

Правительство Российской Федерации

**Федеральное государственное автономное образовательное учреждение
высшего профессионального образования
Национальный исследовательский университет
«Высшая школа экономики»**

Факультет Компьютерных наук
Департамент анализа данных и искусственного интеллекта

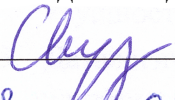
**Программа дисциплины
«Программные системы для обработки и анализа текстов»**

для направления 010400.62 «Прикладная математика и информатика»
подготовки бакалавра

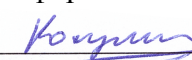
Авторы программы:
Ильвовский Д.А., преподаватель
dilvovsky@hse.ru
Черняк Е.Л., преподаватель
echernyak@hse.ru

УТВЕРЖДАЮ

Одобрена на заседании департамента
анализа данных и искусственного интеллекта
Руководитель департамента С.О. Кузнецов


«28» 08 2014 г.

Академический руководитель
образовательной программы
по направлению 010400
«Прикладная математика и
информатика» А.С. Конушин


«08» 09 2014 г.

Рекомендована Академическим советом образовательной
программы «Прикладная математика и информатика»

«08» 09 2014 г.

Менеджер департамента анализа данных
и искусственного интеллекта Л.И. Антропова



Москва, 2014

*Настоящая программа не может быть использована другими подразделениями
университета и другими вузами без разрешения подразделения разработчика программы*

1. Аннотация

Дисциплина «Программные системы для обработки и анализа текстов» предназначена для подготовки бакалавров направления «Прикладная математика и информатика». Она продолжает цикл дисциплин, связанных с основами анализа данных, информационных технологий и программирования.

В курсе изучаются основные задачи обработки и анализа текстов, методы и инструменты для их решения. Затрагиваются задачи выделения ключевых слов и словосочетаний и определения скрытых тем. Определяются цели и задачи морфологического, лексического и синтаксического анализа. Рассматриваются основы корпусной лингвистики. Приводится обзор целей и основных способов визуализации текстовых данных.

Теоретический материал курса подкрепляется практическими занятиями по использованию популярных инструментов по изучаемой тематике.

2. Область применения и нормативные ссылки

Настоящая программа устанавливает минимальные требования к знаниям и умениям студента и определяет содержание и виды учебных занятий и отчетности.

Программа предназначена для преподавателей, ведущих данную дисциплину, учебных ассистентов и студентов третьего года обучения в бакалавриате по направлению 010400.62 «Прикладная математика и информатика». Дисциплина является факультативной.

Программа разработана в соответствии с:

- Образовательным стандартом ВПО ГОБУ НИУ ВШЭ;
- Образовательной программой подготовки бакалавра по направлению 010400.62 «Прикладная математика и информатика»;
- Рабочим учебным планом подготовки бакалавра по направлению 010400.62, утвержденным в 2013 г.

3. Цели освоения дисциплины

Данная дисциплина ставит своей целью изучение основных задач и методов обработки и анализа текстов, а также освоение программных систем и инструментов, в которых реализованы данные методы. Эти базовые знания и навыки необходимы в профессиональной деятельности специалистов по анализу данных и машинного обучения.

4. Компетенции, формируемые в результате освоения дисциплины

В результате изучения дисциплины студенты должны:

- Знать постановку задач морфологического, синтаксического и лексического анализа методы их решения;
- Владеть основными программными системами для выделения скрытых тем и визуализации текстов;
- Уметь решать задачи выделения ключевых слов и определения тональности с помощью применения существующих программных средств.

В результате изучения дисциплины студент осваивает и развивает следующие компетенции:

| Компетенция | Код по ФГОС/ НИУ | Дескрипторы – основные признаки освоения (показатели достижения результата) | Формы и методы обучения, способствующие формированию и развитию компетенции |
|--|------------------|--|--|
| Умение работать на компьютере, навыки использования основных классов программного обеспечения, работы в компьютерных сетях | ИК-2 | Студент демонстрирует владение интерфейсом программных систем для обработки и анализа текстов | Выполнение домашних заданий, ориентированных на использование программных систем обработки и анализа текстов |
| Способность решать задачи производственной и технологич. деятельности на профессион. уровне, включая разработку математических моделей, алгоритмических и программных решений | ПК-8 | Студент демонстрирует компетентность в выборе той или иной программной системы для решения поставленной перед ним задачи обработки и анализа текстов | Лекции по основным задачам и методам обработки и анализа текстов; решение задач, требующих выбор метода обработки и анализа текстов и программной системы, в которой данный метод реализован |
| Способность применять в профессиональной деятельности современные языки программирования и языки баз данных, операционные системы, электронные библиотеки и пакеты программ и т.п. | ПК-9 | Студент демонстрирует понимание основных методов обработки и анализа текстов, владение основными программными системами обработки и анализа текстов | Лекции по основным задачам и методам обработки и анализа текстов; домашние задания, ориентированные на использование программных систем обработки и анализа текстов |

5. Место дисциплины в структуре образовательной программы

Настоящая учебная дисциплина является факультативной и входит в цикл дисциплин информационных технологий в учебной программе подготовки бакалавра направления 010400.62 «Прикладная математика и информатика».

Изучение курса «Программные системы обработки и анализа текстов» требует базовых знаний по обработке и анализу данных (в объеме бакалаврской программы третьего года обучения по направлению 010400.62). Необходимо также владение базовыми навыками программирования на языке высокого уровня (в объеме курса «Информатика и программирование» первого года обучения указанной бакалаврской программы).

Основные положения дисциплины «Программные системы обработки и анализа текстов» должны быть использованы в дальнейшем при изучении следующих дисциплин программы бакалавра:

- Автоматическая обработка текстов
- Методы машинного обучения и разработки данных,

а также при выполнении курсовых и выпускных квалификационных работ.

6. Тематический план дисциплины «Программные системы обработки и анализа текстов»

| № | Название темы | Всего часов по дисциплине | Аудиторные часы | | Самостоятельная работа |
|---|--|---------------------------|-----------------|-----------------|------------------------|
| | | | Лекции | Сем. и практика | |
| 1 | Введение | 6 | 2 | 2 | 2 |
| 2 | Частотный анализ текстов | 10 | 2 | 2 | 6 |
| 3 | Морфологический анализ | 14 | 2 | 2 | 10 |
| 4 | Выделение ключевых слов и словосочетаний | 20 | 4 | 4 | 12 |
| 5 | Выявление скрытых тем | 12 | 2 | 2 | 8 |
| 6 | Введение в корпусную лингвистику | 12 | 2 | 2 | 8 |
| 7 | Синтаксический анализ | 20 | 4 | 4 | 12 |
| 8 | Визуализация текстов | 14 | 2 | 2 | 10 |
| | Итого | 108 | 20 | 20 | 68 |

7. Формы контроля знаний студентов

Курс «Программные системы обработки и анализа текстов» читается в 3 модуле.

| Тип контроля | Форма контроля | Параметры |
|------------------------------|----------------|--|
| Итоговый контроль в 3 модуле | Устный экзамен | 120 минут, задаются вопросы по билетам |

Критерии оценки знаний

На итоговом контроле студент должен продемонстрировать владение основными понятиями из пройденных тем дисциплины.

Итоговый контроль проводится в форме устного экзамена, включающего несколько вопросов по темам дисциплины.

Порядок формирования оценок по дисциплине

Преподаватель оценивает самостоятельную работу студентов по выполнению домашних работ, выдаваемых на практических занятиях – при этом оценивается правильность выбора метода решения задачи и эффективность его использования. Оценки за домашние задания выставляются в рабочую ведомость, и перед экзаменом модуля за

домашние задания выставляется результирующая оценка по десятибалльной шкале $O_{сам. работа}$.

Оценка итогового контроля выставляется по следующей формуле:

$$O_{дисциплина} = 0,5 \cdot O_{экзамен} + 0,5 \cdot O_{сам. работа}$$

и округляется до целого числа арифметическим способом,

где $O_{экзамен}$ – оценка за работу непосредственно на устном экзамене.

В случае пропусков занятий и домашних заданий студент может сдать все домашние задания не позднее чем за 5 дней до экзамена – в этом случае они учитываются описанным выше способом.

В диплом выставляется **результирующая оценка** $O_{дисциплина}$ по данной учебной дисциплине.

8. Содержание программы по темам

Тема 1. Введение

1. Основные задачи обработки и анализа текстов. Актуальность обработки и анализа текстов.
 2. Краткий исторический экскурс по обработке и анализу текстов.
 3. Обзор существующих систем обработки и анализа текстов. Классификация систем обработки и анализа текстов.
1. Чеповский, А. М. Неразрешимая проблема компьютерной лингвистики // Компьютерра. – 2002. – № 30. – С. 12-18.
 2. Ильвовский, Д. А., Черняк Е. Л. Системы автоматической обработки текстов // Открытые системы. – 2014. – № 1. – С. 51-53.

Дополнительная литература

1. Sparck Jones, K. Natural language processing: a historical review // Current Issues in Computational Linguistics: in Honour of Don Walker. – 1994. – С. 3-16

Тема 2. Частотный анализ текстов

1. Модель мешка слов. Векторное представление текстов. Релевантность в векторной модели. Расширения модели мешка слов.
2. Реализация модели мешка слов в библиотеках Gensim и NLTK.

Основная литература

1. Manning C.D., Schuetze H. Foundations of Statistical Natural Processing / – MIT Press, 1999.
2. Bird, S., Klein, E., Loper, E. Natural Language Processing with Python / – O'Reilly Media, 2009.
3. Řehůřek, R., Sojka, P., Software Framework for Topic Modelling with Large Corpora // Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. – 2010. – С. 45-50.

Дополнительная литература

1. Сузи, Р. Язык программирования Python / – Москва : Интернет-университет информационных технологий (ИНТУИТ), Бином. Лаборатория знаний, 2007.

Тема 3. Морфологический анализ

1. Задача морфологического анализа. Типы языков. Алгоритмы морфологического разбора. Морфологическая разметка. Омонимия.
2. Программные морфологические анализаторы и словари. АОР, mystem, PyMorphy2.

Основная литература

1. Болховитянов, А. В., Чеповский, А. М. Алгоритмы морфологического анализа компьютерной лингвистики. / – Москва : МГУП им. Ивана Федорова, 2013.
2. Jurafsky, D., Martin J. H. Speech and Language Processing / – Pearson Prentice Hall, 2009.

Дополнительная литература

3. Segalovich, I. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. // The 2003 International Conference on Machine Learning; Models, Technologies and Applications. – 2003.
4. Зеленков, Ю. Г., И. В. Сегалович, Титов, В. А. Вероятностная модель снятия морфологической омонимии на основе нормализующих подстановок и позиций соседних слов // Компьютерная лингвистика и интеллектуальные технологии. Труды международного семинара Диалог2005. – 2005. – С. 188-197.

Тема 4. Выделение ключевых слов и словосочетаний

1. Лексический анализ. Словари и тезаурусы. Поиск синонимов.
2. Частотные методы выделения ключевых слов и словосочетаний. Метрики совместной встречаемости. Выделение ключевых словосочетаний по морфологическим шаблонам.
3. Программные средства для выделения ключевых слов: NLTK, Томита-парсер.

Основная литература

1. Manning C.D., Schuetze H. Foundations of Statistical Natural Processing / – MIT Press, 1999.
2. Bird, S., Klein, E., Loper, E. Natural Language Processing with Python / – O'Reilly Media, 2009.
3. Jurafsky, D., Martin J. H. Speech and Language Processing / – Pearson Prentice Hall, 2009.

Дополнительная литература

1. Лукашевич, Н. В. Тезаурусы в задачах информационного поиска / – Москва: Издательство МГУ, 2011

Тема 5. Выявление скрытых тем

1. Модель скрытых тем. Латентное размещение Дирихле (LDA). Параметры модели. Выбор числа скрытых тем. Расширения модели LDA.
2. Программные средства для выделения скрытых тем: Mallet, Gensim.

Основная литература

1. Blei, D.M., Ng, D.M., Jordan, M.I. Latent Dirichlet allocation. // The Journal of Machine Learning Research. – № .3. – 2003. – С. 993-1022.

2. К.В.Воронцов. Лекции по вероятностным тематическим моделям [Электронный ресурс] / – Режим доступа: <http://www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf>, свободный

Дополнительная литература

1. Blei, D. M., Jordan, M. I., Griffiths, T. L., Tenenbaum, J. B. Hierarchical Topic Models and the Nested Chinese Restaurant Process // Advances in Neural Information Processing Systems. – № 16. – 2004.
2. Teh, Y. W., Jordan, M. I., Beal, M. J., Blei, D. M. Hierarchical Dirichlet Processes // Journal of the American Statistical Association. – № 101. – 2006. – С. 1566–158.

Тема 6. Введение в корпусную лингвистику

1. Назначение корпусов. Корпуса русского языка. Методы машинного обучения в корпусной лингвистике.
2. Составление веб-корпусов. HTML-парсеры. Регулярные выражения.

Основная литература

1. Национальный корпус русского языка (НКРЯ) [Электронный ресурс] / – режим доступа: www.ruscorpora.org, свободный.
2. Baroni, M., Bernardini, S., Ferraresi, A., Zanchetta, E. The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora // Journal of Language Resources and Evaluation. – № 43(3). – 2009. – С. 209-226.

Дополнительная литература

1. Sharoff, S. Open-source corpora: using the net to fish for linguistic data. // International Journal of Corpus Linguistics. – № 11(4). – 2006. – С. 435-462.

Тема 7. Синтаксический анализ

1. Задача синтаксического анализа. Деревья синтаксического разбора. Контекстно-свободные грамматики. Деревья зависимостей. Деревья составляющих.
2. Синтаксические шаблоны. Синтаксическое расширения поиска. TreeBank.
3. Программные системы для синтаксического анализа: Томита-парсер, ЭТАПЗ, OpenNLP, StanfordNLP, NLTK.

Основная литература

1. Chomsky, N. Syntactic structures / Walter de Gruyter, 2002.
2. Jurafsky, D., Martin J. H. Speech and Language Processing / – Pearson Prentice Hall, 2009.
3. Mitchell P. M., Marcinkiewicz, M. A., Santorini, B. Building a large annotated corpus of English: The Penn Treebank // Computational linguistics. – № 19(2). –1993. – С. 313-330.

Дополнительная литература

1. Апресян, Ю. Д. Лингвистическое обеспечение системы ЭТАП-2. / Наука, 1989.

Тема 8. Визуализация текстов

1. Принципы визуализации текстов. Визуальный анализ текстовой информации. Способы визуализации текстов.
2. Инструменты визуализации текстов. Онлайн-инструменты. Облака тегов.

Основная литература

1. Manning C.D., Schuetze H. Foundations of Statistical Natural Processing / – MIT Press, 1999.

Дополнительная литература

1. Материалы Workshop on Interactive Language Learning, Visualization, and Interfaces, Baltimore, USA, 2014 [электронный ресурс] / – режим доступа <http://nlp.stanford.edu/events/illvi2014/>, свободный

9. Образовательные технологии

В преподавании данной дисциплины сочетаются:

- лекции в традиционной форме;
- практические занятия, в ходе которых студенты осваивают основные программные системы обработки и анализа текстов;
- домашние практические задания по использованию программных систем обработки и анализа текстов по всем основным темам дисциплины.

10. Оценочные средства для текущего и итогового контроля

Примеры домашних работ

1. Составить собственную коллекцию текстов на русском языке. Составить частотные словари по собственной коллекции: частоты униграмм, биграмм и триграмм, используя возможности библиотеки NLTK.
2. На основе индекса gensim реализовать поиск по запросу по собственной коллекции текстов на русском языке. Поиск должен основываться на вычислении релевантности запроса текстам согласно векторной модели релевантности.
3. Провести морфологический анализ собственной коллекции текстов на русском языке. Посчитать количество существительных и глаголов в размеченной коллекции. Привести примеры полной и частичной омонимии в полученной разметке.
4. Выделить ключевые слова и словосочетания из собственной коллекции текстов на русском языке несколькими методами: с помощью метрик взаимной встречаемости или с помощью грамматических шаблонов. Для выделения ключевых словосочетаний по шаблонам использовать Томита-парсер.
5. Предположить, что в собственной коллекции текстов на русском языке существует некоторое число скрытых тем. Выделить их, используя метод LDA. Объяснить полученные результаты.
6. Составить собственный веб-корпус.
7. Разметить собственную коллекцию текстов или собранный веб-корпус с помощью синтаксических парсеров. Построить деревья зависимостей и деревья составляющих.
8. Построить облако ключевых слов и словосочетаний любого текста.

Вопросы для оценки качества освоения дисциплины

Тема 1.

1. Перечислите основные задачи обработки текстов.
2. Перечислите основные задачи анализа текстов.
3. Назовите несколько библиотек для обработки и анализа текстов.
4. Назовите несколько консольных приложений для обработки и анализа текстов.

Тема 2.

1. Объясните принципы векторной модели (VSM).
2. Как рассчитывается релевантность в векторной модели?

3. Объясните принципы обобщенной векторной модели (gVSM).

Тема 3.

1. Почему возникает необходимость в разборе текста по частям речи?
2. Чем морфологический анализатор отличается от морфологического словаря?
3. Что такое омонимия? Приведите примеры.
4. Перечислите несколько алгоритмов разрешения омонимии.

Тема 4.

1. Как устроен WordNet?
2. Предложите простой алгоритм поиска синонимов.
3. Перечислите несколько метрик совместной встречаемости.
4. Что такое морфологический шаблон?

Тема 5.

1. Объясните принципы модели скрытых тем.
2. Опишите генеративную модель языка, лежащую в основе LDA.
3. Какие расширения модели LDA вы знаете?

Тема 6.

1. Зачем нужны веб-корпуса?
2. Что такое кроулинг?
3. Какие стратегии сбора веб-корпусов вы знаете?

Тема 7.

1. В чем заключается задача синтаксического анализа?
2. Что такое контекстно-свободная грамматика?
3. Чем отличаются деревья разбора от деревьев составляющих?
4. По данному предложению постройте дерево разбора и дерево составляющих.
5. Что такое синтаксический шаблон?
6. Как используются синтаксические шаблоны?
7. В каких задачах возникает потребность в синтаксическом анализе?

Тема 8.

1. Какие способы визуализации текстов вы знаете?
2. Чем отличается удачная визуализация текстов от неудачной?

Базовая литература

1. Jurafsky, D., Martin J. H. Speech and Language Processing / – Pearson Prentice Hall, 2009.
2. Manning C.D., Schuetze H. Foundations of Statistical Natural Processing / – MIT Press, 1999

Основная литература

3. Болховитянов, А. В., Чеповский, А. М. Алгоритмы морфологического анализа компьютерной лингвистики. / – Москва : МГУП им. Ивана Федорова, 2013.
4. Ильвовский, Д. А., Черняк Е. Л. Системы автоматической обработки текстов // Открытые системы. – 2014. – № 1. – С. 51-53.
5. К.В.Воронцов. Лекции по вероятностным тематическим моделям [Электронный ресурс] / – Режим доступа: <http://www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf>, свободный
6. Национальный корпус русского языка (НКРЯ) [Электронный ресурс] / – режим доступа: www.ruscorpora.org, свободный.
7. Чеповский, А. М. Неразрешимая проблема компьютерной лингвистики // Компьютерра. – 2002. – № 30. – С. 12-18.
8. Baroni, M., Bernardini, S., Ferraresi, A., Zanchetta, E. The WaCky Wide Web: A collection of very large linguistically processed Web-crawled corpora // Journal of Language Resources and Evaluation. – № 43(3). – 2009. – С. 209-226.

9. Bird, S., Klein, E., Loper, E. Natural Language Processing with Python / – O'Reilly Media, 2009.
10. Blei, D.M., Ng, D.M., Jordan, M.I. Latent Dirichlet allocation. // The Journal of Machine Learning Research. – № 3. – 2003. – С. 993-1022.
11. Chomsky, N. Syntactic structures / Walter de Gruyter, 2002.
12. Mitchell P. M., Marcinkiewicz, M. A., Santorini, B. Building a large annotated corpus of English: The Penn Treebank // Computational linguistics. – № 19(2). –1993. – С. 313-330.
13. Řehůřek, R., Sojka, P., Software Framework for Topic Modelling with Large Corpora // Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. – 2010. – С. 45-50.

Дополнительная литература

1. Апресян, Ю. Д. Лингвистическое обеспечение системы ЭТАП-2. / Наука, 1989.
2. Зеленков, Ю. Г., И. В. Сегалович, Титов, В. А. Вероятностная модель снятия морфологической омонимии на основе нормализующих подстановок и позиций соседних слов // Компьютерная лингвистика и интеллектуальные технологии. Труды международного семинара Диалог2005. – 2005. – С. 188-197.
3. Лукашевич, Н. В. Тезаурусы в задачах информационного поиска / – Москва: Издательство МГУ, 2011
4. Материалы Workshop on Interactive Language Learning, Visualization, and Interfaces, Baltimore, USA, 2014 [электронный ресурс] / – режим доступа <http://nlp.stanford.edu/events/illvi2014/>, свободный
5. Сузи, Р. Язык программирования Python / – Москва : Интернет-университет информационных технологий (ИНТУИТ), Бинوم. Лаборатория знаний, 2007.
6. Blei, D. M., Jordan, M. I., Griffiths, T. L., Tenenbaum, J. B. Hierarchical Topic Models and the Nested Chinese Restaurant Process // Advances in Neural Information Processing Systems. – № 16. – 2004.
7. Segalovich, I. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. // The 2003 International Conference on Machine Learning; Models, Technologies and Applications. – 2003.
8. Sharoff, S. Open-source corpora: using the net to fish for linguistic data. // International Journal of Corpus Linguistics. – № 11(4). – 2006. – С. 435-462.
9. Sparck Jones, K. Natural language processing: a historical review // Current Issues in Computational Linguistics: in Honour of Don Walker. – 1994. – С. 3-16
10. Teh, Y. W., Jordan, M. I., Beal, M. J., Blei, D. M.. Hierarchical Dirichlet Processes // Journal of the American Statistical Association. – № 101. – 2006. – С. 1566–158.

13. Материально-техническое обеспечение дисциплины

Для лекционных и практических занятий по темам дисциплины используется проектор и компьютеры с инструментальной средой программирования и выходом в сеть Интернет.

Авторы программы: _____ / Ильвовский Д. А./
 _____ / Черняк Е.Л. /