МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ФАКУЛЬТЕТ ПРИКЛАДНОЙ МАТЕМАТИКИ И ИНФОРМАТИКИ

Кафедра математического моделирования и анализа данных

Коляго
Иван Михайлович

АЛГОРИТМЫ АНАЛИЗА И ПРЕДСКАЗАНИЯ РЕЙТИНГОВ ОТЕЛЕЙ НА ОСНОВЕ ТЕКСТОВЫХ ДАННЫХ

Аннотация к дипломной работе

Руководитель от организации:
И. Н. Кравченко
ИООО «ЭПАМ Системз»

Руководитель от кафедры:
канд. физ.-мат. наук,
доцент В. И. Малюгин

Минск, 2016

РЕФЕРАТ

Дипломная работа, 50 с.,11 табл., 14 рис., 20 источников.

Ключевые слова: ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ТЕКСТА, ИНФОРМАЦИОННЫЙ ПОИСК, КЛАССИФИКАЦИЯ ДОКУМЕНТОВ, ИЗВЛЕЧЕНИЕ ИНФОРМАЦИИ, КЛАСТЕРИЗАЦИЯ, АНАЛИЗ ТРЕНДОВ, АНАЛИЗ ТОТАЛЬНОСТИ ТЕКСТА, РАСПОЗНАВАНИЕ ЧАСТЕЙ РЕЧИ, СТЕММИНГ, ТОКЕНИЗАЦИЯ, РАСПОЗНАВАНИЕ ФРАЗ, ИЗВЛЕЧЕНИЕ ОБЪЕКТОВ, ПАРСИНГ

Объект исследований: задачи и проблемы классификации объектов и предсказания соответствующих классов с помощью методов интеллектуального анализа текстовых данных на основе различных моделей.

Цель работы: разработка и исследование моделей и алгоритмов для классификации и предсказания рейтингов отелей на основе текстовых данных.

Методы исследования: методы теории вероятностей и математической статистики, методы численной алгебры, методы программирования, методы анализа больших данных, методы предобработки и классификации текстовых данных, включая: метод ближайших соседей, метод опорных векторов, модель классификации случайного леса (random forest classifier) и модель классификации максимальной энтропии (maximum entropy classifier).

Результаты исследования: разработаны программные средства для классификации и предсказания рейтингов отелей на основе анализа текстовых данных в виде отзывов с использованием методов и алгоритмов интеллектуального текстового анализа.

Область применения: сбор и накопление данных, социологические исследования, маркетинговые исследования.

Актуальность: результаты работы актуальны для маркетинговых исследований, основанных на социальных данных, собранных в сети интернет.

РЭФЕРАТ

Дыпломная праца, 50 с., 11 табл., 14 рыс., 20 крыніц.

Ключавыя словы: ІНТЭЛЕКТУАЛЬНЫ АНАЛІЗ ТЭКСТАЎ, ІНФАРМАЦЫЙНЫ ПОШУК, КЛАСІФІКАЦЫЯ ДАКУМЕНТАЎ, ВЫМАННЕ ІНФАРМАЦЫІ, КЛАСТАРЫЗАЦЫЯ, АНАЛІЗ ТРЭНДАЎ, АНАЛІЗ ТАТАЛЬНАСЦІ ТЭКСТУ, РАСПАЗНАННЕ ЧАСЦІН МОВЫ, СТЕММИНГ, ТОКЕНИЗАЦИЯ, РАСПАЗНАВАННЕ ФРАЗ, ВЫМАННЕ АБ'ЕКТАЎ, ПАРСІНГ

Аб'ект даследаванняў: задачы і праблемы класіфікацыі аб'ектаў і прадказанні адпаведных класаў з дапамогай метадаў інтэлектуальнага аналізу тэкставых дадзеных на аснове розных мадэляў.

Мэта працы: распрацоўка і даследаванне мадэляў і алгарытмаў для класіфікацыі і прадказання рэйтынгаў гатэляў на аснове тэкставых дадзеных.

Метады даследавання: метады тэорыі верагоднасцяў і матэматычнай статыстыкі, метады лікавай алгебры, метады праграмавання, метады аналізу вялікіх дадзеных, метады предобработки і класіфікацыі тэкставых дадзеных, уключаючы: метад бліжэйшых суседзяў, метад апорных вектараў, мадэль класіфікацыі выпадковага лесу (random forest classifier) і мадэль класіфікацыі максімальнай энтрапіі (maximum entropy classifier).

Вынікі даследавання: распрацаваны праграмныя сродкі для класіфікацыі і прадказанні рэйтынгаў гатэляў на аснове аналізу тэкставых дадзеных у выглядзе водгукаў з выкарыстаннем метадаў і алгарытмаў інтэлектуальнага тэкставага аналізу.

Вобласць ужывання: збор і назапашванне дадзеных, сацыялагічныя даследаванні, маркетынгавыя даследаванні.

Актуальнасць: вынікі працы актуальныя для маркетынгавых даследаванняў, заснаваных на сацыяльных дадзеных, сабраных у сеткі інтэрнэт.

THE ABSTRACT

Diploma paper, 50 p., 11 tables, 14 pic., 20 sources

Key words: INTELLECTUAL TEXT ANALYSIS, INFORMATION SEARCH, DOCUMENTS CLASSIFICATION, INFORMATION EXTRACTION, CLUSTERING, TREND ANALYSIS, ANALYSIS OF TEXT TOTALITY, RECOGNITION OF SPEECH PARTS, STEMMING, TOKENIZATION, PHRASE RECOGNITION, OBJECTS RETRIEVING, PARSING

The object of the research: issues and problems of the object classification and predictions of special classes by using methods of intellectual data analysis based on different models.

The purpose of the work: production and research of the models and algorithms for classification and hotel ratings predictions based on text data.

The methods of the research: methods of calculus probability and mathematical statistics, methods of numerical algebra, programming techniques, methods of large data analysis, and methods of pre-processing and text data classification including: the method of nearest neighbors, the method of support vectors, the model of random forest classifier and the model of maximum entropy classifier.

The results of the research: software tools for classification and hotels rating predictions based on the text data analysis in the form of reviews with the usage of methods and algorithms of intellectual text analysis are designed.

The field of application: collection and data storage, sociological research, marketing research.

Actuality: the results of the research are relevant for marketing research based on social data collected on the Internet.