

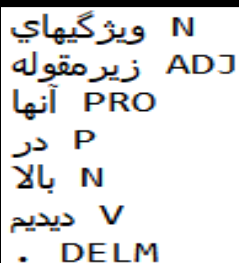
# تمرین دوم پردازش زبان های طبیعی

## فرشاد برجلی زاده

۶۱۰۳۹۹۰۱۵

در این تمرین قصد داریم که با استفاده از الگوریتم ویتربی عملیات part of speech tagging را بروی دیتاست داده شده پیاده سازی کنم.

دادگانی که در اختیار داریم شامل دو فایل Train.txt, Test.txt می شود که در آن ها هر کلمه برچسب گذاری شده است و با یک فاصله رو به روی آن نوشته شده است. نمونه آن را در تصویر مقابل مشاهده می کنید.



تعداد داده های که برچسب گذاری شده اند در هر دو فایل ۲۵۹۷۹۴ کلمه است که از این تعداد در داده آموزشی ۱۸۹۴۹ کلمه منحصر به فرد برچسب گذاری شده وجود دارد، در واقع تعداد واژگان منحصر به فرد ما برابر با همین تعداد است و در کل پیکره متنی ما ۲۳ برچسب متفاوت وجود دارد که در تصویر زیر به نمایش در آمده اند.

Number of tags in corpus: 23

{'DEFAULT', 'PP', 'AR', 'V', 'PS', 'N', 'QUA', 'P', 'CON', 'MS', 'ADV', 'OH', 'OHH', 'DELM', 'IF', 'MQUA', 'PRO', 'IN T', 'SPEC', 'MORP', 'NP', 'DET', 'ADJ'}

الگوریتم ویتربی درواقع برای حل مسئله دوم مارکوف مورد استفاده قرار می گیرد جایی که ما مشاهدات و مدل را داریم و سعی داریم متناسب با این مشاهدات بهترین دنباله حالات را بدست بیاوریم در اینجا هم ما چنین قصدی داریم که به دنباله ای از کلمات که جملات ما را تشکیل می دهند بهترین برچسب را بنیم به همین خاطر به سراغ الگوریتم ویتربی می رویم.

برای این کار نیاز است که احتمال کلمات به شرط برچسب را محاسبه کنیم و هم چنین می بایست احتمال اینکه بعد از یه برچسب ممکن است چه برچسب دیگری را مشاهده کنیم را هم محاسبه کنیم تا بتوانیم از روی آن ها زنجیره مارکوف مدل را ساخته و بر روی آن الگوریتم ویتربی را اجرا کنیم، برای درک بیشتر، ماتریس احتمالات وقوع یک برچسب بعد از یک برچسب دیگر را نمایش می دهیم.

	ADV	INT	MS	PS	DET	MQUA	DELM	AR	MORP	IF	OH	CON	ADJ	N	QUA	NP	P	PRO	V	OHH	SPEC	DEFAULT
ADV	0.022490	0.000000	0.000549	0.000000	0.018925	0.000274	0.041141	0.000000	0.000000	0.005485	0.000000	0.027153	0.172518	0.399890	0.007131	0.000000	0.213933	0.024136	0.065551	0.000000	0.000549	0.000000
INT	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
MS	0.005051	0.000000	0.000000	0.000000	0.000000	0.000000	0.353535	0.000000	0.000000	0.000000	0.000000	0.080808	0.146465	0.232323	0.000000	0.000000	0.090909	0.000000	0.085859	0.000000	0.000000	0.000000
PS	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.419355	0.000000	0.000000	0.000000	0.032258	0.064516	0.000000	0.032258	0.000000	0.000000	0.451613	0.000000	0.000000	0.000000	0.000000	0.000000
DET	0.000000	0.000000	0.000000	0.000000	0.000242	0.000000	0.000969	0.000000	0.000000	0.000000	0.000000	0.012833	0.002906	0.949395	0.000242	0.000000	0.002906	0.001211	0.008717	0.000000	0.019128	0.000000
MQUA	0.000000	0.000000	0.012346	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.012346	0.000000	0.790123	0.074074	0.000000	0.098765	0.012346	0.000000	0.000000	0.000000	0.000000
DELM	0.031458	0.000077	0.002863	0.001006	0.023294	0.000387	0.070887	0.007816	0.000039	0.006655	0.000271	0.093174	0.040435	0.551888	0.008474	0.000000	0.120647	0.022675	0.015942	0.000116	0.000658	0.000580
AR	0.000000	0.000000	0.000000	0.000851	0.000000	0.000000	0.175319	0.797447	0.000000	0.000000	0.000000	0.012766	0.001702	0.008511	0.000000	0.000000	0.000851	0.000851	0.001702	0.000000	0.000000	0.000000
MORP	0.007576	0.000000	0.000000	0.000000	0.000000	0.000000	0.090909	0.000000	0.000000	0.000000	0.000000	0.083333	0.030303	0.303030	0.015152	0.000000	0.333333	0.007576	0.128788	0.000000	0.000000	0.000000
IF	0.020202	0.000000	0.005051	0.000000	0.042929	0.000000	0.002525	0.000000	0.000000	0.002525	0.000000	0.010101	0.055556	0.613636	0.022727	0.000000	0.136364	0.025253	0.060606	0.000000	0.002525	0.000000
OH	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	1.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
CON	0.038744	0.000000	0.000498	0.000136	0.024305	0.001177	0.012311	0.000407	0.000000	0.006427	0.000091	0.039966	0.093012	0.527926	0.017742	0.000045	0.183987	0.024803	0.026885	0.000045	0.000996	0.000000
ADJ	0.011732	0.000000	0.000101	0.000000	0.006604	0.000335	0.080853	0.000000	0.000302	0.000503	0.000000	0.115581	0.085981	0.253419	0.003553	0.000000	0.189528	0.017800	0.233072	0.000000	0.000402	0.000034
N	0.008956	0.000000	0.000787	0.000000	0.008407	0.000210	0.085024	0.000146	0.001098	0.000174	0.000009	0.073250	0.200293	0.361559	0.004117	0.000000	0.124801	0.020401	0.109706	0.000000	0.000842	0.000037
QUA	0.001099	0.000000	0.000000	0.000000	0.015934	0.000000	0.000549	0.000000	0.000549	0.000000	0.000000	0.000000	0.003846	0.814286	0.000549	0.000000	0.116484	0.015934	0.001099	0.000000	0.029670	0.000000
NP	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.400000	0.100000	0.300000	0.000000	0.000000	0.100000	0.000000	0.100000	0.000000	0.000000	0.000000
P	0.006077	0.000000	0.000460	0.000000	0.049652	0.000061	0.004850	0.000092	0.000000	0.000031	0.000000	0.005832	0.024187	0.783456	0.016083	0.000276	0.047729	0.044966	0.014119	0.000000	0.001995	0.000000
PRO	0.022401	0.000000	0.000175	0.000000	0.005775	0.000875	0.034127	0.000175	0.000000	0.000175	0.000000	0.097830	0.089954	0.263213	0.004025	0.000000	0.321666	0.011551	0.145782	0.000000	0.000350	0.000000
V	0.010167	0.000000	0.000045	0.000045	0.004837	0.000179	0.502486	0.000314	0.000000	0.001120	0.000045	0.286335	0.011556	0.070722	0.002508	0.000000	0.069915	0.006494	0.002606	0.000045	0.000179	0.000000
OHH	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.200000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.200000	0.000000	0.000000	0.000000	0.600000	0.000000	0.000000	0.000000	0.000000
SPEC	0.000000	0.000000	0.005714	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.020000	0.965714	0.000000	0.000000	0.002857	0.002857	0.002857	0.000000	0.000000	0.000000
DEFAULT	0.050000	0.000000	0.000000	0.000000	0.000000	0.000000	0.100000	0.000000	0.000000	0.000000	0.000000	0.050000	0.000000	0.650000	0.000000	0.000000	0.050000	0.100000	0.000000	0.000000	0.000000	0.000000

احتمال صفر نشان دهنده این است که بعد از برچسب فعلی برچسب متناظر با ستون موجود در ماتریس نخواهد آمد.

با توجه به این که زمان اجرای الگوریتم بسیار طولانی و طاقت فرسا است به همین خاطر به طور تصادفی ۱۰ جمله از داده train و test را برای ارزیابی الگوریتم مورد ملاک قرار می دهیم که به وسیله آن میزان درستی الگوریتم سنجیده شود.

```
Viterbi Algorithm Accuracy On Train: 98.34710743801654
```

```
Viterbi Algorithm Accuracy On Test: 86.93467336683418
```

همانطور که مشاهده می کنید بر روی داده Train به درستی ۹۸٪ و به درستی ۸۷٪ بر روی داده Test رسیده ایم.