eeqi 词性标注系统

隐马尔科夫模型及其在词性标注中的应用

组名：eeqi

组员：朱昆睿 20151002152

方洁梅 20151002122

周莹莹 20151002247

2018 年 3 月 8 日

一、任务描述

利用词性之间的转移矩阵（如图 1）与从词性到词语的发射矩阵（如图 2）为句子：the bear is on the move 标注词性。（初始状态为 AT：0.2 ， BEZ：0.1，IN：0.1，NN：0.2，VB:0.3 ， PER 0.1）

表 10.3 Brown 语料库中一些标记转移的理想计数。例如，NN在AT后面出现了48636次

第一个标记 \ 第二个标记	AT	BEZ	IN	NN	VB	PERIOD
AT	0	0	0	48 636	0	19
BEZ	1973	0	426	187	0	38
IN	43 322	0	1325	17 314	0	185
NN	1067	3720	42 470	11 773	614	21 392
VB	6072	42	4758	1476	129	1522
PERIOD	8016	75	4656	1329	954	0

图 1

表 10.4 Brown 语料库中一些词和标记同时出现的理想计数。例如，move被标注成NN一共出现了36次

	AT	BEZ	IN	NN	VB	PERIOD
bear	0	0	0	10	43	0
is	0	10 065	0	0	0	0
move	0	0	0	36	133	0
on	0	0	5484	0	0	0
president	0	0	0	382	0	0
progress	0	0	0	108	4	0
the	69 016	0	0	0	0	0
.	0	0	0	0	0	48 809

图 2

二、理论描述:

隐马尔科夫模型由 1.观测序列 2.隐序列 3.转移矩阵 4.发射矩阵 5.初始值 五个部分组成 。

对于词性标注任务，和上面提到的五元组分别对应的内容是：1.分好词的句子（the bear is on the move）2.句子中每个词语对应的词性（问题的输出结果） 3.从一种词性转移到另外一种词性的概率（如图 1） 4.在确定词性的情况下指定某一个词的概率（如图 2： 在词性确定是 AT 的情况下 词语为 the 的情况在语料库中出现了 69016 次，用 69016/表格中的数字之和，就能得到从 AT 到 the 的转移概率）5.词性的初始状态（ AT：0.2 ， BEZ：0.1，IN：0.1，NN：0.2，VB:0.3 ， PER 0.1）

维特比算法用于寻找最可能的隐藏序列：通过列出可能的隐藏状态序列并且计算对于每个组合相应的观察序列的概率来找到最可能的隐藏状态序列。最可能的隐藏状态序列是使下面这个概率最大的组合。

二、算法描述：

1. 平滑

由转移矩阵（图 1）和发射矩阵（图 2）可以看出，图标内有较多次数为零的情况，为了不让概率相乘的结果为零，我们采取 laplace 算法（分子分母同时加上一个常数，这里取 1）把数据平滑。平滑之后的转移矩阵和发射矩阵如下表格（表 3）（表 4）

	AT	BEZ	IN	NN	VB	PERIOD
AT	1	1	1	48637	1	20
BEZ	1974	1	427	188	1	39
IN	43323	1	1326	17315	1	186
NN	1068	3721	42471	11774	615	21393
VB	6073	43	4759	1477	130	1523
PERIOD	8017	76	4657	1330	955	1

（表 3）

	AT	BEZ	IN	NN	VB	PERIOD
bear	1	1	1	11	44	1
is	1	10066	1	1	1	1
move	1	1	1	37	134	1
on	1	1	5485	1	1	1
president	1	1	1	383	1	1
progress	1	1	1	109	5	1
the	69017	1	1	1	1	1
。	69023	10072	5491	543	187	7

（表 4）

将每一个元素除以表格的数值总和得到转移概率矩阵（表 5）和发射概率矩阵（表 6）

	AT	BEZ	IN	NN	VB	PERIOD
AT	2.06E-05	2.06E-05	2.06E-05	0.999506792	2.06E-05	4.11E-04
BEZ	0.750570342	3.80E-04	0.162357414	7.15E-02	3.80E-04	1.48E-02
IN	0.69704917	1.61E-05	0.021334792	0.278591196	1.61E-05	2.99E-03
NN	0.013178352	0.045914464	0.524061598	0.145282693	0.007588658	0.263974236
VB	0.433630846	3.07E-03	0.339807212	0.105462335	9.28E-03	0.108746876
PERIOD	0.533187018	5.05E-03	0.309723331	0.088454376	0.063514233	6.65E-05

（表 5）

	AT	BEZ	IN	NN	VB	PERIOD
bear	1.45E-05	9.93E-05	1.82E-04	2.02E-02	2.34E-01	2.05E-05
is	1.45E-05	0.999305073	1.82E-04	1.84E-03	5.32E-03	2.05E-05

move	1.45E-05	9.93E-05	1.82E-04	6.80E-02	7.13E-01	2.05E-05
on	1.45E-05	9.93E-05	0.998725419	1.84E-03	5.32E-03	2.05E-05
president	1.45E-05	9.93E-05	1.82E-04	0.704044118	5.32E-03	2.05E-05
progress	1.45E-05	9.93E-05	1.82E-04	2.00E-01	2.66E-02	2.05E-05
the	0.999898586	9.93E-05	1.82E-04	1.84E-03	5.32E-03	2.05E-05
。	1.45E-05	9.93E-05	1.82E-04	1.84E-03	5.32E-03	0.999856607

（表 6）

2. 计算

从单词 the 开始计算，由六个词性的初始值和单词 the 到各个词性的发射概率得到第一个单词在各个词性上的初始值。
从第二行开始，计算每个单词各个属性的发射矩阵概率、转移矩阵概率与前一行对应属性的取值三者的成绩，将取值最大的作为该结点的最佳左邻词性，一直到最后一行。
通过比价最后一行各个属性取值的大小找出取值最大的词性作为尾词（move）的词性，再根据最佳左邻推出最佳路径。

三、详例描述：

1.从 the 开始计算，根据各种词性的初始值与发射矩阵计算出各种词性的初始概率：

初始值：0.2*0.9998985859990728AT: 0.19997971719981456
初始值：0.1*9.927529038022436E-5BEZ: 9.927529038022437E-6
初始值：0.1*1.820830298616169E-4IN: 1.8208302986161692E-5
初始值：0.2*0.001838235294117647NN: 3.676470588235294E-4
初始值：0.3*0.005319148936170213VB: 0.0015957446808510637
初始值：0.1*2.0484667226580904E-5PER: 2.0484667226580905E-6

2 利用左邻值和转移概率以及发射概率三个值计算 bear 下各种词性的取值，并且选择最大值作为该节点的取值结果，记录最佳左邻下标：

3. bear-0:
AT 前一结点可能取值：
(取 AT 概率：5.953944703598344E-11,
取 BEZ 概率：1.0795243491136101E-10,
取 IN 概率：1.838792663357369E-10,

取 NN 概率：7.019272050752722E-11,
取 VB 概率：1.0024978502537124E-8,
取 PER 概率：1.5823711505643595E-11,)最大概率： 1.0024978502537124E-8 下标：<4>
BEZ 前一结点可能取值：(
取 AT 概率：4.0798677575813767E-10,
取 BEZ 概率：3.747370068470672E-13,
取 IN 概率：2.9084093291966606E-14,
取 NN 概率：1.675798439955675E-9
,取 VB 概率：4.863959094794769E-10,
取 PER 概率：1.0279011561042992E-12,)最大概率： 1.675798439955675E-9 下标：<3>
IN 前一结点可能取值：(
取 AT 概率：7.482976679191044E-10,
取 BEZ 概率：2.9348287444963713E-10,
取 IN 概率：7.073385999889894E-11,
取 NN 概率：3.508188365497453E-8,
取 VB 概率：9.873371278245249E-8,
取 PER 概率：1.1552402332748974E-10,)最大概率： 9.873371278245249E-8 下标：<4>
NN 前一结点可能取值：(
取 AT 概率：0.004041713127552971,
取 BEZ 概率：1.434950937334377E-8,
取 IN 概率：1.0257243000376943E-7,
取 NN 概率：1.0800373203471155E-6,
取 VB 概率：3.402942204174354E-6,
取 PER 概率：3.663886592825242E-9,)最大概率： 0.004041713127552971 下标：<0>
VB 前一结点可能取值：(
取 AT 概率：9.618331641346582E-7
,取 BEZ 概率：8.834464802058637E-10,
取 IN 概率：6.856605934105906E-11,
取 NN 概率：6.529664820345784E-7,
取 VB 概率：3.466717651928295E-6,
取 PER 概率：3.0450525719158054E-8,)最大概率： 3.466717651928295E-6 下标：<4>
PER 前一结点可能取值：(
取 AT 概率：1.6836965779182337E-9,
取 BEZ 概率：3.015636129398567E-12,
取 IN 概率：1.1162365024843028E-12,
取 NN 概率：1.9880236662790917E-9,
取 VB 概率：3.5547503763187973E-9,
4. 取 PER 概率：2.790779405318953E-15,)最大概率： 3.5547503763187973E-9 下标：<4>

5. 依次类推直到最后一组值 move：

move-2:

AT 前一结点可能取值：(
取 AT 概率：6.23978379952938E-15,
取 BEZ 概率：1.2078089964367506E-18,
取 IN 概率：1.1796481871356655E-15,
取 NN 概率：2.9400911410211183E-15,
取 VB 概率：1.6167334114580266E-17,
取 PER 概率：1.423959111484234E-17,)最大概率： 6.23978379952938E-15 下标：<0>
BEZ 前一结点可能取值：(
取 AT 概率：4.2757355006325417E-14,
取 BEZ 概率：4.1926866081279546E-21,
取 IN 概率：1.8658437468262305E-19,
取 NN 概率：7.019246599684558E-14,
取 VB 概率：7.844131714127557E-19,
取 PER 概率：9.249974106377405E-19,)最大概率： 7.019246599684558E-14 下标：<3>
IN 前一结点可能取值：(
取 AT 概率：7.842222086283976E-14,
取 BEZ 概率：3.2835874091348004E-18,
取 IN 概率：4.53781828585599E-16,
取 NN 概率：1.4694392039308887E-12,
取 VB 概率：1.5922836368406493E-16,
取 PER 概率：1.0395885033282238E-16,)最大概率： 1.4694392039308887E-12 下标：<3>
NN 前一结点可能取值：(
取 AT 概率：1.4247519799392478E-6,
取 BEZ 概率：5.400225669278468E-16,
取 IN 概率：2.213397553566188E-12,
取 NN 概率：1.521656130551655E-10,
取 VB 概率：1.845944228410944E-14,
取 PER 概率：1.1090221597402E-14,)最大概率： 1.4247519799392478E-6 下标：<0>
VB 前一结点可能取值：(
取 AT 概率：3.0698461784653155E-10,
取 BEZ 概率：3.0102196357937056E-17,
取 IN 概率：1.3396182469567038E-15,
取 NN 概率：8.329361843642588E-11,
取 VB 概率：1.702650684236279E-14,
取 PER 概率：8.345187125236465E-14,)最大概率： 3.0698461784653155E-10 下标：<0>
PER 前一结点可能取值：(
取 AT 概率：1.7645280823431016E-13,
取 BEZ 概率：3.373997492560466E-20,
取 IN 概率：7.161037744005562E-18,
取 NN 概率：8.327032671059799E-14,
取 VB 概率：5.732764116484759E-18,
取 PER 概率：2.5113929566582066E-21,)最大概率： 1.7645280823431016E-13 下标：<0>

6. 可见最后一组中的最大概率里面 NN 的取值 1.4247519799392478E-6 下标：<0>

下标：<0> 是最大的，所以 move 的词性取 NN，最佳左邻为 0 所以 move 前一词语 the 取词性 AT，依次类推计算得该句子的词性标注为
the/AT bear/NN is/BEZ on/IN the/AT move/NN

四、程序说明

开发平台：NetBeans8.2
语言：Java
jdk：1.8

scale.java 用于将数据标准化（从频数转换成概率）
词性标注.java 用于计算 the bear is on the move 的词性标注最后以 词语-词性的形式输出结果

五、问题反馈

1. 计算各个属性的取值的过程涉及三个值：发射概率，转移概率与前一属性取值的乘法计算，开始的时候又出现下标混乱的问题倒是结果与预测不一致，后来借助一个维度更少的例子逐个演算、多次调试找出正确的下标取值。

2. 在开始计算的时候又遇到空指针报错的问题，原因在于忽视了边界处理（第一行的词性没有左邻），在调试后问题得到了解决。

3.在标准化转移矩阵和发射矩阵的过程中没有留意到分母的问题：发射矩阵每列的数值之和邓宇一，转移矩阵每行之和等于一