CS215 project report

Group number : 12
Katuri Sai kiran
Ravichandra
Akhil Thatipamula
Project AIM

- Each sentence that is given has a relationship expressed in it between the countries talked about and their attributes.
- We also have world bank data with us which gives the true values of the attributes over time for a given country.
- We should try to find out the relationship that is being talked about in the sentence and also try to give it a confidence score.
Description of what we did

• For function by function description of our code, Please refer to the Readme file written by us.

• We will briefly see the algorithm and ideas that we implemented.
Basic ideas

- We take the input and populate them in “convenient” data structures. (Please check out our efficient use of data structures)

- Now we start with a particular number and for each country in the statement we go through the list of all attributes of that country and enumerate all the attributes for which the value lies within 10% error of the true value of that attribute.

- We find the relative error made and then assign a confidence between 20% and 40% linearly with respect to the relative error made. This is called the basic confidence score.
Going deeper

- We made a list of files that 1 – 11 which contain the keywords for all the 11 categories.
- The keyword also have a certain score from 0-100 which indicates the relevance of that keyword to that particular attribute in which it is present.
- We search the whole sentence to check if the keyword is substring of anyone of the statements. If it is then the values get “points” which is basically the score that is assigned to that particular keyword.
Finding confidence scores

- As already described we have a basic confidence score and apart from that we also have the total points achieved by any number (as previously mentioned points are given for finding a keywords)

- We calculate the contribution of points to confidence scores as follows:
  1. Increase in confidence score falls exponentially with increase in points.
  2. The first 100 points give a credit of 50% to the confidence score.
  3. A total of infinity points will give a contribution of 60% to the confidence scores.
Brainstorming for better ideas

- We have brainstormed a lot and here are a few additions or updates on our code for the better:

- Distance: We first locate the number in the sentence and we move to left and right till we find the country or a keyword whose attribute is suspected to match with the number. We find the distance moved and make this a parameter in assigning confidence score to that number.
• We search for any year in the sentence and if we find it we try to look for that particular year in the world bank data. If such an year doesn't exist we use polynomial or ridge regression to find an estimate of what the value could be on that year and compare it with the values that we have. Similarly we find for words like last decade or last year to find an estimate of the time that the value is related to. We can even detect the tense of the sentence.

• We go through the whole sentence and remove the values which have high probability to be serial numbers or time etc., We can search for : on left or right, we can search for nd or rd etc after the word numbers and so on.....
Challenges faced by us-1

• The serial numbers or the percentages, time etc match with the internet user %, inflation %.
  Though the serial numbers, other irrelevant numbers don't speak anything about the country due to closeness of them with few attributes the code matches them with a reasonably high amount of confidence.

• Similar is the case with years. But in the case of years they obviously give us some useful info about the country and relationship expressed.
Challenges faced by us-2

- Since the population, GDP, Export rate etc., are of the same order, when one of them gets matched, the others also get matched.

- Of course, this is the problem for any code written naively.
Challenges faced by us-3

- This is a more deeper challenge than the above two. We didn't anticipate it during coding but however we found it!!!
- If a statement expresses two different relations and the two different values described fall closely within the two different relations expressed, then if we increase the confidence of one by searching for a keyword, even the other one gets increased
- For example: GDP and export rate and population
- Another: IT.NET.USER, CPI etc
Challenges faced by us-4

- Inability to decode big statements
- The direct techniques used will not be able to decode properly the sentences having a lot of countries at a time.
- This is because in such statements many attributes will be talked about and hence naïve search of keywords result in increase of confidence of the wrong matchings as well.
How we plan to solve the above challenges

- Challenge 1: On testing our output we note that this is the most important source for wrong predictions and confidence.
- An example:
- 8635241 The 2011 budget being unveiled Dec. 7 includes (EURO)4.5 billion in cuts and (EURO)1.5 billion in new taxes -- the fourth emergency budget in three years designed to reverse Ireland's runaway deficits.
- In the above statement 0.7 doesn't speak anything about the relationship, however it matches with internet % users
• This even happens for dates, time, serial numbers etc.,

• To avoid this we used the following scheme:

• If a number is between 0 and 100 and if it is matched to a category $x$, then if there is no keyword corresponding to that category in the statement, then it is considered as a false match and is removed from our list of matches!

• This drastically improved the rate of success predictions

• In fact it decreased the number of predictions by more than 60%
Challenge - 2

- To avoid this challenge we added the concept of keywords.
- We assign a max of 40% confidence for the closeness of matching. The rest 60% is determined by going through sentence and searching for expected words.
- This really helped in assigning the confidence scores accurately.