CS 215 Project

Autumn 2014

Motivation

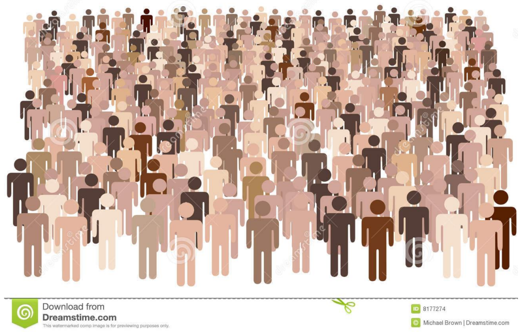
GDP

Goods Exports

India

Population

Internet Users %



Motivation

- Repositories of facts containing this information can be found at many places, like data.worldbank.org, Wikipedia infoboxes etc.

- Countries are popular and finite, finding complete knowledge bases is possible

- What about less popular entities?
  - “What is the population of Arbit Apartments, Powai?”
  - “What is the GDP of Sugarcane Industry of India?”
  - “% of Internet users in Samalkha?”

- What about less popular facts about popular entities?

Motivation

- Good news!
- Web is huge, probably, there is *some* page which contains the information we are looking for.

- The way in which you express a fact about an entity depends **on the fact**, and not the entity.

- We may expect the *sentence structure* to be similar.

- Sentence structure can be captured by POS tag sequences, words, paths in dependency parse tree and so on

  - Population of India is 1.3 billion, making it the second largest country in the world
  - Population of Arbit Apartments, Powai reached 1300

Problem Statement : Introduction

- Given that we know a lot about countries, can we train extractors that run over the web and pull similar facts about other entities?
- Maybe, but we'll need some training data first
- Hand labeling : Pick sentences having numbers and countries, and label with the relation expressed
- 1 million newswire articles have 1.5 million sentences having a country and a number!
- Will take 4166 hours with 10 seconds per sentence, need to automate the process

Problem Statement : Given

- A Knowledge Base (KB) of the form:

/m/0jhd	3313739780	NY.GDP.MKTP.CD
/m/0jhd	3052467460	NY.GDP.MKTP.CD
/m/0jhd	3176749570	NY.GDP.MKTP.CD
/m/07z5n	4.7466025	IT.NET.USER.P2
/m/07z5n	5.082334	IT.NET.USER.P2
/m/07z5n	5.850585	IT.NET.USER.P2
/m/0j1z8	83600	AG.LND.TOTL.K2
/m/0j1z8	83600	AG.LND.TOTL.K2
/m/0j1z8	83600	AG.LND.TOTL.K2
/m/0jgd	2736690	AG.LND.TOTL.K2
/m/0jgd	2736690	AG.LND.TOTL.K2
/m/0jgd	2736690	AG.LND.TOTL.K2

All numbers will be standardized to SI units

Problem Statement : Given

Entity-Id map

/m/035qy	Greece
/m/06tnn	South Georgia and the South Sandwich Islands
/m/0345_	Guatemala
/m/034tl	Guam
/m/036b_	Guinea-Bissau
/m/034m8	Guyana
/m/03h2c	Honduras
/m/03gyl	Haiti
/m/03ryn	Indonesia
/m/03spz	Israel
/m/03t1s	Isle of Man
/m/0168b	British Indian Ocean Territory
/m/0d05q4	Iraq
/m/03_xj	Jersey
/m/03_r3	Jamaica

Problem Statement : Given

- A set of sentences and number – country pairs in the sentence.
- Sentence id [TAB] sentence [TAB] numbers [TAB] entities

- 1 Palau is among the world's smallest countries, with some 20,000 people scattered across 190 square miles (490 square kilometers) of lush tropical landscapes (Palau;190,20000,490)

- 2 After about half of the ballots were counted, 93.1 percent of them were against the bill, which was designed to compensate British and Dutch banking customers for their losses when bank Icesave, an Icelandic Internet bank, collapsed in the fall of 2008, said Iceland's public broadcaster RUV. (Britain, Netherlands, Iceland;2008,93.1)

- 3 We have a three million tourist population that comes in compared to Malaysia’s 22 million (and) Thailand’s 14 million, he said. (Thailand,Malaysia;14000000,22000000)

Problem Statement : Desideratum

- Find out:
  1. Relations expressed in a sentence and
  2. The corresponding confidence score

- SentId [TAB] Country [TAB] Relation [TAB] Number [TAB] Score

- For the sentences shown in the previous slides:

3	Malaysia	SP.POP.TOTL	14000000	SCORE_1
3	Malaysia	SP.POP.TOTL	22000000	SCORE_2
2	NULL	NULL	NULL	NULL

- Multiple Relations per sentence are possible
- Multiple Relations for a given country and number pair are also possible, with different confidence scores

Challenges and Food for Thought

- Basic software setup: reading files, forming maps from country to id
- Ways to assign confidence scores
- Can matchings from multiple sentences be used to calculate the confidence score?
- For a given country and relation, we have several values, each pertaining to different years. Can this fact be used while matching?