نمونه سوالات پایان ترم درس پردازش زبان طبیعی

بخش Text Classification

۱- هدف یک برنامه Classification چیست؟(ورودی و خروجی های این برنامه را توضیح دهید.)
تفاوت Classification به روش Supervised و Unsupervised را توضیح دهید.

۲- ایده اصلی روش Bag of words برای رده بندی اسناد (document classification) را توضیح دهید.

۳- قانون Bayes و کاربرد آن در Classification را توضیح دهید.

۴- دو رابطه زیر، از اصلی ترین رابطه های Classification هستند. هر یک را توضیح دهید و علت تفاوت این دو را شرح دهید:

$$c_{MAP} = \underset{c \in C}{\operatorname{argmax}} P(x_1, x_2, \ldots, x_n \mid c) P(c)$$

$$c_{NB} = \underset{c \in C}{\operatorname{argmax}} P(c_j) \prod_{x \in X} P(x \mid c)$$

۵- روش Laplace (add-1) برای انجام Classification با استفاده از Naïve Bayes از رابطه زیر استفاده میکند:

$$\hat{P}(w_i \mid c) = \frac{count(w_i, c) + 1}{\sum_{w \in V} (count(w, c) + 1)}$$

این رابطه را توضیح دهید و علت علامت های جمع را بررسی کنید.

۶- یک classifier نوشته شده است تا مشخص کند یک متن در مورد ژاپن است یا چین. چهار متن به عنوان داده های آموزش (Training) و یک متن به عنوان تست داده شده است. با استفاده از روش Naïve Bayes و ایده Laplace (add-1) مشخص کنید داده تست، به کدام کلاس تعلق دارد:

	Doc	Words	Class
Training	1	Chinese Beijing Chinese	c
	2	Chinese Chinese Shanghai	c
	3	Chinese Macao	c
	4	Tokyo Japan Chinese	j
Test	5	Chinese Chinese Chinese Tokyo Japan	?

برای محاسبه جواب، روابط زیر را داریم:

$$\hat{P}(c) = \frac{N_c}{N} \qquad \hat{P}(w \mid c) = \frac{count(w, c) + 1}{count(c) + |V|}$$

۷- از روی جدول زیر، Precision، Recall ، Accuracy و F- measure را توضیح دهید:

	correct	not correct
selected	tp	fp
not selected	fn	tn

۸- روش Cross-validation را توضیح دهید.

بخش Maximum Entropy

۱- داده‌های زیر را در نظر بگیرید:

I can can a can

فرض کنید برچسب‌های باینری زیر برای کلمات این جمله انتخاب شده است:

11001

دو توزیع احتمالی مناسب برای الگوریتم‌های generative و discriminative را برای آن تشکیل دهید.

۲- فیچرهای زیر را در نظر بگیرید:

f1(c, d) ≡ [c = PERSON ∧ w-1 = “Mrs.” ∧ isCapitalized(w)]

f2(c, d) ≡ [c = LOCATION ∧ startsWith(w, "A") ^ startOfSentence(w-1)]

f3(c, d) ≡ [c = VERB ∧ endsWith(w, "ed")]

که در آن w کلمه جاری و w-1 کلمه قبلی و isCapitalized به معنی شروع شدن کلمه با حروف بزرگ، startsWith به معنی شروع کلمه با یک رشته خاص و endsWith به معنی پایان یافتن کلمه با یک رشته خاص است. همچنین startOfSentence به معنی اینست که جمله دربرگیرنده کلمه با همان کلمه شروع شود.

In Shiraz, Mrs. Sue viewed Pasargad last year.

فرض کنید ضرایب لامبدای MaxEnt برای سه فیچر بالا به ترتیب برابر ۰.۹، ۰.۶ و ۰.۴ هستند. برای سه کلمه‌ی Shiraz، Sue و Pasargard یکی از تگ‌های LOCATION و PERSON و VERB را محاسبه کنید.

۳- فرض کنید هفت سند داریم که سه تای آن سیاسی و چهار تای آن ورزشی هستند. فرض کنید کلمات زیر در هر کدام از این سندها آمده است و سه سند اول اسناد سیاسی هستند:

سند ۱: باشگاه، مجلس، رییس

سند ۲: مجلس، رییس

سند ۳: مجلس، رییس، پیروزی

سند ۴: باشگاه، رییس

سند ۵: باشگاه، پیروزی

سند ۶: پیروزی، رییس

سند ۷: باشگاه، پیروزی

احتمال سیاسی بودن سند به شرط دیده شدن باشگاه و مجلس و پیروزی را در سند مزبور به دست آورید.

۴- همانطور که میدانید در روش MaxEnt داریم:

$$\log P(C \mid D, \lambda) = \log \prod_{(c,d)\in(C,D)} P(c \mid d, \lambda) = \sum_{(c,d)\in(C,D)} \log \frac{\exp \sum_i \lambda_i f_i(c,d)}{\sum_{c'} \exp \sum_i \lambda_i f_i(c',d)}$$

طبق تعاریف ارائه شده در کلاس درسی مقدار predicted count(f_i, λ) را محاسبه نمایید و در مورد مفهوم آن توضیح دهید.س

راهنمایی (مشتق کسر بالا را نسبت به λ_i محاسبه کنید)

بخش NER

۱- در مورد اصطلاحات زیر توضیح دهید و معادل انگلیسی آن‌ها را نیز بنویسید:

پیکره متنی، تشخیص واحدهای اسمی، فرهنگ لغت

۲- تعدادی از کاربردهای تشخیص واحدهای اسمی را بیان کنید.

۳- برچسب‌های واحدهای اسمی را در جمله زیر به روش IOB تعیین کنید

رحمانی فضلی، سکان‌دار وزارت کشور جمهوری اسلامی ایران تأکید کرد تمام خواسته ما در مراسم اربعین امسال در عراق این بود که امنیت زائرین تأمین شود.

۴- برچسبّ‌های واحدهای اسمی یک جمله به صورت زیر داده شده است. ردیف اول برچسب‌های مشخص شده توسط خبره (golden annotations) و ردیف دوم توسط الگوریتم X مشخص شده است. مقادیر precession و recall و f-measure را برای الگوریتم X تعیین کنید.

Jim	bought	300	shares	of	Acme	Corp.	in	2006	.
B-PER	O	O	O	O	B-ORG	I-ORG	O	B-MISC	O
O	B-PER	O	O	O	O	B-ORG	O	B-MISC	O

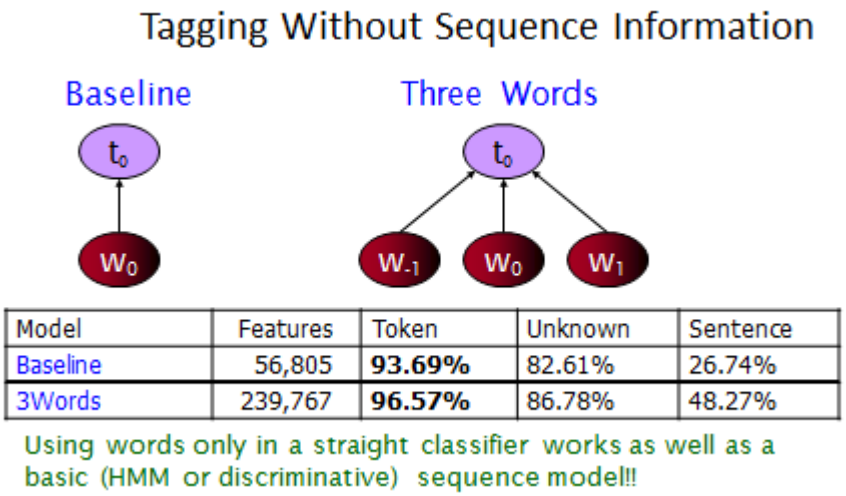
بخش Part-of-speech tagging

۱- چهار POS مختلف نام ببرید و هر کدام را در یک مثال توضیح دهید.

۲- یک کلمه را در سه جمله مختلف طوری بکار ببرید که سه نقش مختلف داشته باشد. توضیح دهید.

۳- هنگام محاسبه کیفیت روش های مختلف POS tagging با استفاده از روش Most freq tag داریم 90% overall / 50% unknown . این روش را توضیح دهید و علت کیفیت بالای آن در حالت overall و کیفیت پایین آن در حالت unknown را مشخص کنید.

۴- تصویر زیر در مورد مقایسه روش های POS tagging بوسیله تک کلمه و سه کلمه است:



Model	Features	Token	Unknown	Sentence
Baseline	56,805	**93.69%**	82.61%	26.74%
3Words	239,767	**96.57%**	86.78%	48.27%

Using words only in a straight classifier works as well as a basic (HMM or discriminative) sequence model!!

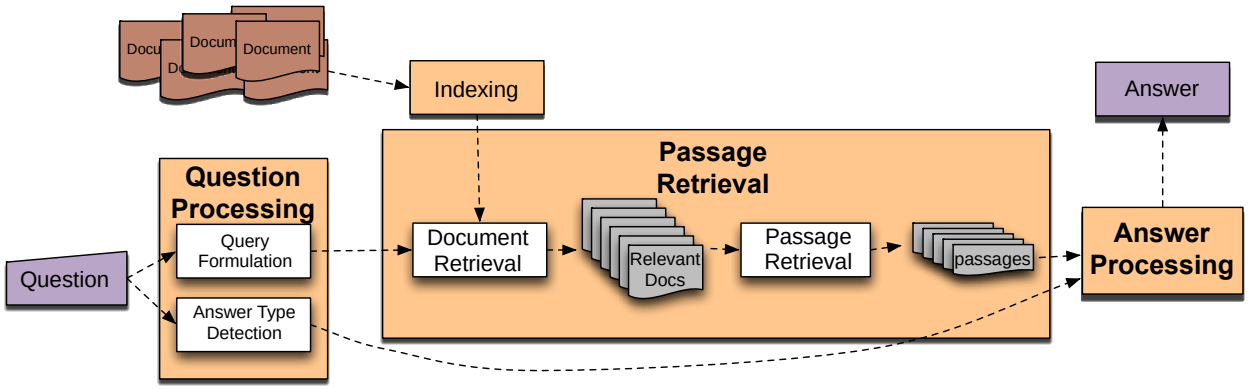
در جدول بالا، علت اختلاف بین دو روش Baseline و 3Words را توضیح دهید. علت اختلاف بسیار زیاد Features چیست؟ علت اختلاف کم در کیفیت های Token و Unknown و علت اختلاف زیاد در کیفیت Sentence چیست؟

بخش Question Answering

۱- منظور از Factoid questions در سیستم های پرسش و پاسخ چیست؟

۲- دو پارادایم مختلف برای سیستم های پرسش و پاسخ وجود دارد: IR-base و Knowledge-base در مورد هریک به اختصار توضیح دهید.

۳- اجزای کلی یک سامانه پرسش و پاسخ به صورت زیر است:



هر یک از این اجزا را توضیح دهید.

۴- یکی از معیارهای ارزیابی سیستم های پرسش و پاسخ روش MRR است که از رابطه زیر بدست می آید:

$$MRR = \frac{\sum_{i=1}^{N} \frac{1}{rank_i}}{N}$$

از روی فرمول این روش را توضیح دهید.

بخش Probabilistic Parsing

۱- تفاوت ساختار وابستگی (dependency structure) و ساختار عبارتی (phrase structure) در چیست؟

۲- آیا هر ساختار عبارتی را می‌توان به ساختار وابستگی تبدیل کرد؟

۳- درخت معادل داده‌های برچسب‌گذاری شده زیر را نمایش دهید:

```
( (S
(NP-SBJ (DT The) (NN move))
(VP (VBD followed)
(NP
(NP (DT a) (NN round))
(PP (IN of)
(NP
(NP (JJ similar) (NNS increases))
(PP (IN by)
(NP (JJ other) (NNS lenders)))
(PP (IN against)
(NP (NNP Arizona) (JJ real) (NN estate) (NNS loans))))))
(, ,)
(S-ADV
(NP-SBJ (-NONE-*))
(VP (VBG reflecting)
(NP
(NP (DT a) (VBG continuing) (NN decline))
(PP-LOC (IN in)
(NP (DT that) (NN market)))))))
(. .)))
```

۴- مفهوم preterminal را در گرامر ساختار عبارتی NLP توضیح دهید.

۵- گرامر احتمالاتی زیر در دسترس است، مقادیر مشخص شده با عبارت X چه هستند؟

S →NP VP	1.0
VP →V NP PP	X1
VP →V NP	0.5
NP →NP NP	0.2
NP →N	0.6
NP →NP PP	X2

PP →P NP X3

N →people 0.4

N →fish 0.3

N →tanks X4

N →rods 0.1

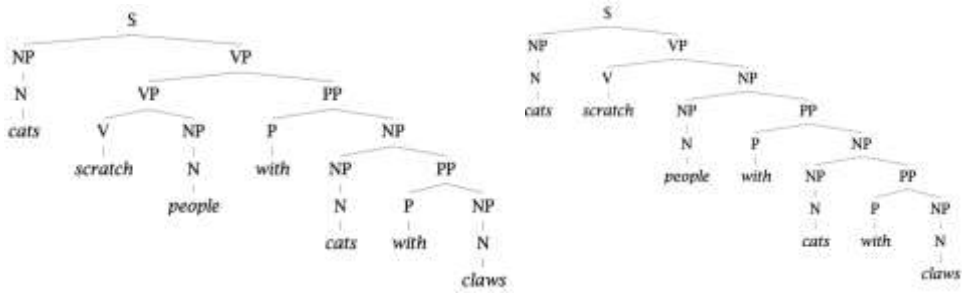
V →people 0.2

V →fish 0.6

V →tanks X5

P →with 1.0

۶- فرض کنید برای جمله cats scratch people with cats with claws تنها دو درخت پارس زیر وجود دارند:



فرض کنید گرامر احتمالاتی استفاده شده به صورت زیر بوده است:

S →NP VP (1.0)

NP → N (0.3) NP →N (0.5) NP →NP PP (0.2)

N →cats (0.4) N →people (0.5) N →claws (0.1)

VP →VP VP (0.3) VP →V NP (0.7)

PP →P NP (1.0)

V →scratch (1.0)

P →with (1.0)

احتمال هر درخت و احتمال جمله مورد نظر را در گرامر بالا محاسبه کنید.

۷- قوانین زیر را در نظر بگیرید:

NP → NNS NP 0.003

NP → NNS NNS 0.015

VP → VB PP	0.042
PP → IN	0.004
VP → VB NP	0.032
NP → NNS PP	0.01
NNS → takes	0.0041
VB → takes	0.002
PP → up	0.3
IN → up	0.0114
NNS → up	0.001

الگوریتم CYK را برای دو کلمه‌ی takes up اجرا کنید.

بخش lexicalized parsing

۱- تفاوت یا تفاوت‌های عمده Lexicalized Parsing در مقایسه با Probabilistic Parsing معمولی چیست؟ نقطه ضعف‌های این روش چیست؟

۲- در مورد روش چارنیاک در عملیات Lexicalized Parsing توضیح دهید.

۳- در مورد Horizontal Markovization و Vertical Markovization با ذکر مثال توضیح دهید.

۴- در مورد روش اسلاو پترف در زمینه برچسب‌های مخفی توضیح دهید.

بخش dependency parsing

۱- درخت Dependency Parsing معادل اطلاعات زیر را ترسیم کنید:

1	این	این	PREM	DEMAJ	2	NPREMOD	_	_
2	میهمانی	میهمانی	N	IANM	12	SBJ	_	_
3	به	به	PREP	PREP	12	ADV	_	_
4	منظور	منظور	N	IANM	3	POSDEP	_	_
5	آشنایی	آشنایی	N	IANM	4	MOZ	_	_
6	هم‌تیمی‌های	هم‌تیمی	N		5	MOZ	_	_
7	او	او	PR	SEPER	6	MOZ	_	_
8	با	با	PREP	PREP	5	NPP	_	_
9	غذاهای	غذا	N	IANM	8	POSDEP	_	_
10	ایرانی	ایرانی	ADJ	AJP	9	NPOSTMOD	_	_
11	ترتیب	ترتیب	N	IANM	12	NVE	_	_
12	داده شد	داد#ده	V	PASS	0	ROOT	_	_
13	.	.	PUNC	PUNC	12	PUNC	_	_

این داده‌ها به فرمت استاندارد Penn Treebank نوشته شده است تنها ستون ششم که مربوط به featureها است از آن حذف شده است تا فضای کمتری را اشغال کند.

آیا درخت بدست آمده Projective است یا Non-projective؟

۲- در روش Malt Parser عملیات‌های زیر را روی جمله زیر اجرا کرده‌ایم. مقدار مجموعه A در انتها چه چیزی خواهد بود؟ (فقط چند عملیات اول نمایش داده شده است)

Fortune sides with him who dares.

Shift, LA $_{compound}$, Shift, Shift, LA $_{case}$, RA $_{nmod}$, Shift

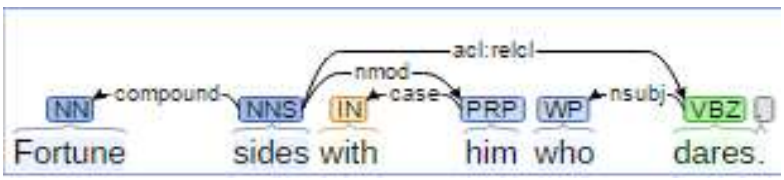
۳- دو مقدار UAS و LAS را برای داده‌های زیر محاسبه کنید:

Gold:

Index	head	word	label
1	2	Fortune	compound
2	0	sides	ROOT
3	4	with	case
4	2	him	nmod
5	6	who	nsubj
6	2	dares	acl:relcl

Parsed:

index	head	word	label
1	2	Fortune	compound
2	0	sides	ROOT
3	5	with	case
4	3	him	nsubj
5	5	who	nmod
6	2	dares	ccomp



بخش Summarization

۱- تفارت خلاصه سازی Query-focused و Generic را توضیح دهید.

۲- تفاوت خلاصه سازی Extractive و Abstractive را توضیح دهید.

۳- مفهوم Snippets را در خلاصه سازی توضیح دهید.

۴- خلاصه سازی از سه بخش تشکیل شده است: content selection ، information ordering ، sentence realization. هر سه بخش را توضیح دهید.

۵- روش ارزیابی خلاصه سازی ROUGE را توضیح دهید. میدانیم این روش به صورت زیر محاسبه میشود:

$$ROUGE-2 = \frac{\sum_{s\in\{RefSummaries\}} \sum_{bigrams\ i\in S} \min(count(i,X), count(i,S))}{\sum_{s\in\{RefSummaries\}} \sum_{bigrams\ i\in S} count(i,S)}$$

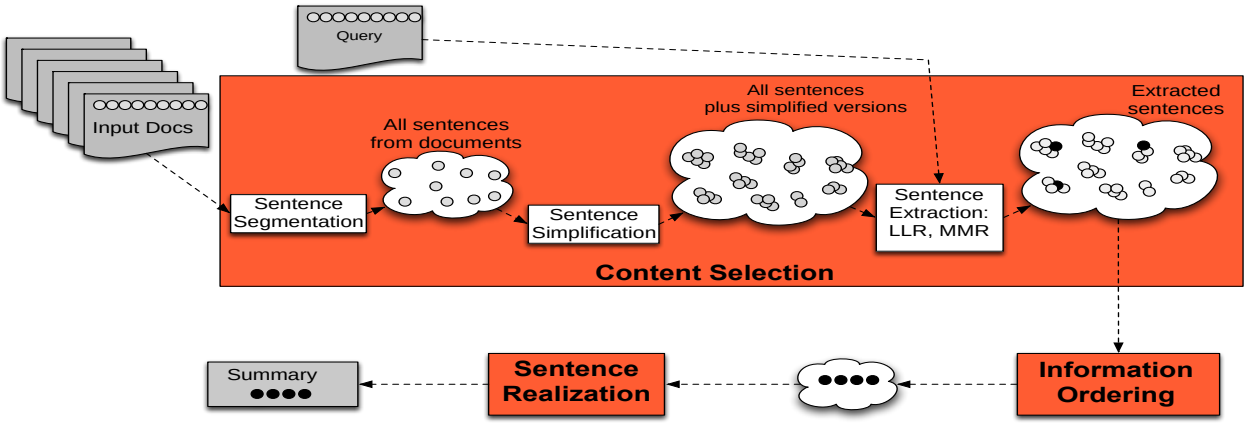
اگر سیستم خلاصه ساز جمله System answer را تولید کرده باشد و دو روش خلاصه سازی انسانی جمله های Human 1 و Human 2 را تولید کرده باشد. کیفیت خلاصه ساز را بوسیله روش ROUGE توضیح دهید.

Human 1: Water spinach is a green leafy vegetable grown in the tropics.

Human 2: Water spinach is a commonly eaten leaf vegetable of Asia.

System answer: Water spinach is a leaf vegetable commonly eaten in tropical areas of Asia.

۶- سیستم خلاصه سازی Query-Focused Multi-Document به صورت زیر است:



این روش را از روی شکل توضیح دهید.

بخش semantic

۱- کلمات Homonymy به چه معنی هستند؟ این کلمات به دسته Homograph و Homophone تقسیم می شوند. هر یک را توضیح داده و برای هر کدام یک مثال بیاورید. همچنین یکی از مشکلاتی که این کلمات در پردازش زبان بوجود می آورند را شرح دهید.

۲- کلمات Polysemy به چه معنی هستند؟ در یک مثال توضیح دهید.

۳- در کلمات sense را توضیح دهید. چطور میتوان فهمید که یک کلمه بیش از یک sense دارد؟

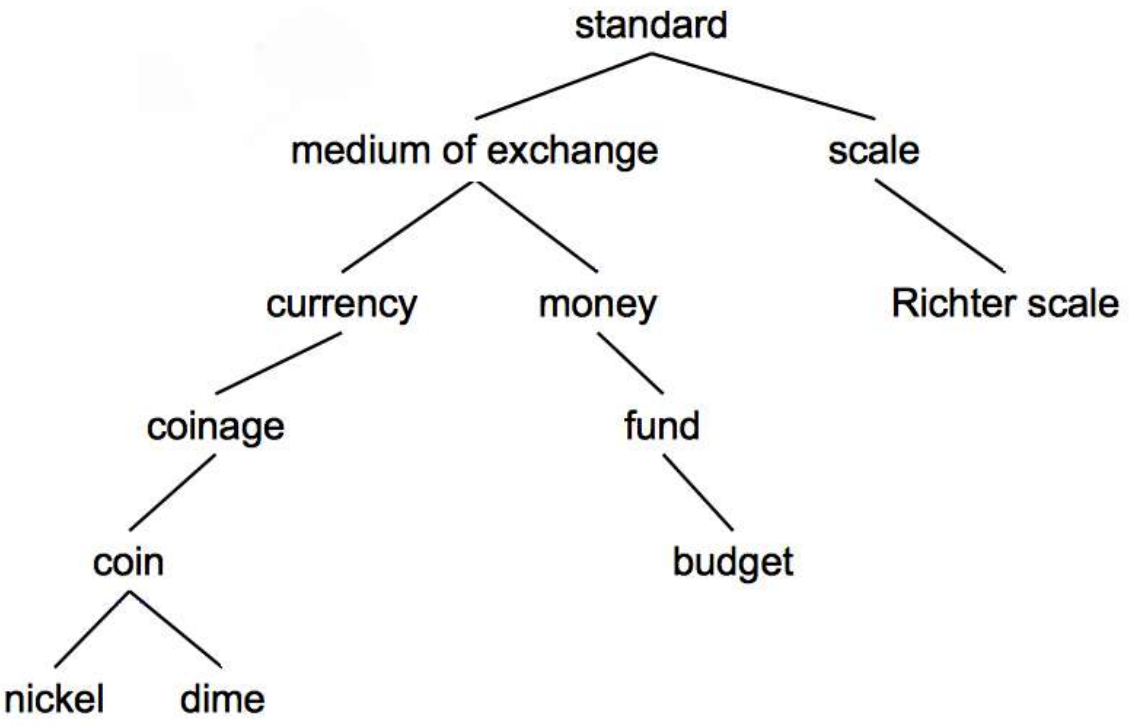
۴- چطور میتوان فهمید که دو کلمه Synonym هستند؟

۵- Hyponymy و Hypernymy را توضیح داده و مثال بزنید.

۶- WordNet چیست و چه تفاوتی با MeSH دارد؟

۷- الگوریتم های یافتن similarity به دو دسته Thesaurus-based و Distributional تقسیم میشون. هر یک را توضیح دهید.

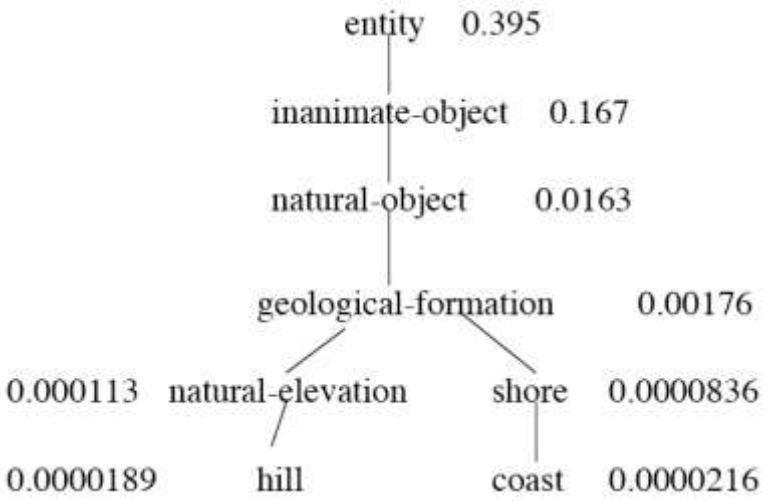
۸- در شکل زیر با توجه به روش path-based similarity مقدار simpath(nickel,money) را حساب کنید.



۹- روش محاسبه شباهت بین دو کلمه به روش Lin به این صورت است:

$$sim_{Lin}(A,B) = \frac{2\log P(LCS(c_1,c_2))}{\log P(c_1)+\log P(c_2)}$$

این رابطه را توضیح دهید و با توجه به شکل پایین خاصل عبارت $sim_{Lin}(\text{hill}, \text{coast})$ را حساب کنید



۱۰- ماتریس term-context زیر را داریم:

	Count(w,context)				
	computer	data	pinch	result	sugar
apricot	0	0	1	0	1
pineapple	0	0	1	0	1
digital	2	1	0	1	0
information	1	6	0	4	0

حاصل عبارت p(w=information,c=data) و pmi(information,data) را حساب کنید.

۱۱- معیار شباهت کسینوسی را تعریف کرده و حاصل عبارت cosine(digital,information) را در جدول زیر محاسبه کنید.

	large	data	computer
apricot	1	0	0
digital	0	1	2
information	1	6	1