Toury

By Senior Drivers
Xinyue Pan
Feng Xiong
Jiaxing Yan
Zeling Zhang

Introduction

We find people may not get essaence of tourism data efficently from internet. Main stream tourism websites also skip data may be meaningful for tourisim. Toruy project is trying to dig information that general tourisim websites do not cover.

Our Approaches

Principal Component Analysis(PCA): Figure out weights of each feature in data. We do not use this for dimensionality reduction.
K-Means Clustering: Decide rank of countries in five aspect via classifying each features.
Rapid Automatic Keyword Extraction (RAKE): Extract keywords by its frequency, phrase length and other features. Rapid and effective for short texts. We use Natural Language Toolkit (NLTK) with it, and bootstrap synthetic results of various settings of RAKE.
N-Gram IDF: Get keywords by its frequency in current context versus the frequency in all documents. It also consider the frequency of its component words. Slow but outperform RAKE on long reviews. It can extract all possible keywords.

Data

Macro Data Analysis: Five data categories of 150 contries, contianing 30 datasets. Download from knoema.com, feature number depends on datasets, ranging from 5 to 8. All values are numerical.
Reviews: 2.43GB comments and account information downloaded from insideairbnb.com. Data are texts, time and numerical in csv format. Data cover 5 cities since 2012.

Experiments and Results

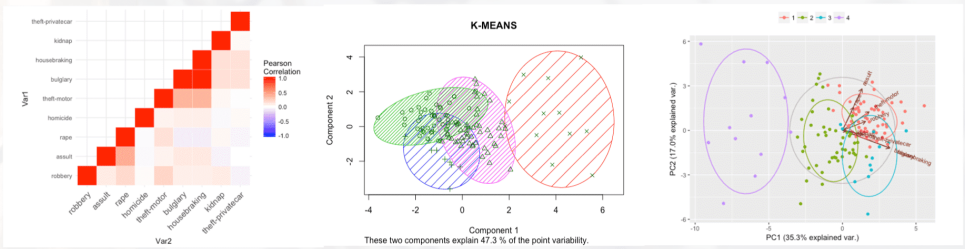

Figure 1-3: Biplot, K-Mean and Correlation Heatmap of crime data of 150 countries.
Results are data for radar charts

Figure 4-6: Radar charts of five aspects of pairs of countries. Macro comprasion shows potetially impactors of torism and travelers decisions.

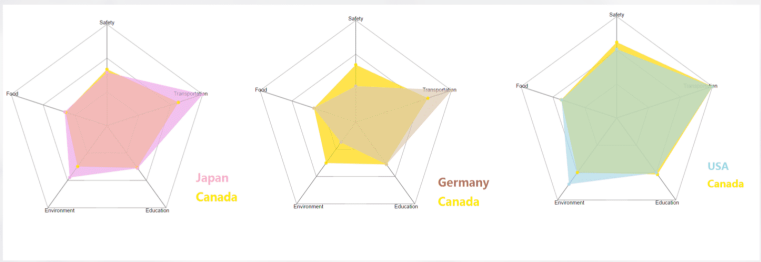

Table 1: Performance comparision of RAKE and N-gram IDF for keywords extraction

Performance comparison

Algorithm name	Runtime(3000 samples / round)	Result variation	Result amount
Rake	34s	Large, unstable	Small
N-gram IDF	6m37s	minor	All possible results

Figure 7-10: Word Cloud of keywords extracted by N-gram IDF, for each season. (2012-2017, Toronto)