CS 215 Data Analysis and Interpretation

Project Report
Team Member

- Ghurye Sourabh Sunil 130050001
  sourabhghurye@cse.iitb.ac.in
- Nikhil Vyas 130050023
  nikhilvyas@cse.iitb.ac.in
- Utkarsh Mall 130050037
  utkarshmall13@cse.iitb.ac.in
Introduction

- The motivation behind the project is to train extractors to pull similar entities over internet.
- Given a previously known knowledge base about countries and their attributes we could extract useful information from internet to get more accurate data.
Overview

- We use our knowledge base to get a certain value (mean and standard deviation) of attributes of country.
- We assume that the value given in the sentences are recent. Since the knowledge base gives us values in chronological order. So giving more weightage to recent values rather than to older will be helpful.
Now we start moving on sentences taking out the numbers and country names, given at end of line.

Since we still do not know about what the attribute is, we try to match the number with every attribute of every country given in sentence.

Finding out confidence value with each of these assuming normal distribution of values.
In finding the confidence values, we also see whether the keyword or related keywords are found in sentence or not, if found then we have more more confidence.

The attribute country pair having maximum confidence value is the required attribute.

While giving final confidence we use every attributes confidence in it, since we are not as sure as earlier if we are comparing it with other values.
While finding confidence we use following:

\[
cf_{final} = \left( \frac{(cf/\sigma)_{max}}{\sum_{\forall i} cf_i/\sigma_i} \right) cf_{max}
\]

To find \( cf_i \), if keyword is there then:

\[
 cf_i = \sqrt{cf_{normal}} \times 100
\]

else:

\[
 cf_i = cf_{normal}
\]
After this we reject the confidence scores less than 10%, because if confidence is below 10% that means that the attribute we are seeking for the value is not what we are getting.

Then we output the attribute in a file with its confidence value.
Major Function & Classes Used

- **init**: Initialises normal array by normal.tsv
- **exist**: Searches for keywords in the given vector of strings.
- **conff**: return confidence given mean, standard deviation, value and whether keywords exist or not.
• Country: stores the different values of attributes, averages and standard deviations for each country. Member Functions:
  1. Constructor: Initializes attribute vectors, mean and standard deviations to zero.
  2. assgn: assigns attribute vector there attribute values and assigns mean and standard deviations to each attribute weighted according to chronology.