ՄԱՐԿՈՎՅԱՆ ՎԻՃԱԿԱՅԻՆ ՄՈԴԵԼՆԵՐԻ ԱՆՑՈՒՄԱՅԻՆ ՀԱՎԱՆԱԿԱՆՈՒԹՅՈՒՆՆԵՐԻ ԾՐԱԳՐԱՅԻՆ ՆԵՐԿԱՅԱՑՄԱՆ ԵՎ ՎԵՐԾԱՆՄԱՆ ՄԵԹՈԴՆԵՐ

ԽՈՒՐՇՈՒԴՅԱՆ ԱՐՄԵՆ
Տնտեսական գիտությունների թեկնածու, դոցենտ
Գավառի պետական համալսարանի դասախոս

ՀԱՅՐԱՊԵՏՅԱՆ ԼԻԱՆՆԱ
Գավառի պետական համալսարանի
բնագիտատնտեսագիտական ֆակուլտետի
համակարգչային ճարտարագիտություն բաժնի
մագիստրատուրայի 2 կուրսի ուսանող
էլփոստ՝ lianna.hayrapetyan03@gmail.com

Սույն հոդվածում ներկայացվում են Մարկովի շղթաների հիմնական սկզբունքները և դրանց գործնական կիրառությունը տեքստի գեներացման խնդրում:

Մարկովի շղթան բնութագրվում է «հիշողության բացակայության» հատկությամբ, ըստ որի՝ յուրաքանչյուր հաջորդ վիճակի (թոքենի) առաջացումը կախված է միայն նախորդ վիճակից: Ներկայացվում է վիճակների և անցումների մոդելավորումը գրաֆային կառուցվածքով, ինչպես նաև անցումների հավանականությունների մատրիցայի ձևակերպումը տվյալների պահպանման և ծրագրային իրականացման համար: Հոդվածում մանրամասն քննարկվում են գործնական իրականացման երկու հիմնական մոտեցում.

1. Պարզ մոդել. Հիմնված է մեկ տոկենից բաղկացած բանալիների վրա, որն ունի թերություններ՝ առաջացնելով քերականորեն և իմաստային ոչ համահունչ տեքստեր:

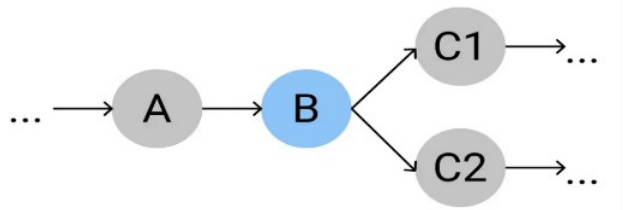
2. Բարդացված մոդել. Օգտագործում է մի քանի տոկեններից բաղկացած (N-գրամ) բանալիներ, ինչը թույլ է տալիս հաշվի առնել ավելի մեծ համատեքստ և բարելավել գեներացված տեքստի բնականությունն ու coherence-ը:

Ընդհանուր առմամբ, Մարկովի շղթայի մոդելը՝ լինելով պարզ և հաշվարկային առումով արդյունավետ, կարող է ծառայել որպես հիմք բնական լեզվի գեներացիայի ավելի զարգացած մոտեցումների համար, հատկապես երբ օգտագործվում են լայնածավալ կորպուսային տվյալներ և համատեքստային երկար բանալիներ:

Բանալի բառեր՝ *Մարկովյան շղթա, հիշողության բացակայություն, անցումների մատրիցա, վիճակների մոդել, տեքստի գեներացիա, տոկենիզացիա, բազմատոկեն բանալի, հավանականական մոդելավորում:*

Մարկովի շղթան վիճակների կամ իրադարձությունների այնպիսի հաջորդականություն է, որի դեպքում յուրաքանչյուր իրադարձության առաջացման հավանականությունը կախված է միայն նախորդ վիճակից և անկախ է դրան նախորդած բոլոր մյուս իրադարձություններից։ Այս հատկանիշը հայտնի է որպես «հիշողության բացակայություն» (*memorylessness*)։

Այլ կերպ ասած՝ եթե համակարգը գտնվում է B վիճակում, ապա հաջորդ՝ C_1 կամ C_2 վիճակների առաջացման հավանականությունը որոշվում է միայն B վիճակով և կախված չէ ավելի վաղ փուլերում տեղի ունեցած A կամ այլ վիճակներից։ Այս հատկության շնորհիվ Մարկովի շղթաները լայնորեն կիրառվում են տարբեր ոլորտներում՝ ֆիզիկայից և տնտեսագիտությունից մինչև ինֆորմատիկա և բնական լեզվի մշակում (Պատկեր 1.)։



Պատկեր 1. Վիճակի անցման հավանականություն

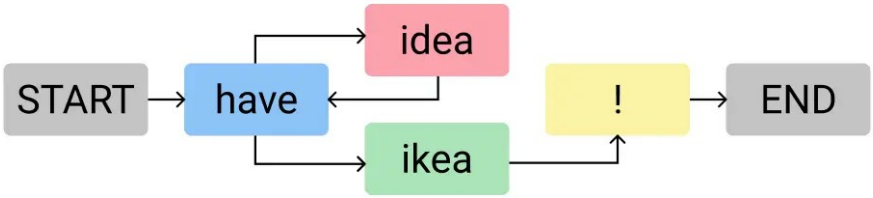
Այդ հատկության շնորհիվ Մարկովի շղթան (այն անվանում են նաև վիճակների մոդել) կարող է ստեղծել բերականորեն ճիշտ կառուցված նախադասություններ, որոնք, սակայն, սովորաբար ունեն թույլ կամ բացակայող իմաստային կապ միմյանց միջև։

Այժմ քննարկենք Մարկովի շղթաների կիրառության հնարավորությունը տեքստի գեներացման համար։ Տեքստի գեներացիայի համատեքստում իրադարձություն կամ վիճակ է համարվում առանձին տոկենը՝ բառը կամ կետադրական նշանը։ Օրինակ, եթե դիտարկենք «have idea have ikea!» արտահայտությունը, այն կարելի է ներկայացնել հետևյալ շղթայի տեսքով.

START → have → idea → have → ikea → ! → END

Այս ներկայացման մեջ յուրաքանչյուր անցում սահմանում է, թե տվյալ տոկենին ինչ կարող է հաջորդել։ Կետադրական նշանների ներառումը կարևոր է, քանի որ դրանք ապահովում են նախադասության կառուցվածքային և բերականական ամբողջականությունը։ Օրինակ, վերջակետը ազդարարում է մեկ նախադասության վերջի և մյուսի սկզբի մասին։

Մարկովի շղթայի վիճակները և դրանց միջև անցումները կարելի է արտապատկերել որպես գրաֆ, որտեղ գագաթները ներկայացնում են վիճակները (թոքենները), իսկ եզրերը՝ դրանց միջև հնարավոր անցումները։ Անցումների հավանականությունները ցույց են տալիս, թե ինչ հաճախականությամբ է կատարվում տվյալ անցումը։ (Պատկեր 2)։



Պատկեր 2. Մարկովյան վիճակների անցման գրաֆ

Անցումների հավանականությունները որոշում են, թե որ ուղղությամբ և ինչ հաճախականությամբ է տեղի ունենում շարժումը։ Օրինակ՝ եթե «have» բառից կարելի է անցնել «idea» կամ «ikea» բառերին հավասար հավանականությամբ, ապա համապատասխան շղթայում կստացվի հավասարակշռված բաշխում։ Եթե որոշ անցումներ ունեն հավասար հավանականություն, ապա գեներացված տեքստը կլինի բազմազան, իսկ եթե որևէ անցում ունի գերակշռող հավանականություն, շղթան հակված է դառնում կրկնությունների և ցիկլերի առաջացման։

START → have → idea → have → idea → have → idea → have → ikea → ! → END

Մարկովի շղթայի ծրագրային իրականացման համար կիրառվում է անցումների հավանականությունների մատրիցան, որի տողերը համապատասխանում են ելքային վիճակներին, իսկ սյունակները՝ հաջորդ վիճակներին։ Յուրաքանչյուր բջիջ պարունակում է տվյալ զույգի անցման հավանականությունը։

Աղյուսակ 1. Անցումների մատրիցան

	START	**have**	**idea**	**ikea**	**!**	**END**
START	0	1	0	0	0	0
have	0	0	0.5	0.5	0	0
idea	0	1	0	0	0	0
ikea	0	0	0	0	1	0
!	0	0	0	0	0	1

Այս աղյուսակում թիվ **0**-ով ներկայացված են **անհնար** անցումները, որոնք երբևէ տեղի չեն ունենում, իսկ թիվ **1**-ով՝ **հավաստ անցումները**, որոնք իրականանում են անպայման։ Նման ներկայացումը հարմար է, օրինակ, մատրիցան **երկչափ զանգվածի (array)** վերափոխելու համար։ Այդ պատճառով առավել արդյունավետ է օգտագործել կոմպակտ ներկայացում, որտեղ պահպանվում են միայն իրականում հնարավոր անցումները՝ օրինակ, ասոցիատիվ զանգվածի կամ բառարանի (dictionary) տեսքով։

Քանի որ աղյուսակի մեծ մասը կազմված է զրոներից՝ այսինքն՝ **անհնար անցումներից**, դրանք կարելի է չպահպանել՝ տեղեկությունը կրճատելով մինչև երկու սյունակ. այս կերպ պահպանվում են միայն այն

վիճակները, որոնց միջև անցումը իրականում հնարավոր է: Այն ներկայացված է Աղյուսակ 2. - ում:

Աղյուսակ 2.

Բանալին	**Հնարավոր հաջորդ իրադարձություններ**
START	→ have
have	→ have
idea	→ ikea
ikea	→ !
!	→ END

Այժմ մենք պահպանում ենք միայն սկզբնական իրադարձությունը և դրան հաջորդող հնարավոր իրադարձությունների ցանկը: Նման աղյուսակը կարող ենք վերածել օբյեկտի, որտեղ բանալին կլինի առաջին սյունակը (սկզբնական իրադարձությունը), իսկ արժեքը՝ երկրորդը (հաջորդող իրադարձությունների ցանկը):

Պատկեր 3. Մարկովյան մոդելի անցումների հավանականությունների ներկայացումը բանալի–արժեք տվյալների կառուցվածքի միջոցով

Բազմատոկեն իրադարձություններ

Վերոնշյալ օրինակով ներկայացված անցումների մատրիցան գործնականում կիրառելի է, սակայն այն բավարար չէ շարահյուսական առումով ճիշտ և բնական լեզվական կառուցվածքներ ստեղծելու համար: Միայն մեկ տոկենից բաղկացած իրադարձությունը պարունակում է չափազանց սահմանափակ տեղեկատվություն իր համատեքստի վերաբերյալ՝ մասնավորապես այն դիրքի մասին, որտեղ տվյալ միավորը

գտնվում է նախադասության կառուցվածքում: Արդյունքում, գեներացված տեքստը հաճախ կարող է լինել ոչ բնական և ոչ համահունչ՝ արտահայտվելով սխալ հոլովներով, նախդիրների ոչ ճիշտ հաջորդականություններով, կամ քերականորեն անհամաձայնված նախադասություններով:

Լեզվական կառուցվածքների ավելի իրատեսական մոդելավորման նպատակով ցանկալի է ապահովել, որ յուրաքանչյուր իրադարձություն որոշ չափով արտացոլի իր համատեքստը: Այս խնդիրը լուծելու համար անհրաժեշտ չէ պահպանել ամբողջական պատմությունը. բավարար է, որ յուրաքանչյուր կոնկրետ տոկեն ներկայացվի այնպիսի ձևով, որը ներառում է իր անմիջական համատեքստային տեղեկությունը: Դրան կարելի է հասնել այն եղանակով, որ մոդելի բանալին (key) ձևավորվի ոչ թե մեկ, այլ մի քանի հարակից տոկեններից:

Օրինակ՝ երկու տոկեններից բաղկացած բանալու պարագայում, նախորդ օրինակի տեքստային շղթան կարող է ներկայացվել հետևյալ անցումների մատրիցայի տեսքով:

Բանալին (2 տոկեն)	**Հնարավոր հաջորդ իրադարձություններ**
START → have	→ idea
have → idea	→ have
idea → have	→ ikea
have → ikea	→ !
ikea → !	→ END
! → END	

Երեք տոկեններից բաղկացած բանալիի դեպքում՝

Բանալին (3 տոկեն)	**Հնարավոր հաջորդ իրադարձություններ**
START → have → idea	→ have
have → idea → have	→ ikea
idea → have → ikea	→ !
have → ikea → !	→ END

Բանալին (3 տոկեն)	Հնարավոր հաջորդ իրադարձություններ
ikea → ! → END	

Տվյալների կառուցվածքը և գեներացման ալգորիթմը մնում են նույնը, սակայն բանալու երկարացման արդյունքում մենք ընդգրկում ենք ավելի շատ տեղեկատվություն յուրաքանչյուր տոկենի շրջակա համատեքստի վերաբերյալ։

Երկար բանալիների դեպքում հնարավոր հաջորդ իրադարձությունների քանակը նվազում է։ Օրինակ՝ վերջին աղյուսակում մենք այլընտրանքներ ընդհանրապես չունենք, բացի սկզբնական նախադասությունը գեներացնելուց։ Սակայն, եթե սկզբնական տվյալների (թոքենների) քանակը բավական մեծ է, այս մոտեցումը հնարավորություն է տալիս գեներացնել տեքստ ոչ թե առանձին «բառերով», այլ ամբողջական «բառակապակցություններով»։ Այս պատճառով ստեղծված տեքստը կունենա ավելի բնական և իրական խոսքին մոտ կառուցվածք։

Սկզբնական տեքստ

Տեքստի գեներացիայի նախնական փուլում անհրաժեշտ է ունենալ **կորպուսային տվյալներ**՝ իրական տեքստեր, որոնց հիման վրա կկառուցվի Մարկովի շղթան։

Սկզբում տեքստը ենթարկվում է **տոկենիզացիայի**, այսինքն՝ բաժանվում է բառերի, կետադրական նշանների և բացակների։ Այնուհետև այդ տոկենների հաջորդականությունից կազմվում է անցումների մատրիցա, որը հիմք է ծառայում նոր տեքստերի գեներացման համար։

Եզրակացություն

Մարկովի շղթաների մոդելը ապահովում է պարզ, բայց արդյունավետ մեխանիզմ՝ պատահական հաջորդականությունների մաթեմատիկական ներկայացման և անալիզի համար։ Այն թույլ է տալիս յուրաքանչյուր հաջորդ իրադարձություն որոշել միայն նախորդ վիճակի հիման վրա, ինչը բնորոշվում է որպես «հիշողության բացակայություն»։

Տեքստի գեներացիայի համատեքստում Մարկովի շղթաները հնարավորություն են տալիս ստեղծել քերականորեն համահունչ, սակայն երբեմն իմաստային կապը թույլ տեքստեր։ Մատրիցային կամ օբյեկտային կառուցվածքներով անցումների ներկայացումը թույլ է տալիս արդյունավետ պահպանել և մշակել տվյալները, բացառելով անհնար անցումները և կրճատելով հաշվարկների ծավալը։

Բացի այդ, **բազմաթիվ տոկեններով բանալիների** օգտագործումը բարելավում է տեքստի բնականությունը՝ ապահովելով համատեքստային տեղեկատվություն յուրաքանչյուր հաջորդող բառի համար և նվազեցնելով իմաստային սխալների հավանականությունը։ Այս մոտեցումը ցույց է տալիս, որ Մարկովի շղթաները կարող են ծառայել որպես հիմք ավելի ճշգրիտ և իրականությանը մոտ տեքստերի գեներացման մոդելների համար, երբ համակցվում են կոնկրետ կորպուսային տվյալների և համատեքստային բանալիների հետ։

Օգտագործված գրականության ցանկ

1. Jurafsky, D., Martin, J. H. *Speech and Language Processing*. 3rd ed., Draft version, Stanford University, 2023.
2. Goldberg, Y. *Neural Network Methods for Natural Language Processing*. Morgan & Claypool, 2017.
3. Mikolov, T., Chen, K., Corrado, G., Dean, J. “Efficient Estimation of Word Representations in Vector Space.” *Proceedings of ICLR*, 2013.
4. https://thecode.media/markov-chain/
5. https://thecode.media/markov-text/

МЕТОДЫ ПРОГРАММНОГО ПРЕДСТАВЛЕНИЯ И ДЕКОДИРОВАНИЯ ПЕРЕХОДНЫХ ВЕРОЯТНОСТЕЙ В МАРКОВСКИХ МОДЕЛЯХ СОСТОЯНИЙ

ХУРШУДЯН АРМЕН
Кандидат экономических наук, доцент
Преподаватель Гаварского государственного университета

АЙРАПЕТЯН ЛИАННА
Студент 2-го курса магистратуры отделения Компьютерная инженерия
факультет Естественных наук и экономики
Гаварский государственный университет
электронная почта: lianna.hayrapetyan03@gmail.com

В данной статье представлены основные принципы цепей Маркова и их практическое применение в задаче генерации текста.

Цепь Маркова характеризуется свойством «отсутствия памяти» (memorylessness), согласно которому возникновение каждого последующего состояния (токена) зависит исключительно от предыдущего. В работе представлено моделирование состояний и переходов с помощью графовой структуры, а также формирование матрицы вероятностей переходов для хранения данных и программной реализации. Подробно рассматриваются два основных подхода к практической реализации:

1. ***Простая модель: Основана на ключах, состоящих из одного токена. Этот подход имеет недостатки, порождая грамматически и семантически несвязные тексты.***
2. ***Усложненная модель: Использует ключи, состоящие из нескольких токенов (N-граммы), что позволяет учитывать более широкий контекст и улучшить естественность и связность (coherence) генерируемого текста.***

В целом, модель цепи Маркова, будучи простой и вычислительно эффективной, может служить основой для более продвинутых подходов к генерации естественного языка, особенно при использовании объемных корпусных данных и длинных контекстных ключей.

Ключевые слова: *цепь Маркова, отсутствие памяти, матрица переходов, модель состояний, генерация текста, токенизация, многотокеновый ключ, вероятностное моделирование.*

METHODS FOR SOFTWARE REPRESENTATION AND DECODING OF TRANSITION PROBABILITIES IN MARKOV STATE MODELS

KHURSHUDYAN ARMEN
Condidate of Economics, Associate Professor
Lecturer of Gavar State University

HAYRAPETYAN LIANNA
2ed-year Master student of Computer Engineering Depertment
Faculty of Natural Sciences and Economics
Gavar State University
e-mail: lianna.hayrapetyan03@gmail.com

This article presents the fundamental principles of Markov chains and their practical application in text generation tasks.

The Markov chain is characterized by the property of "memorylessness," according to which the occurrence of each subsequent state (token) depends solely on the previous state. The article presents the modeling of states and transitions via a graph structure, as well as the formulation of a transition probability matrix for data storage and software implementation. Two main approaches to practical implementation are discussed in detail:

1. ***Simple Model: Based on keys consisting of a single token. This approach has drawbacks, often producing texts that are grammatically and semantically incoherent.***
2. ***Complex Model: Utilizes keys consisting of multiple tokens (N-grams), which allows for considering a larger context and improving the naturalness and coherence of the generated text.***

Overall, the Markov chain model, being simple and computationally efficient, can serve as a basis for more advanced approaches to natural language generation, especially when utilizing large-scale corpus data and long contextual keys.

Keywords: Markov chain, memorylessness, transition matrix, state model, text generation, tokenization, multi-token key, probabilistic modeling.