CS209A Project, Spring 2019

Description

With the increasing popularity of online shopping, it has brought with massive online consumers and the rapid growth of data on merchandise information. Many large-scale e-commerce system or online-shopping sites provide functions for commodity evaluation. A customer can make a review on a product or service that he has purchased and used, or had experience with.

Before buying a product or service, many customers often look at the quantity of a product sold and its evaluating stars, and may also carefully read the evaluating comments made by other customers. So, customer reviews are very important for evaluating commodities.

The goal of the project is data analysis based on customer reviews of some commodities sold online. You are asked to complish the following tasks:

(1) Collect customer reviews of the corresponding commodities on seleted commercial websites;

(2) Preprocess the collected data, including deleting the duplicate data and short trivil reviews, segmenting words and calculating word frequencies.

(3) Show the word segmentation and word frequency statistics on GUI created with JavaFX. Provide an easy way to select some word as classification markers of a certain class.

(4) Classify and quantify the grade of target commodities.

(5) Store the evaluation results of target commodities and update the commodities evaluation periodically according to the setting period.

Project Requirements

(1) You are asked to write a JavaFX graphical user interface displaying your information, including the word frequency statistics and the grades of target commodities and so on. You should also design an intuitive way to guide a user to choose some words as classification markers.

(2) You are asked to write a Web Crawling Program to collect customer reviews and update periodically.

(3) The number of commodities should be at least 3. You may analyze more commodities by category, but the number of categories should be at least 3 and you should also analyze the correlation among them. Correspondingly, you could get higher scores for category analysis instead of individual commodity analysis.

(4) You should store the original data, the evaluation results and responds classification markers into a .json or .xml file(or files). When open your application next time, your last analysis results will be displayed on screen.

(5) Every team should prepare a PPT to explain how you process the data and how to do the data visualization. In your PPT, you should introduce your team members with their contribution rate for each member.

Task 1. Data Collecting (20%)

You are asked to write a Web Crawling Program to collect customer reviews and update periodically from e-commerce websites. JD, Taobao, Amazon, dianping, meituan are good choices. But PAY ATTENTION that that most of EC websites have ANTI-CRAWLER technique, so don't crawl too much data at one time.

Task 2. Data processing (35%)

You are asked to process the data that you crawled.

(1) Preprocessing, including deleting duplicate data and short trivil reviews.

(2) Word segmentation. ICTCLAS is recommended for word segmentation.

(3) You should select some reasonable classification markers and tell why in your PPT.

(4) Design a reasonable method to grade the commodities. If you analyze commodities by category, you should not only analyze the commodity itself, but also the relationship between categories.

Task 3. Data visualization (35%)

(1) At least a TableView, filtering on three different criteria.

(2) At least two charts or an advanced chart.

(3) Animation is a better option.

Here is an excellent example:

https://www.washingtonpost.com/graphics/national/eclipse/?noredirect=on&utm_term=.5a6417cd90f2

Some other examples, please see the images and the animations in the same zip with specification.

Task 4. Data Persistence (10%)

You are asked to store the original data, the evaluation results and responds classification markers into a .json or .xml file(or files).

Reference

语义分析与词频统计相结合的中文文本相似度量方法研究_华秀丽.pdf

Be creative and have fun!