

Лингвистика для математиков

POS-tagging

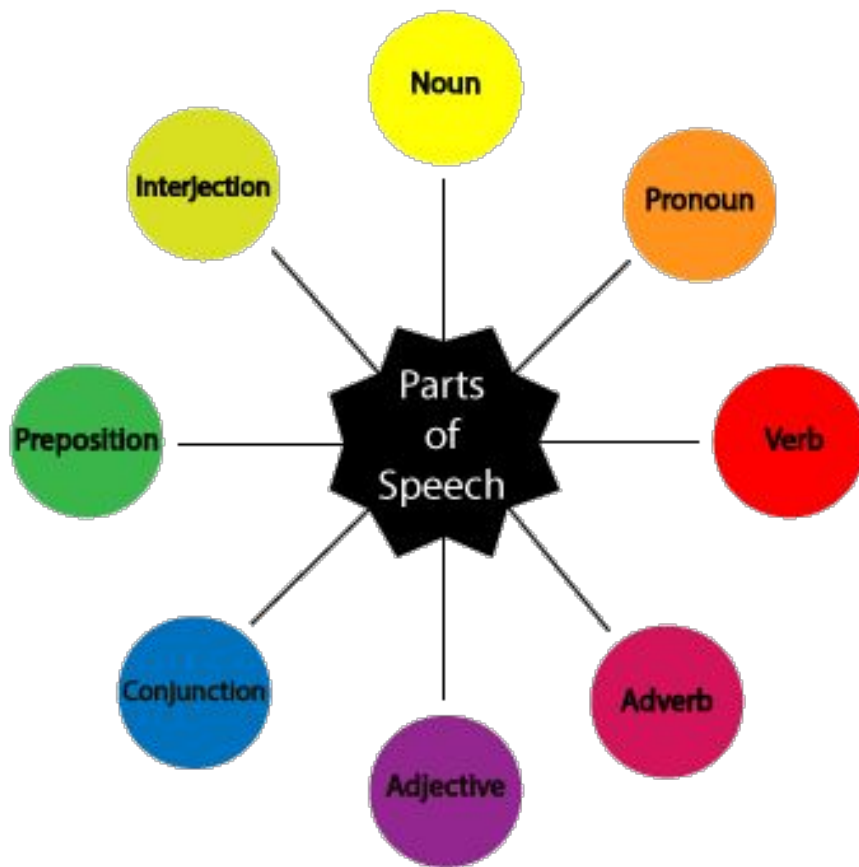
План на сегодня

- Автоматическое выделение частей речи
- Пробный тест по фану без оценок



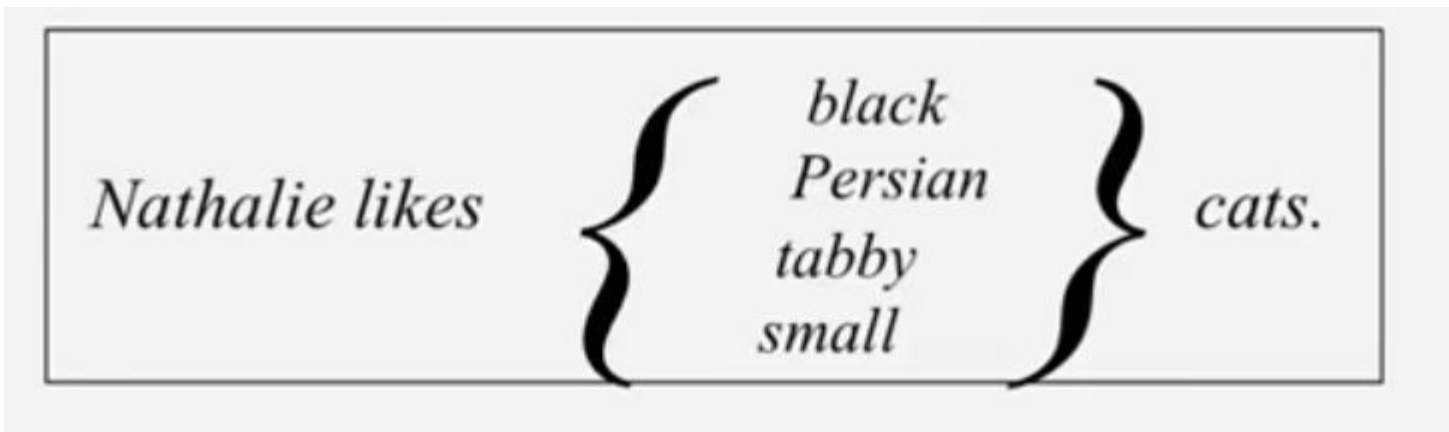
I'M SO ADJECTIVE,
I VERB NOUNS!!!

Какие бывают части речи?



Части речи

Как определить часть речи?



Части речи

Открытые и закрытые

Что это значит?

- **Open class:**
 - nouns, non-modal verbs, adjectives, adverbs
- **Closed class:**
 - prepositions, modal verbs, conjunctions, particles, determiners, pronouns

Глокая куздра штеко будланула бокра и кудрячит бокрѣнка

Части речи

Из Алисы в стране чудес

Lewis Carroll

`Twas brillig, and the slithy toves
Did gyre and gimble in the wabe:
All **mimsy** were the borogoves,
And the **mome raths outgrabe.**

Части речи

Ответы на задачу

- **Wabe, borogoves**
 - Nouns (after "the")
- **brillig**
 - adjective?
 - noun? ("noon")
- **mimsy**
 - adjective
- **slighty toves**
 - adjective+noun?
 - noun+verb? ("the bell tolls")
- **mome raths outgrabe**
 - Adjective+noun+verb?
 - Noun+verb+adverb? ("birds fly outside")

Части речи в разных языках

Вспомним задачу про индонезийский

Части речи в русском

Неоднозначность

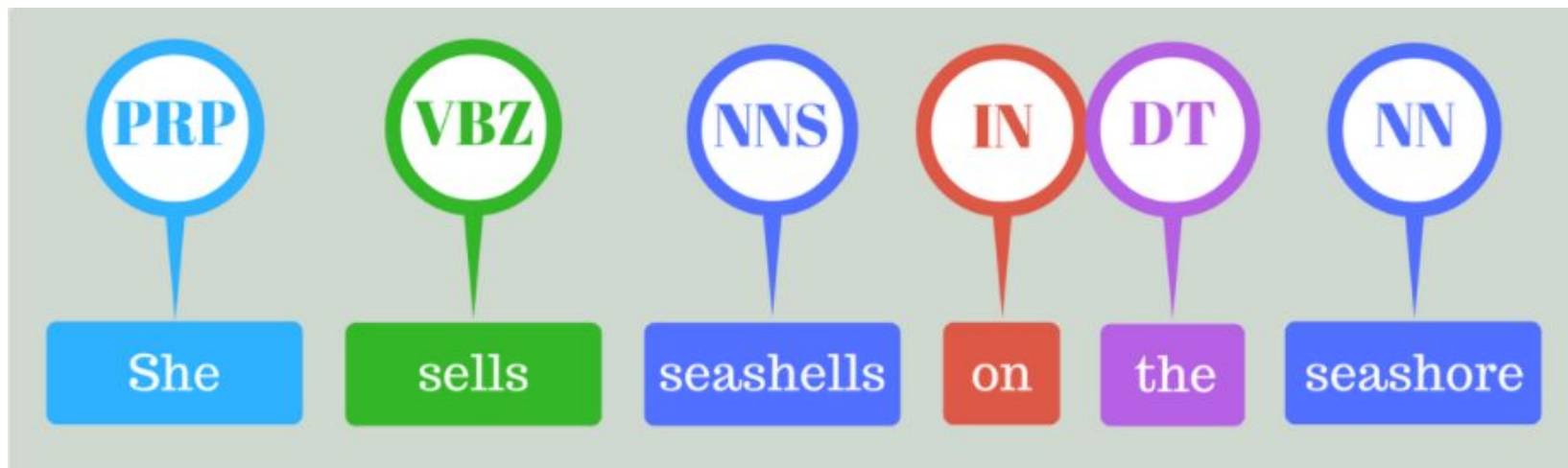
(в английском)

- count (noun) vs. count (verb)
- 11% of all types but 40% of all tokens in the Brown corpus are ambiguous.
- Examples
 - *like* can be tagged as ADP VERB ADJ ADV NOUN
 - *present* can be tagged as ADJ NOUN VERB ADV

Автоматический морфологический анализ

Как автоматически отличить “book that flight” от “hand me this book”?

Нужно провести морфологический анализ



The Penn Treebank tagset

- Университет Пенсильвании. Использовался для ручной разметки корпуса для текстов.

Tag	Description	Example	Tag	Description	Example	Tag	Description	Example
CC	coordinating conjunction	<i>and, but, or</i>	PDT	predeterminer	<i>all, both</i>	VBP	verb non-3sg present	<i>eat</i>
CD	cardinal number	<i>one, two</i>	POS	possessive ending	<i>'s</i>	VBZ	verb 3sg pres	<i>eats</i>
DT	determiner	<i>a, the</i>	PRP	personal pronoun	<i>I, you, he</i>	WDT	wh-determ.	<i>which, that</i>
EX	existential 'there'	<i>there</i>	PRP\$	possess. pronoun	<i>your, one's</i>	WP	wh-pronoun	<i>what, who</i>
FW	foreign word	<i>mea culpa</i>	RB	adverb	<i>quickly</i>	WP\$	wh-possess.	<i>whose</i>
IN	preposition/ subordin-conj	<i>of, in, by</i>	RBR	comparative adverb	<i>faster</i>	WRB	wh-adverb	<i>how, where</i>
JJ	adjective	<i>yellow</i>	RBS	superlatv. adverb	<i>fastest</i>	\$	dollar sign	<i>\$</i>
JJR	comparative adj	<i>bigger</i>	RP	particle	<i>up, off</i>	#	pound sign	<i>#</i>
JJS	superlative adj	<i>wildest</i>	SYM	symbol	<i>+, %, &</i>	“	left quote	<i>' or “</i>
LS	list item marker	<i>1, 2, One</i>	TO	“to”	<i>to</i>	”	right quote	<i>' or ”</i>
MD	modal	<i>can, should</i>	UH	interjection	<i>ah, oops</i>	(left paren	<i>[, (, {, <</i>
NN	sing or mass noun	<i>llama</i>	VB	verb base form	<i>eat</i>)	right paren	<i>],), }, ></i>
NNS	noun, plural	<i>llamas</i>	VBD	verb past tense	<i>ate</i>	,	comma	<i>,</i>
NNP	proper noun, sing.	<i>IBM</i>	VBG	verb gerund	<i>eating</i>	.	sent-end punc	<i>. ! ?</i>
NNPS	proper noun, plu.	<i>Carolinas</i>	VBN	verb past part.	<i>eaten</i>	:	sent-mid punc	<i>: ; ... --</i>

Figure 8.1 Penn Treebank part-of-speech tags (including punctuation).

Universal dependencies

- Этот набор тегов используется в большинстве современных корпусов
- Используется для большого количества языков
- Можно сравнивать разные языки и делать разборы более однообразными
- + синтаксический парсинг

Точность

- базовый алгоритм: если слово неоднозначно, присваиваем ему ту часть речи, которая чаще всего встречается в корпусе (для этого слова) --- 90% точность
- более сложные алгоритмы (скрытые марковские модели, машинное обучение и т.д.) --- 97% точность
- человек --- 98% точность

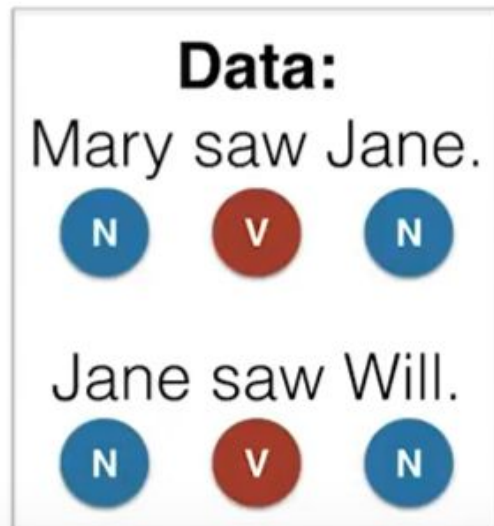
Какими методами мы можем воспользоваться?

- на основе сета правил
- стохастические (с помощью машинного обучения, с помощью марковских моделей)

Первый метод: сверяемся с таблицей

	N	V
Mary	1	0
saw	0	2
Jane	2	0
Will	1	0

Mary saw Will.



Первый метод: сверяемся с таблицей

Lookup Table

	N	V	M
Mary	2	0	0
see	0	3	0
Jane	2	0	0
Will	2	0	3



Mary will see Will.
N M V M

Data:
Mary will see Jane.
N M V N
Will will see Mary
N M V N
Jane will see Will.
N M V N

Второй метод: n-граммы

Bigrams

	N-M	M-V	V-N
mary-will	1	0	0
will-see	0	3	0
see-jane	0	0	1
will-will	1	0	0
see-mary	0	0	1
jane-will	1	0	0
see-will	0	0	1



Mary will see Will.



Data:

Mary will see Jane.



Will will see Mary



Jane will see Will.



Второй метод: n-граммы

Jane will spot Mary --- эта пара (биграмм) не встретится в таблице. Как мы тогда присвоим ему частотность/вероятность?



Mary



Jane



Will



Spot

Data:

Mary Jane can see Will.

Spot will see Mary.

Will Jane spot Mary?

Mary will pat Spot

Скрытые марковские модели

Будущее зависит от прошлого только через настоящее

Это называется марковской цепью

Скрытые марковские модели

Сначала классический пример про погоду и настроение

<https://www.youtube.com/watch?v=kqSzLo9fenk>

до 11 минуты

Скрытые марковские модели

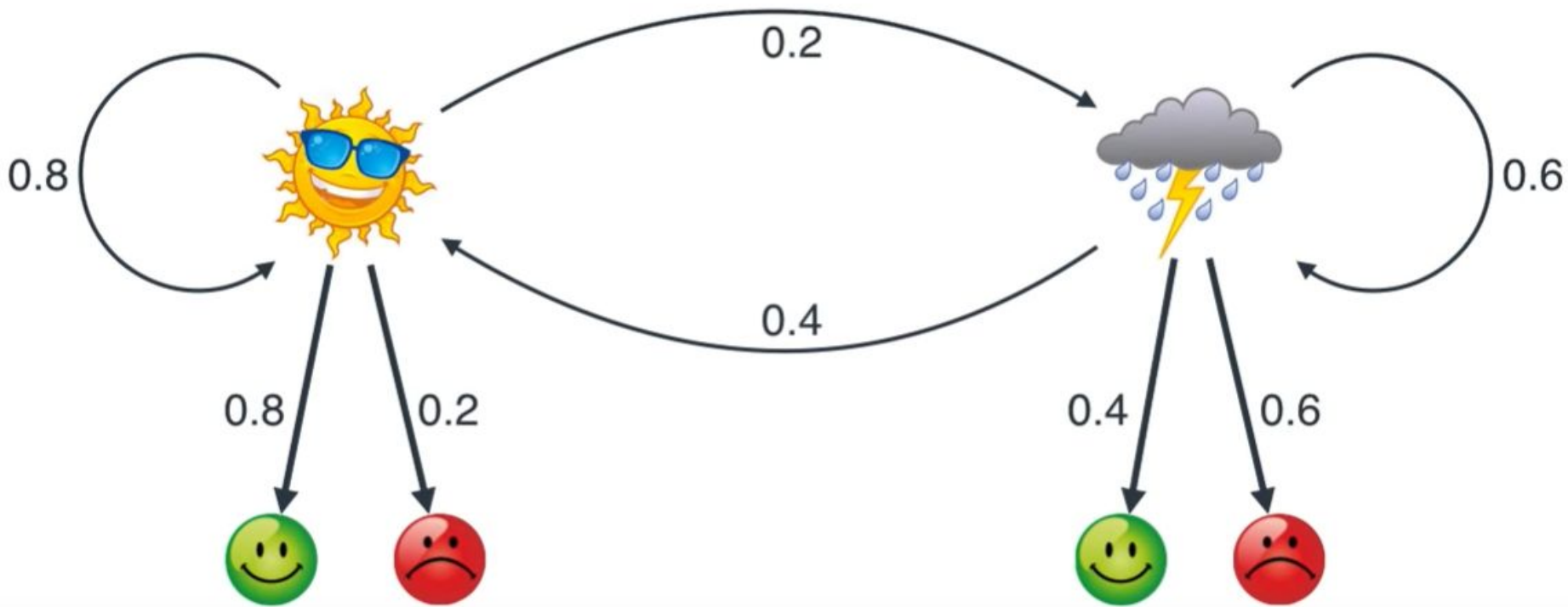
Нам нужна последовательность наблюдений. Событий и каких-то зависимых от них событий

Два типа вероятностей:

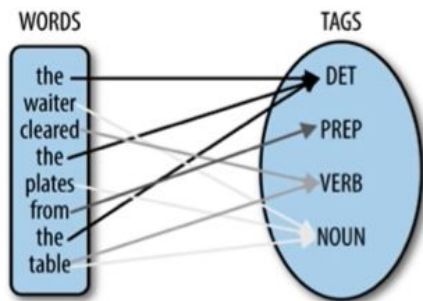
- вероятность перехода из одного состояния в другое
- вероятность того, что при условии, что есть одно состояние, то ему соответствует какое-то событие

Наша первая марковская модель

Как это соотноситься с языком?



Применения скрытых марковских моделей



Part of Speech Tagging



Robot Localization

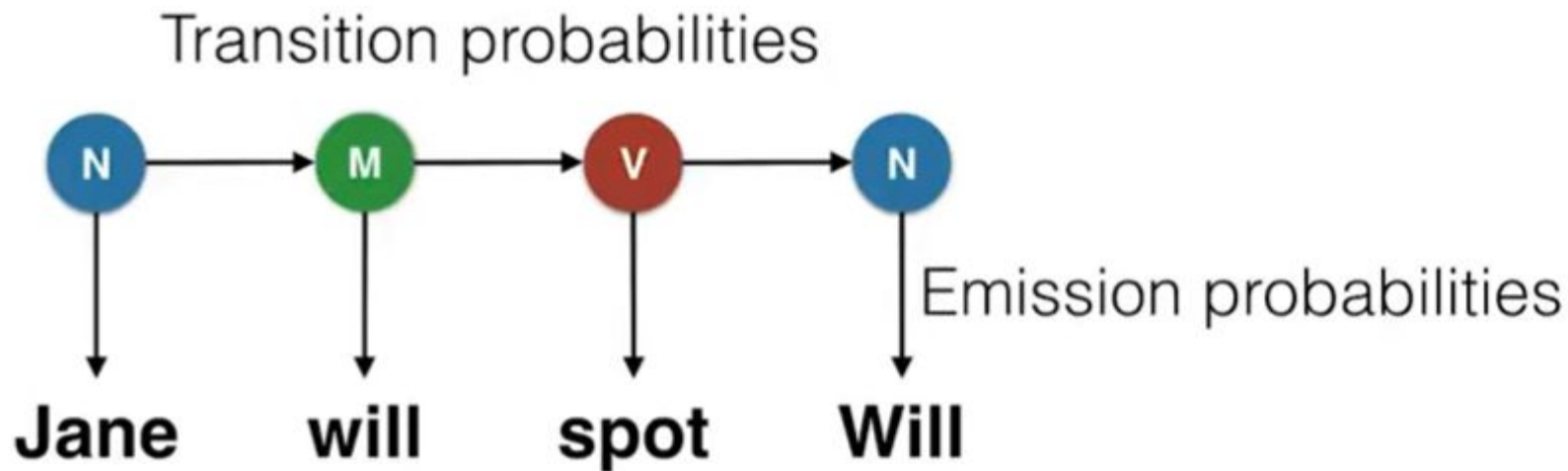


Genetics



Speech Recognition

Скрытые марковские модели



Скрытые марковские модели

Emission Probabilities

	N	M	V
Mary	4	0	0
Jane	2	0	0
Will	1	3	0
Spot	2	0	1
Can	0	1	0
See	0	0	2
Pat	0	0	1


Mary Jane can see Will.


Spot will see Mary.


Will Jane spot Mary?


Mary will pat Spot

Скрытые марковские модели

Emission Probabilities

	N	M	V
Mary	4/9	0	0
Jane	2/9	0	0
Will	1/9	3/4	0
Spot	2/9	0	1/4
Can	0	1/4	0
See	0	0	1/2
Pat	0	0	1/4

N N M V N

Mary Jane can see Will.

N M V N

Spot will see Mary.

M N V N

Will Jane spot Mary?

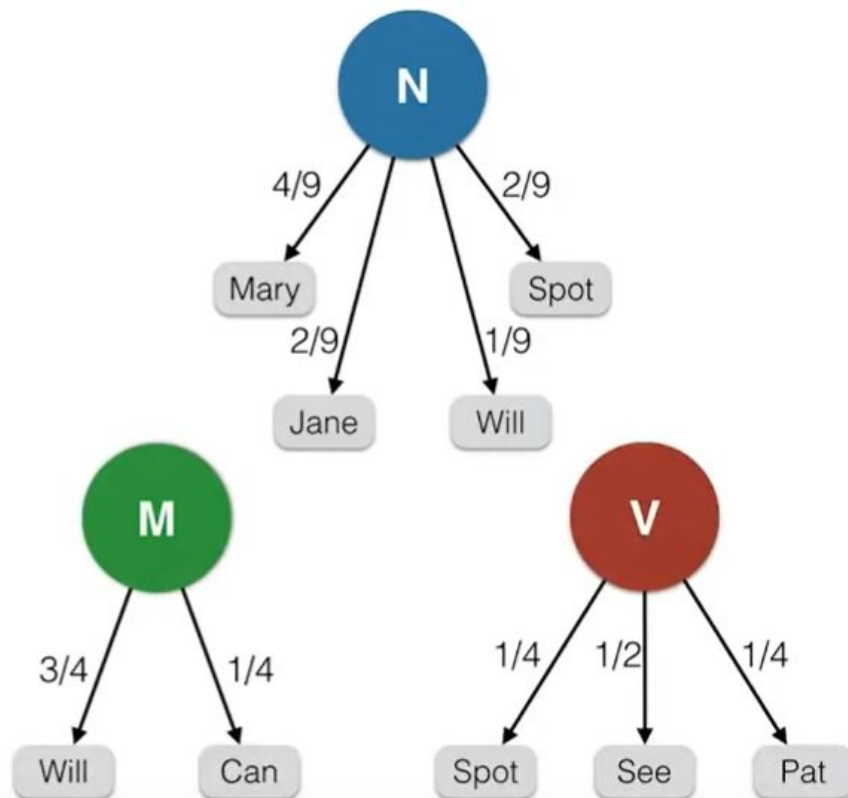
N M V N

Mary will pat Spot

Скрытые марковские модели

Emission Probabilities

	N	M	V
Mary	$4/9$	0	0
Jane	$2/9$	0	0
Will	$1/9$	$3/4$	0
Spot	$2/9$	0	$1/4$
Can	0	$1/4$	0
See	0	0	$1/2$
Pat	0	0	$1/4$



Скрытые марковские модели

Transition Probabilities

	N	M	V	<E>
<S>	3/4	1/4	0	0
N	1/9	1/3	1/9	4/9
M	1/4	0	3/4	0
V	1	0	0	0

<S> N N M V N <E>
Mary Jane can see Will.

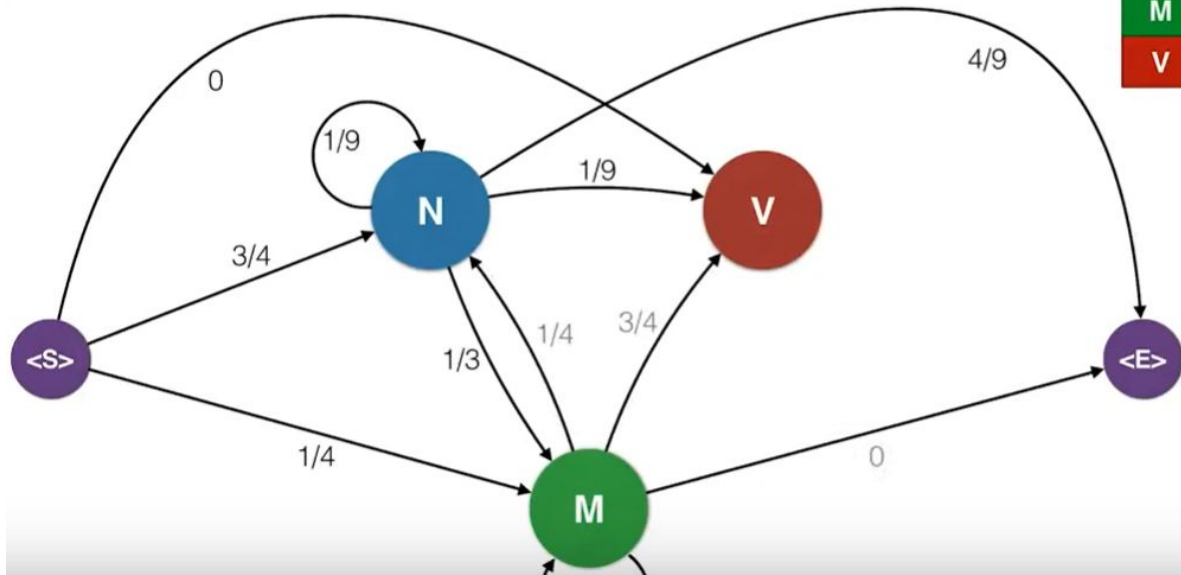
<S> N M V N <E>
Spot will see Mary.

<S> M N V N <E>
Will Jane spot Mary?

<S> N M V N <E>
Mary will pat Spot

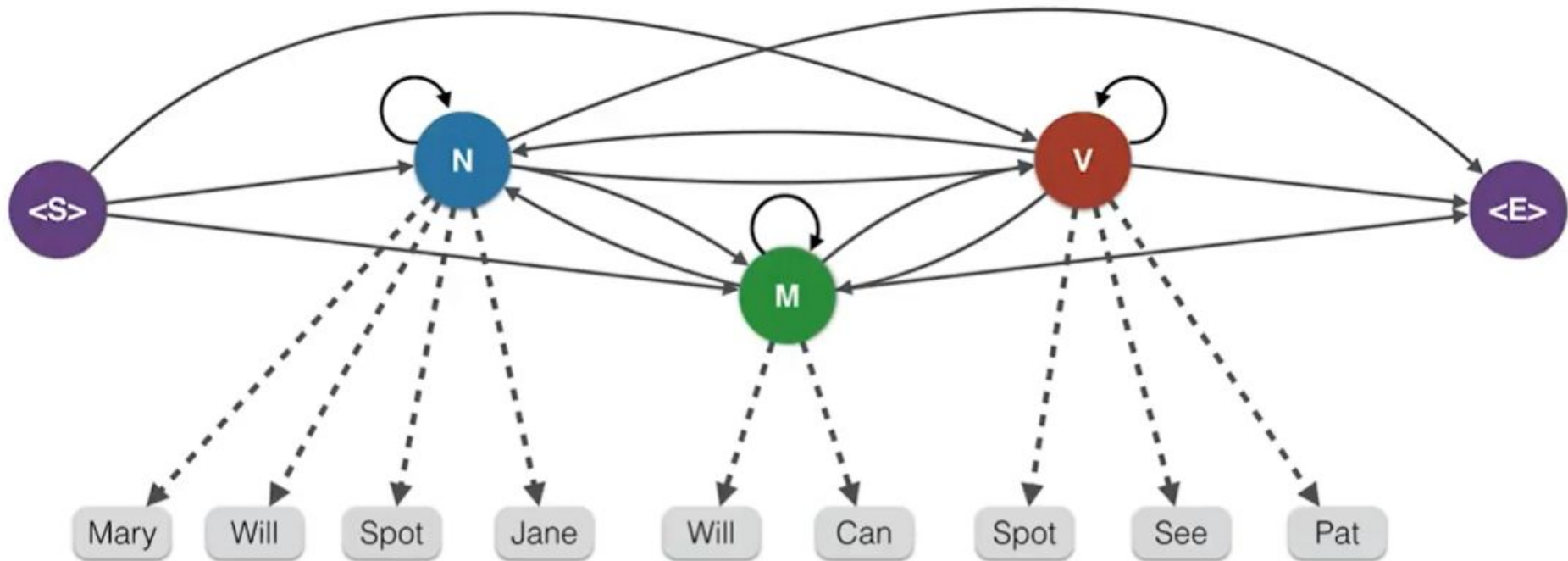
Скрытые марковские модели

Transition Probabilities



	N	M	V	<E>
<S>	$3/4$	$1/4$	0	0
N	$1/9$	$1/3$	$1/9$	$4/9$
M	$1/4$	0	$3/4$	0
V	1	0	0	0

Скрытые марковские модели



Скрытые марковские модели

https://www.youtube.com/watch?v=ZDXIExZIVMs&list=PLC0PzjY99Q_U5bba7gYJicCxIufrFmlTa&index=7

Скрытые марковские модели

Задача:

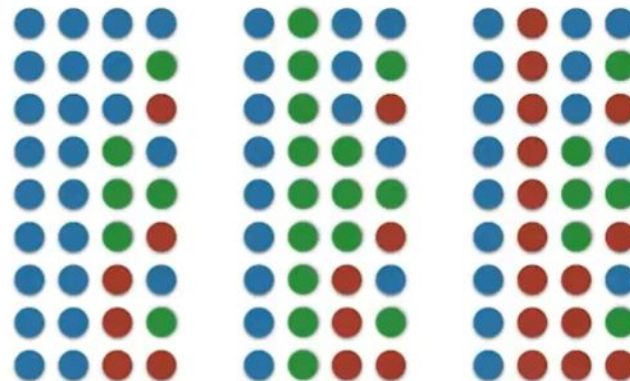
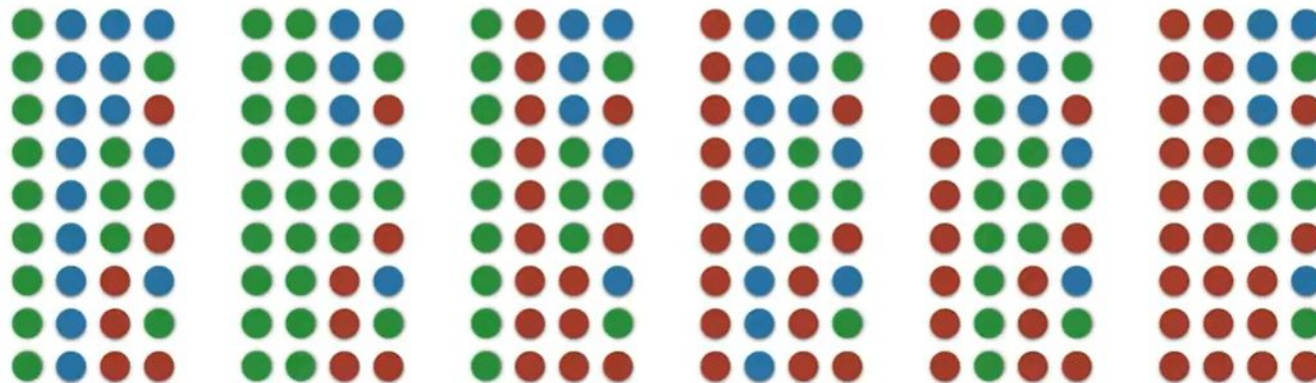
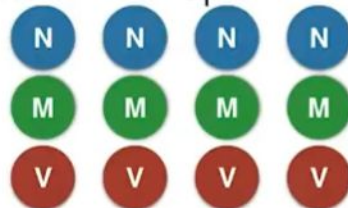
У нас есть 3 части речи: modal verb, verb, noun. Сколько возможных цепочек частей речи нужно проверить скрытой марковской модели для выбора наиболее вероятной для предложения

Jane will spot Will

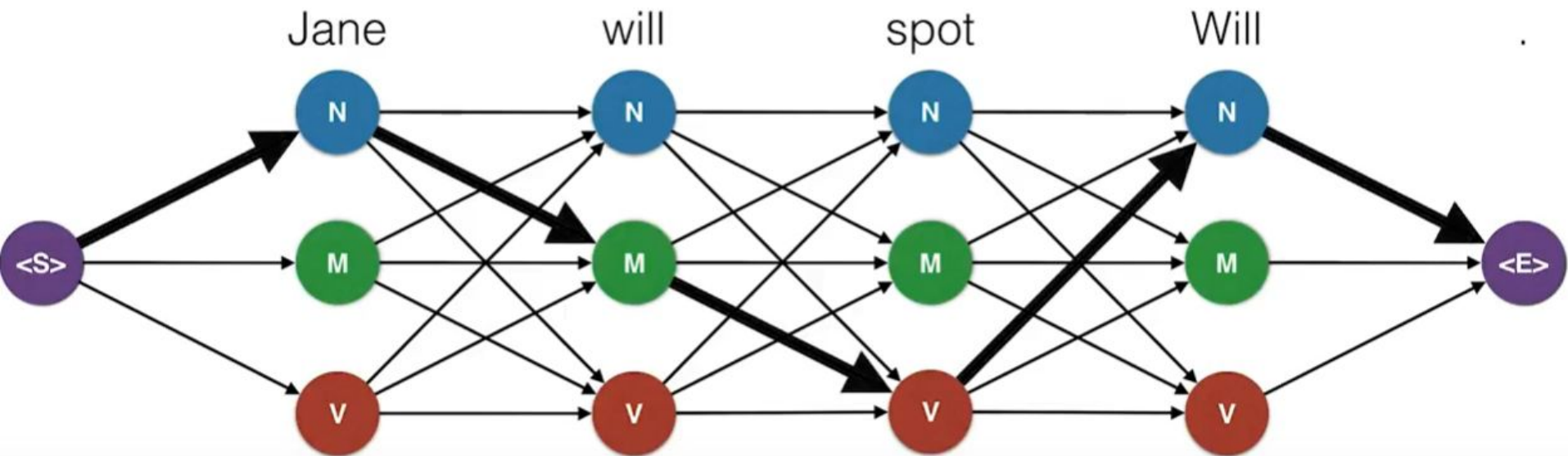
Скрытые марковские модели

Answer: 81 Possibilities

Jane will spot Will.



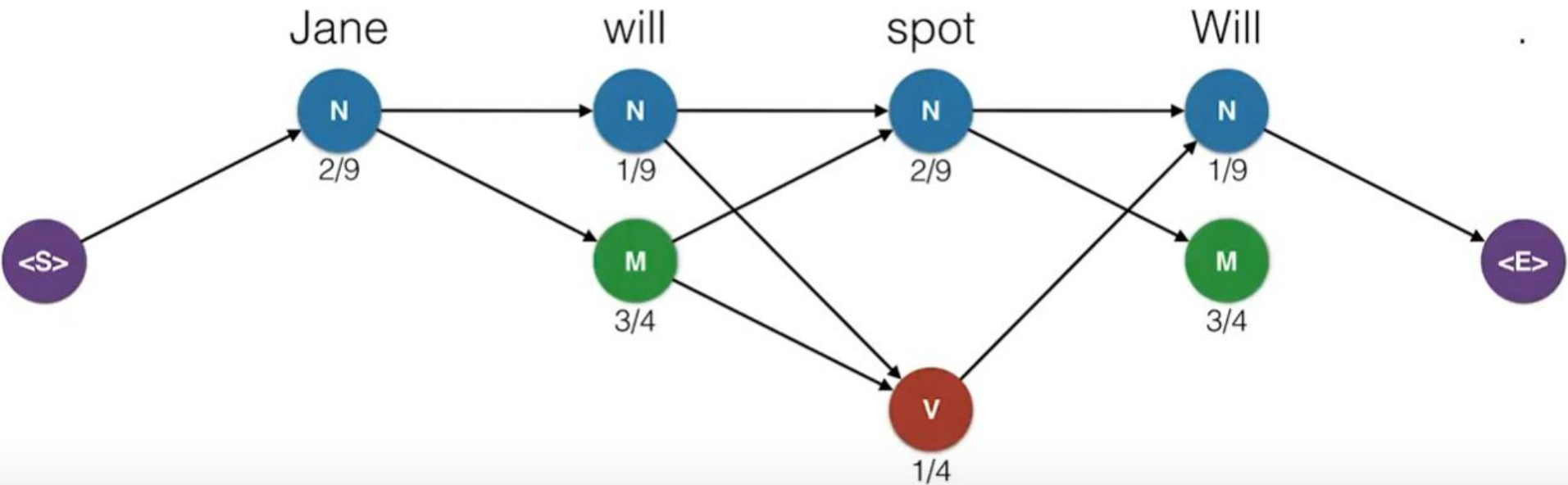
Скрытые марковские модели



Скрытые марковские модели

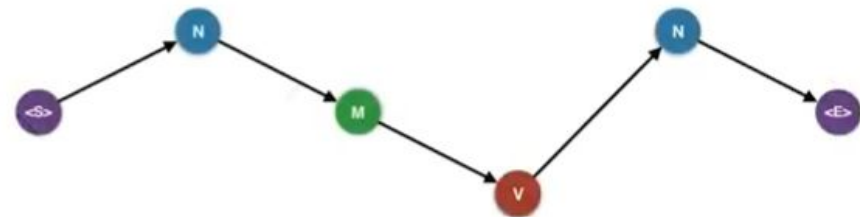
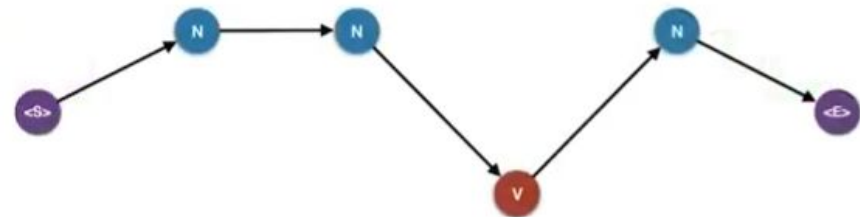
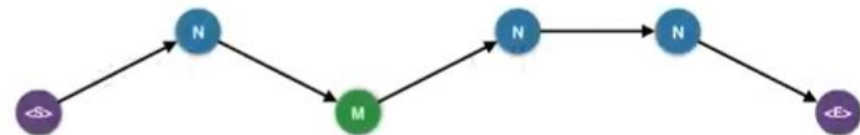
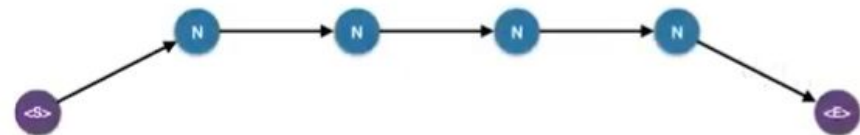
Сколько путей нам нужно проверить теперь?

Что мы удалили?

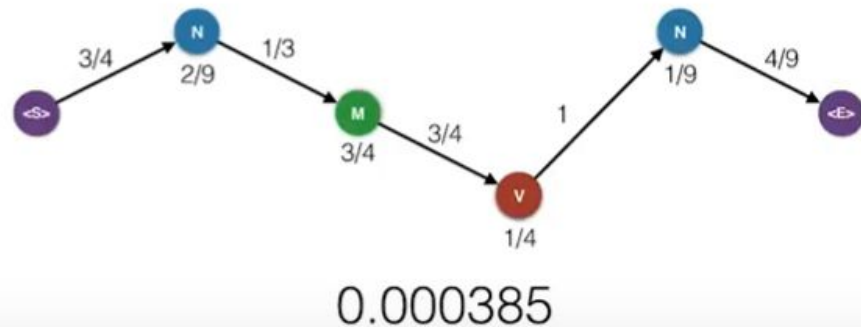
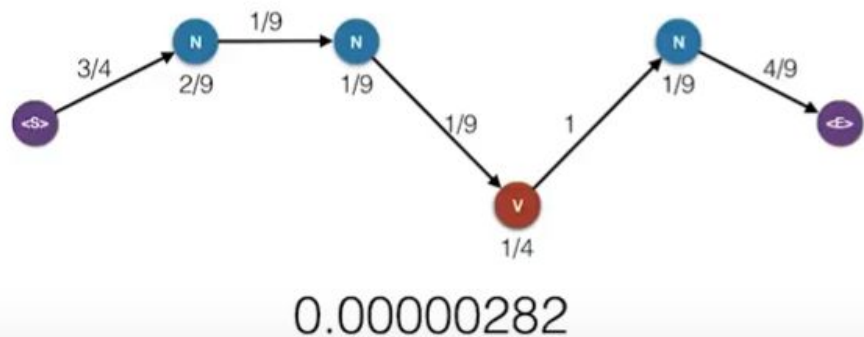
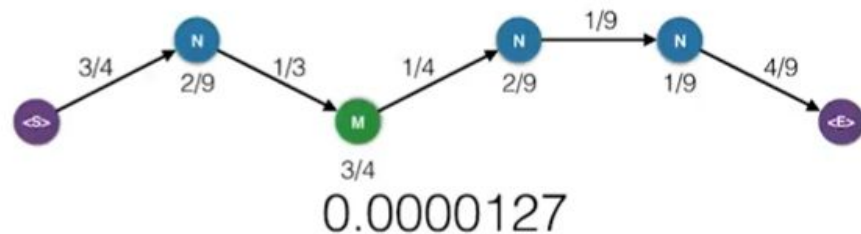
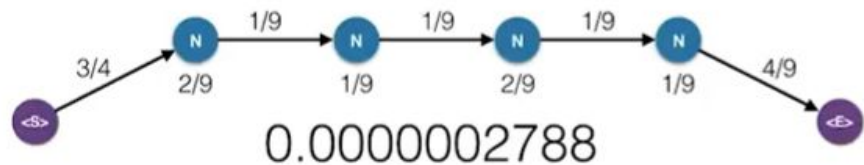


Скрытые марковские модели

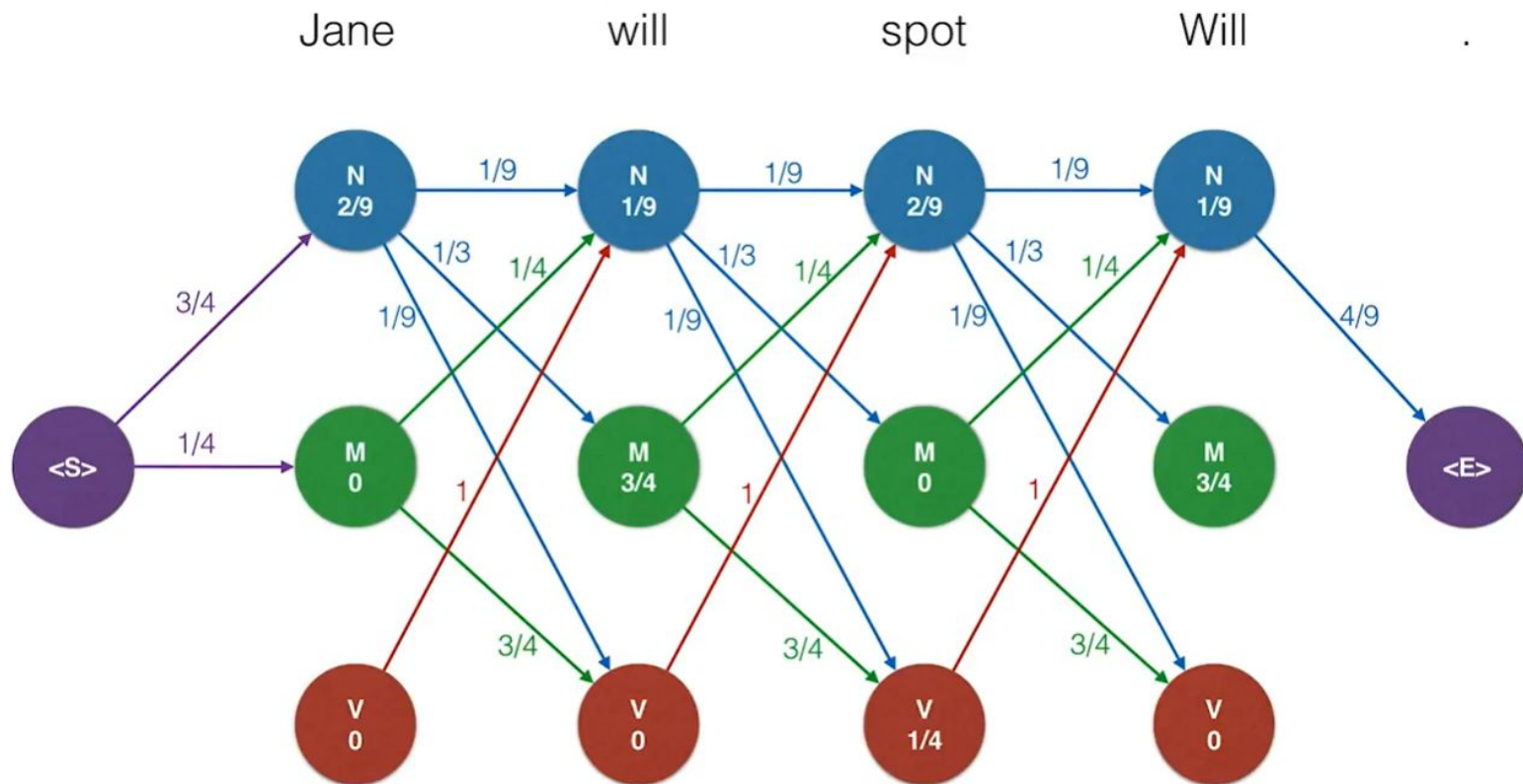
Ответ: 4



Скрытые марковские модели



Алгоритм Витерби



Алгоритм Витерби

https://www.youtube.com/watch?v=mHEKZ8jv2SY&list=PLC0PzjY99Q_U5bba7gYJicCxlufrFmITa&index=13

Задача на марковские процессы

В процессе опроса владельцев автомобилей трех американских марок: марки А, марки В, марки С, им был задан вопрос о том, какую торговую марку они бы выбрали для следующей покупки.

Среди владельцев автомобилей марки А 20% сказали что выберут опять эту же марку, 50% сказали, что они бы перешли на марку В, а 30% заявили, что предпочли бы марку С.

Среди владельцев автомобилей марки В 20% сказали, что перейдут на марку А, в то время как 70% заявили, что приобрели бы опять автомобиль марки В, а 10% заявили, что в следующий раз предпочли бы марку С.

Среди владельцев автомобилей марки С 30% ответили, что перешли бы на марку А, 30% сказали, что перешли бы на марку В, а 40% заявили, что остались бы верны той же марке С.

Задача на марковские процессы

Вопрос 1 : Если некто приобрел автомобиль марки А, то какова вероятность, что его второй машиной будет автомобиль марки С?

Некоторый fun

Задача:

Даны фразы из биографии французской актрисы Эммануэль Беар, приведённой на сайте «Каталог биографий известных актёров».

Некоторый fun

1. Режиссерам привзглянулась нежная красота Беар, и без ролей она не сидела.
2. Но «своего» режиссера Эммануэль порадовалось встретить лишь в 1992 году.
3. Обрелась невероятно тонкая и красивая картина (не в последнюю очередь благодаря Беар), которая обрела «Сезара» как оптимальный кинофильм того года.
4. Она настолько ладно сыграла метания героини между двумя супругчинами, что Даниэль Отёй, который был супругом Беар в кинофильме и в жизни, выбирал не приезжать на съемки, когда там снимались сцены с любовником героини Эммануэль.

Некоторый fun

5. Своих детей и свою личную жизнь артистка ревностно оберегает от внимания газетчиков, но папарацци очень любят Беар, видимо, позже что она очень фотогенична.

6. Много лет Эммануэль Беар была «лицом» известной фирмы «Christian Dior», но не так давно ее на этом посту поменяла российская манекенщица Крп правда Семеновская.

Задание 1. Отметьте слова, которые вам показались странными.

Задание 2. Объясните их появление в этом тексте.

Некоторый fun

Программки, которые порождают похожие слова, как правило именуют синонимайзерами, и их довольно ценят те, кто наваривает созданием и «раскруткой» вебсайтов. Дело в том, собственно что поисковые системы давным-давно научились отсеивать сайты-близнецы с схожим или же подобным содержанием, и задача скорого сотворения оригинальных слов на заданную тему довольно популярна. В онлайне просто присутствуют почти все 10-ки этих сервисов. Вот итог обработки сего абзаца одним из них.

Некоторый fun



«Детский сад № 7 «Семицветик»

ВИДЯЩИХ



Фотоальбом

[котлеты рецепт приготовления](#) > Подobie словарь

Подobie словарь

Словари – одним из наиболее стародавних и наиболее знатных достижений славной лингвистики.

Но в какой мере общераспространенные соображения о словарях подходят реальности? Как это становилось предварительно и что видоизменилось в новую, компьютерную эпоху? Всё ли располагать информацией словари – а коль скоро нет, то кто располагать сведениями точнее их?

Всегда ли заслуживает верить словарям, позволительно ли встать вовсе без них и что ожидает словари в будущем? Виноградова РАН, учитель учреждения лингвистики РГГУ, звание факультета филологии высочайшей средние учебные заведения экономики. Читает лекцию кандидатура филологических уроков боб Леонидович Иомдин, старый ученый служащий установления советского слога им. первейшие близости словарей

Некоторый fun

- Подумайте, как NLP помогает отсеивать такие сайты?
- Как вы думаете насколько давно придумали эту задачу?

Задача на языковые модели

Попробуйте описать образование глагольных основ в языке **йоулумни** (индейский язык где-то в Северной Америке). Запишите регулярками 3 основы

основа	герундий	дуратив
saw “кричать”	saw-inay	sawaa-ʔaa-n
cuim “разрушать”	cuim-inay	cuimuu-ʔaa-n
hooyo “называть”	hooy-inay	hooyo-ʔaa-n
diiyl “охранять”	diiyl-inay	diiyil-ʔaa-n
ʔilk “петь”	ʔilk-inay	ʔiliik-ʔaa-n
hiwiit “гулять”	hiwt-inay	hiwiit-ʔaa-n

Глагольные формы в языке йоулумни

Спасибо за внимание!

Литература