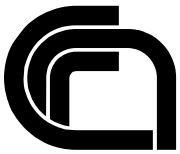

Consiglio Nazionale delle Ricerche

A-ZARS: A Software for Data Analysis

A. Minissale, A. Lo Duca, A. Marchetti

IIT B4-08/2020

Nota Interna

Ottobre 2020

Istituto di Informatica e Telematica

A-ZARS

A Software for Data Analysis

Alessia Minissale*, Angelica Lo Duca+, Andrea Marchetti+

*Università di Pisa

a.minissale@studenti.unipi.it

+ Istituto di Informatica e Telematica, Consiglio Nazionale delle Ricerche

angelica.loduca@iit.cnr.it

andrea.marchetti@iit.cnr.it

Table of Contents

Abstract

A-Zars is a software completely written in Python for the analysis of reviews extracted from Online Travel Agencies (OTAs) through the ZARS software. A-ZARS allows you to search for a service within the review using Machine Learning algorithms, perform sentiment analysis of the reviews and add the Geonames code to the locations of origin of the authors of the reviews. A-Zars is released under the GNU General Public License v 3.0.

Keywords

Data Analysis, Machine Learning, Sentiment Analysis

1. Introduction

The software implemented for data analysis, A-ZARS (Zars for Analysis)[1] aims to analyze the reviews extracted from OTA. Specifically, this software takes in the reviews in the format extracted from ZARS[2] and allows you to perform the following analyzes:

1) Search for a service within the review using Machine Learning algorithms
2) Sentiment analysis of the reviews
3) Adding the Geonames code to the locations of origin of the authors of the reviews.

2. Audience

The use of the software can be intended for two types of users:

- Accommodation facilities: the execution of the software can be useful when you intend to improve the services offered according to the preferences of the traveler. Furthermore, investing in services increases *brand reputation*, increases perceived value and triggers word of mouth. It has significant strategic implications as it allows to improve the effectiveness of the segmentation and, consequently, the effectiveness of the operational marketing action.
- Operators in the tourism sector: the software has been designed and implemented for data collection and, above all, to study and predict through data analysis, the behavior and interest of the customer when choosing the structure in which to stay.

[1] https://github.com/alessiamns/A-ZARS
[2] https://github.com/alessiamns/ZARS

3. System requirements

- Python 3 (https://www.python.org/downloads/)
- Xampp (https://www.apachefriends.org/download.html)
- Chrome Driver (https://chromedriver.chromium.org/)
- Jupyter (https://jupyter.org/)

Table 1. Required Libraries

Package Name	Link	License
Selenium	https://www.selenium.dev/downloads/	Free
MySQL connector	https://www.mysql.com/it/products/connector/	Free
Pandas	https: // pandas.pydata.org/	Free
Numpy	https://numpy.org/	Free
Scikit-learn	https://scikit-learn.org/	Free
Googletrans	https://py-googletrans.readthedocs.io/en/latest/	Free
Langdetect	https://pypi.org/project/langdetect/	Free

4. Architecture

The software (see Figure 1) consists of a series of files informat .json and .csv extracted from the SQL server where the database is located (1). The extracted data is displayed in the form of tables (2) and, subsequently, the procedure applied is the following: language recognition, translation of reviews into English (3), identification of the nation and association of the unique code geonames (4), training of the model (5) and final output (6).

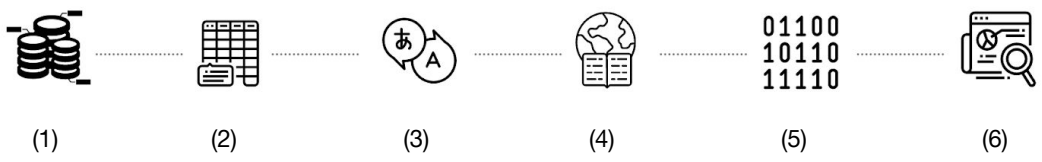

Figure 1. Software architecture

This data is analyzed and processed in notebooks with extension `.ipynb`. The A-ZARS software consists of the following modules:

1) *Language Analysis*: this module allows you to first associate the language with each review and, subsequently, implements the automatic translation of the reviews and services in English
2) *Dataset Enrichment*: allows you to enrich the starting nationalities with the unique code extracted from Geonames
3) *Data Analysis*: allows you to extract a service of your choice from the reviews. In the example, breakfast is considered as a service (`Breakfast_analysis.ipynb`)

5. License

The software is released under the GNU General Public License v3.0[3]. The GNU General Public License is a free copyleft license for free software. It guarantees end users, such as organizations, businesses or individuals, to use, share and even modify the software.

6. Using the software

6.1 Installation

- Download zip from GitHub (https://github.com/alessiamns/A-ZARS)
- Starting Jupyter and uploading the notebooks

6.2 Running the software

It is necessary to download the folder from the zip archive which will contain the files in `.ipynb` to start with `Jupyter Notebook`. In detail, the operations performed by the scripts follow.

DatasetEnrichment.ipynb

[3] https://www.gnu.org/licenses/gpl-3.0.html

The script has been implemented on the Geonames site[4]. This is an automated operation (web scraping) that allows you to search for the name of the country and extract the Geonames code associated with it (see Figure 2), and then insert it in the list created (column `country` containing the country associated with the review) .

	country	geonames_id
0	Regno Unito	2635167
1	Italia	3175395
2	Spagna	2510769
3	Germania	2921044
4	Stati Uniti	6255150
...	...	...
121	Kosovo	831053
122	Palestina	6254930
123	Bahamas	3572887
124	Leicester	2644668
125	Azerbaigian	587116

Figure 2. Geonames code association

The creation of the table is followed by the updating of the reviews table, with the creation of a new column `geonames_id` (see Figure 3) which will contain, respectively, the IDs associated with the corresponding countries (the condition to be satisfied is equality between the nation contained in the country column and the nation of each review: the corresponding ID will be associated with this).

	Name	City	Rating	Review	Hometown	Date_of_stay	Trip_type	geonames_id
0	Delle Vittorie Luxury Rooms & Suites	Palermo	5	Excellent facilities in a spacious luxurious l...	Regno Unito	settembre 2019		2635167
1	Bed and Breakfast Il Rifugio	Caltanissetta	5	location molto comoda...Ti permette di postegg...	Italia	settembre 2019	Ha viaggiato con amici	3175395
2	Delle Vittorie Luxury Rooms & Suites	Palermo	5	estuvimos alojados 4 noches y no puedo poner n...	Spagna	febbraio 2020		2510769
3	Garibaldi R&B	Messina	5	Graziosissimo R&B con vista sullo stretto. Ott...	Italia	gennaio 2020	Ha viaggiato per affari	3175395
4	Bed and Breakfast Il Rifugio	Caltanissetta	5	Arrivato in questo B&B quasi X caso e cioè gra...	Italia	giugno 2019	Ha viaggiato da solo	3175395
...	...	...	...	...	...	...	...	...
45343	Sabbinirica	Ragusa	5	It is a very nice hotel. The hotel is very new...	Germania	novembre 2017	Ha viaggiato con la famiglia	2921044
45344	Itria Palace	Ragusa	4	Il mio fidanzato ed io abbiamo pernottato in q...	Italia	agosto 2020	Ha viaggiato in coppia	3175395
45345	Itria Palace	Ragusa	5	Partiamo dalla gentilezza e la disponibilità d...	Italia	agosto 2020	Ha viaggiato in coppia	3175395
45346	Itria Palace	Ragusa	4	Bellissimo hotel, grande gentilezza e posizion...	Italia	luglio 2020		3175395
45347	Itria Palace	Ragusa	5	Una struttura accogliente con annesso parchegg...	Italia	agosto 2020	Ha viaggiato in coppia	3175395

Figure 3. Reviews table updated with geonames codes

[4] https://www.geonames.org

LanguageAnalysis.ipynb

The next step identifies the language of the reviews. Through the use of library langdetect Python, which supports 55 languages (ISO 639-1 codes[5]), it was possible to identify the language of the reviews, giving each review the respective language identification code. The text of the review is taken as input and, through langdetect, the language in which it is written is identified. In total, 24 languages were identified. The result is an update of the reviews table (see Figure 4) with the addition of the column Language which contains, respectively, the ISO 639-1 codes associated with the language of each review.

	Name	City	Rating	Review	Hometown	Date_of_stay	Trip_type	geonames_id	Language
0	Delle Vittorie Luxury Rooms & Suites	Palermo	5	Excellent facilities in a spacious luxurious l...	Regno Unito	settembre 2019		2635167	en
1	Bed and Breakfast Il Rifugio	Caltanissetta	5	location molto comoda...Ti permette di postegg...	Italia	settembre 2019	Ha viaggiato con amici	3175395	it
2	Delle Vittorie Luxury Rooms & Suites	Palermo	5	estuvimos alojados 4 noches y no puedo poner n...	Spagna	febbraio 2020		2510769	es
3	Garibaldi R&B	Messina	5	Graziosissimo R&B con vista sullo stretto. Ott...	Italia	gennaio 2020	Ha viaggiato per affari	3175395	it
4	Bed and Breakfast Il Rifugio	Caltanissetta	5	Arrivato in questo B&B quasi X caso e cioè gra...	Italia	giugno 2019	Ha viaggiato da solo	3175395	it
...	...	...	...	...	...	...	...	...	...
45343	Sabbinirica	Ragusa	5	It is a very nice hotel. The hotel is very new...	Germania	novembre 2017	Ha viaggiato con la famiglia	2921044	en
45344	Itria Palace	Ragusa	4	Il mio fidanzato ed io abbiamo pernottato in q...	Italia	agosto 2020	Ha viaggiato in coppia	3175395	it
45345	Itria Palace	Ragusa	5	Partiamo dalla gentilezza e la disponibilità d...	Italia	agosto 2020	Ha viaggiato in coppia	3175395	it
45346	Itria Palace	Ragusa	4	Bellissimo hotel, grande gentilezza e posizion...	Italia	luglio 2020		3175395	it
45347	Itria Palace	Ragusa	5	Una struttura accogliente con annesso parchegg...	Italia	agosto 2020	Ha viaggiato in coppia	3175395	it

Figure 4. Review table update with Language column

For the subsequent data analysis, including a comparison with the services, it was necessary to translate all the reviews available in English. The reviews extracted are written in 24 languages: these are stored in a table in which each review corresponds to an ID, the nationality of origin, the capital where the accommodation is located, the rating, the type of trip and the date of living room. The text of the reviews is translated into English using the Python library `Googletrans`, which implements the Google Translate API. The reviews in English will be saved in a new DataFrame (see Figure 5) which will contain, in correspondence with the review in the original language, the same translated into English, saved in a new column (`review_en`).

[5] https://en.wikipedia.org/wiki/List_of_ISO_639-1_codes

	index	lang	review	review_en	hometown	rating
0	1	it	location molto comoda...Ti permette di postegg...	very convenient location ... It allows you to ...	Italia	5
1	3	it	Graziosissimo R&B con vista sullo stretto. Ott...	Very nice R&B overlooking the strait. Excellen...	Italia	5
2	4	it	Arrivato in questo B&B quasi X caso e cioè gra...	Arriving at this B&B almost X chance, that is,...	Italia	5
3	6	it	Colorato e curato nei minimi dettagli. Le stan...	Colorful and with attention to the smallest de...	Italia	5
4	7	it	Al centro di Trapani c'è questa bellissima str...	In the center of Trapani there is this beautif...	Italia	5
5	9	it	durante il ns tour ci siamo fermati per una no...	during our tour we stopped for a night in this...	Italia	4
6	12	it	Ho soggiornato per la seconda volta. Una confe...	I stayed for the second time. A confirmation: ...	Italia	5
7	15	it	Abbiamo alloggiato qui su consiglio di amici. ...	We stayed here on the recommendation of friend...	Italia	5
8	17	it	Insieme alla mia famiglia, abbiamo passato una...	Together with my family, we spent one night in...	Italia	5
9	18	it	Bellissima struttura in centro a Trapani, che ...	Beautiful structure in the center of Trapani, ...	Italia	5

Figure 5. Example table with reviews translated into English

The table created manually for the annotations (see Figure 6) consists of the following columns:

- `index`: ID of the review (extracted from the original table)
- `lang`: original language of the review (ISO-codes)
- `review`: text of the review translated into English
- `amenity_1,..., 6`: hotel services found (annotated manually)
- `hometown`: corresponding origin (extracted from the original table)
- `rating`: score assigned to the review

	index	lang	review	amenity_1	amenity_2	amenity_3	amenity_4	amenity_5	amenity_6	hometown	rating
69	145	it	B&B located near the port and convenient to re...	Breakfast available	parking area					Italia	4
70	147	it	Brand new rooms, satisfactory cleaning, brand ...	Wifi	Air conditioning	parking area				Italia	5
71	149	it	I have been for three days on busy occasions. ...	Breakfast available						Italia	5
102	3	it	Very nice R&B overlooking the strait. Excellen...	Wifi	Breakfast available					Italia	5
103	7	it	In the center of Trapani there is this beautif...	Breakfast available	Roof terrace					Italia	5
104	17	it	Together with my family, we spent one night in...	Coffee bar	Breakfast available					Italia	5
105	31	it	With my girlfriend we decided to visit Messina...	Air conditioning	Breakfast available	Check-in 24 hours				Italia	4
106	35	it	Structure located in the historic center of Tr...	Coffee bar	Breakfast available					Italia	5
107	56	it	I only stayed one night at this property and I...	parking area		Roof terrace				Italia	4
108	58	it	We stayed on the banks of the strait staying a...	Buffet breakfast	Wifi					Italia	5

Figure 6. Excerpt from annotation table

DataAnalysis.ipynb

An open source Machine Learning library was used for the Python programming language, Scikit-learn: it contains classification, regression and clustering algorithms and support vector machines, logistic regression, Bayesian classifier, k-mean and DBSCAN, and is designed to work with the NumPy and SciPy libraries.

A standard practice when training supervised learning algorithms is to separate some data from the training set to evaluate the accuracy of the classifier. Scikit-learn includes a handy feature to randomly split training data into subsets (`train_test_split`).
As previously mentioned, the dataset consists of 300 reviews, of which 75% will be used for training and the remaining 25% for testing.

Train and test split

```
from sklearn.model_selection import train_test_split

X = df['review'].to_list()
y

X_train, X_test, y_train, y_test = train_test_split(X,y, test_size=0.25)
```

To learn the classification rules, the sklearn function is applied to the reviews `CountVectorizer()`. The CountVectorizer class produces a *bag-of-words* from a string or file. By default, CountVectorizer converts characters in documents to lowercase, tokenizes documents, and builds a vocabulary. Tokenization is the process of splitting a string into tokens or meaningful sequences of characters. Tokens are often words, but they can also be shorter sequences that include punctuation characters and affixes. The CountVectorizer class tokenizes text using a regular expression that splits strings on spaces and extracts character sequences two or more characters long.

CountVectorizer

```
from sklearn.feature_extraction.text import CountVectorizer

vect = CountVectorizer()
X_train_vect = vect.fit_transform(X_train)
X_train_vect
```

```
<225x2915 sparse matrix of type '<class 'numpy.int64'>'
        with 15227 stored elements in Compressed Sparse Row format>
```

This representation is stored in a SciPy sparse matrix, where each row corresponds to a document and each column to a word from our *training vocabulary*.

`X_train_vect` is the matrix on which Naïve Bayes Bernoulli model is trained.

Naïve Bayes methods are a set of supervised learning algorithms based on the application of Bayes' theorem with the 'naïve' assumption of the conditional independence between each pair of characteristics given the value of the class variable.[6]

Unlike the Multinomial model, which takes into account the frequency of words in the document, the Bernoulli model not only does not care about the frequency, but shows how common the word or sequence of words is in the entire collection.

The assumption is that the data follow a multivariate Bernoulli distribution, in which each characteristic is a binary characteristic, that is, it is considered whether the word is present or not present. Usually, it is quite common in text documents to use the multinomial Naïve Bayes, but there are cases where one tends to use Bernoulli, especially if one wants to somehow say that the frequency is irrelevant and is only the presence or absence of a word is the thing that interests most.

Model

```
from sklearn.naive_bayes import BernoulliNB

model = BernoulliNB()
model.fit(X_train_vect,y_train)
```

```
BernoulliNB()
```

After training the model, you can predict the label for a new dataset using the *predict function*. To the function `predict` the test data that has been obtained are passed as parameters.

Predict

```
X_test_vect = vect.transform(X_test)

y_pred = model.predict(X_test_vect)
```

The model is then applied to the 4000 reviews.

[6] https://scikit-learn.org/stable/modules/naive_bayes.html

Predict for 4000 reviews

```
df = pd.read_json('data/pred/reviews_for_predict.json')

X = df['review_en']
X_vect = vect.transform(X)

y_pred = model.predict(X_vect)
```

The generated DataFrame will be composed of the following columns:

- `index`: index that identifies the review within the original table
- `nation`: nationality of origin of the traveler
- `Breakfast Y/N`: binary value that corresponds to the presence/absence of the service within the review
- `Rating`: rating attributed to the review

Result

```
dataset = pd.DataFrame({'nation' : df['hometown'],
                        'Breakfast Y/N':y_pred,
                        'Rating': df['rating']})
dataset.reset_index()
```

	index	nation	Breakfast Y/N	Rating
0	11191	Italia	1	4
1	11192	Italia	1	4
2	11193	Italia	1	3
3	11194	Italia	1	1
4	11195	Italia	0	5
...	...	...	...	...
3995	7697	Argentina	1	5
3996	7702	Argentina	0	4
3997	7703	Argentina	1	5
3998	7707	Argentina	1	4
3999	7708	Argentina	1	4

4000 rows × 4 columns