问题描述：

通过 nlp001 的测试，发现测试集合如果出现定位不准确的情况，就会导致问题。

比如首先用了如下测试集合，主要看测试语句和主题

序号 测试语句 主题 View

1 本月top3的销售是谁？ 销售TopN 销量表

2 上月top3的销售是谁？ 销售TopN 销量表

3 这个月哪三个销售卖的最好？ 销售TopN 销量表

4 今年3月的销售头五名是谁？ 销售TopN 销量表

5 今年帽子卖得最好的销售是谁? 销售TopN 销量表

6 给我一张销售报表 产品销量 销量表

7 今年的总销量是多少钱 产品销量 销量表

8 给我现在的总销量信息 产品销量 销量表

9 给我三月四月的销售总额 产品销量 销量表

10 2017年2月的销售总额 产品销量 销量表

11 马总要现在的销售总额 产品销量 销量表

12 现在帽子销售了多少总额 产品销量 销量表

13 这个月帽子总销量是？ 产品销量 销量表

14 鞋子这个月卖了多少钱 产品销量 销量表

15 上个月围巾的总销量? 产品销量 销量表

16 给我一张关于帽子的销售报表 产品销量 销量表

17 要销售报表，今年的。 产品销量 销量表

18 要销售报表，当月的。 产品销量 销量表

19 要销售报表，4月份和5月份的总和 产品销量 销量表

20 看销售报表，上半年总和 产品销量 销量表

21 看一下销售报告，马上 产品销量 销量表

22 我想看看销售情况，给我现在的销售状况 产品销量 销量表

这里只有两个主题，一个是销售TopN, 一个是产品销售

将训练集当成测试集输入后得到的结果全是产品销量，如下图 Y是希望的结果，predicted 是实际测试结果

Y= ['销售TopN', '销售TopN', '销售TopN', '销售TopN', '销售TopN', '产品销量', '产品销量', '产品销量', '产品销量', '产品销量', '产品销量', '产品销量', '产品销量', '产品销

量', '产品销量', '产品销量', '产品销量', '产品销量', '产品销量', '产品销量', '产品销量', '产品销量']

predicted= ['产品销量' '产品销量' '产品销量' '产品销量' '产品销量' '产品销量' '产品销量' '产品销量' '产品销量' '产品销量'

'产品销量' '产品销量' '产品销量' '产品销量' '产品销量' '产品销量' '产品销量' '产品销量' '产品销量' '产品销量'

'产品销量' '产品销量']

感觉 销售TopN 完全被忽略了。 为什么呢？ 后面逐步分析。

修改训练集测试一

序号 测试语句 主题 View

1 本月top3的销售是谁？ 01 销量表

2 上月top3的销售是谁？ 02 销量表

3 这个月哪三个销售卖的最好？ 03 销量表

4 今年3月的销售头五名是谁？ 04 销量表

5 今年帽子卖得最好的销售是谁? 05 销量表

6 给我一张销售报表 06 销量表

7 今年的总销量是多少钱 07 销量表

8 给我现在的总销量信息 08 销量表

9 给我三月四月的销售总额 09 销量表

10 2017年2月的销售总额 10 销量表

11 马总要现在的销售总额 11 销量表

12 现在帽子销售了多少总额 12 销量表

13 这个月帽子总销量是？ 13 销量表

14 鞋子这个月卖了多少钱 14 销量表

15 上个月围巾的总销量? 15 销量表

16 给我一张关于帽子的销售报表 16 销量表

17 要销售报表，今年的。 17 销量表

18 要销售报表，当月的。 18 销量表

19 要销售报表，4月份和5月份的总和 19 销量表

20 看销售报表，上半年总和 20 销量表

21 看一下销售报告，马上 21 销量表

22 我想看看销售情况，给我现在的销售状况 22 销量表

将测试结果的主题全部设置为不一样。 测试结果就完全符合预期

Y= ['01', '02', '03', '04', '05', '06', '07', '08', '09', '10', '11', '12', '13', '14', '15', '16', '17', '18', '19', '20', '21', '22']

predicted= ['01' '02' '03' '04' '05' '06' '07' '08' '09' '10' '11' '12' '13' '14' '15'

'16' '17' '18' '19' '20' '21' '22']

这说明主题相对独立的时候，结果非常好。

修改训练集测试二

序号 测试语句 主题 View

1 本月top3的销售是谁？ grpB 销量表

2 上月top3的销售是谁？ 02 销量表

3 这个月哪三个销售卖的最好？ 03 销量表

4 今年3月的销售头五名是谁？ 04 销量表

5 今年帽子卖得最好的销售是谁? 05 销量表

6 给我一张销售报表 06 销量表

7 今年的总销量是多少钱 07 销量表

8 给我现在的总销量信息 08 销量表

9 给我三月四月的销售总额 09 销量表

10 2017年2月的销售总额 10 销量表

11 马总要现在的销售总额 grpA 销量表

12 现在帽子销售了多少总额 grpA 销量表

13 这个月帽子总销量是？ 13 销量表

14 鞋子这个月卖了多少钱 14 销量表

15 上个月围巾的总销量? 15 销量表

16 给我一张关于帽子的销售报表 16 销量表

17 要销售报表，今年的。 17 销量表

18 要销售报表，当月的。 18 销量表

19 要销售报表，4月份和5月份的总和 19 销量表

20 看销售报表，上半年总和 20 销量表

21 看一下销售报告，马上 21 销量表

22 我想看看销售情况，给我现在的销售状况 grpB 销量表

有意识的修改 grpB 和 grpA ,不过不是符合实际意义。 发现测试结果集完全偏向 grpB 和 grpA

Y= ['grpB', '02', '03', '04', '05', '06', '07', '08', '09', '10', 'grpA', 'grpA', '13', '14', '15', '16', '17', '18', '19', '20', '21', 'grpB']

predicted= ['grpB' 'grpB' 'grpA' 'grpB' 'grpA' 'grpA' 'grpA' 'grpA' 'grpA' 'grpA'

'grpA' 'grpA' 'grpA' 'grpA' 'grpA' 'grpA' 'grpA' 'grpA' 'grpA' 'grpA'

'grpA' 'grpB']

这里不是很理解为什么那些比如02~21 的主题都偏向了 grpA or grpB , 需要研究一下

修改训练集测试三

序号 测试语句 主题 View

1 本月top3的销售是谁？ 销售TopN 销量表

2 上月top3的销售是谁？ 销售TopN 销量表

3 这个月哪三个销售卖的最好？ 销售TopN 销量表

4 今年3月的销售头五名是谁？ 销售TopN 销量表

5 今年帽子卖得最好的销售是谁? 销售TopN 销量表

6 给我一张销售报表 产品销量 销量表

7 今年的总销量是多少钱 产品销量 销量表

8 给我现在的总销量信息 产品销量 销量表

9 给我三月四月的销售总额 产品销量 销量表

10 2017年2月的销售总额 产品销量 销量表

11 马总要现在的销售总额 某个产品销量 销量表

12 现在帽子销售了多少总额 某个产品销量 销量表

13 这个月帽子总销量是？ 某个产品销量 销量表

14 鞋子这个月卖了多少钱 某个产品销量 销量表

15 上个月围巾的总销量? 某个产品销量 销量表

**16 给我一张关于帽子的销售报表 好吧 销量表**

17 要销售报表，今年的。 好吧 销量表

18 要销售报表，当月的。 好吧 销量表

19 要销售报表，4月份和5月份的总和 好吧 销量表

20 看销售报表，上半年总和 好吧 销量表

21 看一下销售报告，马上 nishuo 销量表

22 一个全新的主题 新主题 销量表

这次将主题相对分散，看看结果。

Y= ['销售TopN', '销售TopN', '销售TopN', '销售TopN', '销售TopN', '产品销量', '产品销量', '产品销量', '产品销量', '产品销量', '某个产品销量', '某个产品销量', '某个产品销

量', '某个产品销量', '某个产品销量', '好吧', '好吧', '好吧', '好吧', '好吧', 'nishuo', '新主题']

predicted= ['销售TopN' '销售TopN' '销售TopN' '销售TopN' '销售TopN' '产品销量' '产品销量' '产品销量' '产品销量'

'产品销量' '某个产品销量' '某个产品销量' '某个产品销量' '某个产品销量' '某个产品销量' '**产品销量'** '好吧' '好吧' '好吧'

'好吧' '好吧' '产品销量']

发现了有意思的现象，看第16句，原来的主题应该是 好吧， 结果被分配成 某个产品销量，这个数据还是很有意思的，因为原话中 16 句和 12,13 句似乎更接近，所以被分配到那个主题群了。

结论

SVC 应该是工作的，不过对训练集中的主题必须相对独立，干扰因素必须足够小才可以带来比较大的预期。