GLOBAL RESEARCH NETWORK
ONLINE RESEARCH HUB

AMERICAN JOURNAL OF SOCIAL AND HUMANITARIAN RESEARCH

ISSN: 2690-9626
Vol.3, No 5, 2022

Software Technologies for Research and Development of Linguistic Models

Karimov Jasur Khasanboyevich; Zokirov Sanjar Ikromjon Ugli, Ph.D
Senior Lecturer, Department of IES, Fergana Polytechnic Institute, Fergana, Uzbekistan

ABSTRACT: The article discusses the study of the morphological system of the Uzbek language using the Python programming system and its NLTK library as one of the software technologies aimed at solving linguistic problems.

KEYWORD: linguistic model, software technologies, Python, NLTK, stop-words, body text

Introduction

The use of information and communication technologies in scientific research and in the process of visualizing results is one of the most effective research methods. This is the result of the fact that modern information technologies make it possible to obtain and save a large amount of data, perform fast data processing, automate technical calculations, obtain results in several versions, visualize the results and predict situations [1].

Today, there are a number of software technologies for creating, researching, and using computer-oriented language models of natural language. These include a variety of technologies for processing natural language, such as programming systems, various programming libraries that include linguistic models, a linguistic processor and linguistic databases, and morphological and semantic-syntactic analyzers.

Based on recent research on automatic processing of linguistic data, Python stands out among programming systems such as Java, Perl, C ++, C #, Ruby, VB .NET, Scala and R for its very simple instructions and very large information capacity [2].

In particular, Python and its SpaCy, gensim, and NLTK packages are now being used effectively around the world as natural language text processing methods [3].

However, modern methods of studying the problems of Uzbek linguistics with the help of today's automated information systems have not yet been developed, both theoretically and practically. Although preliminary research is being conducted on modeling word groups based on the syntactic structure of Uzbek sentences [4] and improving the linguistic base for morphological analysis of words [5], but the formation of formal grammar based on natural Uzbek grammar The creation of various software technologies for the study of the Uzbek language on the basis of modern methods, based on these linguistic models and existing technologies, remains a topical issue.

ISSN 2690-9626 (online), Published by "Global Research Network LLC"
under Volume: 3 Issue: 5 in May-2022 https://grnjournals.us/index.php/AJSHR

NLTK and its modules. NLTK (Natural Language ToolKit) is a package of programs and libraries for symbolic and statistical processing of natural language, which can be used effectively in the fields of computer linguistics, machine learning and information retrieval.

The NLTK library is an open source system with a dynamic feature as a component module of the Python programming system, which is constantly replenished with new programming modules. Therefore, even if it is installed on a computer, it is possible to download a new version of it online and constantly update the existing database.

All software modules in this package, the first version of which was first released in 2001, are written in Python (look at the Table 1).

1-Table NLTK modules and their functional functions (abbreviated)

Modules	The role of language in processing	Methods of functionalization
nltk.corpus	Accessing corpora	Standardized interfaces for corpus and lexicons
nltk.tokenize, nltk.stem	String processing	Word and speech tokenizers, stemmers
nltk.collocations	Collocation discovery	t-criteria, xi-square, point information
nltk.tag	Part-of-speech tagging	n-gram, backoff, Brill, HMM, TnT
nltk.classify, nltk.cluster	Classification	Solution tree, maximum entropy, Baes method, k-mean method.
nltk.chunk	Chunking	Regular expressions, n-grams, named objects
nltk.parse	Parsing	Table, character separation, unification, probability, interdependence
nltk.sem, nltk.inference	Semantic interpretation	Lambda-calculation, first-order predicates

Stop-words. It is known that for words that do not have a clear grammatical meaning in the given sentence or are almost not used in search engines, the term "stop-words" is accepted in English, and "stop-slova" or "shumovye slova" in Russian. For example, the words belonging to the category "stop-slova" in Russian are understood as [6]:

- connectors (or, but, so that, then, then, only just);
- pronouns (he, we, him, you, you, you, her, what, who, them, all, they, me, all, me, me, so);
- prepositions (for, on, on, with, from, from, before, without, over, under, for, with, after, in);
- downloads (not, same, then, would, total, total, even, yes, no);
- pronouns (oh, oh, oh, bravo, hello, thank you, sorry);
- numbers and figures (1, 2, 3 one, two, three first, second, third, ...);
- punctuation marks and special characters (., - _ = + /!;:%? *);
- introductory words (say, maybe, let's say, to be honest, for example, in fact, however, in general, in general, probably);
- separate letters (a, b, c, ...);
- indefinite prepositions, forms (something, some, somewhere, somehow, further, closer, earlier, later, sometime), etc.

ISSN 2690-9626 (online), Published by "Global Research Network LLC"
under Volume: 3 Issue: 5 in May-2022 https://grnjournals.us/index.php/AJSHR

These words, specific to each language, are organized in the form of a separate list through a text file, which is usually placed in the text body as an internal module of the NLTK library [7].

Some of the above modules of the NLTK library can also be used effectively in modeling elements of Uzbek grammar. Of course, this process cannot be done directly, of course. To do this, it is necessary to carry out certain computer modeling work on Uzbek linguistic objects.

Experimental part. Below we get acquainted with the text of the program, which produces Uzbek words of the stop-words type, formed by the authors as a linguistic base using the nltk.corpus module of the NLTK library.

1. First we run the Python program. If the program does not have an NLTK library installed, type the following command on the command line to install it:

>>> import nltk

>>> nltk download ()

The information capacity of the NLTK library is around 4 GB, and its installation is done online, which takes some time.

2. So, once the NLTK library is loaded, we import words of type stop-words from its nltk.corpus module:

>>> from nltk.corpus import stopwords

3. After this command, words of the stop-words type, which are available in more than 30 languages, such as English, Russian, German, French, Turkish, Tajik, Kazakh and so on, are loaded into the memory. Stop-words in each language are organized in the form of a separate text file, to which we add a file of stop-words in the Uzbek language, and then type the following code text:

>>> set (stopwords.words ('uzbek'))

The result is as follows (Fig. 1):

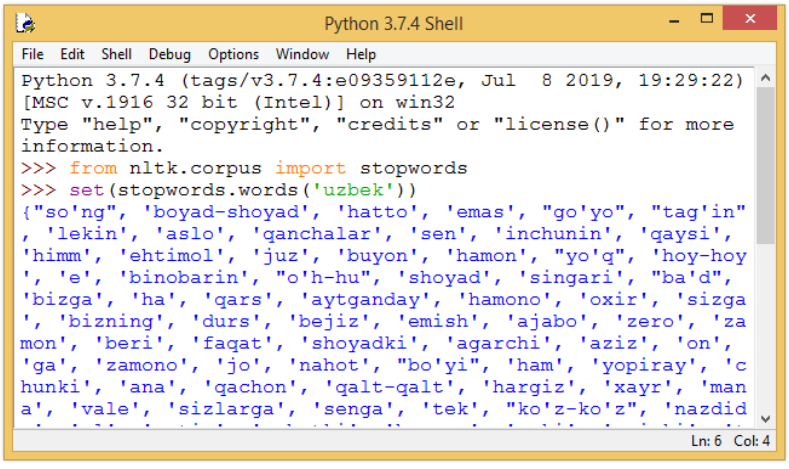

Fig. 1. Screening of Uzbek words of stop-words type.

You can also use the tokenization module of the NLTK package (nltk.tokenize) for words and phrases in Uzbek. For example, let us be given the following text:

"It's me. I am a programmer. I love my job."

ISSN 2690-9626 (online), Published by "Global Research Network LLC"
under Volume: 3 Issue: 5 in May-2022 https://grnjournals.us/index.php/AJSHR

Each sentence, each word, and the punctuation in this text can be divided into separate blocks. To do this, consider the following program text:

```
>>> from nltk.tokenize import sent_tokenize, word_tokenize
>>> MyText = "It's me. I'm a programmer. I love my profession."
>>> print (sent_tokenize (MyText))
['This is me.', 'I am a programmer.', 'I love my profession.']
>>> print (word_tokenize (MyText))
['This', 'I', '.', 'I am a programmer', '.', 'I love my profession', '.']
```

SpaCy Library. SpaCy is an open source free library designed to work with NLP in Python, which is actively used in linguistic research. In particular, SpaCy is one of the most effective service technologies for word processing, retrieving the necessary information from the text, developing intellectual functions, and working with NLP problems in general.

Using the SpaCy library, more than a dozen different intellectual problems can be solved on text in natural language (look at the Table 2).

2-Table. Modules and basic functions of the SpaCy library

Module name	Description
Tokenization	Performs the function of segmenting words, punctuation and other elements in the text.
Part-of-speech **(POS)** Tagging	A noun or verb replaces words specific to a word group in a way that is specific to the flow of speech.
Dependency Parsing	Just as an object or a subject, it represents the syntactic connections that represent the relationship between each individual token.
Lemmatization	Forms the lemma of the word, that is, the basic form, for example, the lemma of the word "was" is like "be", the lemma of the word "rats" is like "rat".
Sentence Boundary Detection **(SBD)**	Performs searching and segmentation of individual sentences.
Named Entity Recognition **(NER)**	Names of people are used to identify and highlight named objects associated with companies and geographic areas.
Similarity	Comparison of words, text, documents, and similar objects in terms of their mutual similarity.
Text Classification	Used to perform a classification operation on the whole or a certain part of the text.
Rule-based Matching	Find a sequence of tokens similar to a regular expression based on text and linguistic annotation.
Training	Update statistical forecast models.
Serialization	Save items in file or byte rows.

When working with the SpaCy technology of the Python programming system, it is possible to refer to 15 (currently) language models in 8 languages (see Table 3).

ISSN 2690-9626 (online), Published by "Global Research Network LLC"
under Volume: 3 Issue: 5 in May-2022 https://grnjournals.us/index.php/AJSHR

Table-3 Linguistic models of the SpaCy library

Natural languages	Linguistic models
English	en_core_web_sm en_core_web_md en_core_web_lg en_vectors_web_lg
German	de_core_news_sm de_core_news_md
French	fr_core_news_sm fr_core_news_md
Spanish	es_core_news_sm es_core_news_md
Greek	el_core_news_sm el_core_news_md
Portuguese	pt_core_news_sm
Italian	it_core_news_sm
Dutch	nl_core_news_sm

These models can be installed separately online or with the SpaCy library during the installation process. To do this, use the following commands, for example:

pip install –U spacy

python –m spacy download en_core_web_sm

However, it is also possible to install these files from a local folder by downloading them from the appropriate site, using the following command:

pip install d: \ en_core_web_sm.tar.gz

The SpaCy library was released in 2015 and can be considered as a new modification of NLTK.

In general, in addition to Python, it is possible to work with hundreds of other technologies for processing texts in different natural languages, for example, using Pattern, one of the most popular modules of Python program, grammatically, morphologically, text in English, Spanish, German, French, Italian and Dutch. spelling and syntactic analysis is possible.

Conclusion. The software technologies mentioned above and all the language models studied using them actually serve as a basis for the machine translation procedure. We will still need many formal models to study various aspects of the Uzbek language on the basis of these technologies.

In this article, we have formed a database of Uzbek-type stop-words and studied it as an object in the NLTK library, which can serve as a basis for the creation of automatic search engines for Uzbek text.

References:

[1]. L.K. Mamadilieva, S.I. Zokirov. “Automation problems of finding the optimal coordinates of a photocell in a selective radiation photothermogenerator.” IJARSET, Vol. 6, Issue 9, Sep 2019

[2]. Bird S., Klein E., Loper E. Natural language processing with Python: analyzing text with the natural language toolkit. – " O'Reilly Media, Inc.", 2009.

ISSN 2690-9626 (online), Published by “Global Research Network LLC"
under Volume: 3 Issue: 5 in May-2022 https://grnjournals.us/index.php/AJSHR

[3]. Okhunov, M., & Minamatov, Y. (2021). Application of Innovative Projects in Information Systems. European Journal of Life Safety and Stability (2660-9630), 11, 167-168.

[4]. Рахматова О. К., Рахимов А. Р. ПРАГМАЛИНГВИСТИЧЕСКИЕ КОНЦЕПЦИИ ФЕНОМЕНА РЕЧЕВОГО ПОВЕДЕНИЯ И РЕЧЕВОГО ДИСКУРСА //Подписано в печать: 19.11. 2021 Дата выхода в свет: 22.11. 2021 Формат 70х100/16. – 2021.

[5]. Холматова Д. А., Рахматова О. К. ТЕОРЕТИЧЕСКИЕ ОСНОВЫ РАЗРАБОТКИ УЧЕБНЫХ ПОСОБИЙ //Вопросы науки и образования. – С. 30.

[6]. Kadirjanovna R. O. Pragmalinguistic Concepts of the Phenomenon of Speech Behavior and Speech Discourse //International Journal of Multicultural and Multireligious Understanding. – 2021. – Т. 8. – №. 5. – С. 495-500.

[7]. Абдуллаева М. Х., Башарова Г. Г., Рахматова О. К. Преимущества индивидуального подхода в образовательном процессе //Проблемы современной науки и образования. – 2019. – №. 12-1 (145). – С. 88-90.

[8]. Холматова Д. А., Рахматова О. К., Косимова Д. Р. Этнографическая терминология и ее лингвистический анализ (на материалах русского и узбекского языков) //Вестник науки и образования. – 2019. – №. 19-3 (73). – С. 40-42.

[9]. Рахматова О. К., Косимова Д. Р. Актуальные проблемы преподавания русского языка в технических вузах //Проблемы современной науки и образования. – 2019. – №. 12-2 (145). – С. 127-129.

[10]. Кучкарова Д. Т. ЭНЕРГОСБЕРЕГАЮЩИЕ СИСТЕМЫ УПРАВЛЕНИЯ МАШИН И АГРЕГАТОВ ШЕЛКОМОТАНИЯ //ББК 1 Р76. – 2021. – С. 92.

[11]. Кучкарова Д. Т. Анализ энергосберегающих режимов перекачивающих машин и агрегатов на промышленных предприятиях //Проблемы современной науки и образования. – 2020. – №. 1 (146).

[12]. Shamsunovna N. A. THE CONCEPT, ESSENCE, FEATURES OF THE METHODS AND TECHNIQUES USED IN TEACHING FOREIGN LANGUAGES. – 2022.

[13]. Nabievna K. B. The study of quantitatively in linguistics //ACADEMICIA: An International Multidisciplinary Research Journal. – 2021. – Т. 11. – №. 3. – С. 1848-1854.

[14]. Nabievna K. B. MANIFESTATION OF QUANTITATIVELY AT THE LEXICAL LEVEL. – 2022.

[15]. Nabiyev M. Moisture Accumulation and Durability of Panel Walls in Aggressive Environment //Eurasian Journal of Engineering and Technology. – 2022. – Т. 5. – С. 40-44.

[16]. Yuldashev N. K. et al. The effect of mechanical deformation on the photovoltaic properties of semiconductor polycrystalline film structures CdTe: Sn //Scientific-technical journal. – 2019. – Т. 23. – №. 3. – С. 9-14.

[17]. Сулаймонов Х. М. и др. Фотоэлектрические свойства полупроводниковых поликристаллических пленочных структур CdTe: Sn при статических механических деформациях //Известия Ошского технологического университета. – 2019. – №. 3. – С. 180-186.

[18]. Сулаймонов Х. М. ВЛИЯНИЕ ЦИКЛИЧЕСКИХ ДЕФОРМАЦИЙ НА ЭЛЕКТРОПРОВОДНОСТЬ КОМПОЗИТНЫХ ПЛЕНОК(BixSb1− x) 2Te3 В

ISSN 2690-9626 (online), Published by "Global Research Network LLC"
under Volume: 3 Issue: 5 in May-2022 https://grnjournals.us/index.php/AJSHR

ЗАВИСИМОСТИ ОТ ЧАСТОТЫ ПЕРЕМЕННОГО ТОКА //Знание. – 2016. – №. 2-3. – С. 24-26.

[19]. Сулаймонов Х. М. ОПТИЧЕСКИЕ СВОЙСТВА ПОЛИКРИСТАЛЛИЧЕСКИХ ПЛЕНОК PbSe В ИК ОБЛАСТИ СПЕКТРА //Oriental renaissance: Innovative, educational, natural and social sciences. – 2021. – Т. 1. – №. 11. – С. 828-836.

[20]. Sulaymonov K. M. et al. EDGE ABSORPTION SPECTRA OF HEAVILY DOPED POLYCRYSTALLINE PBTE: PB AND PBTE: TE FILMS //Scientific-technical journal. – 2020. – Т. 24. – №. 2. – С. 22-26.

[21]. Abduqaxxorovich O. S. et al. Development and research of heterostructures with an internal thin layer based on p-type silicon //European science review. – 2018. – №. 9-10-1. – С. 183-185.

[22]. Kasimakhunova A. M., Zokirov S. I., Norbutaev M. A. Development and Study of a New Model of Photothermogenerator of a Selective Radiation with a Removable Slit //Development. – 2019. – Т. 6. – №. 4.

[23]. Kasimakhunova A. M. et al. Photo Thermal Generator of Selective Radiation Structural and Energetic Features //Journal of Applied Mathematics and Physics. – 2019. – Т. 7. – №. 06. – С. 1263.

[24]. Kasimaxunova A. M., Norbutaev M., Baratova M. Thermoelectric generator for rural conditions //Scientific progress. – 2021. – Т. 2. – №. 6. – С. 302-308.

[25]. Müller A. C., Guido S. Introduction to machine learning with Python: a guide for data scientists. – " O'Reilly Media, Inc.", 2016.

[26]. Hakimov M. H. Modeli obrabotki russkogo jazyka po tehnologii mnogojazykovogo modeliruemogo komp'juternogo perevodchika //Dostizhenija nauki i obrazovanija. – 2019. – №. 3 (44).

[27]. Abdurakhmonova N. Modeling analytic forms of verb in Uzbek as stage of morphological analysis in machine translation // Iranian Journal of Social Sciences and Humanities Research, 2017, vol. 5, issue 3, pp. 97-107.

[28]. Termin: Stop-slova. (URL: https://promopult.ru/)Stop-words with NLTK. (URL: https://pythonprogramming.net/).

[29]. Stop-words with NLTK. (URL: https://pythonprogramming.net/).

[30]. Зокиров С. И. У., Норбутаев М. А. СОЛНЕЧНЫЙ ТРЕКЕР ДЛЯ ФОТОТЕРМОГЕНЕРАТОРА СЕЛЕКТИВНОГО ИЗЛУЧЕНИЯ //Universum: технические науки. – 2021. – №. 4-5 (85). – С. 9-13.

ISSN 2690-9626 (online), Published by "Global Research Network LLC"
under Volume: 3 Issue: 5 in May-2022 https://grnjournals.us/index.php/AJSHR