Country Prediction on Social Media using Machine Learning

Taha Küçükkatırcı, Yaşar Fatih Enes Yalçın
Advisor: Ali Taylan Cemgil

Introduction & Motivation

The aim of our project is **to predict home country** of social media users who do not specify it explicitly. We did our work with a basic greedy approach and two probabilistic approaches.

On the figure, we can make inferences that the user is Turkish. The cues are **screen name** and the **language** of the bio of the user.

Knowing locations of users may help organizations to focus on more promising target audience. This problem **can be extended** to other prediction problems with **similar missing information** structure.

Figure 1: Example Twitter account

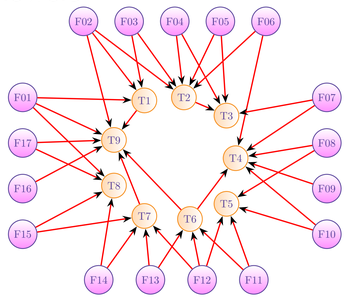
Data Format

We pursue to estimate the location of **target users**. For each user in training data, we store only the following information as node features:

- Unique ID & Screen name
- Location string
- Language
- Time zone
- Followers list (only for t.u.)

In our training set, there are approximately 10 million users with above-mentioned features. In our test set, there are 550 people which are followers of the OpenMaker Twitter account.

Figure 2: Target Users & Followers



Deterministic Greedy Approach

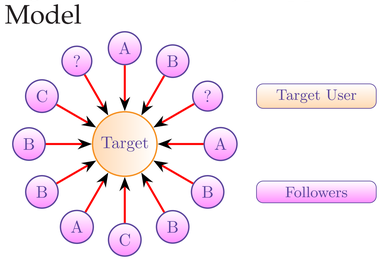
This algorithm traverses all followers of a target user and determines the most frequent country among followers.

Frequencies of countries:

- A = 3
- B = 5
- C = 2

So that, the algorithm picks the country B as a prediction.

Figure 3: Greedy Approach Model



Naïve Bayes

By the **naïve assumption of independence**, the simplified version of the Bayesian relation is:

$$P(y|x_1,...,x_n) \propto P(y)\prod_{i=1}^{n} P(x_i|y) \quad (1)$$

Applying Naïve Bayes to our problem

We determined three key properties of users as features:

- Language (l)
- Name (n)
- Timezone (t)

So, mathematically the problem looks like the following:

$$P(c \mid l,n,t) \propto P(c)P(l \mid c)P(n \mid c)P(t \mid c) \quad (2)$$

where c denotes the country to predict.

The **likelihood probabilities** in Eq. 2 are found with the help of our user database not including the target users or their followers.

$$P(l_i \mid c) = \frac{count(l_i,c)}{\sum_{j=1}^{n_l} count(l_j,c)} \quad (3)$$

where l_i denotes the language whose probability given country c is computed and n_l denotes the total number of languages. The other two likelihood probabilities are computed similarly.

Gibbs Sampling

Gibbs Sampling is an **MCMC** algorithm. In this approach, we used information of neighbors. For example, if a neighbor is from *Country A*, the target user is most probably from *Country A* but he might also be from other countries too. We have created a **compatibility matrix** which indicates the probability of a country given the neighbors' country.

$$M_{ij} = \begin{cases} kp, & \text{if i = j for some k} \geq 1 \\ p, & \text{otherwise} \end{cases}$$

Now we have everything to **iterate over the network** and **assign drawn countries** to target users.

$$P(c_i|tu) = \psi(c_i|tu_{lang}, tu_{timezone}) \prod_{neighbors_{tu}:n} \phi(c_i|n_{country}) \quad (4)$$

On the RHS of Eq. 4, ψ is just Naïve Bayes and the product comes from compatibility matrix. For a target user, for each country, we have a probability. We draw a country from these probabilities and assign the country to the target user, hoping that iterating and updating the network will converge to the true values.

Results

Results of the Deterministic Greedy Approach are in the table below.

Deterministic Greedy Approach Results

	Found in 1 Guess	in 2 Guesses	in 3 Guesses
Accuracy	0.683	0.801	0.852

Results of the Naïve Bayes approach are in the table below. Plus symbols (+) in the table mean that corresponding feature is considered, minus symbols (−) means otherwise.

Naïve Bayes Results

L	N	T	Accuracy in 1 Guess	Accuracy in 2 Guesses	Accuracy in 3 Guesses
+	-	-	0.32	0.508	0.672
+	+	-	0.489	0.667	0.742
+	-	+	0.546	0.651	0.731
+	+	+	**0.618**	**0.734**	**0.796**

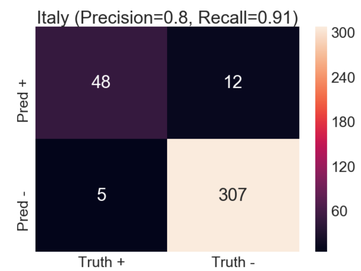
Since English is a very widespread language on the Internet, it distorts our model by giving more probability to the USA than it should be. Best accuracy results with multiple guesses without English using users are **0.789**, **0.88** and **0.891**, respectively.

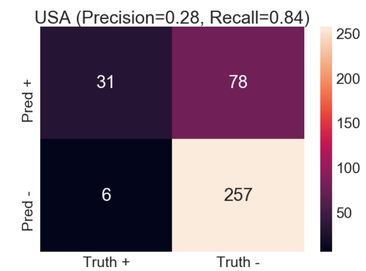
Gibbs Sampling Results

# of Iterations	Accuracy
10	0.634
100	0.632
200	0.634
320	**0.642**
400	0.637
500	0.629
600	0.626

where k is set empirically to 1.30. It is clear that Gibbs sampling algorithm does not show any improvement and we can conclude it is not a good algorithm for this problem.

Here is the confusion matrices for Italy and the USA below.





Conclusion & Future Work

Our work revealed some significant outcomes about the problem, however it is still very open to improvement. If you are interested in this work, you might reach us for more detailed information.