دانشکده مهندسی کامپیوتر

دکتر بهروز مینایی

بهار ۱۴۰۱

تمرین سری سوم

پردازش زبان و گفتار

ثمین حیدریان

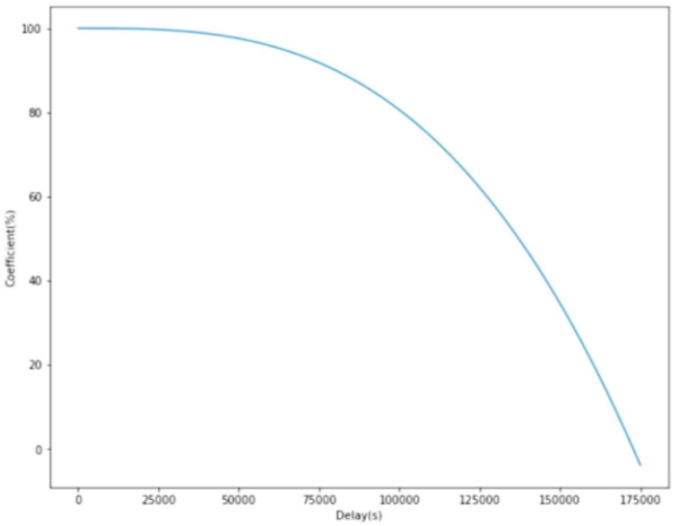
توحید عابدینی

تاریخ تحویل: جمعه ۱۳ خرداد ساعت ۲۳:۵۹:۵۹

قوانین:

- سوالات این تمرین از مبحث «تجزیه نحوی و برچسب زنی اجزای[1] سخن » می‌باشد و برای پاسخ به [2]سوالات آن نیاز به دانش نسبی درمورد این مبحث دارید.
- این تمرین شامل ۴ سوال می‌باشد. ۳ سوال تئوری و نوشتاری هستند و ۱ سوال عملی و شامل پیاده سازی است.
- درصورت وجود هرگونه سوال، در کلاس درس و یا در گروه تلگرامی درس بپرسید. (لطفا پی‌وی پیام ندهید.)
- هرگونه ایده گرفتن از تمرین دیگران و کدهای موجود در اینترنت که موجب تشابه غیرعادی و بالای کد شما با دیگری شود، تقلب محسوب می‌شود. درصورت مشاهده‌ی تقلب، نمره‌ی تمرین برای هر دو دانشجوی متخلف **صفر** منظور خواهد شد.
- لطفا برای انجام تمرین، زمان مناسب اختصاص دهید و انجام آن را به روزهای پایانی موکول نکنید. دقت کنید تمرین به هیچ عنوان تمدید نخواهد شد.
- پاسخ ارسالی شما باید علاوه بر کدهای مربوط به هر سوال، شامل یک گزارش در قالب یک فایل PDF باشد که محتوای گزارش مربوطه توضیحات تکمیلی شما درخصوص هر سوال و اسکرین‌شات از نتیجه اجرای کدهای شما باشد.
- تمامی فایل‌های موردنیاز برای تمرین را به صورت یک فایل ZIP با فرمت شماره دانشجویی_نام و نام خانوادگی_HW3 نام‌گذاری کرده و ارسال کنید. (برای مثال HW3_NameFamily_98000000)
- تاخیر در ارسال تمرین ها بر اساس نمودار زیر محاسبه خواهد شد. محور افقی نمودار، مقدار تاخیر به ثانیه و محور عمودی، ضریب اعمالی در نمره تمرین است.

[1] Constituency Parsing
[2] POS tagging


100
80
60
40
20
0
Coefficient(%)
0
25000
50000
75000
100000
125000
150000
175000
Delay(s)

سوالات تئوری:

1. در این قسمت میخواهیم PoS tagging را با استفاده از روش HMM[3] مطابق با ۴ جمله زیر انجام دهیم:

- Mark can watch.
- Will can mark watch.
- Can Tom watch?
- Tom will mark watch.

برچسب اجزای سخن (Part of Speech) زیر را طبق مراحلی که در ادامه نوشته شده اند مشخص کنید (فرض کنید سه تگ Verb، Noun و Modal را داریم).

- **Can Tom mark watch?**

مرحله ۱:

ابتدا جدولی مانند جدول زیر تشکیل دهید که Emission probability ها را تشکیل میدهد.

کلمات	Noun	Modal	Verb
Tom	2/6	0	0

مرحله ۲:

در این مرحله دو برچسب به ابتدا و انتهای جملات اضافه میشوند. **<S>** به ابتدای جمله و **<E>** به انتهای جمله اضافه میشود.

- <S> Mark can watch. <E>
- <S> Will can mark watch. <E>
- <S> Can Tom watch?
- <S> Tom will mark watch. <E>
-

سپس جدول زیر را تشکیل دهید و احتمال وقوع دو برچسب با یکدیگر (Transition probability) را محاسبه کنید.

[3] Hiddem Markov Model

	Noun	Modal	Verb	<E>
<S>				
Noun				
Modal				
Verb				

مرحله ۳:

در انتها گراف جملات را رسم کنید و با محاسبه احتمالات مسیر، برچسب صحیح کلمات را به دست آورید. در این مرحله راس ها و یال هایی که احتمال صفر دارند باید حذف شوند. همچنین راس هایی که به نقطه پایانی نمیرسند باید حذف شوند.

توجه: برای اطلاعات بیشتر به لینک زیر مراجعه کنید.

https://www.mygreatlearning.com/blog/pos-tagging/

2 . مطابق با گرامر داده شده و با استفاده از الگوریتم CKY، عملیات تجزیه و تحلیل نحوی را با کشیدن جدول برای جمله زیر انجام دهید.

- **John eats pie with cream**

S → NP VP 0.8
S → S conj S 0.2
NP → Noun 0.2
NP → Det Noun 0.4
NP → NP PP 0.2
NP → NP conj NP 0.2
VP → Verb 0.4
VP → Verb NP 0.3
VP → Verb NP NP 0.1
VP → VP PP 0.2

PP → P NP	1.0
Noun → John	0.2
Noun → Jack	0.3
Noun → pie	0.1
Noun → cream	0.3
Noun → cake	0.1
Verb → eat	0.2
Verb → eats	0.3
Verb → drinks	0.5
P → with	0.6
P → by	0.4
Det → a	0.3
Det → the	0.7
conj → and	0.8
conj → or	0.2

John	eats	pie	with	cream	
					John
					eats
					pie
					with
					cream

3 . محدودیت های PCFGs را با استفاده از مثال توضیح دهید. برای هر مورد توضیح دهید که lexicalized grammer چگونه محدودیت مربوطه را کاهش میدهد؟

سوال عملی:

در این سوال باید یک PoS tagger برای زبان انگلیسی پیاده سازی نمایید. در فایل نوت‌بوک پیوست شده با نام PoS_Tagger کتابخانه هایی که امکان استفاده از آن را دارید آورده شده است و همچنین مجموعه داده مورد استفاده در آن نوشته شده است.

برای پیاده سازی PoS tagger باید از درخت تصمیم[4]استفاده نمایید. با استفاده از دانش خود و مطالبی که در کلاس گفته شده و همچنین جست و جو در اینترنت سعی کنید ویژگی هایی کاربردی در این زمینه را استخراج نموده و در آموزش رده بند[5]درخت تصمیم از آنها استفاده کنید.

تفکیک مجموعه داده به دو داده آموزش[6]و آزمون[7]باید به صورت ۸۰-۲۰ باشد.

در صورتی که به مشکل حافظه برخوردید فقط از ۱۰۰۰۰ توکن اول در دیتاست خود (ترتیب مهم نیست.) استفاده نموده و دقت رده بند حداقل باید ٪۸۵ باشد.

موفق باشید

[4] Decision Tree
[5] Classifier
[6] Train
[7] Test