نمونه سوالات درس پردازش زبان طبیعی

بخش Background and Overview

۱- چهار مورد از کاربردهای پردازش زبان طبیعی را نام‌برده و هر مورد را به اختصار توضیح دهید.

۲- ابهام در زبان طبیعی چیست؟ سطوح ابهام را نام برده و توضیح دهید. (برای سهولت می‌توانید از مثال استفاده کنید)

۳- رویکرد رده‌بندی (Classification) در مسائل پردازش زبان طبیعی در چه صورت قابل استفاده است؟ دو مثال ذکر کنید که می‌توان از این رویکرد در حل آن‌ها سود برد؟

۴- رویکرد رده‌بندی (Classification) در برخی از مسائل پردازش زبان طبیعی قابل استفاده نیست. چهار نمونه از این مسائل را ذکر کنید. توضیح دهید که چرا در این مسائل نمی‌توان از رده‌بندی استفاده کرد.

۵- جمله‌ی زیر و دو برداشت مختلف از آن را در نظر بگیرید.

“At last, a computer that understands you like your mother”

1- It understands you as well as your mother understands you
2- It understands (that) you like your mother

ابهام به وجود آمده مربوط به کدام یک از سطوح ابهام است؟ توضیح دهید.

بخش Basic Text Processing

۱- به طور معمول در پردازش زبان طبیعی نیازمند پیش پردازش متون هستیم. چهار نمونه از این پیش‌پردازش‌ها را نام برده و به طور مختصر توضیح دهید.

۲- یکسان‌سازی متون (Normalization) را تعریف کنید. برای یکسان‌سازی متون انگلیسی و فارسی شش مثال ذکر کنید. (دو مثال برای انگلیسی و چهار مثال برای فارسی) به نظر شما یکسان‌سازی در کدام یک از دو زبان انگلیسی و فارسی از اهمیت بیشتری برخوردار است.

۳- عبارات منظم خواسته شده را بنویسید:

الف) تمام رشته‌های قابل تولید با الفبای انگلیسی

ب) تمام رشته‌های قابل تولید با حروف بزرگ انگلیسی که به حرف C ختم می‌شوند.

ج) آدرس ایمیل (فرض کنید تنها کاراکترهای کوچک و بزرگ انگلیسی و کاراکتر نقطه - . - مجاز هستند)

۴- مشخص کنید تعاریف زیر مربوط به کدام اصطلاح هستند؟

- تعداد کلمات موجود در واژه‌نامه‌ی یک زبان (token-type)

- تعداد کل کلمات موجود در یک پیکره‌ی متنی (token-type)
- افعال مختلف از یک ریشه با زمان‌ها و اشخاص مختلف (wordform-lemma)
- ریشه‌ی تمامی افعال با زمان‌ها و اشخاص مختلف (wordform-lemma)

۵- فرض کنید برای تشخیص مجموعه‌ی خاصی از کلمات در یک متن عبارت منظمی نوشته‌ایم. این عبارت منظم تمامی کلماتی که به دنبال آن بوده‌ایم را برای ما مشخص می‌کند، ولی تعدادی از کلمات که شرایط مورد انتظار را ندارند را نیز جزو آن‌ها در نظر می‌گیرد. عبارتی که نوشته‌ایم در معیار Precision عملکرد بهتری دارد یا Recall؟ توضیح دهید.

۶- فرض کنید برای تشخیص مجموعه‌ی خاصی از کلمات در یک متن عبارت منظمی نوشته‌اید. این عبارت منظم نمی‌تواند تمامی کلماتی که شما به دنبال آن بوده‌اید را برای شما مشخص کند، ولی اگر کلمه‌ای را تشخیص داد، شرایط مورد انتظار را دارد. عبارتی که نوشته‌ایم در معیار Precision عملکرد بهتری دارد یا Recall؟ توضیح دهید.

۷- یکی از راه‌های نرمال‌سازی متن در زبان انگلیسی تبدیل تمامی حروف بزرگ به حروف کوچک است. به نظر شما این کار در ترجمه‌ی ماشینی نتایج خوبی به دنبال دارد؟ توضیح دهید.

۸- در تبدیل کلمه‌ی automation به automate و automat به ترتیب کدام یک از کارهای stemming و lemmatization انجام شده‌است؟

بخش Language Modeling

۱- بیشترین و کمترین مقدار ممکن برای سرگشتگی (perplexity) را محاسبه کنید.

۲- اعداد ذکر شده در جدول تعداد رخداد دو-گرم‌ها در یک متن خام است. واژگان ذکر شده در ستون اول سمت چپ، اولین واژه و واژگان ذکر شده در ردیف اول دومین واژهٔ هر دو-گرم هستند.

	i	want	to	eat	chinese	food	lunch	spend
i	5	827	0	9	0	0	0	2
want	2	0	608	1	6	6	5	1
to	2	0	4	686	2	0	6	211
eat	0	0	2	0	16	2	42	0
chinese	1	0	0	0	0	82	1	0
food	15	0	15	0	1	4	0	0
lunch	2	0	0	0	0	1	0	0
spend	1	0	1	0	0	0	0	0

تعداد رخداد هر کدام از واژه‌ها نیز به صورت زیر است:

i	want	to	eat	chinese	food	lunch	spend
2533	927	2417	746	158	1093	341	278

با در نظر گرفتن موارد زیر:

P(i | <s>) = 0.25, P(</s> | food) = 0.68

احتمال جملات زیر را محاسبه کنید.

الف) <s> i want chinese food </s>

ب) <s> i want to eat chinese food </s>

ج) <s> i want to spend food </s>

۳- اگر مدل زبانی به ما در تشخیص بین دو جمله‌ی high winds tonight و large winds tonight کمک کند، به نظر شما کاربرد مورد نظر چه کاربردی بوده‌است؟ توضیح دهید. (مثلا بازشناسی گفتار، ترجمه‌ی ماشینی، تصحیح املایی و ...)

۴- فرض کنید می‌خواهیم از روی یک مجموعه‌ی داده یک مدل زبانی آموزش دهیم. به نظر شما قبل از آموزش مدل باید کلمات توقف (stop words) را از متن حذف کنیم؟ چرا؟

۵- فرض کنید در یک زبان با اندازه‌ی واژگان ۱۰۰۰۰ کلمه یک مدل زبانی ۳-gram از روی یک مجموعه‌ی داده با تعداد کلمات (token)۱۰۰۰۰۰ کلمه آموزش داده‌ایم و مقدار سرگشتگی را روی یک مجموعه‌ی آزمون به دست آورده‌ایم. اگر یک مدل ۴-gram روی همان مجموعه‌ی داده آموزش دهیم، انتظار دارید مقدار سرگشتگی افزایش داشته باشد یا کاهش؟ توضیح دهید.

۶- فرض کنید یک زبان از کلمات {A, B, C, D} تشکیل شده و می‌خواهیم یک مدل زبانی بایگرام برای این زبان آموزش دهیم. اطلاعات زیر را از مجموعه‌ی آموزشی در اختیار داریم:

count (<s> A) = 2
count (<s> B) = 1
count (<s> C) = 2
count (<s> D) = 0

count (A A) = 0
count (A B) = 5
count (A C) = 1
count (A D) = 4
count (A </s>) = 1

count (B A) = 0
count (B B) = 0
count (B C) = 6
count (B D) = 2
count (B </s>) = 0

count (C A) = 6
count (C B) = 1
count (C C) = 0
count (C D) = 8
count (C </s>) = 1

count (D A) = 3
count (D B) = 1
count (D C) = 7
count (D D) = 1
count (D </s>) = 3

count (<s>) = 5
count (A) = 11
count (B) = 8
count (C) = 16
count (D) = 15

یک مدل زبانی روی این مجموعه‌ی داده آموزش داده و آن را به روش gooe turing هموار کرده‌ایم. سرگشتگی این مدل را روی جمله‌ی B C A B A حساب کنید. (تگ‌های شروع و پایان را نیز به جمله اضافه نمایید)

بخش Edit Distance

۱- با استفاده از الگوریتم Levenshtein فاصله و alignment بین دو کلمه‌ی correct و coorecct را بیابید.

بخش Spelling Correction

۱- فرض کنید کلمه‌ی acress را به عنوان یک غلط املایی تشخیص داده‌ایم و با استفاده از مدل کانال نویزی و اطلاعاتی که به صورت جدول زیر به دست آورده‌ایم می‌خواهیم کلمه‌ی درست را تشخیص دهیم. کدام کلمه انتخاب می‌شود؟

Candidate Correction	Correct Letter	Error Letter	x|w	P(x|word)	P(word)
actress	t	-	c|ct	.000117	.0000231
cress	-	a	a|#	.00000144	.000000544
caress	ca	ac	ac|ca	.00000164	.00000170
access	c	r	r|c	.000000209	.0000916
across	o	e	e|o	.0000093	.000299
acres	-	s	es|e	.0000321	.0000318
acres	-	s	ss|s	.0000342	.0000318

بخش Text Classification

۱- هدف یک برنامه رده‌بندی (Classification) چیست؟ (ورودی و خروجی این برنامه را توضیح دهید.)

۲- نحوه ایجاد بردار واژگان با استفاده از روش سبد واژگان (Bag of Words) را با مثال توضیح دهید.

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

۴- دو رابطه زیر، از اصلی‌ترین روابط رده‌بندی هستند. هر یک را توضیح داده و علت تفاوت آن‌ها را بیان کنید.

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

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

۵- در روش Laplace (add-1) با Naïve Bayes از رابطه زیر استفاده می‌شود:

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

این رابطه و علت علامت‌های جمع را توضیح دهید.

۶- رده‌بندی برای بررسی متون نوشته شده است. این رده‌بند دارای دو کلاس ژاپن و چین است. چهار متن به عنوان داده‌های آموزش و یک متن به عنوان داده آزمایش موجود است. با استفاده از روش Naïve Bayes و ایده Laplace (add-1)، کلاس داده آزمایش را پیدا کنید.

	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 از روابط زیر استفاده کنید.

$$\hat{P}(c) = \frac{N_c}{N}$$

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

۷- فرض کنید هفت سند داریم. سه سند اول سیاسی (اسناد یک تا سه) و چهار سند بعدی ورزشی (اسناد چهار تا هفت) هستند. کلمات زیر در هر کدام از سندها آمده است:

سند ۱: باشگاه، مجلس، رئیس

سند ۲: مجلس، رئیس

سند ۳: مجلس، رئیس، پیروزی

سند ۴: باشگاه، رئیس

سند ۵: باشگاه، پیروزی

سند ۶: پیروزی، رئیس

سند ۷: باشگاه، پیروزی

احتمال سیاسی بودن سند به شرط دیده شدن باشگاه، مجلس و پیروزی را محاسبه کنید.

۸- با استفاده از جدول زیر دقت، فراخوانی، صحت و معیار F را توضیح دهید.

	correct	not correct
selected	tp	fp
not selected	fn	tn

۹- Cross-Validation را توضیح دهید.

بخش Maximum Entropy

۱- جمله زیر را در نظر بگیرید:

I can can a can

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

11001

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

۲- ویژگی‌های 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")]

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

In Shiraz, Mrs. Sue viewed Pasargad last year.

فرض کنید ضرایب مدل بیشینه آنتروپی برای سه ویژگی بالا به ترتیب برابر ۰,۹، ۰,۶و ۰,۴ است. برای سه واژهٔ Shiraz، Sue و Pasargad یکی از برچسب‌های LOCATION، PERSON و VERB را محاسبه کنید.

۳- در روش بیشینه آنتروپی:

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

مقدار predicted count(f_i, λ) را محاسبه کرده و در مورد مفهوم آن توضیح دهید. (راهنمایی: مشتق کسر بالا را نسبت به λ_i محاسبه کنید.)

بخش POS 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 برچسب ادات سخن مختلف داشته باشد و نوع برچسب هرکدام را مشخص نمایید.

۳- با استفاده از روش Most Frequent Tag برای POS Tagging به طور معمول ۹۰ درصد صحت به طور کلی و ۵۰ درصد صحت برای کلمات ناشناخته بدست می‌آید. این روش را توضیح داده و علت صحت بالای آن در حالت کلی و صحت پایین آن برای کلمات ناشناخته را بیان کنید.

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:

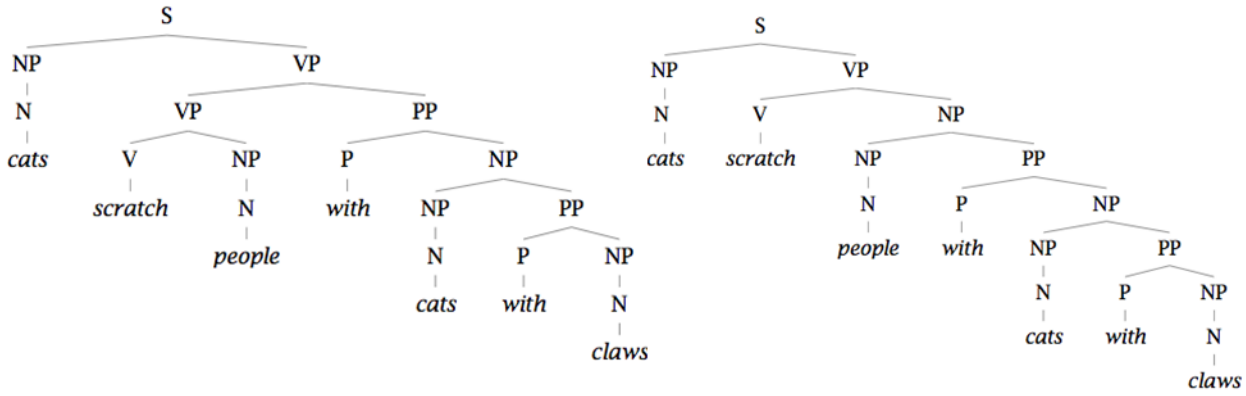
```
( (S
    (NP-SBJ (DT The) (NN move))
    (VP (VBD followed)
      (NP
        (NP (DT a) (NN round))
        (PP (IN of)
          (NP
            (NP (JJ similar) (NNS increases))
            (PP (IN by)
              (NP (JJ other) (NNS lenders)))
            (PP (IN against)
              (NP (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- 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

۵- 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.

۶- 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 CKY 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 وابستگی افکنشی و غیر افکنشی چیست؟ با مثال توضیح دهید.

۲- 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 (ستون ویژگی‌ها) حذف شده است.

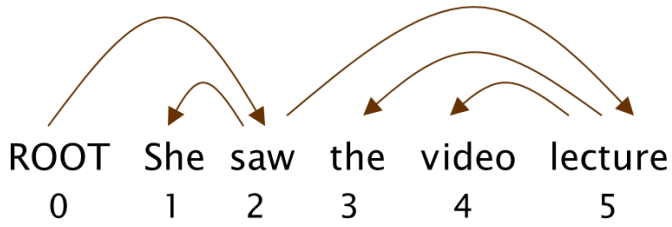
درخت بدست آمده Projective است یا Non-Projective؟

۳- 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

۴- دو معیار UAS و LAS را توضیح داده و با توجه به موارد داده شده، هر دو را محاسبه کنید:



Gold			
1	2	She	nsubj
2	0	saw	root
3	5	the	det
4	5	video	nn
5	2	lecture	dobj

Parsed			
1	2	She	nsubj
2	0	saw	root
3	4	the	det
4	5	video	nsubj
5	2	lecture	ccomp

بخش **Word Meaning and Similarity**

۱- کلمات Homonymy را توضیح دهید. (این کلمات به چند دسته تقسیم می‌شوند؟ هر کدام را با ذکر مثال توضیح دهید.)

۲- یکی از مشکلات کلمات Homonymy در پردازش زبان طبیعی را شرح دهید.

۳- کلمات Polysemy را با مثال توضیح دهید.

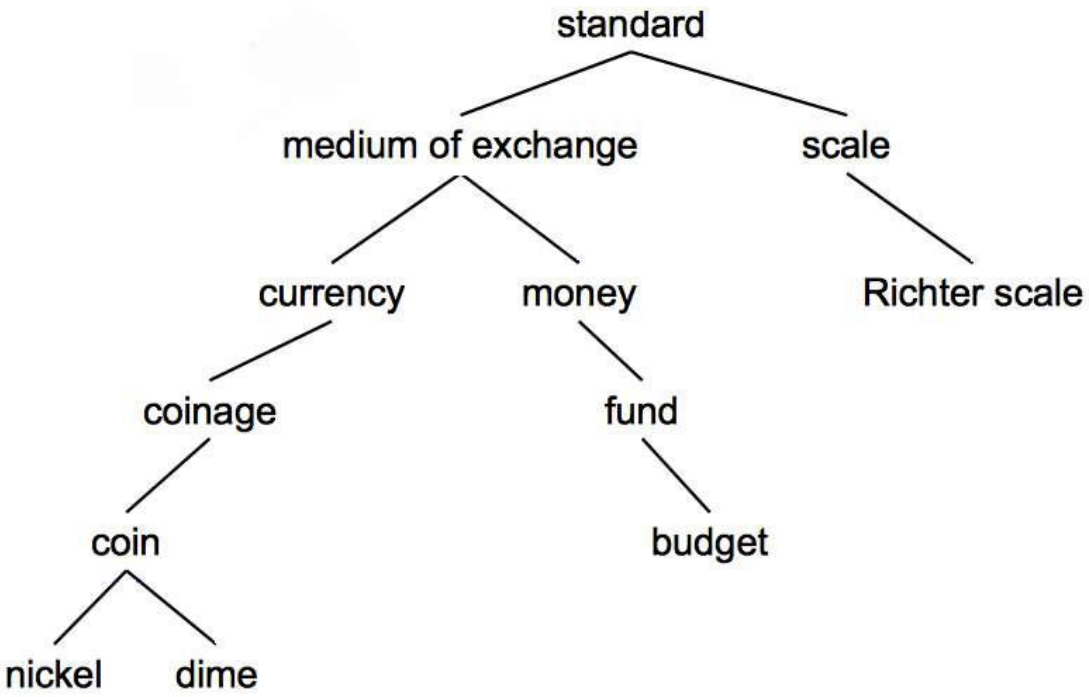
۴- sense چیست؟ چطور می‌توان دریافت یک کلمه بیش از یک sense دارد؟

۵- چطور می‌توان دریافت دو کلمه Synonym هستند؟

۶- Hyponymy و Hypernymy را با ذکر مثال توضیح دهید.

۷- دسته‌بندی الگوریتم‌های یافتن Similarity را نام برده و توضیح دهید.

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


standard
medium of exchange
scale
currency
money
Richter scale
coinage
fund
coin
budget
nickel
dime