Федеральное государственное бюджетное учреждение науки
Институт систем информатики им. А.П. Ершова
Сибирского отделения Российской академии наук
(ИСИ СО РАН)

УТВЕРЖДАЮ
Директор ИСИ СО РАН

_____
« 1 » сентября 2015 г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

«Математическая лингвистика и обработка текстов на естественном языке»

Направление подготовки: 09.06.01 «Информатика и вычислительная техника»

Специальность: 05.13.11 «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей»

Уровень образования: подготовка кадров высшей квалификации

Квалификация выпускника: Исследователь. Преподаватель-исследователь

Составители рабочей программы

Н.с. лаб.моделирования сложных систем, к.ф.-м.н. _____ Батура Т.В.
(должность, ученое звание, ученая степень) (подпись) (ФИО)

Рабочая программа утверждена на заседании Ученого совета Института
«07» июля 2015 г., протокол № 5-2015

Председатель Ученого совета _____ Марчук А.Г.
(подпись) (ФИО)

СОГЛАСОВАНО:

Зам. директора по науке
к.ф.-м.н. _____ Мурзин Ф.А.
(подпись) (ФИО)

Зав. аспирантурой _____ Воронко Н.Ф.
(подпись) (ФИО)

1. Цели освоения дисциплины

Целями освоения дисциплины «Математическая лингвистика и обработка текстов на естественном языке» являются ознакомление с основными методами математической логики, теории вероятностей и математической статистики, элементов автоматизированной обработки текстов: формальных методов анализа текстов, алгоритмов семантического поиска и извлечения информации, элементов теории речевых действий, особенностей построения тезаурусов, знакомство с основами корпусной лингвистики. Также уделяется внимание практическому применению принципов обработки текстовой информации в поисковых системах и системах автоматического определения авторства, при анализе социальных сетей и спам-сообщений, необходимых для самостоятельной работы в научно-исследовательской сфере.

Для достижения поставленных целей выделяются следующие задачи:

- изучение математических основ наиболее интересных и важных для приложений алгоритмов из теории информации, обработки текстов на естественном языке.
- ознакомление с нестандартными методами обработки информации: нейрокомпьютерный подход, методы кластеризации, нечеткая логика Заде.
- ознакомление с методами обработки текстовой информации: алгоритмами морфологического и синтаксического анализа, методами классификации и кластеризации, алгоритмами поиска ключевых слов и др.

2. Место дисциплины в структуре основной профессиональной образовательной программы послевузовского профессионального образования (аспирантура)

Данная дисциплина «Математическая лингвистика и обработка текстов на естественном языке» (Б1.В.ДВ.4) относится к группе дисциплин по выбору аспиранта вариативной части по специальности 05.13.11.

3. Требования к уровню подготовки аспиранта, завершившего изучение данной дисциплины

Аспиранты, завершившие изучение данной дисциплины, должны:

- **знать:** содержание программы курса, принципы функционирования автоматизированных лингвистических систем, формулировки задач, условия применимости и характеристики рассмотренных в курсе методов;
- **уметь:** применять методы математической лингвистики для анализа текстовой информации на естественном языке;
- **владеть:** методами математической лингвистики для анализа текстовой информации на естественном языке.

Компетенции, формируемые у обучающихся, в соответствии с ООП по направлению 09.06.01 «Информатика и вычислительная техника» и профилю (специальности) 05.13.11 «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей»:

Универсальные компетенции: УК1, УК3, УК5.
Общепрофессиональные компетенции: ОПК1, ОПК2, ОПК3, ОПК6

Профессиональные компетенции: ПК1, ПК2, ПК5

4. Объем дисциплины и виды учебной работы

Общая трудоемкость дисциплины составляет __2___ зачетных единицы ____72___ часов.

№ п/п	Наименование тем и разделов	ВСЕГО (часов)	Аудиторные занятия (часов), в том числе			Самостоятельная работа (часов)
			Лекции	Семинары	Лаб. работы	
1	Общие принципы построения систем автоматизированной обработки текстов.	4	2			2
2	Синтаксическая структура предложения. Методы синтаксического анализа.	8	3			5
3	Принципы построения синтаксических анализаторов. Взаимодействие синтаксического и фрагментационного анализа.	4	1			3
4	Семантический анализ текстов. Лексические функции. Валентности слов.	4	1			3
5	Теоретико-множественные модели языка.	4	2			2
6	Теория речевых действий. Классификация речевых действий.	4	1			3
7	Представление знаний для компьютерной обработки. Тезаурусы и онтологии. Общие принципы построения.	4	2			2
8	Семантические сети. Фреймы. Формальные логические модели.	6	2			4
9	Корпусная лингвистика. Частотные методы в компьютерной лингвистике.	4	1			3

10	Классификация и кластеризация. Иерархические и вероятностные подходы.	6	3			3
11	Автоматические системы извлечения информации. Алгоритмические основы.	4	2			2
12	Формальные методы атрибуции текстов.	8	2			6
13	Анализ социальных сетей. Направления и методы исследований.	8	2			6
14	Методы обнаружения спама.	4	2			2
	ИТОГО:	72	26			46

5. Разделы дисциплины и виды занятий

№ п/п	Название раздела дисциплины	Объем часов / зачетных единиц					
			из них				
			лекции	семинары	практ. занятия	КСР	
1	Раздел 1. Общие принципы и задачи построения систем автоматической обработки текстовой информации	13	13				
2	Раздел 2. Семантический анализ текстов	13	13				
	ИТОГО часов	26	26				46

6. Содержание дисциплины

(*Раздел, тема учебного курса, содержание лекции*)

6.1. Новизна курса (научная, содержательная; сравнительный анализ с подобными курсами в России и за рубежом)

Первый раздел служит основой и знакомит студентов с общими принципами и задачами построения систем автоматической обработки текстовой информации. Он подготавливает студентов к дальнейшему обсуждению возникающих проблем и их решению. Во втором разделе рассматриваются довольно сложные вопросы семантического анализа текстов, под-

ходы и методы, недостаточно и разрозненно освещенные в литературе. В третьем разделе рассматриваются способы практического применения алгоритмов обработки текстов в различных системах поиска и извлечения информации, системах определения авторства, при анализе текстового контента социальных сетей.

6.2. Содержание разделов и тем курса

1. Основные понятия математической лингвистики. Устройство систем автоматизированной обработки текстов. Основные этапы построения. Графематический и морфологический анализ.
2. Методы задания синтаксической структуры предложений. Системы составляющих. Деревья подчинения. Проблемы синтаксического анализа.
3. Принципы построения синтаксических анализаторов. Проект "Автоматическая Обработка Текста" (АОТ). Синтаксический парсер LINK. Применение морфологического и синтаксического анализа в поисковых системах.
4. Методы теоретического исследования семантики текстов. Лексические функции. Валентности слов. Теоретико-модельный подход.
5. Теоретико-множественные модели языка. Основные определения: отмеченные последовательности, контексты, дистрибутивные классы и др. Формализация понятий «часть речи» и «синтаксический тип». Формализация понятия «грамматический род». Формализация «категории падежа».
6. Представление знаний для компьютерной обработки. Тезаурусы и онтологии. WordNet. Общие принципы построения. Меры семантической близости.
7. Семантические сети. Фреймы. Формальные логические модели. Искусственные языки и нотации, применяемые в компьютерной лингвистике.
8. Корпусная лингвистика. Частотные методы в компьютерной лингвистике.
9. Теория речевых действий. Классификация речевых действий.
10. Модели и методы автоматической классификации и кластеризации текстовой информации. Иерархические и вероятностные подходы. Интеллектуальный анализ данных.
11. Автоматические системы извлечения информации. Алгоритмические основы. Принципы обработки неструктурированной и плохо структурированной информации. Тематическая индексация текстов.
12. Формальные методы определения автора текста. Лингвостатистические параметры. Статистические методы атрибуции. Авторский инвариант и лингвистические спектры. Применение методов кластеризации и классификации для установления авторства текстов.
13. Социальные сети. Направления исследований. Графовые модели анализа социальных сетей. Понятие центральности. Методы обнаружения сообществ и анализ связных подгрупп. Модели динамики сети.
14. Методы обнаружения спама: вероятностные и статистические, байесовский классификатор.

6.3. Перечень примерных вопросов и заданий для самостоятельной работы

1. Перечислить направления компьютерной лингвистики.
2. Сформулировать общие принципы построения автоматизированных систем обработки текстов.
3. Разъяснить принципы работы графематического и морфологического анализаторов.

4. Перечислить методы задания синтаксической структуры предложений.
5. Разъяснить принципы работы фрагментационного и синтаксического анализаторов. Описать принцип их взаимодействия.
6. Изложить основные идеи подхода И. Мельчука к семантическому анализу.
7. Привести примеры мер семантической близости.
8. Дать определения отмеченных последовательностей, контекста, дистрибутивных классов.
9. Дать формальные определения частей речи, грамматического рода и категории падежа в терминах модели языка, предложенной С. Маркусом.
10. Изложить основные идеи теории речевых действий.
11. Привести классификацию речевых действий.
12. Суть теоретико-модельного подхода к исследованию семантики текстов.
13. Сформулировать принципы построения тезаурусов и онтологий. Сходства и отличия.
14. Дать определения семантических сетей, фреймов.
15. Неточные рассуждения. Что такое логика Заде?
16. Привести примеры искусственных языков и нотаций.
17. Что такое корпусная лингвистика?
18. Применение частотных методов в компьютерной лингвистике. Перечислить, описать, привести примеры.
19. В чем отличие между классификацией и кластеризацией текстов?
20. Перечислить методы классификации и кластеризации текстовой информации. Сформулировать основные принципы.
21. Разъяснить принципы работы автоматических систем извлечения информации.
22. Сформулировать принципы обработки неструктурированной и плохо структурированной информации. Индексация текстов.
23. Перечислить формальные методы атрибуции текстов.
24. Дать определения лингвостатистических параметров, авторского инварианта и лингвистических спектров.
25. Привести примеры использования методов кластеризации и классификации для определения авторства текстов.
26. Перечислить основные направления исследований социальных сетей.
27. Дать определения центральностей разного типа.
28. Описать методы анализа социальных сетей.
29. Перечислить основные методы обнаружения спам-сообщений. Привести примеры.
30. Пояснить принцип работы байесовского классификатора.

6.4. Перечень примерных тем рефератов и докладов

1. Проблемы автоматизации синтаксического анализа предложений.
2. Проблемы обнаружения кореференций и анафор в текстах на ЕЯ.
3. Применение алгоритмов и методов обработки текстовой информации в технике.
4. Применение алгоритмов и методов обработки текстовой информации в медицине.
5. Применение алгоритмов и методов обработки текстовой информации в системах безопасности.
6. Возможности программных приложений для анализа социальных сетей.
7. Проблемы автоматической идентификации авторов текстов.
8. Методы автоматического построения онтологий.

9. Сравнение алгоритмов обнаружения и исправления ошибок и опечаток.
10. Сравнение алгоритмов морфологического анализа.

7. Самостоятельная работа аспирантов

7.1. Изучение основной и дополнительной литературы по вопросам программы.

7.2. Аспирантам может быть дано задание написать реферативный обзор или подготовить устный доклад.

8. Учебно-методическое и информационное обеспечение дисциплины

8.1. Основная и дополнительная литература

а) основная литература:

1. Батура Т.В. Методы анализа компьютерных социальных сетей // Вестник НГУ. Серия: Информационные технологии. Новосибирск, 2012. Том 10, Вып. 4. С. 13–28.
2. Батура Т.В. Методы определения авторского стиля текстов и их программная реализация // Программные системы и вычислительные методы. М.: НБ-Медиа, 2014. № 2. С. 197–216. DOI: 10.7256/2305-6061.2014.2.11705. http://www.nbpublish.com/library_read_article.php?id=-30093
3. Кобзарева Т. Ю. В поисках синтаксической структуры: автоматический анализ русского предложения с опорой на сегментацию. М.: РГГУ. 2015. 371 с.
4. Щипицина Л.Ю. Информационные технологии в лингвистике: Учебное пособие. М.: Флинта: Наука, 2013. 128 с.
5. S. Abramsky, M. Sadrzadeh. Semantic Unification // Lecture Notes in Computer Science, 2014, V. 8222, pp. 1-13. URL: http://arxiv.org/pdf/1403.3351v1.pdf
6. Чатуев М.Б., Чеповский А.М. Частотные методы в компьютерной лингвистике: Учебное пособие. – М.: МГУП, 2011. – 88с.
7. Маркус С. Теоретико-множественные модели языков. – М.: Наука, 1970. – 332 с.
8. Мельчук И.А. Опыт теории лингвистических моделей «Смысл-Текст» // М.: Школа «Языки русской культуры», 1999. – 346 с.
9. Charu C. Aggarwal Social network data analytics. 2011. 520 p.
10. ACL Anthology — A Digital Archive of Research Papers in Computational Linguistics http://anthology.aclweb.org/

б) дополнительная литература

1. R. Socher et al. Semantic Compositionality through Recursive Matrix-Vector Spaces. 2013. URL: http://nlp.stanford.edu/pubs/SocherHuvalManningNg_EMNLP2012.pdf
2. L.S. Moss, H.-J. Tiede, Applications of modal logic in linguistics, in: Handbook on Modal Logics, Elsevier, Amsterdam, 2007, pp. 299-341. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.88.1863&rep=rep1&type=pdf
3. D. Jurafsky, J. H. Martin. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. 2008.

1024 p. URL: http://www.deepsky.com/~merovech/voynich/voynich_manchu_reference_materials/PDFs/jurafsky_martin.pdf
4. The Stanford Natural Language Processing Group http://nlp.stanford.edu/
5. Апресян Ю. Д. Идеи и методы современной структурной лингвистики. М.: Просвещение, 1966. 305 с.
6. Ануреев И.С., Батура Т.В., Боровикова О.И., Загорулько Ю.А., Кононенко И.С., Марчук А.Г., Марчук П.А., Мурзин Ф.А., Сидорова Е.А., Шилов Н.В. Модели и методы построения информационных систем, основанных на формальных, логических и лингвистических подходах // Моногр. / Институт систем информатики им. А.П. Ершова СО РАН. – Новосибирск: Изд. СО РАН, 2009.
7. Batura Tatiana, Murzin Feodor, Proskuryakov Alexey, Trelevich Jennifer. Some Approaches to Detection of Spam and Senders of Spam // Восьмая междунар. конф. памяти акад. А.П. Ершова, "Перспективы систем информатики", Рабочий семинар "Наукоемкое программное обеспечение", Новосибирск 2011. – С. 1-6.
8. Заде Л. Понятие лингвистической переменной и его применение к принятию приближеных решений. – М., 1976. – 166 с.
9. Батура Т.В., Мурзин Ф.А. Машинно-ориентированные логические методы отображения семантики текста на естественном языке// Моногр. / Институт систем информатики им. А.П. Ершова СО РАН. – Новосибирск: Изд. НГТУ, ISBN 978-5-7782-1138-4, 2008. – 248с.
10. Труды международной конференции по компьютерной лингвистике "Диалог" http://www.dialog-21.ru/

8.2. Перечень вопросов и заданий (аттестации) и/или тем рефератов

Перечень вопросов и заданий, тем рефератов и докладов совпадает с расшифровкой к пунктам 6.3 и 6.4.

9. Материально-техническое обеспечение дисциплины

Для лекций используется класс, оснащённый мультимедийным проектором и имеющий в составе программное обеспечение MS Office и Acrobat Reader, MS Visual C++, Maple 5.4., Matlab 7.0, графические редакторы. Литература из основного и вспомогательного списков доступна в электронно-библиотечной системе ИСИ СО РАН и в Мемориальной библиотека А.П. Ершова (каб. 265). Для самостоятельной работы используются персональные компьютеры слушателей курса.

ДОПОЛНЕНИЯ И ИЗМЕНЕНИЯ В РАБОЧЕЙ ПРОГРАММЕ

за 2015 / 2016 учебный год

В рабочую программу Математическая лингвистика и обработка текстов на естественном языке

(наименование дисциплины)

Для специальности (тей) 05.13.11

(номер специальности)

Вносятся следующие дополнения и изменения:

Дополнения и изменения внес _____

должность, ФИО, подпись)

Рабочая программа пересмотрена и одобрена на заседании Ученого совета Института

Председатель Ученого совета _____ _____

(подпись) (ФИО)