1.wav2vec 聚类

1.1聚类结果

```
   1 44 34 5 7 16 5 27 27 4 2 5 40 40 34 39 37 27 0 39 35 11 16 26 1 4 46 2 32 4 24 10
30 27 33 36 36 46 3 19 10 48 2 34 41 41 40 5 37 32 24 33 39 35 42 24 0 47 47 8 2 40 5
17 35 0 19 30 11 0 17 35 21 37 19 30 14 29 33 30 25 23 2 10 34 41 40 34 3 19 30 27 28
12 28 18 33 1 17 11 26 33 35 0 9 47 19 5 9 35 19 1 36 2 41 41 30 37 3 9 18 33 36 19 10
48 2 41 40 39 36 45 9 11 0 33 37 32 3 33 39 36 45 45 29 4 24 29 9 2 5 25 13 29 36 4 31
36 31 40 6
   2 44 34 7 27 27 26 9 27 16 49 7 33 39 16 19 10 30 48 23 5 21 3 33 30 37 22 19 10 49
48 0 1 17 33 39 37 3 14 4 33 39 17 13 11 26 41 41 34 28 7 20 20 33 1 16 2 5 41 46 34 39
7 45 28 42 3 33 15 26 9 47 3 39 12 12 26 49 30 4 24 0 28 35 39 37 26 33 30 12 39 4 29
36 25 39 36 29 29 31 29 29 49 31 6 6
   3 34 34 34 7 19 37 27 19 17 16 19 36 21 32 2 10 2 40 34 42 12 28 20 36 4 26 39 15 15
9 32 36 4 4 24 19 48 19 17 18 33 17 35 39 36 4 26 33 37 19 10 30 48 0 49 25 45 28 42 12
46 13 41 40 34 39 37 45 26 17 22 21 25 2 40 5 39 47 43 19 5 48 39 36 8 29 39 36 25 29
33 29 47 47 29 39 25 29 17 13
   4 44 34 7 7 19 1 16 27 27 26 27 2 40 40 41 5 5 34 7 9 4 3 28 19 30 22 33 1 37 9 9 2
5 41 41 41 34 34 23 27 9 47 0 33 30 43 28 35 33 39 22 33 49 36 22 9 4 28 42 0 9 14 43 2
41 5 34 39 22 3 9 47 22 22 36 4 9 22 48 43 19 10 43 43 19 48 29 31
   5 44 34 7 16 5 1 42 27 15 3 49 14 27 19 5 16 33 42 0 33 30 25 26 2 41 41 34 36 7 33
39 23 3 19 30 48 0 33 14 23 33 4 10 36 24 9 43 28 23 3 42 23 26 33 10 49 30 48 29 36 43
2 40 41 40 36 4 14 45 19 30 4 24 1 9 1 4 0 33 37 23 29 30 25 23 2 41 40 39 36 12 33 30
36 36 4 29 10 39 25 2 40 41 46 34 49 35 19 49 14 45 45 28 23 3 12 45 48 47 29 9 47 29
49 37 25 26 48 25 2
```

```
    1 t ou5 f uɜ n ɑuɜ ph i2 n ts.h ɑɜ ŋ k uo5 l ə1 n ɑ5 f a5 ŋ f a5 ŋ ts. ə5 ŋ ts. ə5
ŋ p aiɜ p aiɜ n ə5 n n ə5 n t ə1 t ou5 f u2 kh uai5 ərɜ ts.h i.5 k uo5 x ou5 m aɜ n kh
ou2 ʲ yɜ ɕ iɑ5 ŋ
    2 ts. i.5 ʲ yɜ x ongɜ t ou5 l ɑ5 s i̪5 kh ueiɜ ər2 s. i.5 ts ɑuɜ j i2 t i5 ŋ x ɑu2 j
i5 yɛ5 ər5 s. i.ɜ tɕh i5 ʐ i.5 j ɑu5 tɕh y5 tsh a5 n tɕ iɑ5 k ai5 k uoɜ ɕ i5 n ts ongɜ
th ong2 t ə4 tɕ iou5 ts. i.ɜ j iɜ s. i.5
    3 t a5 n s. i.5 th ɑ5 n iɜ ŋ yæ5 n ʐ a5 ŋ ts i̪5 tɕ i2 x oɜ tɕh i5 ts i̪2 ərɜ n y2
ts.h i.5 kh ɑ5 ŋ j iɛ5 n tsh ai5 j iɛ2 p u5 ʐ a5 ŋ w ɑɜ ŋ s. i.5 ʲ yəɜ n l ɑu2 ʐ əɜ n
ts.h i.5 kh u2
    4 k a5 n s. u2 th iɛ5 n oɜ ts ai5 w oɜ s. ə2 ŋ t ə4 k w ɑ5 ts. ong2 ts. i.ɜ tɕh y5
t ou5 j iouɜ s uo2 f ɑ5 s. ə5 ŋ j i5 p a5 n j i5 n iɛɜ n f ɑ5 s. ə5 ŋ l iɑ2 ŋ t ai5 j
iɜ yu2 ŋ ts ai5 th u2 ts. ong5 yɛ5 t ong5
    5 ʲ yɜ s. i.5 ts. ong5 tɕh iou5 t ə2 j iou5 j iou5 yɛ5 s o5 ts. ong5 p iɛ5 n j iou2
j i5 l y2 s. u5 t a5 n t ə1 ɕ iɑ5 ŋ ʲ yə5 n l y5 j i2 ŋ j i5 ŋ j i5 ʲ yɜ w o2 t ə4 l iɜ
ŋ x w uəɜ n w o2 t ə4 ɕ i5 n m ə5 ŋ ts. ong5
```

1.2、长度对比

```
    行号 聚类后 原始标记
    1 91  65
    2 98  66
```

```
 3 86  62
 4 105 71
 5 102 76
 6 119 86
 7 90  69
 8 93  61
 9 120 83
10 121 82
```

2.生成MT训练数据

train.wrd

```
1 豆 腐 脑 品 尝 过 了 那 方 方 正 正 白 白 嫩 嫩 的 豆 腐 块 儿 吃 过 后 满 口 余 香
2 至 于 洪 都 拉 斯 奎 尔 是 早 已 定 好 一 月 二 十 七 日 要 去 参 加 该 国 新 总 统 的 就 职 仪 式
3 但 是 他 宁 愿 让 自 己 和 妻 子 儿 女 吃 糠 咽 菜 也 不 让 王 世 云 老 人 吃 苦
4 甘 薯 天 蛾 在 我 省 地 瓜 种 植 区 都 有 所 发 生 一 般 一 年 发 生 两 代 以 蛹 在 土 中 越 冬
5 于 是 中 秋 的 幽 幽 月 色 中 便 有 一 缕 疏 淡 的 香 韵 绿 影 映 逸 于 我 的 灵 魂 我 的 新 梦 中
```

train.phn

```
1 BU BK AI BZ BD AK BD BM AS AB BJ BM BD BF BP BA BD AK BH BH AD BR BR BR AK BM AG BN BV AG BP AX BJ BP BV AK BP AX BV BC BJ BP AK AK BT AK BT AB AK AX AB BX BT BJ BJ AX BA AG BF BP BM BI BH AD BR BR BK BZ BG AI BJ BN BD BM BA BF AK BH BF BC BP BA BF AV AU AL BS BH AD AH
2 BK BP AI AI AR AW AW BD AQ AK BN AK BD AU BZ BD AD BR BK BK AG BM AI BT AU BG BZ BP AF AY AE BE BJ BP BC BM AY BE AV BS AF AK AB AC AS AD BK BR BR BR BR BK BS AF AE AS AU AG BM AF AZ BC AC BS AF BJ BJ BP AF AZ AK AQ AE AU AL AS BC BN BV BV BP BA BV AK BX BF BJ AS BA BF BG AO AO BZ BG BH AG AH
3 BU BK AI AU AG BZ AR AG BR BK BM AI BE AR BE AR BD BJ AC AX AB AC AS AT AU AW AX AB BZ BZ AC BL AU BZ AX BX BT BC BE AV BJ AC AX BJ BM AX BE BS BT AB AU BM BH AD AG BR BR BQ BK BE BS AB BP AQ AK AF AY AK BD AU BG BE AO BF AK BH BF BF BZ BG BF BJ BP BA AD AH
4 BU AL BK AI AU AL BD AR BP AR BD BD AU AP AF AK AQ AU BZ BD BV BQ BL AB BP BI AZ BJ BN AQ BJ BG BY AU BG AV BQ BP BN AU BZ BN AY BP AF AU AX AD BQ AG BQ BK BL AB BP AP AF AB BC BE BL BL BW BS AM AU BZ AP AF AZ BG AK AM AF AW AL AP AZ AD AG BW BW BW BE BL BS AY AD BQ BZ AP AZ BP BN BA BP BN BA BF AV AM BJ BA AD AH
5 BU AI AU AG AR AR BK BK BR BK BK BP AI BV BC AC BD AK BD BS BD AE BS BD BE AV AM BZ BN BX AB BP BA AD BC AD BR BR BK BP BL BX AE BL AK AV AG BN AQ BP AF AZ AB BX AZ BZ AS AP AF BV BE AN AB AK AN AV BL BC BE AT AV AK BI AB BX AZ AK AO AO AW BI BI AZ BF AD BK BR BK BA BF BX AU AG AS AO AK BA BF BJ BG BA AD
```

3.生成词典

lexicon.txt

```
 5 一 ii i1
 6 一一 ii i1 ii i1
 7 一丁点 ii i4 d ing1 d ian3
 8 一万 ii i2 uu uan4
 9 一下子 ii i2 x ia4 z iy5
10 一专多能 ii i4 zh uan1 d uo1 n eng2
11 一世 ii i2 sh ix4
12 一丘之貉 ii i1 q iu1 zh ix1 h e2
13 一丝一毫 ii i4 s iy1 ii i4 h ao2
14 一丝不挂 ii i4 s iy1 b u2 g ua4
15 一丝不苟 ii i4 s iy1 b u4 g ou3
16 一个 ii i2 g e4
17 一个劲 ii i2 g e4 j in4
```

lexicon_generate.txt

```
 1 一 BE
 2 一 一 BE
 3 一 丁 点 BT BE BS
 4 一 万 BE BE BE AK
 5 一 下 子 AP BE
 6 一 专 多 能 BT AK BE BN
 7 一 世 BE
 8 一 丘 之 貉 BG BE 丘 BU
 9 一 丝 一 毫 BE BE AU BA
10 一 丝 不 挂 AG BE AU BI
11 一 丝 不 苟 AG 苟 BE AU
12 一 个 BE
13 一 个 劲 BL
14 一 个 半 BR BE
15 一 举 BE BL AC AV
16 一 举 两 得 AC AB BE AF
17 一 了 百 了 AK AK BE AK
18 一 事 无 成 BY BM BE AQ
19 一 二 AF
20 一 五 一 十 AQ AU BE BE
```