университета(ТПУ). В рамках2-х факторной дисперсионной моделиВТ (3-х уровневый фактор «ГОД» и 7-и уровневый фактор «ИНСТИТУТ») исследовано влияние факторов «ГОД» и «ИНСТИТУТ» на ВТ. Результаты ВТ ТПУ являются высоко значимо неоднородными по институтам за счет значимого размахасредних баллов-результатов ВТ институтов от 1,4до2,5. Для каждого года выделены однородные (различающиеся незначимо) группы институтов. Результаты ВТ математических знаний студентов ТПУ имеют положительную динамику. Различие результатов ВТ ТПУ по годам оценивается как высоко значимое ($p< 0,0005$) за счет высоко значимого отличия результатов ВТ ТПУ 2011г. (средний балл 1,724) от 2012-2013гг. при незначимом ($p > 0,10$) различии результатов ВТ ТПУ 2012г. (средний балл 2,261) и 2013г. (средний балл 2,268).Оценена динамика качества приема на 1 курс по институтам в широком диапазоне значимости: незначимая в период 2012-2013гг. для одного из семи институтов, для остальных институтов в период 2011-2012гг. значимая положительная, но в период 2012-2013гг. три института имеют положительную незначимую динамику, один имеет отрицательную незначимую динамику, а два – отрицательную статистически значимую динамику.Результаты проведенного статистического анализа могут быть учтены в рамках проходящей реформы высшего образования.

ANALYSIS OF VARIANCE OF OUTCOMES OF ENTERING TESTING OF MATHEMATICAL KNOWLEDGE IN TECHNICAL COLLEGE

Mihalchuk A.A., Arefyev V.P., Filipenko N.M.

National research Tomsk polytechnic university, Tomsk, Russia
(634050, Tomsk, Lenin's avenue, 30), e-mail: aamih @tpu.ru

The statistical analysis of outcomes of entering testing (ET) on the mathematician of a gang 2011-2013 in which students-resident of the first course of seven institutes of Tomsk polytechnic university (TPU) participated is spent. Within the limits of 2 factor dispersing models ET (3 levelof the factor «YEAR» and 7 levelof the factor «INSTITUTE») influence of factors «YEAR» and «INSTITUTE» on BT is investigated. Outcomes of ET TPU are highly significantly inhomogeneous on institutes at the expense of significant scope of mean scores of outcomes ET of institutes from 1,4 to 2,5. For each year are selected homogeneous (differing not significant) groups of institutes. Outcomes ET of mathematical knowledge of studentsTPU have positive dynamics. Distinction of outcomes of ET TPU on years is estimated as highly significant ($p<0,0005$) at the expense of highly significant difference of outcomes of ET TPU 2011г. (mean score 1,724) from 2012-2013гг. at insignificant ($p>0,10$) distinction of outcomes of ET TPU2012г. (mean score 2,261) and 2013г. (mean score 2,268). Dynamics of quality of reception on 1 course on institutes in a wide range of the importance is estimated: insignificant in phase 2012-2013гг. for one of seven institutes, for remaining institutes in phase 2011-2012гг. significant positive, but in phase 2012-2013гг. three institutes have the positive notsignificant dynamics, one has negative not significant dynamics, and two - negative statistically significant dynamics. Outcomes of the spent statistical analysis can be considered within the limits of passing reform of higher education.

ОПЕРЕДЕЛЕНИЕ РЕГИОНА АВТОРА ПО ДАННЫМ ЖИВОГО ЖУРНАЛА

Морозов Е.В., Богданова Д.Н.

ООО «АбиИнфоПоиск», Москва, Россия (127273, Москва, ул. Отрадная, 2Б.6),
e-mail:eugene_m@abbyy.com

В настоящей работе представлен корпус записей русскоязычных блогов с информацией о местоположении автора, а также проведено исследование методов машинного обучения для автоматического определения региона автора. Для создания корпуса использовалась коллекция текстов блоговой платформы Живой Журнал (http://livejournal.com). Регионы авторов были приведены к единому виду, после чего из них были выбраны регионы с наибольшим количеством текстов. Корпус был очищен от выбросов – текстов, не представляющих интереса с точки зрения данного исследования. В данном исследовании были изучены различные наборы признаков, размеры обучающих коллекций и методы машинного обучения. Проведённые эксперименты показали, что большая часть текстов не содержит достаточно информации для определения региональной привязки, однако имеется существенная часть текстов, пригодных для региональной классификации.

GEOGRAPHIC LOCATION PREDICTION IN BLOGS

Morozov E.V., Bogdanova D.N.

LLC «AbiInfoPoisk», Moscow, Russia (127273, 2B.6, Otradnaya, Moscow), e-mail:eugene_m@abbyy.com

This paper presents research on geographical lexical variation detection. We present a new corpus of Russian blogs labeled with geographical information. The corpus was extracted from LiveJournal (http://livejournal.com). Only those blogs that contained enough information about the author were used in the experiments. We have performed outlier detection, and thus, removed spam and other irrelevant data. We have studied various feature sets and performed classification based on Support Vector Machine and Naïve Bayes algorithms. The obtained results show that geographic location prediction is a hard task, and many of the blogs do not contain enough information to determine location of their authors, even though in certain cases accurate classification is possible.