ФГБОУ ВО «БАШКИРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

ФАКУЛЬТЕТ РОМАНО-ГЕРМАНСКОЙ ФИЛОЛОГИИ

Утверждено:
на заседании кафедры
протокол№_7_ от «27»_января__2021_ г.
Зав. кафедрой _____/Морозкина Е.А.

Согласовано:
Председатель УМК факультета /института
_____/Мазунова Л.К.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ (МОДУЛЯ)

дисциплина___Лингвистические компоненты информационных электронных систем_
(наименование дисциплины)

Часть формируемая участниками образовательных отношений
Дисциплина по выбору
(указать часть (обязательная часть или часть, формируемая участниками образовательных отношений, факультатив))

Программа бакалавриата

Направление подготовки
45.03.03 Фундаментальная и прикладная лингвистика
(указывается код и наименование направления подготовки (специальности))

Направленность (профиль) подготовки
Языковые технологии
(указывается наименование направленности (профиля) подготовки)

Квалификация
Бакалавр
(указывается квалификация)

Разработчик (составитель) ____доц., к. филол.н._____ (должность, ученая степень, ученое звание)	_____/_Мифтахова Р.Г.___ (подпись, Фамилия И.О.)

Для приема: 2021

Уфа 2021 г.

Составитель / составители: Мифтахова Р.Г.

Рабочая программа дисциплины *утверждена* на заседании кафедры протокол от «_27_» __января_____ 2021 г. № _7__

Заведующий кафедрой _____/ Морозкина Е.А.

Список документов и материалов

1. Перечень планируемых результатов обучения по дисциплине, соотнесенных с установленными в образовательной программе индикаторами достижения компетенций
2. Цель и место дисциплины в структуре образовательной программы
3. Содержание рабочей программы (объем дисциплины, типы и виды учебных занятий, учебно-методическое обеспечение самостоятельной работы обучающихся)
4. Фонд оценочных средств по дисциплине

   4.1. Перечень компетенций и индикаторов достижения компетенций с указанием соотнесенных с ними запланированных результатов обучения по дисциплине. Описание критериев и шкал оценивания результатов обучения по дисциплине.

   4.2. Типовые контрольные задания или иные материалы, необходимые для оценивания результатов обучения по дисциплине, соотнесенных с установленными в образовательной программе индикаторами достижения компетенций. Методические материалы, определяющие процедуры оценивания результатов обучения по дисциплине.
5. Учебно-методическое и информационное обеспечение дисциплины

   5.1. Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины

   5.2. Перечень ресурсов информационно-телекоммуникационной сети «Интернет» и программного обеспечения, необходимых для освоения дисциплины, включая профессиональные базы данных и информационные справочные системы
6. Материально-техническая база, необходимая для осуществления образовательного процесса по дисциплине

1. Перечень планируемых результатов обучения по дисциплине, соотнесенных с установленными в образовательной программе индикаторами достижения компетенций

По итогам освоения дисциплины обучающийся должен достичь следующих результатов обучения:

Категория (группа) компетенций[1] (при наличии ОПК)	Формируемая компетенция (с указанием кода)	Код и наименование индикатора достижения компетенции	Результаты обучения по дисциплине
	ПК-9 - Способен пользоваться лингвистически ориентированными программными продуктами	*ПК-9* Знает типы, характеристики и особенности основных доступных в Интернете лингвистических ресурсов.	*Знание* типов, характеристики и особенности основных доступных в Интернете лингвистических ресурсов.
		ПК-9 Умеет сравнивать данные, полученные с использованием различных электронных лингвистических ресурсов и систем; применять методы математического анализа и моделирования в профессиональной деятельности	*Умение* сравнивать данные, полученные с использованием различных электронных лингвистических ресурсов и систем; применять методы математического анализа и моделирования в профессиональной деятельности
		ПК-9 Владеет практическим опытом тестирования электронных лингвистических ресурсов, систем и компонентов	*Владение* практическим опытом тестирования электронных лингвистических ресурсов, систем и компонентов
	ПК-12 - Способен проводить квалифицированное тестирование	*ПК-12 Знает* различные способы и инструментарий тестирования	*Знание* различных способов и инструментарий тестирования

[1] Указывается только для УК и ОПК (при наличии).

	лингвистически ориентированных программных продуктов, электронных ресурсов, лингвистически ориентированных систем и лингвистических компонентов интеллектуальных и информационных электронных систем	лингвистически ориентированных программных продуктов, электронных ресурсов, лингвистически ориентированных систем и лингвистических компонентов интеллектуальных и информационных электронных систем	лингвистически ориентированных программных продуктов, электронных ресурсов, лингвистически ориентированных систем и лингвистических компонентов интеллектуальных и информационных электронных систем
		ПК-12 Умеет квалифицированно тестировать лингвистически ориентированные программные продукты, электронные ресурсы, лингвистически ориентированные системы и лингвистические компоненты интеллектуальных и информационных электронных систем.	*Умение* квалифицированно тестировать лингвистически ориентированные программные продукты, электронные ресурсы, лингвистически ориентированные системы и лингвистические компоненты интеллектуальных и информационных электронных систем.
		ПК-12 Владеет навыками, необходимыми для тестирования различных ресурсов и продуктов электронной среды, в том числе лингвистически ориентированных программных продуктов, электронных ресурсов, лингвистически ориентированных систем и лингвистических компонентов интеллектуальных и информационных электронных систем	*Владение* навыками, необходимыми для тестирования различных ресурсов и продуктов электронной среды, в том числе лингвистически ориентированных программных продуктов, электронных ресурсов, лингвистически ориентированных систем и лингвистических компонентов интеллектуальных и

			информационных электронных систем.

2. Цель и место дисциплины в структуре образовательной программы

Дисциплина «*Лингвистические компоненты информационных электронных систем*» относится к обязательной части.

Дисциплина изучается на __2_ *курсе(ах)* в __3____ семестр*е(ах).*

Целью обучения программированию для лингвистов как виду профессиональной деятельности являются знакомство с инструментарием языка Python, базовых понятий, структур данных и алгоритмов, необходимых для изучения прикладных методов квантитативной лингвистики, а также для самостоятельного решения некоторых задач компьютерной лингвистики. Для освоения дисциплины необходимы компетенции, сформированные в рамках изучения следующих дисциплин «Компьютерные методы обработки языковой информации», «Иностранный язык», «Информационные технологии в лингвистике», «Лингвистические компоненты информационных электронных систем» и пр.

3. Содержание рабочей программы (объем дисциплины, типы и виды учебных занятий, учебно-методическое обеспечение самостоятельной работы обучающихся)

Содержание рабочей программы представлено в Приложении № 1.

4. Фонд оценочных средств по дисциплине

4.1. Перечень компетенций и индикаторов достижения компетенций с указанием соотнесенных с ними запланированных результатов обучения по дисциплине. Описание критериев и шкал оценивания результатов обучения по дисциплине.

Код и формулировка компетенции: ПК-9 Способен пользоваться лингвистически ориентированными программными продуктами;

Код и наименование индикатора достижения компетенции	**Результаты обучения по дисциплине**	**Критерии оценивания результатов обучения**	
		«Зачтено»	**«Не зачтено»**
ПК-9.1. Знает: типы, характеристики и особенности основных доступных в Интернете лингвистических ресурсов.	Знать: типы, характеристики и особенности основных доступных в Интернете лингвистических ресурсов	Обучающийся знает типы, характеристики и особенности основных доступных в Интернете лингвистических ресурсов;	Обучающийся не знает типы, характеристики и особенности основных доступных в Интернете лингвистических ресурсов;
ПК-9.2. Умеет: сравнивать	*Уметь:* сравнивать данные, полученные	Обучающийся умеет сравнивать данные,	Обучающийся не умеет сравнивать данные,

данные, полученные с использованием различных электронных лингвистических ресурсов и систем; применять методы математического анализа и моделирования в профессиональной деятельности.	с использованием различных электронных лингвистических ресурсов и систем; применять методы математического анализа и моделирования в профессиональной деятельности	полученные с использованием различных электронных лингвистических ресурсов и систем; применять методы математического анализа и моделирования в профессиональной деятельности	полученные с использованием различных электронных лингвистических ресурсов и систем; применять методы математического анализа и моделирования в профессиональной деятельности
ПК-9.3. Имеет практический опыт тестирования электронных лингвистических ресурсов, систем и компонентов	*Владеть:* практическим опытом тестирования электронных лингвистических ресурсов, систем и компонентов	Обучающийся владеет практическим опытом тестирования электронных лингвистических ресурсов, систем и компонентов	Обучающийся не владеет практическим опытом тестирования электронных лингвистических ресурсов, систем и компонентов

Код и формулировка компетенции: *ПК-12* Способен проводить квалифицированное тестирование лингвистически ориентированных программных продуктов, электронных ресурсов, лингвистически ориентированных систем и лингвистических компонентов интеллектуальных и информационных электронных систем

Код и наименование индикатора достижения компетенции	**Результаты обучения по дисциплине**	**Критерии оценивания результатов обучения**	
		«Зачтено»	**«Не зачтено»**
ПК-12.1 Знает: различные способы и инструментарий тестирования лингвистически ориентированных программных продуктов, электронных ресурсов, лингвистически ориентированных систем и лингвистических компонентов интеллектуальных и информационных электронных систем	Знать: различные способы и инструментарий тестирования лингвистически ориентированных программных продуктов, электронных ресурсов, лингвистически ориентированных систем и лингвистических компонентов интеллектуальных и информационных электронных систем	Обучающийся знает различные способы и инструментарий тестирования лингвистически ориентированных программных продуктов, электронных ресурсов, лингвистически ориентированных систем и лингвистических компонентов интеллектуальных и информационных электронных систем	Обучающийся не знает способы и инструментарий тестирования лингвистически ориентированных программных продуктов, электронных ресурсов, лингвистически ориентированных систем и лингвистических компонентов интеллектуальных и информационных электронных систем

ПК-12.2 Умеет: квалифицированно тестировать лингвистически ориентированные программные продукты, электронные ресурсы, лингвистически ориентированные системы и лингвистические компоненты интеллектуальных и информационных электронных систем	*Уметь:* квалифицированно тестировать лингвистически ориентированные программные продукты, электронные ресурсы, лингвистически ориентированные системы и лингвистические компоненты интеллектуальных и информационных электронных систем	Обучающийся умеет квалифицированно тестировать лингвистически ориентированные программные продукты, электронные ресурсы, лингвистически ориентированные системы и лингвистические компоненты интеллектуальных и информационных электронных систем	Обучающийся не умеет квалифицированно тестировать лингвистически ориентированные программные продукты, электронные ресурсы, лингвистически ориентированные системы и лингвистические компоненты интеллектуальных и информационных электронных систем
ПК-12.3 Владеет: навыками, необходимыми для тестирования различных ресурсов и продуктов электронной среды, в том числе лингвистически ориентированных программных продуктов, электронных ресурсов, лингвистически ориентированных систем и лингвистических компонентов интеллектуальных и информационных электронных систем	*Владеть:* навыками, необходимыми для тестирования различных ресурсов и продуктов электронной среды, в том числе лингвистически ориентированных программных продуктов, электронных ресурсов, лингвистически ориентированных систем и лингвистических компонентов интеллектуальных и информационных электронных систем	Обучающийся владеет навыками, необходимыми для тестирования различных ресурсов и продуктов электронной среды, в том числе лингвистически ориентированных программных продуктов, электронных ресурсов, лингвистически ориентированных систем и лингвистических компонентов интеллектуальных и информационных электронных систем	Обучающийся не владеет навыками, необходимыми для тестирования различных ресурсов и продуктов электронной среды, в том числе лингвистически ориентированных программных продуктов, электронных ресурсов, лингвистически ориентированных систем и лингвистических компонентов интеллектуальных и информационных электронных систем

4.2. Типовые контрольные задания или иные материалы, необходимые для оценивания результатов обучения по дисциплине, соотнесенных с установленными в образовательной программе индикаторами достижения компетенций. Методические материалы, определяющие процедуры оценивания результатов обучения по дисциплине.

Код и наименование индикатора достижения компетенции	Результаты обучения по дисциплине	Оценочные средства
ПК-9.1. Знает: типы, характеристики и особенности основных доступных в Интернете лингвистических ресурсов.	*Знать*: типы, характеристики и особенности основных доступных в Интернете лингвистических ресурсов.	групповой опрос, индивидуальное задание, контрольная работа
ПК-9.2. Умеет: сравнивать данные, полученные с использованием различных электронных лингвистических ресурсов и систем; применять методы математического анализа и моделирования в профессиональной деятельности.	*Уметь* сравнивать данные, полученные с использованием различных электронных лингвистических ресурсов и систем; применять методы математического анализа и моделирования в профессиональной деятельности.	индивидуальное задание контрольная работа
ПК-9.3. Имеет практический опыт тестирования электронных лингвистических ресурсов, систем и компонентов	*Владеть* практическим опытом тестирования электронных лингвистических ресурсов, систем и компонентов.	индивидуальное задание контрольная работа
ПК-12.1 Знает: различные способы и инструментарий тестирования лингвистически ориентированных программных продуктов, электронных ресурсов, лингвистически ориентированных систем и лингвистических компонентов интеллектуальных и информационных электронных систем	*Знать* различные способы и инструментарий тестирования лингвистически ориентированных программных продуктов, электронных ресурсов, лингвистически ориентированных систем и лингвистических компонентов интеллектуальных и информационных электронных систем.	групповой опрос, индивидуальное задание.
ПК-12.2 Умеет: квалифицированно тестировать лингвистически	*Уметь* квалифицированно тестировать лингвистически ориентированные программные продукты,	Групповой опрос, индивидуальное задание, контрольная работа

ориентированные программные продукты, электронные ресурсы, лингвистически ориентированные системы и лингвистические компоненты интеллектуальных и информационных электронных систем	электронные ресурсы, лингвистически ориентированные системы и лингвистические компоненты интеллектуальных и информационных электронных систем.	
ПК-12.3 Владеет: навыками, необходимыми для тестирования различных ресурсов и продуктов электронной среды, в том числе лингвистически ориентированных программных продуктов, электронных ресурсов, лингвистически ориентированных систем и лингвистических компонентов интеллектуальных и информационных электронных систем	*Владеть* навыками, необходимыми для тестирования различных ресурсов и продуктов электронной среды, в том числе лингвистически ориентированных программных продуктов, электронных ресурсов, лингвистически ориентированных систем и лингвистических компонентов интеллектуальных и информационных электронных систем.	Групповой опрос, индивидуальное задание, контрольная работа

Критериями оценивания при модульно-рейтинговой системе являются баллы, которые выставляются преподавателем за виды деятельности (оценочные средства) по итогам изучения модулей (разделов дисциплины), перечисленных в рейтинг-плане дисциплины (текущий контроль – максимум 50 баллов; рубежный контроль – максимум 50 баллов, поощрительные баллы – максимум 10). Шкалы оценивания: зачтено – от 60 до 110 рейтинговых баллов (включая 10 поощрительных баллов), не зачтено – от 0 до 59 рейтинговых баллов.

Рейтинг – план дисциплины *(при необходимости)*

_____Лингвистические компоненты информационных электронных систем_____
(название дисциплины согласно рабочему учебному плану)
направление/специальность_____45.03.03 Фундаментальная и прикладная лингвистика __
курс _____2_____, семестр____3____

Виды учебной деятельности студентов	Балл за конкретное задание	Число заданий за семестр	Баллы	
			Минимальный	Максимальный
Модуль 1 Regular Expressions, Text Normalization, Edit Distance				
Текущий контроль				
1. Аудиторная работа (групповой опрос)	10	1	0	10
2. Индивидуальное задание	10	1	0	10
Рубежный контроль				
1. Контрольная работа	10	1	0	10
Итого				30
Модуль 2 N-gram Language Models. .				
Текущий контроль				
1. Аудиторная работа (групповой опрос)	10	1	0	10
2. Индивидуальное задание	10	1	0	10
Рубежный контроль				
1. Индивидуальное задание	20	2	0	20
				40
Модуль 3 Naive Bayes and Sentiment Classification.				
Текущий контроль				
1. Аудиторная работа (групповой опрос)	**10**	**1**	**0**	**10**
2. Индивидуальное задание	**10**	**1**	**0**	**10**
1. Аудиторная работа (групповой опрос)				
Рубежный контроль				
1. Контрольная работа	**10**	**2**	**0**	**10**
				30
Поощрительные баллы				
1. Участие в научных конференциях и фестивалях, культурных мероприятиях, публикация статей				**3**
2. Публикация статей				**5**
3. Работа со школьниками (кружок, конкурсы, олимпиады)				**2**
Посещаемость (баллы вычитаются из общей суммы набранных баллов)				
1. Посещение практических занятий			**0**	**–6**
Всего				**110**
1. **Итоговый контроль** Зачет				-

1. Задания для группового опроса.

По модулю 1 Regular Expressions, Text Normalization, Edit Distance

1. Write regular expressions for the following languages.
a. the set of all alphabetic strings;
b. the set of all lower case alphabetic strings ending in a b;
c. the set of all strings from the alphabet a,b such that each a is immediately preceded by and immediately followed by a b;.
2. Compute the edit distance (using insertion cost 1, deletion cost 1, substitution cost 1) of "leda" to "deal". Show your work (using the edit distance grid).
3. Figure out whether drive is closer to brief or to divers and what the edit distance is to each. You may use any version of distance that you like.
4. Now implement a minimum edit distance algorithm and use your hand-computed results to check your code.
5. Write regular expressions for the following languages. By "word", we mean an alphabetic string separated from other words by whitespace, any relevant punctuation, line breaks, and so forth.
a. the set of all strings with two consecutive repeated words (e.g., "Humbert Humbert" and "the the" but not "the bug" or "the big bug");
b. all strings that start at the beginning of the line with an integer and that end at the end of the line with a word;
c. all strings that have both the word grotto and the word raven in them (but not, e.g., words like grottos that merely contain the word grotto);
d. write a pattern that places the first word of an English sentence in a register. Deal with punctuation.

По модулю 2 N-gram Language Models.

1. Write out the equation for trigram probability estimation. Now write out all the non-zero trigram probabilities for the given test corpus.
2 Calculate the probability of the sentence i want chinese food. Assume the additional add-1 smoothed probabilities P(i|<s>) = 0.19 and P(</s>|food) = 0.40.
3 Which of the two probabilities you computed in the previous exercise is higher, unsmoothed or smoothed? Explain why.
4 We are given the following corpus, modified from the one in the chapter:
<s> I am Sam </s>
<s> Sam I am </s>
<s> I am Sam </s>
<s> I do not like green eggs and Sam </s>
Using a bigram language model with add-one smoothing, what is P(Sam |am)? Include <s> and </s> in your counts just like any other token.
5 Suppose we didn't use the end-symbol </s>. Train an unsmoothed bigram grammar on the following training corpus without using the end-symbol </s>:
<s> a b
<s> b b
<s> b a
<s> a a
Demonstrate that your bigram model does not assign a single probability distribution across all sentence lengths by showing that the sum of the probability
of the four possible 2 word sentences over the alphabet {a,b} is 1.0, and the
sum of the probability of all possible 3 word sentences over the alphabet {a,b}

is also 1.0.
6 Suppose we train a trigram language model with add-one smoothing on a given corpus. The corpus contains V word types. Express a formula for estimating P(w3|w1,w2), where w3 is a word which follows the bigram (w1,w2), in terms of various N-gram counts and V. Use the notation c(w1,w2,w3) to denote the number of times that trigram (w1,w2,w3) occurs in the corpus, and so on for bigrams and unigrams.
7 We are given the following corpus, modified from the one in the chapter:
<s> I am Sam </s>
<s> Sam I am </s>
<s> I am Sam </s>
<s> I do not like green eggs and Sam </s>
If we use linear interpolation smoothing between a maximum-likelihood bigram model and a maximum-likelihood unigram model with $\lambda 1 = \frac{1}{2}$ and $\lambda 2 = 1/2$, what is P(Sam|am)? Include <s> and </s> in your counts just like any other token.
8 Write a program to compute unsmoothed unigrams and bigrams.
9 Run your n-gram program on two different small corpora of your choice (you might use email text or newsgroups). Now compare the statistics of the two corpora. What are the differences in the most common unigrams between the two? How about interesting differences in bigrams?
10 Add an option to your program to generate random sentences.
11 Add an option to your program to compute the perplexity of a test set.
12 You are given a training set of 100 numbers that consists of 91 zeros and 1 each of the other digits 1-9. Now we see the following test set: 0 0 0 0 0 3 0 0 0 0. What is the unigram perplexity?

По модулю 3 Naive Bayes and Sentiment Classification.

1. Given the following short movie reviews, each labeled with a genre, either comedy or action:
a. fun, couple, love, love comedy
b. fast, furious, shoot action
c. couple, fly, fast, fun, fun comedy
d. furious, shoot, shoot, fun action
e. fly, fast, shoot, love action
and a new document D:
fast, couple, shoot, fly
compute the most likely class for D. Assume a naive Bayes classifier and use add-1 smoothing for the likelihoods.
2. Train two models, multinomial naive Bayes and binarized naive Bayes, both with add-1 smoothing, on the following document counts for key sentiment words, with positive or negative class assigned as noted.
doc "good" "poor" "great" (class)
d1. 3 0 3 pos
d2. 0 1 2 pos
d3. 1 3 0 neg
d4. 1 5 2 neg
d5. 0 2 0 neg
Use both naive Bayes models to assign a class (pos or neg) to this sentence:
A good, good plot and great characters, but poor acting.
Recall from page 60 that with naive Bayes text classification, we simply ignore (throw out) any word that never occurred in the training document. (We don't throw out words that appear in some classes but not others; that's what add-one smoothing is for.) Do the two models agree or disagree?
3. Assume the following likelihoods for each word being part of a positive or

negative movie review, and equal prior probabilities for each class.

	pos	neg
I	0.09	0.16
always	0.07	0.06
like	0.29	0.06
foreign	0.04	0.15
films	0.08	0.11

What class will Naive bayes assign to the sentence “I always like foreign films.”?

Критерии оценки (в баллах) результатов для группового опроса:

5 балла выставляется студенту, показавшему умение применять знания теории межкультурной коммуникации на практике, свободно и аргументировано обосновывать решение конкретных задач;

3 балл выставляется студенту, показавшему не достаточно глубокое знание теории межкультурной коммуникации, не умеющему в полной мере свободно и аргументировано обосновать решение конкретных задач;

0 баллов выставляется студенту, который не понимает поставленной задачи и не способенее верно решить.

Индивидуальные задания

Верны ли следующие высказывания. Почему:

• Many language processing tasks can be viewed as tasks of classification.
• Text categorization, in which an entire text is assigned a class from a finite set, includes such tasks as sentiment analysis, spam detection, language identification, and authorship attribution.
• Sentiment analysis classifies a text as reflecting the positive or negative orientation (sentiment) that a writer expresses toward some object.
• Naive Bayes is a generative model that makes the bag of words assumption (position doesn’t matter) and the conditional independence assumption (words are conditionally independent of each other given the class)
• Naive Bayes with binarized features seems to work better for many text classification tasks.
• Classifiers are evaluated based on precision and recall.
• Classifiers are trained using distinct training, dev, and test sets, including the use of cross-validation in the training set.
• Statistical significance tests should be used to determine whether we can be confident that one version of a classifier is better than another.
• Designers of classifiers should carefully consider harms that may be caused by the model, including its training data and other components, and report model characteristics in a model card.

Индивидуальные задания выполняются в письменной форме и являются частью текущего контроля. Оцениваются в соответствии с рейтиг-планом дисциплины.

Контрольная работа по Модулю 1.

1. Define a string s = 'colorless'. Write a Python statement that changes this to “colourless” using only the slice and concatenation operations.
2. We can use the slice notation to remove morphological endings on words. For example, dogs'[:-1] removes the last character of dogs, leaving dog. Use slice
notation to remove the affixes from these words (we’ve inserted a hyphen to indicate the affix

boundary, but omit this from your strings): dish-es, run-ning, nationality, un-do, pre-heat.
3. We saw how we can generate an IndexError by indexing beyond the end of a string. Is it possible to construct an index that goes too far to the left, before the start of the string?
4. We can specify a "step" size for the slice. The following returns every second character within the slice: monty[6:11:2]. It also works in the reverse direction: monty[10:5:-2]. Try these for yourself, and then experiment with different step values.
5. What happens if you ask the interpreter to evaluate monty[::-1]? Explain why this is a reasonable result.
6. Describe the class of strings matched by the following regular expressions:
a. [a-zA-Z]+
b. [A-Z][a-z]*
c. p[aeiou]{,2}t
d. \d+(\.\d+)?
e. ([^aeiou][aeiou][^aeiou])*
f. \w+|[^\w\s]+
Test your answers using nltk.re_show().
7. Write regular expressions to match the following classes of strings:
a. A single determiner (assume that a, an, and the are the only determiners)
b. An arithmetic expression using integers, addition, and multiplication, such as 2*3+8
8. ○ Write a utility function that takes a URL as its argument, and returns the contents of the URL, with all HTML markup removed. Use urllib.urlopen to access the contents of the URL, e.g.:
raw_contents = urllib.urlopen('http://www.nltk.org/').read()

Контрольная работа по Модулю 3.

1. Save some text into a file corpus.txt. Define a function load(f) that reads from the file named in its sole argument, and returns a string containing the text of the file.
a. Use nltk.regexp_tokenize() to create a tokenizer that tokenizes the various kinds of punctuation in this text. Use one multiline regular expression inline comments, using the verbose flag (?x).
b. Use nltk.regexp_tokenize() to create a tokenizer that tokenizes the following kinds of expressions: monetary amounts; dates; names of people and organizations.
2. ○ Rewrite the following loop as a list comprehension:
```
>>> sent = ['The', 'dog', 'gave', 'John', 'the', 'newspaper']
>>> result = []
>>> for word in sent:
... word_len = (word, len(word))
... result.append(word_len)
>>> result
[('The', 3), ('dog', 3), ('gave', 4), ('John', 4), ('the', 3), ('newspaper', 9)]
```
3. Define a string raw containing a sentence of your own choosing. Now, split raw on some character other than space, such as 's'.
4. Write a for loop to print out the characters of a string, one per line.
5. What is the difference between calling split on a string with no argument and one with ' ' as the argument, e.g., sent.split() versus sent.split(' ')? What happens when the string being split contains tab characters, consecutive space characters, or a sequence of tabs and spaces? (In IDLE you will need to use '\t' to enter a tab character.)
6. Create a variable words containing a list of words. Experiment with words.sort() and sorted(words). What is the difference?
7. Explore the difference between strings and integers by typing the following at a Python prompt: "3" * 7 and 3 * 7. Try converting between strings and integers using int("3") and str(3).
8. Earlier, we asked you to use a text editor to create a file called test.py, containing the single

line monty = 'Monty Python'. If you haven't already done this (or can't find the file), go ahead and do it now. Next, start up a new session with the Python

Учебно-методическое и информационное обеспечение дисциплины

5.1. Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины

Основная литература:

1. Елагина, Ю.С. Практикум по устному переводу: учебное пособие / Ю.С. Елагина; Министерство образования и науки Российской Федерации, Оренбургский Государственный Университет. - Оренбург: ОГУ, 2017. - 107 с. - Библиогр.: с. 95-98 - ISBN 978-5-7410-1648-0; То же [Электронный ресурс]. - URL: http://biblioclub.ru/index.php?page=book&id=481754 .
2. Терехова, Е.В. Двусторонний перевод общественно-политических текстов (с элементами скорописи в английском языке): учебное пособие / Е.В. Терехова. - 3-е изд., стер. - Москва: Издательство «Флинта», 2017. - 320 с. - Библиогр. в кн. - ISBN 978-5-89349-955-1; То же [Электронный ресурс]. - URL: http://biblioclub.ru/index.php?page=book&id=115136

Дополнительная литература:

1. Захарова, Т.В. Практические основы компьютерных технологий в переводе : учебное пособие /Т.В. Захарова, Е.В. Турлова ; Министерство образования и науки Российской Федерации, Оренбургский Государственный Университет. - Оренбург : Оренбургский государственный университет, 2017. - 109 с. : табл., граф., ил. - Библиогр.: с. 104. - ISBN 978-5-7410-1736-4; То же [Электронный ресурс]. - URL: http://biblioclub.ru/index.php?page=book&id=481823
2. Михеева, Е.В. Информационные технологии в профессиональной деятельности : учебное пособие / Е.В. Михеева. - Москва: Проспект, 2014. - 448 с. - Библиогр. в кн. - ISBN 978-5-392- 12318-6; То же [Электронный ресурс]. - URL: http://biblioclub.ru/index.php?page=book&id=251602).
3. Бовтенко, М.А. Язык пользователя персонального компьютера: учебное пособие / М.А. Бовтенко, Е.В. Кугаевская. – Новосибирск: НГТУ, 2011. – Ч. 2. – 75 с. – ISBN 978-5-7782-1873-4 – ЭВИ в ЭБС «Универс. библ. онлайн» – URL: http://biblioclub.ru/index.php?page=book&id=228749
4. Бойченко, Г.Н. Информационные сервисы Интернет в профессиональной деятельности педагога: учебное пособие / Г.Н. Бойченко. – Новокузнецк: Кузбасская государственная педагогическая академия, 2008. – 106 с. – ISBN 978-5-85117-320-2 – ЭВИ в ЭБС «Универс. библ. онлайн» –URL: http://biblioclub.ru/index.php?page=book&id=88672
5. Гафурова, Н.В. Методика обучения информационным технологиям. Теоретические основы: учебное пособие / Н.В. Гафурова, Е.Ю. Чурилова. – Красноярск: СибФУ, 2012. – 111 с. – ISBN 978-5-7638-2234-2 – ЭВИ в ЭБС «Универс. библ. онлайн» – URL: http://biblioclub.ru/index.php?page=book&id=229302 .
6. Гунина, Н.А. Компьютер для работы и досуга = ComputerforWorkandLeisure: учебное пособие / Н.А. Гунина, И.В. Шеленкова, А.А. Шиповская. – Тамбов: Издательство

ФГБОУ ВПО «ТГТУ», 2012. – 137 с. – ЭВИ в ЭБС «Универс. библ. онлайн» – URL:http://biblioclub.ru/index.php?page=book&id=277672

5.2. Перечень ресурсов информационно-телекоммуникационной сети «Интернет» и программного обеспечения, необходимых для освоения дисциплины

1. Научно-образовательный портал «Лингвистика в России: ресурсы для исследователей»:http://uisrussia.msu.ru/linguist/_B7_komp_tehn_v_prepodavanii_jazykov.jsp
2. Универсальная научно-популярная онлайн-энциклопедия. http://www.krugosvet.ru/enc/gumanitarnye_nauki/lingvistika/GERMANSKIE_YAZIKI.html
3. Свободная энциклопедия Википедия: http://ru.wikipedia.org/wiki/Германские_языки
4. Библиотека Русского гуманитарного интернет-университета: http://www.i-u.ru/biblio
5. Лингвоинфо: интернет-журнал: http://www.lingvoinfo.com
6. Онлайн-энциклопедия http://www.krugosvet.ru
7. Русский филологический портал Philology.ru.: http://philology.ru/linguistics1.html1
8. Центр лингвистической документации: http://www.mccme.ru/ling/ index.htm
9. Windows 8 Russian. Windows Professional 8 Russian Upgrade. Договор № 104 от 17.06.2013 г. Лицензии бессрочные
10. Microsoft Office Standard 2013 Russian. Договор № 114 от 12.11.2014 г. Лицензии бессрочные

6.Материально-техническая база, необходимая для осуществления образовательного процесса по дисциплине

Наименование специализированных аудиторий, кабинетов, лабораторий	***Вид занятий***	***Наименование оборудования, программного обеспечения***
1	***2***	***3***
Учебная аудитория для проведения занятий семинарского типа: аудитория №31(мультимедийный класс), (ул. Коммунистическая, д. 19, лит. А, А1)	Практические занятия	**Аудитория №31** Учебная мебель, учебно-наглядные пособия, доска, моноблоки − 13 шт. с выходом в Интернет, обеспечивающие доступ к электронной информационно-образовательной среде (ЭИОС) вуза, мультимедиа-проектор MitsubishiEX320U XGA, экран настенный Classic Norma 244*183, ноутбук ASUSX51RL (место хранения деканат ФРГФ, ауд.№ 6а) **Аудитория №4** Учебная мебель, учебно-наглядные пособия, доска, моноблоки − 12 шт. с выходом в Интернет, обеспечивающие доступ к электронной информационно-образовательной среде (ЭИОС) вуза **Аудитория №11** Учебная мебель, учебно-наглядные пособия, доска
Аудитория для текущего контроля и промежуточной аттестации: Аудитория №4, Аудитория № 11 (ул. Коммунистическая, д. 19, лит. А, А1)	Групповые и индивидуальные консультации, текущий контроль и промежуточная аттестация	

Помещения для самостоятельной работы: аудитория №13 (читальный зал) (ул. Коммунистическая, д. 19, лит. А, А1)	Самостоятельная работа	**Аудитория №13** Учебная мебель, учебно-наглядные пособия, доска, учебно-методическая литература, многофункциональное устройство – 1 шт., моноблоки − 2 шт. с выходом в Интернет, обеспечивающие доступ к электронной информационно-образовательной среде (ЭИОС) вуза, книжный фонд читального зала ФРГФ Windows 8 Russian. Windows Professional 8 Russian Upgrade. Договор № 104 от 17.06.2013 г. Лицензии бессрочные Microsoft Office Standard 2013 Russian. Договор № 114 от 12.11.2014 г. Лицензии бессрочные.

Приложение № 1

ФГБОУ ВО «БАШКИРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ»

ФАКУЛЬТЕТ РОМАНО-ГЕРМАНСКОЙ ФИЛОЛОГИИ

СОДЕРЖАНИЕ РАБОЧЕЙ ПРОГРАММЫ[2]

дисциплины ____Лингвистические компоненты информационных электронных систем___ на ___3___ семестр
(наименование дисциплины)
____очная_____
форма обучения

Вид работы	Объем дисциплины
Общая трудоемкость дисциплины (з.е. / часов)	2/72
Учебных часов на контактную работу с преподавателем:	18,2
лекций	-
практических/ семинарских	18
лабораторных	-
других (групповая, индивидуальная консультация и иные виды учебной деятельности, предусматривающие работу обучающихся с преподавателем) (ФКР)	0,2
из них, предусмотренные на выполнение курсовой работы / курсового проекта[3]	
Учебных часов на самостоятельную работу обучающихся (СР)	53,8
из них, предусмотренные на выполнение курсовой работы / курсового проекта[4]	-
Учебных часов на подготовку к экзамену/зачету/дифференцированному зачету (Контроль)	-

Форма(ы) контроля:
экзамен____-____ семестр
зачет_____1_____ семестр
курсовая работа / курсовой проект __-__ семестр

[2] *Количество часов/з.е. указывается в соответствии с учебным планом, таблицы заполняются отдельно по каждой форме обучения (очной, очно-заочной, заочной).*
[3] *Контактных часов – 2*
[4] *Количество часов на самостоятельную работу указывается на усмотрение разработчика, но* ***не более 20 часов***

№ п/п	Тема и содержание	Форма изучения материалов: лекции, практические занятия, семинарские занятия, лабораторные работы, самостоятельная работа и трудоемкость (в часах)				Задания по самостоятельной работе студентов	Форма текущего контроля успеваемости (коллоквиумы, контрольные работы, компьютерные тесты и т.п.)
		ЛК	ПР/СЕМ	ЛР	СР		
1	2	3	4	5	6	7	8
1.	Regular Expressions, Text Normalization, Minimum Edit Distance of Levenshtein, Damerau-Levenshtein. Spelling Correction. Real word-errors. Non-word errors. Basics of Formal Grammars.	-	6	-	18	Основная литература 1	Групповой опрос, индивидуальные задания, контрольная работа
2.	N-gram Language Models. Markov's Model. Chain Rule. Good-Turing Smoothing. Laplace Smoothing. Perplexity.	-	6	-	18	Основная литература 1	Групповой опрос, индивидуальные задания
3.	Naïve Bayes and Sentiment Classification/ Precision. Recall. F-measures. Statistical Significance Testing	-	6	-	18.8	Основная литература 1, 2	Групповой опрос, индивидуальные задания, контрольная работа
	Всего часов:		18		53.8		