Федеральное государственное бюджетное учреждение науки

**Институт систем информатики им. А.П. Ершова**

Сибирского отделения Российской академии наук

УТВЕРЖДАЮ

Директор ИСИ СО РАН

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

« 2 » сентября 2015 г.

**РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ**

**«Методы обработки текстовой информации»**

**Направление подготовки:** 09.06.01 «Информатика и вычислительная техника»

**Специальность:** 05.13.17 «Теоретические основы информатики»

**Уровень образования:** подготовка кадров высшей квалификации

**Квалификация выпускника:** Исследователь. Преподаватель-исследователь

Составители рабочей программы

С.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 «Методы обработки текстовой информации» являются:

* ознакомление аспирантов с математическими основами алгоритмов обработки текстовой информации;
* формирование и развитие знаний о формальных моделях и методах анализа текстов: основных методах математической логики, теории вероятностей и математической статистики, применяемых для семантического анализа, поиска и извлечения информации;
* освоение аспирантами принципов обработки текстовой информации в системах автоматического определения авторства и при анализе данных из социальных сетей.

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 методами обработки текстовой информации: алгоритмами морфологического, синтаксического и семантического 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 «Методы обработки текстовой информации» (Б1.В.ДВ.4) относится к группе дисциплин по выбору аспиранта вариативной части по специальности 05.13.17.

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

Аспиранты, завершившие изучение данной дисциплины, должны:

* **знать:** общие принципы построения систем автоматической обработки текстов, формулировки задач, характеристики и условия применимости изучаемых методов.
* **уметь:** применять методы математической лингвистики для анализа текстовой информации на естественном языке.
* **владеть:** применять методы обработки текстовой информации при разработке и тестировании программных продуктов.

Компетенции, формируемые у обучающихся, в соответствии с ООП по направлению 09.06.01 «Информатика и вычислительная техника» и профилю (специальности) 05.13.17 «Теоретические основы информатики»:

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 | Основные этапы построения систем автоматической обработки текстовой информации | 4 | 2 |  |  | 2 |
| 2 | Методы задания синтаксической структуры предложений. Деревья составляющих. Деревья зависимостей. | 8 | 3 |  |  | 5 |
| 3 | Порождающие грамматики. Распознающие грамматики. Регулярные выражения и конечные автоматы | 4 | 1 |  |  | 3 |
| 4 | Методы теоретического исследования семантики текстов. Лексические функции. Валентности слов. | 4 | 1 |  |  | 3 |
| 5 | Теоретико-модельный подход к исследованию семантики | 4 | 2 |  |  | 2 |
| 6 | Теоретико-множественные модели языка в семантическом анализе | 4 | 2 |  |  | 2 |
| 7 | Анализ тональности текстов. Теория речевых действий. Классификация речевых действий. | 4 | 1 |  |  | 3 |
| 8 | Представление знаний для компьютерной обработки. Тезаурусы. Семантические сети. Фреймовые модели. Онтологические модели | 6 | 2 |  |  | 4 |
| 9 | Методы измерения семантического расстояния. Семантические роли и семантические ограничения. | 4 | 2 |  |  | 2 |
| 10 | Прецедентный анализ. Корпусная лингвистика. Частотные методы в компьютерной лингвистике. | 4 | 1 |  |  | 3 |
| 11 | Классификация и кластеризация. Машинное обучение с учителем и без учителя. | 6 | 3 |  |  | 3 |
| 12 | Автоматические системы извлечения информации. Алгоритмические основы. | 4 | 2 |  |  | 2 |
| 13 | Формальные методы атрибуции текстов | 8 | 2 |  |  | 6 |
| 14 | Методы анализа данных из социальных сетей | 8 | 2 |  |  | 6 |
|  | 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 | СРС |
| 1 | Раздел 1. Принципы и задачи автоматической обработки текстовой информации | 13 | 13 |  |  |  |  |
| 2 | Раздел 2. Методы семантического анализа 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).

Отличительной особенности курса по сравнению с другими подобными курсами является сравнительно большой объем регулярно обновляемого материала в соответствии с требованием в быстро меняющейся области профессиональной деятельности IT.

В процессе изучения курса у аспирантов развиваются такие методы мышления как выдвижение гипотез и формулирование проблем, анализ, синтез, индукция, дедукция, абстрагирование, конкретизация, обобщение, ограничение, аналогия, противоположность.

## 6.2. Содержание разделов и тем курса

1. Основные понятия математической лингвистики. Основные этапы построения систем автоматизированной обработки текстов. Графематический анализ. Алгоритм Витерби. Приемы морфологического анализа.
2. Фрагментационный анализ. Методы задания синтаксической структуры предложений. Деревья составляющих. Деревья зависимостей.
3. Синтаксические анализаторы. Проблемы синтаксического анализа. Применение морфологического и синтаксического анализа в поисковых системах.
4. Порождающие грамматики. Распознающие грамматики. Применение регулярных выражений и конечных авоматов при обработке текстов.
5. Методы теоретического исследования семантики текстов. Лексические функции. Актанты и валентности слов. Теоретико-модельный подход. Исчисление предикатов первого порядка.
6. 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».
7. Анализ тональности текстов. Теория речевых действий. Классификация речевых действий.
8. Представление знаний для компьютерной обработки. Концептуальные диаграммы зависимостей. Тезаурусы. Семантические сети. Фреймовые и онтологические модели. Общие принципы построения.
9. Методы измерения семантического расстояния. Семантические роли и семантические ограничения.
10. Прецедентный анализ текста. Корпусная лингвистика. Частотные методы в компьютерной лингвистике.
11. Модели и методы автоматической классификации и кластеризации текстовой информации. Машинное обучение с учителем и без учителя. Регрессионные, метрические, линейные, вероятностные методы классификации. Иерархические и центроидные методы. Интеллектуальный анализ данных.
12. Автоматические системы извлечения информации. Алгоритмические основы. Принципы обработки неструктурированной и плохо структурированной информации. Тематическая индексация текстов. Задачи распознавания именованных элементов и разрешения конференций.
13. Формальные методы определения авторства текстов. Лингвостатистические параметры текста. Методы энтропийного кодирования, алгоритм Хаффмана. Статистические методы атрибуции.
14. Анализ социальных сетей. Графовые модели социальных сетей. Анализ отношения предпочтения. Моделирование психологических операций. Понятие центральности. Методы обнаружения сообществ и анализ связных подгрупп. Модели динамики сети.

## 6.3. Перечень примерных вопросов для зачета

1. Перечислить направления компьютерной лингвистики.
2. Перечислить основные этапы 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.
6. Перечислить проблемы автоматизации синтаксического анализа.
7. Объяснить принципы использования регулярных выражений в поисковых системах.
8. Изложить основные идеи подхода И. Мельчука к семантическому анализу.
9. Изложить основную идею теоретико-модельного подхода к исследованию семантики текстов.
10. Дать определения отмеченных последовательностей, контекста, дистрибутивных классов.
11. Дать формальные определения частей речи, грамматического рода и категории падежа в терминах модели языка, предложенной С. Маркусом.
12. Изложить основные идеи теории речевых действий.
13. Привести классификацию речевых действий.
14. Сформулировать принципы построения систем представления знаний для компьютерной обработки.
15. В чем отличие между тезаурусами, семантическими сетями и онтологиями?
16. Дать определения семантических сетей и фреймовых моделей.
17. Привести примеры мер семантической близости.
18. Что такое корпусная лингвистика?
19. Применение частотных методов в компьютерной лингвистике. Перечислить, описать, привести примеры.
20. В чем отличие между классификацией и кластеризацией текстов?
21. Перечислить методы классификации и кластеризации текстовой информации. Сформулировать основные принципы.
22. Пояснить принцип работы наивного Байесовского классификатора.
23. Разъяснить принципы работы автоматических систем извлечения информации.
24. Сформулировать принципы обработки неструктурированной и плохо структурированной информации. Индексация текстов.
25. Перечислить формальные методы атрибуции текстов.
26. Дать определения лингвостатистических параметров, авторского инварианта и лингвистических спектров.
27. Привести примеры использования методов кластеризации и классификации для определения авторства текстов.
28. Перечислить основные направления исследований социальных сетей.
29. Дать определения центральностей разного типа.
30. Описать методы анализа социальных сетей.

## 6.4. Перечень примерных тем рефератов и докладов

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

# 7. Самостоятельная работа аспирантов

7.1. Повторение учебного материала с целью закрепления, самостоятельное ознакомление с литературой по данному курсу. Во время самостоятельной работы аспиранты должны усвоить пройденный материал, ознакомиться с дополнительной литературой с целью более глубокого понимания изучаемых вопросов и расширения кругозора.

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

# 8. Учебно-методическое и информационное обеспечение дисциплины

## 8.1. Основная и дополнительная литература

а) основная литература

1. Батура Т.В. Методы определения авторского стиля текстов и их программная реализация // Программные системы и вычислительные методы. М.: НБ-Медиа, 2014. № 2. C. 197–216. DOI: 10.7256/2305-6061.2014.2.11705.
2. Батура Т.В., Мурзин Ф.А., Перфильев А.А., Шманина Т.В. Методы повышения эффективности поиска информации на основе синтаксического анализа // Моногр. / Институт систем информатики им. А.П. Ершова СО РАН. Новосибирск: Изд-во СО РАН, 2014. ISBN 978-5-7692-1398-4. 76 с.
3. Jurafsky D., Martin J. Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition. 2008. 1024 p.
4. Кобзарева Т. Ю. В поисках синтаксической структуры: автоматический анализ русского предложения с опорой на сегментацию. М.: РГГУ. 2015. 371 с.
5. Charu C. Aggarwal Social network data analytics. – 2011. – 520 p.
6. ACL Anthology — A Digital Archive of Research Papers in Computational Linguistics http://anthology.aclweb.org/.
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. Abramsky S., Sadrzadeh M. Semantic Unification // Lecture Notes in Computer Science, 2014, V. 8222, pp. 1-13. URL: http://arxiv.org/pdf/1403.3351v1.pdf
10. Болховитянов А.В., Чеповский А.М. Алгоритмы морфологического анализа компьютерной лингвистики: Учебное пособие. – М.: МГУП имени Ивана Федорова, 2013. – 198 с.

б) дополнительная литература

1. Sajadi, A., Borujerdi, M. Machine Translation Based on Unification Link Grammar // Journal of Artificial Intelligence Review. DOI=10.1007/s10462-011-9261-7, 2013, pp. 109–132.
2. Лукашевич Н.В. Модели и методы автоматической обработки неструктурированной информации на основе базы данных онтологического типа: Диссертация на соискание ученой степени доктора технических наук. Москва, 2014. 312 с.
3. Рубашкин В.Ш. Семантический компонент в системах понимания текста // КИИ-2006. Десятая национальная конференция по искусственному интеллекту с международным участием. Труды конференции. – М.: Физматлит, 2006. – С. 455-463.
4. Socher R. et al. Semantic Compositionality through Recursive Matrix-Vector Spaces. 2013. URL: http://nlp.stanford.edu/pubs/SocherHuvalManningNg\_EMNLP2012.pdf
5. Moss L.S., Tiede H.-J. 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
6. 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.
7. The Stanford Natural Language Processing Group http://nlp.stanford.edu/
8. 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.
9. Щипицина Л.Ю. Информационные технологии в лингвистике: Учебное пособие. М.: Флинта:Наука, 2013. 128 с.
10. 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.

в) программное обеспечение и Интернет-ресурсы

Стандартные компьютеры типа Pentium с выходом в Интернет, программное обеспечение (через "/" указаны допустимые варианты): операционная система Linux/Windows/Solaris/Ubuntu; среда разработки приложений: Eclipse/NetBeans/Visual Studio/IntelliJ IDEA; пакеты прикладных программ любых версий: Maple, Matlab, Adobe Photoshop/GIMP/Corel Draw/Inkscape, MS Office.

## 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

Персональные компьютеры слушателей курса и преподавателя.

**ДОПОЛНЕНИЯ И ИЗМЕНЕНИЯ В РАБОЧЕЙ ПРОГРАММЕ**

за \_\_\_\_\_/\_\_\_\_\_ учебный год

В рабочую программу Методы обработки текстовой информации

(наименование дисциплины)

Для специальности (тей) 05.13.17

(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)