پردازش زبان طبیعی

آزمونک اول – پاسخ نامه

زمان: ۴۰ دقیقه

۱- در جدول زیر در سمت چپ عبارات منظم نوشته شده است و در سمت راست رشته هایی که ممکن است توسط این عبارات منظم پذیرفته شوند. مقابل هر عبارت منظم شماره رشته ای که می پذیرد را بنویسید.(یا وصل کنید) (جمعا ۵ نمره - هر یک ۰,۵ نمره)

شماره	رشته	عبارت منظم	
1	ABODE	AB?B?C?A	4
2	BBC6	AB?C+E	3
3	ACE	B*C[2-6]	2
4	ABBA	A..C	5
5	ABDC	^A[BC]OD[EF]$	1
6	DAB	ABC[0-9]	9
7	ABCD	(A|B)(B|C)[^D](C|D)$	7
8	CAB	[A-C][-AC][A-C]	8
9	ABC1	[^A-C][-AC][A-C]	6
10	ABCDE	A+B+C+D+E+	10

۲- در مورد perplexity به سوالات زیر پاسخ دهید:

a. معیار perplexity یا سرگشتی یک مدل زبانی برابر با ۷۸۲ است، این عدد یعنی چه؟ مدل زبانی که perplexity بیشتری دارد مدل بهتری است یا مدلی که perplexity آن کمتر است؟ مقدار ایده آل برای perplexity چند است؟ (۵ نمره)

Perpleixity در واقع همان weighted branching factor است و بیان می کند بعد از یک کلمه به طور میانگین چند کلمه مناسب می تواند بیاید. بنابراین هر چه میزان perplexity کمتر باشد به این معنی است که مدل توانسته است انتخاب های بعدی را محدود کند و در نتیجه مدل با perplexity کمتر مدل بهتری است. همچنین مقدار ایده آل برایperplexity برابر یک است که هیچگاه قابل دسترس نیست.

b. به نظر شما perplexity زبان فارسی بیشتر است یا انگلیسی؟ چرا؟(۵ نمره)

از یک جهت انگلیسی میتواند بیشتر باشد، چون تعداد کلمات زبان انگلیسی خیلی بیشتر از فارسی است و در نتیجه تنوع ظهور کلمات بیشتر و per pl exi ty بیشتر میشود. از طرف دیگر، فارسی چونfree word order و همچنین highly inflectional است، به این دو دلیل میتواند تنوع کلمات بیشتر دیده شود. در مجموع بنظر فارسی Per pl ei xit y بیشتری دارد.

پردازش زبان طبیعی

آزمونک اول – پاسخ نامه

زمان: ۴۰ دقیقه

۳- یکی از مشکلات اصلی سامانه های خطایاب املایی، عدم وجود اسامی همه اشخاص در دیکشنری آن است. دراینصورت در خطایابی یک نامه (یاخبر) اتفاقی که می افتد، آن است که اسامی اشخاصی که در آن نامه (یا خبر)به عنوان خطا تشخیص میدهد. چه روشی پیشنهاد میدهید؟ (۵ نمره)

استفاده از handling UNK و همچنین استفاده از Adaptive LM که در ان یکبار از روی کل سند تست نیز یک unigram اموزش می یابد و به دیکشنری اصلی با یک وزنی اضافه میشود