Программа дисциплины «Анализ данных для лингвистов»

Утверждена
Академическим советом ОП
Протокол № 15 от «28» июня 2018 г.

Автор	Г. А. Мороз
Число кредитов	3
Контактная работа (час.)	36
Самостоятельная работа (час.)	78
Курс	3, 4
Формат изучения дисциплины	без использования онлайн-курса

I. ЦЕЛЬ И РЕЗУЛЬТАТЫ ОСВОЕНИЯ ДИСЦИПЛИНЫ

Задачей курса «Анализ данных для лингвистов» является продолжение знакомства с различными методами анализа данных. Курс разделен на несколько тематических блоков: первый связан с применением байесовских статистических методов (байесовский апдейт, байесовский доверительный интервал, байесовский фактор, байесовкская эмпирическая оценка), второй связан с методами уменьшения размерности (PCA, LDA, CA, MCA), третий блок связан с методами кластеризации (k-means, иерархическая кластеризация, смешанные модели) и последний блок будет посвящен проблемам применения регрессионного анализа (регрессия со смешанными эффектами, обобщённая аддитивная модель).

II. СОДЕРЖАНИЕ УЧЕБНОЙ ДИСЦИПЛИНЫ

1. Работа с распределениями
2. Байесовские статистические методы
3. Кластеризация и смешанные модели
4. Уменьшение размерности
5. Проблемы применения регрессионного анализа

III. ОЦЕНИВАНИЕ

Результирующая оценка выставляется по следующей формуле:

$$O_{\text{итоговая}} = 14 \times \int_{-\frac{1}{10} \times O_{\text{накопленная}}}^{\frac{1}{10} \times O_{\text{экзамен}}} \frac{\exp(\frac{1}{2}(x-1)^2)}{\sqrt{2\pi}} dx + O_{\text{дополнительный балл}}$$

Оценка за курс складывается из оценок за домашние работы ($O_{\text{накопленная}}$), экзамен ($O_{\text{экзамен}}$), а также дополнительный балл ($O_{\text{дополнительный балл}}$), который присуждается студенту, первым указавшим на ошибку в формуле оценивания.
Способ округления всех оценок: арифметический.

IV. ПРИМЕРЫ ОЦЕНОЧНЫХ СРЕДСТВ

- Посчитайте значение правдоподобия распределения $\mathcal{N}(\mu = 22,\ \sigma^2 = 6)$ для двух наблюдений 57 и 43.
- Проведите байесовский апдейт данных $Beta(11, 34)$, используя априорное распределение $Beta(23, 45)$, и посчитайте симметричный 95% байесовский доверильный интервал и 95% интервал максимальной плотности.

V. РЕСУРСЫ

5.1 Основная литература

- Albert, J. Bayesian computation with R / J. Albert. – 2nd ed. – Heidelberg; Dordrecht; London; New York: Springer, 2009. – 298 с. – (Use R!) . ISBN 978-0-387-92297-3.
- Fox, J. An R companion to applied regression / J. Fox, S. Weisberg. – 2nd ed. – Los Angeles [etc.]: SAGE Publications, 2011. – 449 с. ISBN 978-1-412-97514-8.

5.2 Дополнительная литература

- Greenacre, M. Correspondence analysis in practice / M. Greenacre. – 2nd ed. – Boca Raton; London; New York: Chapman & Hall/CRC, 2007. – 280 с. – (Interdisciplinary statistics series) - ISBN 978-1-584-88616-7.
- Doing Bayesian data analysis: a tutorial with R, JAGS, and Stan / J. K. Kruschke. – 2nd ed. – Amsterdam [etc.]: Elsevier, 2015. – 759 с. ISBN 978-0-12-405888-0.
- Gries, S. T. Statistics for linguistics with R: a practical introduction / S. T. Gries. – 2nd rev. ed. – Berlin; Boston: De Gruyter Mouton, 2013. – 359 с. ISBN 978-3-11-030728-3.
- Gries S. T. Ten lectures on quantitative approaches in cognitive linguistics: corpus-linguistic, experimental, and statistical applications / S. T. Gries. – Leiden; Boston: Brill, 2017. – 211 с. – (Distinguished lectures in cognitive linguistics) . ISBN 9789004336216.

5.3 Программное обеспечение

№ п/п	Наименование	Условия доступа
1	R	Свободно распространяемое ПО
2	Rstudio	Свободно распространяемое ПО

5.4 Профессиональные базы данных, информационные справочные системы, интернет-ресурсы (электронные образовательные ресурсы)

№ п/п	Наименование	Условия доступа
1	Страница курса	https://agricolamz.github.io/2019_data_analysis_for_linguists/
2	Introduction to attractor landscapes	https://ncase.me/attractors/

5.5 Материально-техническое обеспечение дисциплины

Учебные аудитории для лекционных занятий по дисциплине обеспечивают использование и демонстрацию тематических иллюстраций, соответствующих программе дисциплины в составе:

- ПЭВМ с доступом в Интернет (операционная система, офисные программы, антивирусные программы);
- мультимедийный проектор с дистанционным управлением.

Учебные аудитории для лабораторных и самостоятельных занятий по дисциплине оснащены ПЭВМ, с возможностью подключения к сети Интернет и доступом к электронной информационно-образовательной среде НИУ ВШЭ.