PROJECT REPORT: CS 215

Group: 06

130050012- Siddharth Bulia

130050013- Animesh Baranawal

130050032- Amit Malav

Instructor: Ganesh Ramakrishnan, Aman Madaan

- ## Basic Software Setup:

Language Used: Python3

Addition:

The content of country_id_map.txt is copied to a newly created file country_id_map_orig.txt with added nationality names like Israeli, Britain, etc. A number of important indicators have been added after reading various articles on WEB to predict relation more confidently, in selected_indicators file.

Data Structure:

Two classes for the program: - Country and Keyword.

Country: It has two data members, "lists" which is a list of lists and country id. “lists” is used to store the tuples consisting of a target and a number from the kb-facts-train_SI.txt file. "id" is used to store the country id.

Keyword: It has two data members, "name" which is used to store name of the target (like AG.LND.TOTL.K2). "lists" is used to store the indicators/keywords associated with it.

Taking Input:

All the country and nationality names are mapped to the country code using the modified country_id_map.txt and mapped back to country name from the country code using newly created country_id_map_orig.txt file. A dictionary type data structure is used for this.

Reading facts from kb-facts-train_SI.txt uses the country code to find the country associated with it. Each data number is stored in a list whose index is corresonding to the number of digits in the number (using the log and ceiling function) along with its attribute . This forms the main content of the "lists" data member of the country class. These can be accessed using dictionary country_facts .

All the Keywords(indicators) are then initialised using "target-relations.tsv" file and "selected-indicators" file.

- ## Assigning Confidence Score:

  First of all, only those numbers are assigned confidence which are within 30% of the statement number.
  We gave the confidence using a **two-step procedure**.
  Assigning confidence scores takes part in two phases:

  a) The initial confidence score is assigned on the **basis of the closeness** of the statement number to the data member for a given pair.

  b) The confidence scores then change as per the **number of keywords** found in the statement corresponding to the attribute of the data number.
  The keywords of the attributes are stored in selected_indicators file.

  Functions used in assigning confidence scores:
  Considering that no country-number relation can achieve a confidence score of 1.0 (100%), we have used exponential and normal distribution functions for assigning the confidence score to a relation.
  A normal distribution function of the form **Ae^(-x^2/2B)** is used with appropraite A and B such that at x=0 we have confidence score = 0.5 and at x=0.3 we have confidence score of 0.05 .

  Further, the confidence is increased using a exponential function **{ 1 - (1 - (init c.s.)*2^(- count)) }** where init c.s. is the initial confidence score and count is the number of keywords matched from the kyword list of the appropriate attribute. Also if no keywords are matched, then the confidence score reduces by two-third(0.66).
  This same ideology is followed for all the country-number pairs found in a particular sentence.

- ## Output and its format:

  The output consists of 6 parts:

  a) Statement id
  b) Country
  c) Statement number
  d) Data number (to which it is matched)
  e) Data number attribute (to which it is matched)
  f) Confidence score

  Between outputs of every two statements there is an empty line.
  If no relation is found for a sentence then "no country matching found" is displayed.

- **Improvements which can be made further:**

a) For many relations in a sentence, it’s observed that the confidence score is about 0.3-0.4. This is because though the number has been found within a 30% range but no keyword has been found in the sentence corresponding to the attribute. This can be improved using a more sophisticated keyword list which consists of more keywords of the particular attribute.

b) Many statements fall within no range, i.e. the number-country relation is of a very different field like sports etc. This can be improved by adding more categories in the attribute list along with the data of the country.

c) The keyword list is very naive. This can be improved by sophisticating it i.e. different keywords increase the confidence score by different amounts. If a keyword which is common to multiple categories increases the confidence by a less amount however a keyword very specific to a category increases the confidence score significantly.

d) The model used can also be improved. Instead of checking closeness to every data number, a distribution type of model can be used i.e. for a country and an attribute, a distribution can be made for the data corresponding to these. The statement data can then be tried to fit in between and the better it fits, the more confidence score it gets.