For office use only

T1 ____________

T2 ____________

T3 ____________

T4 ____________

Team Control Number

88022

Problem Chosen

B

For office use only

F1 ____________

F2 ____________

F3 ____________

F4 ____________

2018
MCM/ICM
Summary Sheet

How many languages

Abstract

This paper establishes the XXX model to predict the trends of global languages and provide the best recommendations about the location of international office for a multinational company.

Firstly, in order to predict how languages of the world may vary over time, we build a force model and consider the influence a country gives to the other as a force, after getting the resultant force of this model, we can find how the ratio of languages of a country will change in the future. Using this model, we predict the trends of native speakers and total language speaks in the next 50 years and find that ... .

Secondly, based on the result our model produces, we use K-means algorithm to help us locate the best place for the company's global offices, using data collected from Twitter sampling. And we compare our recommendations with the global office chosen by world top 500 to verify our method and get great results.

Thirdly, we studied how would our model's results change with the type of our client company....

Keywords: Force model, K-means clustering

Contents

1 Introduction

1.1 Background

Half of the world's population speak one of ten languages as their native language, although there are nearly 7,000 languages spoken on the earth. But with the influence of government, culture, economy and the impact of globalization, popularity of languages may change over the time.

A COO of a multinational service company wants to know the trends of global languages and locations for this company's new international offices. So we designed a multiple regression model and applied a fuzzy algorithm to help this company make the choice.

1.2 Restatement of the Problem

We are required to build a model to predict the trends of global languages, including the number of speakers and the geographic distribution change of the top-10 languages. Then we need to decide where the new international offices of the company should be located, or take efforts to reduce the number of offices.

2 Basic Assumptions

2.1 Assumption 1.

We assume that the land area of every country doesn't change during the period of time we study.

2.2 Assumption 2.

We ignore unpredictable or low-probablity events that may cause great impact to languages trends.

3 Analysis of the Problem

We consider the distribution of languages as the output of a function related to multiple factors, such as GDP, immigrants, population, imports, exports, and etc. These factors not only affect a country's language distribution, but also have an effect on other countries' languages.

So we build the model like this, assume ...

4 Models and Methodology

4.1 Time Series Prediction

In order to how the factors we mentioned above will change in the next 50 years, we use the well-known ARIMA model to predict them. ARIMA model, i.e autoregressive integrated moving average model, is a widely used method for predicting time series. Considering these factors change over the time, and in order to simplify this problem, we regard them as factors only related to time. The ARIMA model can be represented as following form.

$$\left(1-\sum_{i=1}^{p}\sigma_i L^i\right)(1-L)^d X_t = \left(1+\sum_{i=1}^{q}\theta_i L^i\right)\epsilon_t \tag{1}$$

In the equation above, L stands for Lag Operator, d∈Z, d>0.

Considering these factors we use vary randomly and are related to many other factors, so they are not stationary variables and can not be directly used int ARIMA model. So we calculate the difference of factors we study and apply ARIMA model on them. After getting the forecast result we calculate the accumulation to restore prediction result. Take the GDP prediction for an example, we collect the GDP of countries from 2002 to 2016, calculate the difference of adjacent years and use the $auto.arima$ model of R language to help us choose the best parameters for ARIMA model. After this ARIMA model gives us the prediction of how the difference of a country's GDP will change in the next 50 years, we calculate the accumulation of this prediction values and regard the final sum as our prediction on this country's GDP.

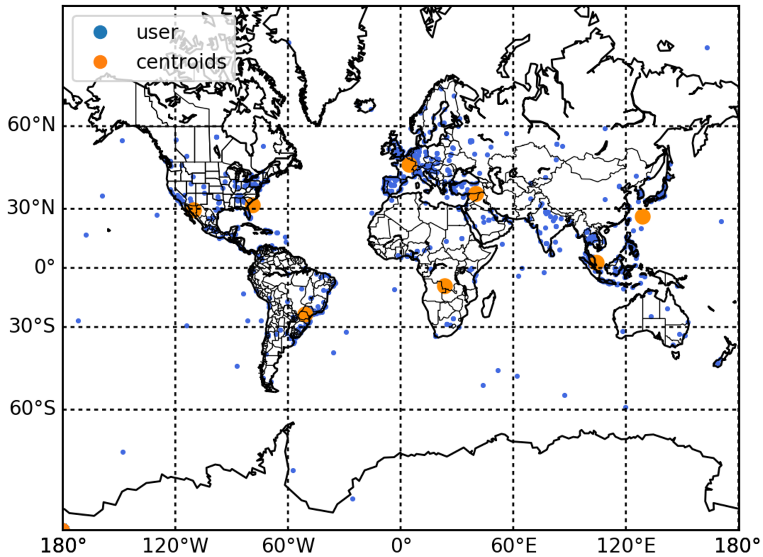
4.2 K-Means

4.2.1 Short Term Recommendation

In order to get the best locations for this company's six new offices, we take the power of social network into consideration. Concretely, we use data sampled in Twitter for 24 hours, which we think can represent the social life of people around the world well. The data sampled from Twitter API gives us the information about the locations of uploaders and the language of tweets and origin language of uploaders when he or she registered Twitter.

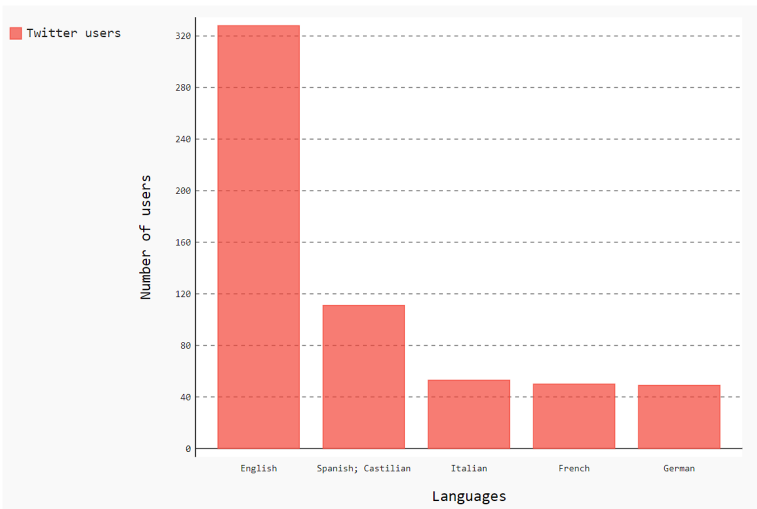
Considering this company is a multinational service company and needs to be more multinational, the new offices of it should be close to its clients or information. So we mark the locations of tweets uploaders on a world map and use K-means algorithm to find centroids of different clusters of uploaders.

Because of the need for multi-language speakers of this company, it is important to consider the language environment of these new offices. To simplify this problem, we regard the uploaders whose tweets use different language from his or her registration language as a potential polyglot, so the language environment around them might be more open and attactive. Based on this assumption we pick those "polyglots" from our data and apply K-means algorithm on them. After setting K to be equal to 8 (This company already has two offices on Shanghai and New York, so the total offices it has will be 8), we got the following results.



Our result shows all eight locations that is appropriate to set up new offices. Apart from the two offices that already exist in Shanghai and New York (which may not be exactly on its true place on this map, but in consideration of sampling errors, we think it is still reasonable to regard the points lies on east coast of North America and China to be New York and Shanghai), we can still see the other six new offices. They are located in San Francisco, São Paulo, Paris, Jerusalem, Kinshasa and Singapore. It is shown that they are nearly located in seperated continents on earth, which is in accordence with the need to get a more open language environment.

In terms of what languages will be spoken in these six new offices, we still use the data from Twitter to help us decide. We collect tweets uploaders near the six new offices and count the types of languages of these tweets. Let us use the office in Paris as an example, we calculate what types of languages are used by uploaders around Paris and how many tweets are uploaded using corresponding language. The result can be shown as follows.



So languages spoken in the office in Paris might be English, Spanish, Italian, French and German. And here are the languages spoken in other new offices.

- Office in San Francisco: English, Spanish, French, Japanese and Portuguese.
- Office in São Paulo: Portuguese, Spanish, English, Japanese and Finnish.
- Office in Jerusalem: Turkish, English, Russian, Arabic and Japanese.
- Office in Kinshasa: English, Spanish, French, Portuguese and Afrikaans.
- Office in Singapore: Chinese Mandarin, English, Indonesian, Thai and Spanish.

The K-means model we use is from $Python$. It uses an effective k-means++ algorithm to speed up the total clustering process. The intuition of k-means++ algorithm can be interpreted as follows: when choosing the next centroid, always choose the centroid that is far from all centroids that have already been chosen. This meets with the desire to be more international because it disperses the new offices and make them distributed globally.

4.2.2 Long Term Recommendation

4.2.3 Less Offices

5 The Model Results

6 Validating the Model

7 Conclusions

8 A Summary

9 Strengths and weaknesses

9.1 Strengths

- **Applies widely**
  This system can be used for many types of airplanes, and it also solves the interference during the procedure of the boarding airplane,as described above we can get to the optimization boarding time.We also know that all the service is automate.
- **Improve the quality of the airport service**
  Balancing the cost of the cost and the benefit, it will bring in more convenient for airport and passengers.It also saves many human resources for the airline.
-

9.2 Weaknesses

- ...
- ...

References

[1] D. E. KNUTH The TEXbook the American Mathematical Society and Addison-Wesley Publishing Company , 1984-1986.

[2] Lamport, Leslie, LATEX: “ A Document Preparation System ”, Addison-Wesley Publishing Company, 1986.

Appendices

Appendix A First appendix

Here are simulation programmes we used in our model as follow.

Appendix B Second appendix