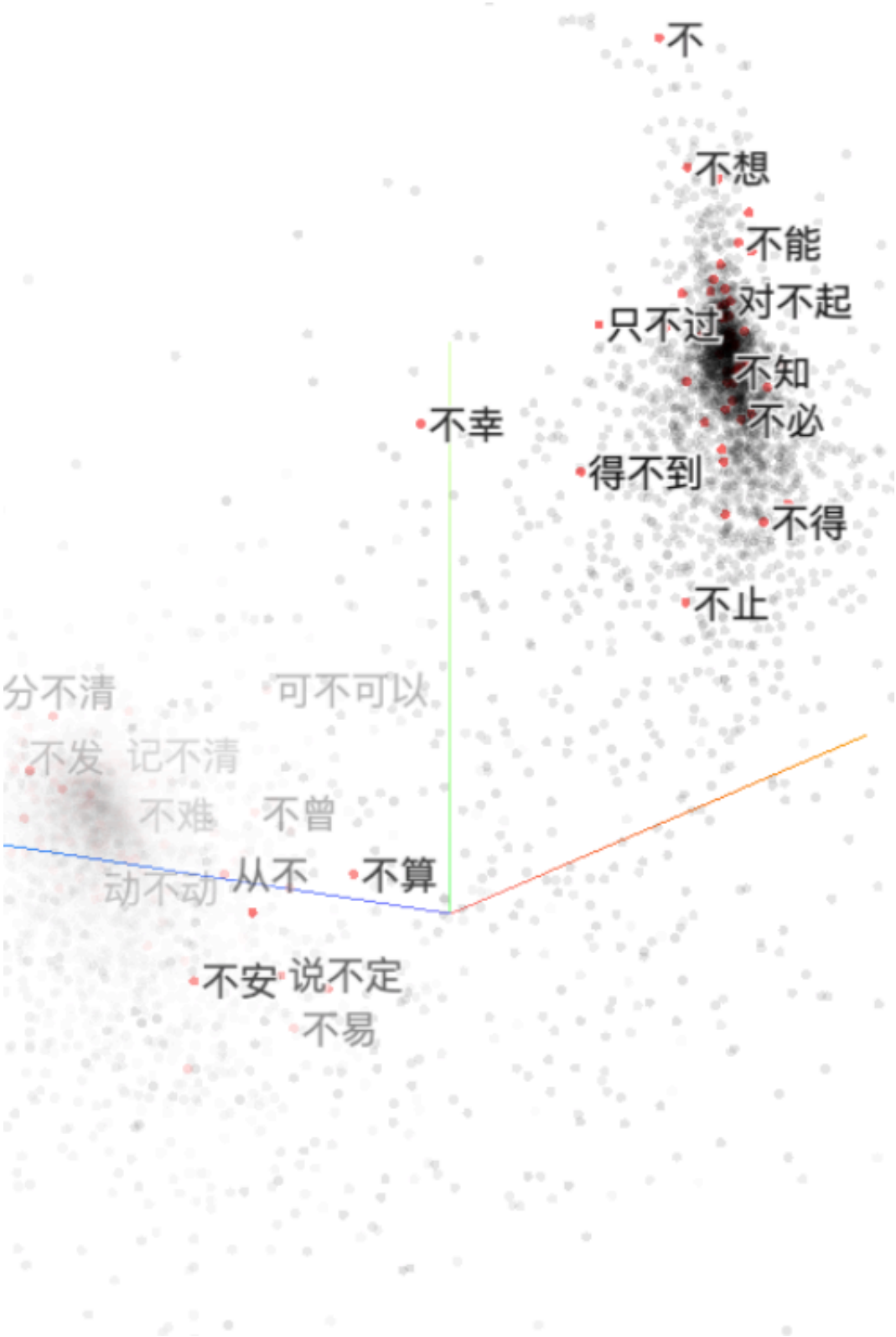
Word: 谈恋爱
('谈恋爱', 1.4901161193847656e-08)
('全世界', 0.03813180289993796)
('忍不住', 0.03961395016568274)
('勇气', 0.03970118886930373)
('心理', 0.03981430330030662)
('有个', 0.04013793531801365)
('没人', 0.04062415781231337)
('后面', 0.04116532058651564)
('结局', 0.04140081463627532)

Word: 分手
('长时间', 2.634178031930877e-09)
('棒棒', 0.0342738597570093)
('一趟', 0.034814792122923946)
('难吃', 0.03531841661105828)
('媒体', 0.03538535704406029)
('沮丧', 0.03549423475991918)
('不肯', 0.0356898363555036)
('醒过来', 0.035916840486831476)

Word: 醒过来
('老家', 1.862645149230957e-09)
('king', 0.03415536775107703)
('墙裂', 0.03417191064796669)
('最心水', 0.03421160094164291)
('前两天', 0.03442687528181577)
('人待', 0.034829550572338996)
('吃个', 0.034844217221763105)
('对得起', 0.034845155732380693)
('这颗', 0.03494241814149202)
('爷爷奶奶', 0.035168132198980594)

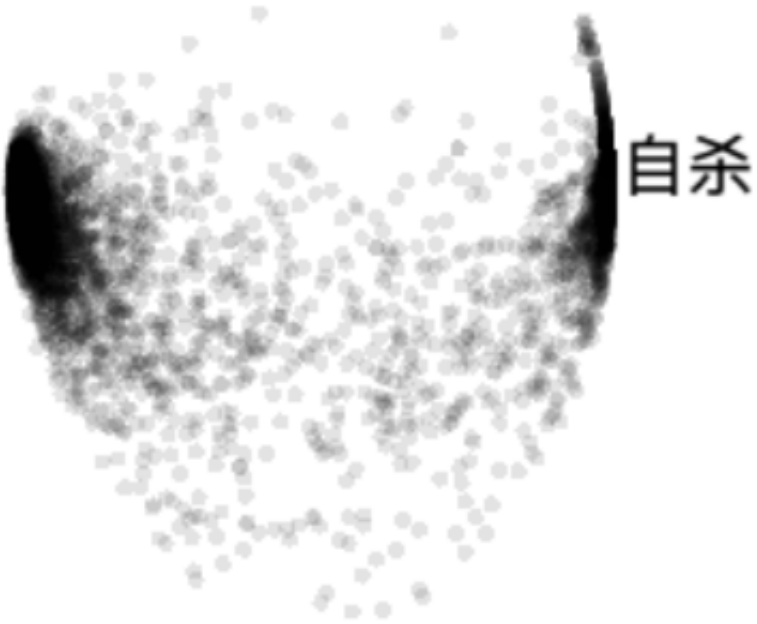
可视化


DATA
5 tensors found
Word2Vec 10K
Edit by
Tag selection as
Load
Publish
Download
Label
Sphereize data
Checkpoint: vector.tsv
Metadata: label.tsv
Points: 9986
Dimension: 100
Selected 101 points
痛苦
label 痛苦
好像
以为
东西
特别
那个
每次
害怕
痛苦
想要
有人
这种
最后
心里
永远
难受
出来
说
更
记得
睡
写
比
一次
逼
Show All Data
Isolate 101 points
Clear selection
Search
自杀
by
neighbors
100
distance
COSINE
EUCLIDEAN
Nearest points in the original space:
以后 0.000
最后 0.000
不好 0.000
特别 0.000
这种 0.000
那个 0.000
只有 0.000
虽然 0.000
努力 0.000
以前 0.000
快乐 0.000
挺 0.000
买 0.000
害怕 0.000
东西 0.000
想要 0.000
怕 0.000
出来 0.000
突然 0.000
发现 0.000
真 0.000
明白 0.000
总是 0.000
到底 0.000
话 0.000
该 0.000
活着 0.000
啥 0.000
更 0.000
一切 0.000
的话 0.000
所有 0.000



http://projector.tensorflow.org

使用PCA降维，可以看见词语集中于两个部分，右边一侧为敏感词



6. 平均句子词向量，观察离哪个近（optional）