

Report for CS215 Project made in \LaTeX

Anmol Arora (130050027), Aman Goel (130050041), Rohan Rathod (130050002) - Group Number 24

Email anmolarora@cse.iitb.ac.in , amangoel@cse.iitb.ac.in ,
rohanrathod@cse.iitb.ac.in

Abstract This report describes about our experience with our CS215 Project work

1 Problem Statement

In this problem we are given the map of countries to their codes, a set of sentences (facts or news articles) along with country code and some numbers which appear in the sentence. We have also been given a list of possible relations which can exist for a particular country and figure. We are supposed to read in the sentence and predict what is the relation of the figures with the countries in that sentence. Since we are predicting the relation, it is reasonable to give a confidence score for it. There may exist multiple relations and the confidence scores for each of these relations.

This can be thought as a smaller scaled down, basic version of a bigger machine learning problem.

2 Tools used

We have primarily used *Python* as our base programming language. We have used various additional modules of python such as *math* and *scipy*. The data-structure that forms the core of our database is *Python dictionary*.

3 Timeline of Project

We started the project work on November 5th, 2014. This was primarily the discussion part wherein the team members made attempts to interpret the problem statement. However due to end - semester exams, we could not progress much with the coding part and so we decided to continue the project work after the end - semester exams. The proper coding related part was resumed on November 13th, 2014 and the entire project was successfully completed by November 16th, 2014.

4 Phases of project

4.1 Phase 1

Phase 1 primarily involved reading the files provided and creating the internal database. We read the files - countries-id-map.txt , kb-facts-train-SI.tsv , sentences.tsv. Other files that were useful for use were - selected-indicators , target-relations.tsv For reading the tsv files, we used the python functionality of tab as a delimiter which made the file reading pretty easy

4.2 Phase 2

Phase 2 involved creation of the dictionary database from the files read in Phase 1 We made 3 dictionaries - facts, country-map which formed the core of our database facts consists of a dictionary where the key is the country code and the value is another dictionary whose key is the relationship and the value is the sample list similarly country-map was a simple dictionary where the key value was the country code and the value was country name

4.3 Phase 3

Phase 3 consisted of the main algorithm We read each sentence For each of the numbers in the third field, we checked for each of the countries in the fourth field to check if the value lies somewhere near the sample points in some relation field for that particular country. We basically took the value in the relationship field

which was closest to the number encountered in the sentence and calculated the relative error. We rejected those values for which the relative error was more than 50 percent. Apart from this we also defined a dictionary of keywords for each of the relationship values. Whenever a value was close enough to the number encountered in the sentence, we checked if that sentence contained any keyword for that particular field. If it did, it was accepted as the output, otherwise it was rejected.

4.4 Phase 4

Phase 4 involved coming up with an algorithm to assign a proper confidence score. We finally ended up deciding that 100 - relative error will serve as a good estimate of confidence score

5 Output

```

2040406 China FP.CPI.TOTL.ZG 4.400000 SCORE-92.63
2056179 Israel BN.KLT.DINV.CD 2600000000.000000 SCORE-94.77
2182716 China FP.CPI.TOTL.ZG 2.700000 SCORE-98.21
2182716 China FP.CPI.TOTL.ZG 2.700000 SCORE-98.21
2185244 Hungary FP.CPI.TOTL.ZG 5.500000 SCORE-96.39
2185244 Hungary FP.CPI.TOTL.ZG 25.000000 SCORE-93.29
2185244 Hungary FP.CPI.TOTL.ZG 5.500000 SCORE-96.39
2185244 Hungary FP.CPI.TOTL.ZG 3.000000 SCORE-88.55
2195967 USA SP.DYN.LE00.IN 1388534400.000000 SCORE-63.06
2195967 Afghanistan SP.DYN.LE00.IN 1388534400.000000 SCORE-99.19
2205376 Belgium AG.LND.TOTL.K2 31000000512.000000 SCORE-97.62
2205376 Belgium AG.LND.TOTL.K2 32115853312.000000 SCORE-93.94
2225954 Indonesia FP.CPI.TOTL.ZG 5.000000 SCORE-97.41
2233413 Togo SP.POP.TOTL 6500000.000000 SCORE-99.57
2235556 Myanmar AG.LND.TOTL.K2 599999971328.000000 SCORE-91.85
2235556 Thailand AG.LND.TOTL.K2 599999971328.000000 SCORE-82.56
2239574 Indonesia SP.POP.TOTL 230000000.000000 SCORE-99.58
2306385 Israel SP.POP.TOTL 7600000.000000 SCORE-99.69

```

2404777	Afghanistan SP.POP.TOTL	28000000.000000	SCORE-98.95
2521472	Israel SP.POP.TOTL	5500000.000000	SCORE-99.19
2564223	India BX.GSR.MRCH.CD	10000000000.000000	SCORE-100.00
2564223	India BN.KLT.DINV.CD	10000000000.000000	SCORE-100.00
2619357	Sweden SP.POP.TOTL	9000000.000000	SCORE-99.93
2623971	China SP.POP.TOTL	1300000000.000000	SCORE-99.71
2639408	Singapore SP.POP.TOTL	5080000.000000	SCORE-99.93
2721574	Poland FP.CPI.TOTL.ZG	3.500000	SCORE-98.40
2721574	Poland FP.CPI.TOTL.ZG	2.300000	SCORE-98.06
2863572	Brazil FP.CPI.TOTL.ZG	4.500000	SCORE-92.44
2863572	Brazil FP.CPI.TOTL.ZG	2012.000000	SCORE-96.92
2939803	Malawi SP.POP.TOTL	13000000.000000	SCORE-99.42
2982919	China SP.POP.TOTL	1000000000.000000	SCORE-99.38
3012226	Indonesia SP.POP.TOTL	230000000.000000	SCORE-99.58
3017135	Russia FP.CPI.TOTL.ZG	7.400000	SCORE-92.10
3046241	Iceland IT.NET.USER.P2	0.200800	SCORE-100.00
3046241	Netherlands IT.NET.USER.P2	0.200800	SCORE-60.05
3046906	Israel BN.KLT.DINV.CD	2500000000.000000	SCORE-91.12
3066658	Slovenia SP.POP.TOTL	2200000.000000	SCORE-93.06
3111992	Niger SP.POP.TOTL	7000000.000000	SCORE-98.69
3163088	Bangladesh FP.CPI.TOTL.ZG	6.860000	SCORE-98.60
3163088	Bangladesh FP.CPI.TOTL.ZG	8.430000	SCORE-99.67
3191187	China IT.NET.USER.P2	60.000000	SCORE-58.16
3191187	USA IT.NET.USER.P2	60.000000	SCORE-97.93
3191187	China IT.NET.USER.P2	20.000000	SCORE-88.50
3191187	China FP.CPI.TOTL.ZG	20.000000	SCORE-93.26
3191187	USA IT.NET.USER.P2	20.000000	SCORE-92.52
3191187	USA FP.CPI.TOTL.ZG	20.000000	SCORE-51.95
3217383	Hungary FP.CPI.TOTL.ZG	5.700000	SCORE-99.90
3217383	Latvia FP.CPI.TOTL.ZG	5.700000	SCORE-92.12
3217383	Hungary EP.PMP.DESL.CD	2.800000	SCORE-53.40
3217383	Hungary FP.CPI.TOTL.ZG	2.800000	SCORE-82.65
3217383	Latvia FP.CPI.TOTL.ZG	2.800000	SCORE-94.28
3240014	Australia FP.CPI.TOTL.ZG	4.500000	SCORE-99.45
3240014	Australia FP.CPI.TOTL.ZG	4.500000	SCORE-99.45

3242488 Uganda SP.POP.TOTL 33000000.000000 SCORE-99.59
3350727 Brazil SP.POP.TOTL 190000000.000000 SCORE-100.00
3453722 Israel BN.KLT.DINV.CD 4000000000.000000 SCORE-90.85
3477006 Ukraine FP.CPI.TOTL.ZG 4.500000 SCORE-86.88
3477006 Ukraine FP.CPI.TOTL.ZG 10.000000 SCORE-94.54
3654745 Albania SP.POP.TOTL 3200000.000000 SCORE-99.88
3662632 Bulgaria SP.DYN.LE00.IN 2493050368.000000 SCORE-93.70
3662632 Sudan SP.DYN.LE00.IN 2493050368.000000 SCORE-72.30
4018247 Malawi SP.DYN.LE00.IN 1893456000.000000 SCORE-90.36
4018247 Malawi SP.DYN.LE00.IN 1262304000.000000 SCORE-99.55
4448115 Azerbaijan FP.CPI.TOTL.ZG 1.500000 SCORE-96.95
4448115 Azerbaijan FP.CPI.TOTL.ZG 2009.000000 SCORE-79.14
4448115 Azerbaijan FP.CPI.TOTL.ZG 20.799999 SCORE-99.96
4474980 Russia SP.POP.TOTL 148000000.000000 SCORE-99.90
4474980 Russia SP.POP.TOTL 142000000.000000 SCORE-99.96
4488807 Zimbabwe SP.DYN.LE00.IN 1483207168.000000 SCORE-99.46
4495781 Japan BN.KLT.DINV.CD 10800000000.000000 SCORE-92.00
4504237 USA BN.KLT.DINV.CD 10720000000.000000 SCORE-92.80
4504237 USA BN.KLT.DINV.CD 10720000000.000000 SCORE-92.80
4571337 Israel FP.CPI.TOTL.ZG 4.600000 SCORE-99.95
4571337 Israel FP.CPI.TOTL.ZG 6.300000 SCORE-96.90
4649425 India SP.POP.TOTL 1100000000.000000 SCORE-99.43
4821044 China SP.POP.TOTL 1300000000.000000 SCORE-99.71
4821372 China SP.POP.TOTL 1300000000.000000 SCORE-99.71
5074748 Poland FP.CPI.TOTL.ZG 2.000000 SCORE-94.75
5317722 Pakistan SP.POP.TOTL 170000000.000000 SCORE-99.94
5363166 China IT.NET.USER.P2 10.000000 SCORE-95.03
5363166 China IT.NET.USER.P2 -18.000000 SCORE-100.00
5378999 USA FP.CPI.TOTL.ZG 9.000000 SCORE-98.56
5378999 USA FP.CPI.TOTL.ZG 2.700000 SCORE-99.15
5378999 USA FP.CPI.TOTL.ZG 1.600000 SCORE-99.12
5379111 Australia EN.ATM.CO2E.KT 299999985664.000000 SCORE-99.30
5711021 China EN.ATM.CO2E.KT 299999985664.000000 SCORE-69.25
5711021 China EN.ATM.CO2E.KT 750000013312.000000 SCORE-97.20
5771674 Mozambique BN.KLT.DINV.CD 200000000.000000 SCORE-69.56

5771674 Rwanda BN.KLT.DINV.CD 200000000.000000 SCORE-74.86
5771674 Rwanda BN.KLT.DINV.CD 230000000.000000 SCORE-56.08
5847940 India FP.CPI.TOTL.ZG 10.000000 SCORE-97.93
6153628 Switzerland SP.POP.TOTL 7780000.000000 SCORE-99.53
7057792 Peru AG.LND.TOTL.K2 1290000007168.000000 SCORE-99.22
7057792 Peru SP.POP.TOTL 28220000.000000 SCORE-99.62
7067063 China BX.GSR.MRCH.CD 10000000000.000000 SCORE-100.00
7067063 China BN.KLT.DINV.CD 10000000000.000000 SCORE-100.00
7254786 Chile SP.DYN.LE00.IN 2335262464.000000 SCORE-99.81
7254786 Cuba SP.DYN.LE00.IN 2335262464.000000 SCORE-99.95
7254786 Chile SP.DYN.LE00.IN 2493050368.000000 SCORE-99.94
7254786 Cuba SP.DYN.LE00.IN 2493050368.000000 SCORE-99.92
7316439 Bangladesh FP.CPI.TOTL.ZG 8.120000 SCORE-99.92
7316439 Bangladesh FP.CPI.TOTL.ZG 8.120000 SCORE-99.92
7531135 USA BX.GSR.MRCH.CD 10000000000.000000 SCORE-100.00
7531135 USA BN.KLT.DINV.CD 10000000000.000000 SCORE-100.00
7531135 India BX.GSR.MRCH.CD 10000000000.000000 SCORE-100.00
7531135 India BN.KLT.DINV.CD 10000000000.000000 SCORE-100.00
7531135 USA BX.GSR.MRCH.CD 10000000000.000000 SCORE-100.00
7531135 USA BN.KLT.DINV.CD 10000000000.000000 SCORE-100.00
7531135 India BX.GSR.MRCH.CD 10000000000.000000 SCORE-100.00
7531135 India BN.KLT.DINV.CD 10000000000.000000 SCORE-100.00
7628811 Bangladesh SP.POP.TOTL 144000000.000000 SCORE-99.40
7628811 Bangladesh SP.POP.TOTL 65000000.000000 SCORE-99.57
7677683 Turkey BN.KLT.DINV.CD 8000000000.000000 SCORE-94.35
7688228 Russia SP.DYN.LE00.IN 1893456000.000000 SCORE-93.07
7710134 China EP.PMP.DESL.CD 1.500000 SCORE-82.81
7710134 China FP.CPI.TOTL.ZG 1.500000 SCORE-97.48
7710134 China EP.PMP.DESL.CD 1.500000 SCORE-82.81
7710134 China FP.CPI.TOTL.ZG 1.500000 SCORE-97.48
7712949 Guinea-Bissau SP.POP.TOTL 1500000.000000 SCORE-98.94
8015781 Australia SP.POP.TOTL 21000000.000000 SCORE-99.17
8041360 Slovenia SP.POP.TOTL 2200000.000000 SCORE-93.06
8527292 China FP.CPI.TOTL.ZG 3.000000 SCORE-98.09
8547868 Cuba NY.GDP.MKTP.CD 10000000000.000000 SCORE-100.00

8625484	Philippines	BN.KLT.DINV.CD	4100000000.000000	SCORE−54.51
8625484	Philippines	BN.KLT.DINV.CD	3000000000.000000	SCORE−93.54
8625484	Philippines	BN.KLT.DINV.CD	1100000000.000000	SCORE−84.45
8644726	Japan	SP.POP.TOTL	127000000.000000	SCORE−99.90
8644726	Japan	SP.POP.TOTL	90000000.000000	SCORE−97.30
9706798	India	SP.POP.TOTL	1200000000.000000	SCORE−99.53

6 Citations

We used various online tutorials for learning pyhton syntax. We also used online tutorials to look up for LaTeX