УДК: 004.43(045)

Жумабаева А.Н., Таалайбеков Н. Т.
Н.Исанов атындагы КМКТАУ
Компьютердик лингвистика жана маданий аралык байланыштар кафедрасынын ага окутуучусу,
Н.Исанов атындагы КМКТАУ
Компьютердик лингвистика жана маданий аралык байланыштар кафедрасынын студенти
Бишкек, Кыргыз Республикасы
Жумабаева А.Н., Таалайбеков Н. Т.
Старший преподаватель кафедры компьютерная лингвистика и межкультурная коммуникация
КГУСТА им. Н. Исанова
Студент кафедры компьютерная лингвистика и межкультурная коммуникация
КГУСТА им. Н. Исанова
Бишкек, Кыргызская Республика
A.N. Zhumabaeva, N.T. Taalaibekov
Senior Lecturer of the Department of Computational Linguistics and Intercultural Communication
KSUCTA n.a. N. Isanov,
Student of the Department of Computational Linguistics and Intercultural Communication
KSUCTA n.a. N. Isanov,
Bishkek, Kyrgyz Republic

ТАБИГЫЙ ТИЛДИ ИШТЕТҮҮДӨГҮ PYTHON КИТЕПКАНАЛАРЫ

БИБЛИОТЕКИ PYTHON ДЛЯ ОБРАБОТКИ ЕСТЕСТВЕННОГО ЯЗЫКА

PYTHON LIBRARIES FOR NATURAL LANGUAGE PROCESSING

Аннотация: Табигый тилди иштетүүнүн (ТТИ) көйгөйлөрүн чечүү үчүн азыркы учурда көптөгөн китепканалар түзүлүүдө. ТТИ долбоору үчүн программалоо тилин тандоодо Pythondун көптөгөн касиеттери абдан ыңгайлуу кылат, айрыкча табыгий тилди иштетүүгө келгенде. Pythondун бул өзгөчө касиеттерине биринчи кезекте тилдин жөнөкөй синтаксиси жана тунук семантикасы кирет. Мындан тышкары, иштеп чыгуучулар башка тилдер жана куралдар менен мыкты интеграциялык колдоосун пайдалана алышат. Python иштеп чыгуучуларга ТТИге байланыштуу маселелерди чечүү үчүн кеңири куралдар жана китепканалар менен камсыз кылат.

Аннотация: В настоящее время создаются множество библиотек для решения проблем обработки естественного языка (ОЕЯ). Многие свойства Python делают удобным вариантом при выборе языка программирования для проекта, особенно если речь идет об обработке естественного языка. К этим особенностям относятся прежде всего, простой синтаксис и понятная семантика языка. Кроме того, разработчики могут воспользоваться отличной поддержкой интеграции с другими языками и инструментами. Python предоставляет разработчикам широкий спектр инструментов и библиотек для решения проблем, связанных с ОЕЯ.

Annotation: Many libraries are currently being created to solve natural language processing (NLP) problems. Many of the properties of Python make it a convenient choice when choosing a programming language for a project, especially when it comes to natural language processing. These features include, first of all, simple syntax and clear semantics of the language. In addition, developers can take advantage of excellent support for integration with other languages and tools. Python provides developers with a wide range of tools and libraries for solving NLP-related problems.

Өзөктүү сөздөр: Табигый тилди иштетүү, китепканалар, куралдар, машиналык окутуу, жасалма интеллект, иштеп чыгуучулар

Ключевые слова: обработка естественного языка, библиотеки, инструменты, машинное обучение, искусственный интеллект, разработчики

Key words: natural language processing, libraries, tools, machine learning, artificial intelligence, developers

Табигый тилди иштетүү (ТТИ) – бул тексттик маалыматтарды талдоо үчүн лексикалык жана машиналык окутуу алгоритмдерин колдонуу менен жасалма интеллект тармагы. ТТИнин максаты компьютерди кадимки жана рационалдуу аныкталган көрсөтмөлөрдөн тышкары, адамдын оозеки тилин жана тексттин маанисин түшүнүүгө кантип, кандайча мажбурлоосун аныктоо.

Сүйлөөнү таануу, документтерди жалпылоо, аты аталган объектилерди таануу, суроолорго жооп берүү, авто-толтуруу, текстти болжолдуу (предиктивдүү) киргизүү ж.б. у.с. системаларын түзүү үчүн табигый тилди иштетүүнү колдонобуз.

Бүгүнкү күндө, көпчүлүгүбүздө сүйлөгөн сөзүбүздү тааныган смартфондор бар. Биздин сөзүбүздү түшүнүү үчүн аларда ТТИ колдонулат. Ошондой эле, операциондук системасында сүйлөө речти таануу мүмкүнчүлүгү бар ноутбуктарды да колдонобуз.

ТТИ өнүккөн эсептөө көндүмдөрүнө (навыкам) таянгандыктан, иштеп чыгуучулар мыкты, жеткиликтүү курамдарга (инструменты) муктаж. Табигый тил менен иштей турган кызматтарды (сервистерди) түзүү үчүн бул шаймандар ТТИнин ыкмаларынан жана алгоритмдеринен максималдуу пайда өндүрүп чыгарууга жардам бериши керек.

ТТИ жасалма интеллектин бир бөлүгү болгондуктан көпчүлүк учурда машиналык окутууга таянат. ТТИ процесси төмөнкүдөй:

- текст жана үн түрүндө болгон адамдын сөзүн киргизүүдөгү жазуулар;
- үн берилиштерди текстке айландыруу;
- берилиштердин маанисин аныктоо үчүн грамматика, талдоо ыкмалары колдонуу менен текстти иштетүү;
- иштелип чыккандарды экрандан көрсөтүү жолу менен же үн чыгаруу аркылуу адамга жеткирүү.

Мурун табигый тилин иштеп чыгуу боюнча долбоорго эксперттер гана катыша алышчу. Мындай иш математика, машиналык окутуу жана лингвистика боюнча мыкты билимдерди талап кылган. Эми иштеп чыгуучулар текстти алдын-ала иштетүүнү жөнөкөйлөтүү үчүн даяр куралдарды колдоно алышат. Бул аларга машиналык окутуунун моделин түзүүдө басым кылууга мүмкүнчүлүк берет. ТТИ нин көйгөйлөрүн чечүү үчүн көптөгөн китепканалар түзүлгөн.

ТТИ долбоору үчүн программалоо тилин тандоодо Pythondун көптөгөн касиеттери абдан ыңгайлуу кылат, айрыкча табыгий тилди иштетүүгө келгенде. Pythondун бул өзгөчө касиеттерине биринчи кезекте тилдин жөнөкөй синтаксиси жана тунук семантикасы кирет. Мындан тышкары, иштеп чыгуучулар башка тилдер жана куралдар менен мыкты интеграциялык колдоосун пайдалана алышат. Бул машиналык окутуу сыяктуу ыкмалар үчүн керектелет. Python иштеп чыгуучуларга ТТИге байланыштуу маселелерди чечүү үчүн кеңири куралдар жана китепканалар менен камсыз кылат. Азыр эң маанилүү китепканаларды карап чыгайлы.

1. NLTK, Naturalm Language Toolkit деген сөздүн кыскартылышы. Pythondo жазылган табыгий тилди символдук жана статистикалык иштетүүгө арналган

китепканалардын жана программарадын топтому. Python үчүн табигый тилди иштетүүдөгү эң мыкты китепканалардын бири.Көптөгөн тилдерде, анын ичинде орус тилинде иштөөнү колдойт. Ал 100 корпустан ашуун жана аларга байланыштуу WordNet, WebText Corpus, NPS Chat, SemCor, FrameNet сыяктуу лексикалык ресурстарга ээ. Windows, Mac OS и Linux операционндук системалар үчүн көптөгөн колдонмолору бар акысыз китепкана.

Бүгүнкү күндө ал ТТИ жана машиналык окутууну жаңы үйрөнө баштаган иштеп чыгуучулар үчүн билим алууда негизгиси болуп саналат.

1. NLTK китепканасы Пенсильвания университетинде Стивен Берд жана Эдвард Лаупер тарабынан иштелип чыккан. Ал ТТИ иизилдөөлөрүндө маанилүү ролду ойногон.

Башка Python китепканалары жана куралдары менен бирге NLTK дүйнө жүзүндөгү университеттерде окуу программасында колдонулат.

Китепкана ар тараптуу, бирок табигый тилди иштетүүдө колдонуу кыйын. NLTK бир топ жай болуп, өндүрүштө тездик менен өнүгүп жаткан талаптарга жооп бербейт. Ресурстар: NLTK Book, Погрузитесь в NLTK.

2. Textblob. Тажрыйбасын Pythonдогу ТТИден баштаган иштеп чыгуучулар үчүн Textblob милдеттүү болуп саналат. Textblob үйрөнчүктөргө ТТИ менен негизги маселелерин өздөштүрүүгө жардам берүү үчүн жөнөкөй интерфейс менен камсыз кылат. Прототиптерди долбоорлоордо өтө пайдалуу. Бирок анда да NLTKнын негизги кемчиликтери бар.

Ал ар кандай табигый тилди иштетүү китепканасы үчүн бир нече төмөнкүдөй стандарттуу функцияларды камсыз кылат: кептин бөлүгүн белгилөө; сезимди талдооо; классификациялоо; n-грамм; сөз өзгөрүүлөр; WordNet интеграциясы; Google Translate негизинде тилдин котормосу жана аныктамасы; сөздөрдүн жана сөз айкалыштардын жыштыгы; талдоо; орфографиялык оңдоо. Ресурстар: Документация TextBlob, Основы обработки естественного языка с помощью TextBlob.

3. CoreNLP. Бул китепкана, Стэнфорд университетинде Java тилинде жазылып иштелип чыккан. Көп тилдер, анын ичинде Python үчүн оболочкалар менен жабдылган. Pythondo табыгий тилди иштетүүдө өз күчүн сынап көрүүнү каалаган иштеп чыгуучулар үчүн пайдалуу. Иштеп чыгаруу чөйрөсүндө китепкана тез жана жакшы иштейт. Кээ бир CoreNLP компоненттерин NLTK менен интеграциялоого болот, бул NLTKнын эффективдүүлүгүн сөзсүз түрдө жогорулатат. Ал бир нече лингвистикалык функцияларды сунуш кылат. Бул функциялар лемматизацияга, сүйлөөнү (речти) жана морфологиялык тилдерди бөлүүгө, токенизацияга которулат.

Эгерде сиз ар дайым жаңыланып туруучу жана жогорку сапаттагы аналитикалык сунуш кылган грамматикалык анализ куралдарынын кеңири спектри бар заманбап жана ишенимдүү ТТИ инструменттерин издеп жатсаңыз, анда бул эң сонун тандоо. CoreNLP ийкемдүү, башка тилдер менен жакшы интеграцияланышы сиздин муктаждыктарыңыз үчүн ыңгайлуу кеңейе турган жана функцияналдуу вариант. Ресурстары: Документация CoreNLP, Список оболочек Python для CoreNLP.

4. Gensim бул Питондо жазылган тематикалык моделдөө, окшоштуктарын изилдөө жана табигый тилди иштетүү китепканасы. 2009- жылы иштелип чыккан. Ал чоң текстти, массивдерди иштете алат. Gensim эки нерседен артык болууга умтулат, бири – табигый тилди иштетүү, экинчиси – маалымат издөө.

Gensimдин колдонуунун милдеттүү шарты - бул аны иштетүү үчүн NumPy жана Scipy пакеттеринин бар болушу. Ресурстары: gensim Documentation

5. SpaCy – өндүрүштө колдонууга арналган салыштырмалуу жаш китепкана. Ошол себептен, NLTK сыяктуу башка питонТТИ китепканаларына караганда кыйла жеткиликтүү.

SpaCy бүгүнкү күндө эң тез синтаксистик талдоочуну сунуштайт. Мындан тышкары, инстументарий Cython до жазылгандыктан, ал дагы тез жана эффективдүү. Бирок бир да куралдар кемтиксиз иштебейт. Буга чейин карап чыккан китепканаларга салыштырмалуу SpaCy эң аз (жети) тилди колдойт.

Машиналык окутуунун популярдуулугунун өсүшү менен жана SpaCy негизги китепкана экендигин билдиргендиктен бул курал жакын арада дагы көп программмалоо тилдерин камтый башташы мүмүкүн. SpaCy чоң маселелердин аткарууда аз убакыт коротууга арналган ашыкчылыгы жок китепкана. Ресурсы: Документация spaCy, Введение в НЛП с SpaCy.

6. Polyglot. Бул китепкана анча белгилүү эмес, бирок ар тараптуу талдоону жана көптөгөн тилдерди кодогондуктан көп колдонулган китепканалардын бири болуп саналат. NumPутин жардамы менен ал абдан тез иштейт. SpaCyy колдобогон тилдерге байланыштуу долбоорлорго эффективдүү жана жөнөкөй мыкты вариант.

7. Pattern. Табигый тилдер менен иштөө үчүн Python иштеп чыгуучулары колдонгон NLP китепканаларындагы дагы бир байлык. Pattern сөз бөлүгүн белгилөө *part-of-speech tagging*), сезимдерди, вектордук мейкиндикти анализдөө, моделдөө (SVM), классификация, кластерлөө, n-граммдык издөө жана WordNet куралдары менен камсыз кылат. Pattern – биринчи кезекте берилиштерди интеллектуалдык анализдөө, табигый тилди иштетуу, машиналык окутуу жана тармактык анализ сыяктуу ар кандай көптөгөн максаттарга керектерлеген куралдар бар Python учун веб-майнинг модулу. Ал 350дөн ашык бирдик тесттери жана 50дөн ашык мисалдар менен документтештирилген. Веб-API колдоосу Patternттин функцияналдуулугун кеңейтүү үчүн башка программалоо тилдери менен оңой интеграцияланышын камсыз кылат.

8. Scikit–learn. Иштеп чыгуучуларга машиналык окутуунун моделин түзүү үчүн алгоритмдердин кенири спектрин сунуштайт. Анын иштеши текстти классификациялоо маселелерин чечүүгө арналган объекттерди жаратуу үчүн “сөз баштыкча” ыкмасын («мешок слов», *bag-of-words model*) колдонууга мүмкүндүк берет. Бул китепкананын күчтүү жагы класстардын интуитивдүү методдору болуп саналат. Мындан тышкары, scikit-learn программасы иштеп чыгуучуларга алардын мүмкүнчүлүктөрүн колдонууга жардам берген мыкты документтерге ээ. Бирок, китепканада текстти алдын-ала иштеп чыгуу үчүн нейрон тармактары колдонулбайт

Каалаган система менен өз ара аракеттенишүүсү негизги иш-аракеттердин бири болуп саналат жана колдонуучулар үчүн бул процессти мүмкүн болушунча жеңилдетүүгө ар дайым көңүл буруу керек. Сүйлөшүү системалары бара-бара көнүмүш адатка айлангандыктан, биздин күнүмдүк сүйлөөбүздү таануу үчүн биздин чечимдерге болгон талап өсүүдө. Табигый тилди иштетүү бизге бул маселени компьютерлер үчүн жөнөкөйлөтүүгө мүмкүндүк берди. Дайыма бат өзгөрүлүп турган мезгилде, Python ылайыкташууга жөндөмдүү, инновация киргизүү жана көп сандаган заманбап эсептөөчү көйгөйлөрдүн чечимдерин сунуштай ала турганын далилдеди. Табигый тилди иштетүүгө келгенде, Python эң мыкты технология болуп калды.

Колдонулган адабияттар:

1. Хобсон Лейн , Ханнес Хапке, Коул Ховард, . Обработка естественного языка в действии. – СПБ.: Питер, 2020. – 576 с.

2. Steven Bird, Ewan Klein, Edward Loper, Jason Baldridge “Proceedings of the Third Workshop on Issues in Teaching Computational Linguistics: Multidisciplinary Instruction with the Natural Language Toolkit ”. Columbus, Ohio, USA, June 2008. – P.62 –70.
3. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : учеб. пособие / Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. — М.: МИЭМ, 2011.
4. Steben Bird, Ewan Klein, Edward Loper “Natural Language Processing with Python”. O’Reilly, 2009.
5. 5. Python Libraries for Natural Language Processing. An Overview Of popular python libraries for Natural Language Processing. Claire D. Costa Apr 28, 2020
6. https://python-school.ru/
7. ttps://pythonist.ru/

Рецензент: к.филол.н., доцент Жумалиева Г.Э.