Министерство образования и науки РФ
Федеральное государственное автономное образовательное учреждение высшего образования
«Новосибирский национальный исследовательский государственный университет» (Новосибирский государственный университет, НГУ)
Гуманитарный факультет

Программа рассмотрена
на заседании кафедры
фундаментальной и прикладной
лингвистики
29.08.2014
_____
Зав. кафедрой, проф. М.К. Тимофеева

Утверждаю

_____
декан гуманитарного
факультета, профессор
1.09.2014
Л.Г. Панин

Основная образовательная программа
высшего образования

Направление подготовки
035800 – Фундаментальная и прикладная лингвистика

Квалификация (степень) выпускника –
бакалавр

ПРОГРАММА УЧЕБНОГО КУРСА «АНАЛИЗ СИМВОЛЬНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ»

(54 часа, 1 з.е.)

1. **Наименование дисциплины**

ПРОГРАММА УЧЕБНОГО КУРСА «АНАЛИЗ СИМВОЛЬНЫХ ПОСЛЕДОВАТЕЛЬНОСТЕЙ»

Программа дисциплины «Анализ символьных последовательностей» составлена в соответствии с требованиями к обязательному минимуму содержания и уровню подготовки дипломированного бакалавра по направлению 035800 «Фундаментальная и прикладная лингвистика» в целях обеспечения реализации учебного процесса в НГУ..

Автор __Мирошниченко Любовь Александровна, к.т.н., с.н.с. ИМ СО РАН

2. Перечень планируемых результатов обучения по дисциплине, соотнесенных с планируемыми результатами освоения образовательной программы.

Цель освоения дисциплины

Дисциплина «Анализ символьных последовательностей» имеет своей целью:

познакомить слушателей с разнообразием алгоритмов и структур данных, предназначенных для анализа символьных последовательностей.

В рамках указанной цели ставятся следующие задачи:

- рассмотреть спектр проблем и методов анализа символьных последовательностей произвольной языковой природы;
- продемонстрировать разнообразие алгоритмов и структур данных, предназначенных для решения каждой проблемы;
- проследить аналогии между методами анализа символьных последовательностей в области анализа текстов и в биологии;

В результате освоения дисциплины обучающийся должен:

Иметь представление о разнообразных алгоритмах анализа символьных последовательностей, используемых, в частности, в широко распространенных текстовых редакторах и системах сжатия информации

Знать структуры данных, с помощью которых реализуются алгоритмы анализа символьных последовательностей

Уметь среди множества изученных алгоритмов выбирать тот, который оптимальным образом подходит к решению конкретной задачи и промоделировать работу выбранного (или заданного) алгоритма на тестовом примере.

Перечисленные результаты образования являются основой для формирования следующих общекультурных и общепрофессиональных компетенций:

***а) общекультурными* (*ОК*)**

– способностью применять методы математического анализа и моделирования в профессиональной деятельности — ОК-10

***б) профессиональными (ПК)*:**

в области производственно-практической деятельности и проектной деятельности:

- умением использовать лингвистические технологии для проектирования систем анализа и синтеза естественного языка, в том числе лингвистических компонентов интеллектуальных и информационных электронных систем ПК-17
- способностью работать в междисциплинарной команде ПК-26
- способностью общаться с экспертами в других областях знаний ПК-27

3. Место дисциплины в структуре образовательной программы.

Курс «Анализ символьных последовательностей» является частью блока Б2 ДВ1 «Математический и естественнонаучный цикл» «Дисциплины по выбору» и используется в других курсах этого раздела.

Программа дисциплины составлена с учетом связей и соотношения учебных дисциплин, преподаваемых на отделении фундаментальной и прикладной лингвистики гуманитарного факультета НГУ.

Изучение курса «Анализ символьных последовательностей» требует предварительных знаний в области информатики и дискретной математики. Необходимые сведения из биологии будут сообщены в ходе курса. Желательно знание основ молекулярной биологии и генетики в объеме курса средней школы.

Преподавание дисциплины базируется на знаниях и умениях, полученных при изучении курсов «Информатика и основы программирования», «Математические методы в филологии», связано с курсами «Математические модели языка», «Корпусная лингвистика», «Базы данных и информационный поиск», «Автоматический анализ текста», «Введение в прикладную лингвистику».

Освоение дисциплины необходимо как предшествующее при специализации в области компьютерных методов фундаментальной и прикладной лингвистики.

4. Объем дисциплины в зачетных единицах с указанием количества академических, выделенных на контактную работу обучающихся с преподавателем (по видам учебных занятий) и на самостоятельную работу обучающихся.

Общая трудоемкость дисциплины составляет 1 зачетная единица, 54 часа. Из них на контактную работу с преподавателем 32 часа (лекции), на самостоятельную работу студентов – 18 часа Интерактивных занятий – 1 час.

Вид учебной работы	*Семестр*
	8
Общая трудоемкость дисциплины	54
Аудиторные занятия, в том числе:	32
Лекции	32
Семинары	
Контрольные работы	
Самостоятельная работа	18
Виды промежуточного контроля	зачёт

5. Содержание дисциплины “Анализ символьных последовательностей”, структурированное по темам с указанием отведенного на них количества астрономических часов и видов учебных занятий

№ п/п	Раздел дисциплины	Семестр	Неделя семестра	Виды учебной работы, включая самостоятельную работу студентов и трудоемкость (в часах)	Формы текущего контроля успеваемости *(по неделям семестра)* Форма промежуточной аттестации *(по семестрам)*
				Лекция	
1	**Тема 1.** Вводная лекция	8	1	2	
2	**Тема 2.** Формальные языки и грамматики	8	2	2	
3	**Тема 3.** Алгоритмы поиска образца в тексте	8	3-5	6	
4	**Тема 4.** Сортировка	8	6	2	
5	**Тема 5.** Классификация повторов		7	2	
6	**Тема 6.** Алгоритмы выявления повторов	8	8-11	8	
8	**Тема 7.** Расстояния и меры сходства последовательностей	8	12-15	8	
9	**Тема 8.** Сложность текста	8	16	1	
10	**Дискуссия по тематике курса**			1	
					Зачёт
	. Итого:			**32**	

6. Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине

Методические рекомендации для самостоятельной работы с литературой и Интернет-ресурсами (перечень вопросов для самостоятельной проработки для каждой темы).

Методы анализа символьных последовательностей развиваются как в прикладных исследованиях, так и теоретических. Основными прикладными исследованиями являются библиографический поиск, дешифровка знаменных песнопений, анализ песенных мелодий, построение систем автоматического распознавания речи, анализ последовательностей ДНК и других биологических последовательностей (белковых, последовательностей генов в геноме и др.). Среди теоретических исследований следует особо отметить теорию автоматов, теорию формальных языков и грамматик. В 80-е годы прошлого столетия были осознаны связи между методами анализа символьных последовательностей в разных областях и сделаны первые попытки вычленить общие их основы.

Изучение аналогий между проблемами и методами анализа символьных последовательностей из различных прикладных областей является весьма продуктивным. В частности, для анализа текстов на естественных языках, которые являются одним из главных объектов анализа в информационных технологиях, полезно знакомство с

методами и подходами к анализу символьных последовательностей, выработанными в биоинформатике. Это связано как с аналогиями между текстами различной языковой природы, так и с тем, что разработка методов анализа последовательностей в последние годы наиболее интенсивно происходила именно в биоинформатике.

Тема 1. Вводное занятие. Примеры символьных последовательностей из различных языковых систем.

Тема 2. Формальные языки и грамматики. Определение формального языка и формальной грамматики. Классификация грамматик. Конечные автоматы.

Тема 3. Алгоритмы поиска образца в тексте. Алгоритм Кнута-Морриса-Пратта, алгоритм Бойера-Мура, битовый параллелизм, алгоритм Карпа-Рабина. Обобщения задачи поиска образца. Множественный поиск. Алгоритм Ахо – Корасик. Поиск по частично-специфицированному запросу. Образцы с переменными.

Тема 4. Сортировка. Алгоритмы числовой и лексикографической сортировки.

Тема 5. Классификация повторов. Примеры повторов в разных языковых системах. Классификация повторов. Полный частотный спектр

Тема 6. Алгоритмы выявления повторов. Сортировка, хеширование, *l* граммные деревья. Дерево префикс-идентификаторов. Алгоритм Мартинеца. Алгоритм Вайнера. Определение суффиксного дерева. Алгоритм Мак-Крейга. Алгоритм Укконена. Задачи, решаемые с помощью суффиксного дерева. Суффиксный автомат. Ациклический граф слова..

Тема 7. Расстояния и меры сходства последовательностей. Расстояние Хэмминга. Редакционное (эволюционное) расстояние. Схема динамического программирования. Максимально длинная общая подпоследовательность. Ранговые меры близости. Сложностное и трансформационное расстояние. Инверсионное расстояние.

Тема 8. Сложность текста. Колмогоровская сложность. Энтропия. Лингвистическая и комбинаторная сложность. Грамматическая сложность. Сложность по Лемпелю и Зиву. Сложностные разложения. Поиск периодичностей в тексте. Сложность и случайность. Сложность и сжатие.

Виды внеаудиторной самостоятельной работы студентов по дисциплине “Анализ символьных последовательностей”

1) работа над лекционным материалом;
2) работа с литературой;
3) работа с Интернет-источниками;
4) практическое использование изучаемых алгоритмов, программ и методов;
5) написание рефератов;
6) подготовка к зачёту.

7. Фонд оценочных средств для промежуточной аттестации обучающихся по дисциплине.

Вопросы для обсуждения, презентаций, проверки знаний:

Промоделировать работу алгоритмов поиска образца на тестовом примере.

Построить пример текста, на котором число операций алгоритма Бойера-Мура минимально при заданном образце.

Промоделировать работу алгоритма Ахо-Корасик на тестовом примере.

Построить дерево префикс-идентификаторов для заданного текста.
Использование суффиксного дерева для поиска межтекстовых повторов.
Поиск палиндромов с помощью суффикного дерева.
Поиск образцов с помощью суффиксного дерева.
Построить схему динамического программирования для двух заданных текстов.
Построить максимально длинную общую подпоследовательность.
Для каких последовательностей можно определить инверсионное расстояние.
Построить сложностное разложение заданного текста.

8. Перечень основной и дополнительной учебной литературы, необходимой для освоения дисциплины (литература имеется в библиотеке Института математики СО РАН).

1. Дэн Гасфилд. Строки, деревья и последовательности в алгоритмах. Информатика и вычислительная биология // Невский диалект, 2003
2. Maxime Crochemore, Wojciech Rytter. Text Algorithms// Oxford University Press, 1994
3. Maxime Crochemore, Wojciech Rytter. Jewels of Stringology // World Scientific, 2002

Дополнительная литература

1. Колмогоров А.Н. Три подхода к определению понятия "количество информации" // Проблемы передачи информации. - Т.1, вып. 1, 1965. С. 3-11.
2. Гусев В.Д. Сложностные профили символьных последовательностей // В сб. Методы обработки символьных последовательностей и сигналов (Вычислительные системы, вып. 132).- Новосибирск, 1989, 35-63
3. Bafna V. and Pevzner P.A. Sorting by Reversals: Genome Rearrangements in Plant Organelles and Evolutionary History of X Chromosome // Molecular Biology and Evolution, V. 12, № 2, 1995, 239-246.
4. Bell T.C., Cleary J.G., Witten I.H. Text compression, 1990, Prentice Hall, Jnc.
5. Benson, G. Tandem repeats finder: a program to analyze DNA sequences // NAR (1999), V. 22, 573-580.
6. Bernaola-Galvan, P., Roman Roldán, R and Oliver, J.L. Compositional segmentation and long-range fractal correlation in DNA sequences // Phys. Rev. E, v 53, 1996, 5181-5189.
7. A. Bolshoy. DNA sequence analysis linguistic tools: contrast vocabularies, compositional spectra and linguistic complexity, Applied Bioinformatics, 2003, 2(2), 103-112.
8. Blumer A., Blumer J., et al Building the minimal DFA for the set of all subwords of a word on_line in Linear Time // Lect/ Notes in Comput. Sci. -- 1984. -- Vol. 172. -- P.109-118.
9. Crochemore M., Verin R. Zones of low entropy in genomic sequences // Computers and Chemistry. -- 1999. -- Vol. 23. -- P. 275--282.
10. Ebeling W., Jiménes-Montaño M.A. On Grammars, Complexity, and Information Measures of Biological Macromolecules // Math. Biosci, 1980. - № 52. 53-71.
11. Hancock J.M., Armstrong J.S. SIMPLE34: an improved and enhanced implementation for VAX and Sun computers of the SIMPLE algorithm for analysis of clustered repetitive motifs in nucleotide sequences // Comput. Appl. Biosci., 1994. - Vol. 10, 67-70.
12. Donald E. Knuth. The art of computer programming. Vol.2. Seminumerical Algorithms. Addison-Wesley Publishing Company, 1969.
13. Lempel, A. and Ziv, J. On the complexity of finite sequences, IEEE Trans. Inform. Theory, 1976, vol. IT-22, no. 1, pp. 75–81.

14. Li W. The complexity of DNA: the measure of compositional heterogeneity in DNA sequences and measures of complexity //Complexity, V.3, 1997, 33-37.
15. M. Lothaire. Combinatorics on Words. Addison-Wesley, Reading, MA, 1983.
16. A. Bolshoy. DNA sequence analysis linguistic tools: contrast vocabularies, compositional spectra and linguistic complexity, Applied Bioinformatics, 2003, 2(2), 103-112.
17. R. Román-Roldán, P. Bernaola-Galván and J.L. Oliver. Sequence compositional complexity of DNA through an entropic segmentation method, Physical review letters, V.80, 1998, 1344-1347.
18. Ukkonen E. Constructing suffix trees on-line in linear time, Algorithmica, V.14,1995, 249-260.
19. H.Herzel, Complexity of Symbol Sequences, Syst. Anal.Model. Simul., V.5, № 5, 1988, 435-444.
20. J.-S.Varré, J.-P.Delahaye, E. Rivals: Transformation distances: a family of dissimilarity measures based on movements of segments. // Bioinformatics 15(3): 194-202 (1999)
21. Wagner R.A., Fisher M.J. The string-to-string correction problem // J.ACM, Jan.1974.- Vol.21, No.1, 168-173.
22. Wootton J.C., Federhen S., Statistics of local complexity in amino acid sequences and sequence databases // Comput. Chem., V.17, 1993, 149-163.

9. Методические указания для обучающихся по дисциплине.

Итоговый зачёт ставится при положительной оценке трёх составляющих:
работы в течение семестра,
подготовке сообщений с презентациями;
ответ на один из зачётных вопросов.

10. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине.

Ноутбук, медиапроектор, мультемидийные презентации