Knowledge and the Web: Homework 3

Katrien Laenen Gust Verbruggen Ward Schodts

November 15, 2015

This document thoroughly discusses three research questions, along with the knowledge bases needed to solve them. When talking about ontologies, relations will be *emphasised* and classes `monospaced`.

1 Which European politicians have *a high chance* of receiving a Nobel Prize?

We emphasise *a high chance* because that could use some clarification. How do we compare probabilities of receiving Nobel Prizes between arbitrary politicians?

First, we build a profile of a politician. This should not only cover their performance in the European Parliament, other achievements and activities are probably far more interesting. Where are they from? What subjects have they worked on before or talked about in public? Have they received other awards? Perhaps they've written a book or performed research activities?

Next, we attempt to build an analogous profile for Nobel Prize winners. Even though prizes in Peace or perhaps Literature and Economic Sciences are far more likely to be won by a politician, we'll build the model for all winners.

The final step is comparing the profiles we built. Do we find particular similarities between a specific Nobel Prize winner and politician? Or, for example, we could try learning a model for a Nobel Prize winner trough machine learning and classify the politicians using this model.

Knowledge bases

Aside from the ToE knowledge base, we need to gather as much information as possible about the people we'll build profiles for.

DBpedia (http://wiki.dbpedia.org/) is always a good starting point. Coming straight from Wikipedia, it is updated daily and should thus be up-to-date all the time. For many projects, Wikipedia is *the* go-to data source, because it contains complete information about relevant topics to our subject. Although it can be edited by anyone, data correctness is verified thoroughly by a team of content moderators.

Nobelprize.org (http://data.nobelprize.org/) is the most complete data source concerning nobel prize winners. From 1901 on, it contains all nobel prizes that have been won. Moreover, it is maintained by the official organisation responsible for the Nobel Prize and should thus be complete and correct. It makes use of the FOAF and DBpedia ontologies and more importantly, laureates have a *owl:sameAs* relation with DBpedia entries for easy lookup of background information.

2 Which politicians have a high chance of being murdered?

Politicians get a lot of criticism and sometimes they even get murdered. Of course, this can be for political reasons as well as for non-political reasons. The goal of this question is to predict which politicians have a high probability of being killed by comparing them with the profiles of murdered politicians. We expect that the subjects they talk about and their political beliefs are the factors which have the biggest influence on this probability. On top of that, for politicians that live in places with high murder rates or that visit these kinds of places often, it is likely that this probability will even increase.

Knowledge bases

Aside from the ToE knowledge base, we also need the following knowledge bases.

European Union Open Data Portal (https://open-data.europa.eu/) contains datasets about homicides and other criminal activities e.g. homicide offences.This data is necessary to determine if location influences the chance of being murdered. The European Commission maintains this website to keep this information timely and accurate.

2 x homicide?

Freebase (https://www.freebase.com/) This is an openly available knowledge base specialised among others in well-known people. Because most politicians are well-known, this is an interesting knowledge base. Furthermore, Freebase contains the property *cause_of_death* which we can use to find murders or homicides among the politicians. As such, this database is highly relevant and complete regarding this research question. Like Wikipedia, Freebase can be edited by anyone. But this also means that the information can be corrected by anyone and people make incremental improvements all the time. Besides this, there is a group of trusted Freebase experts which are responsible for solving such issues. Thus, there will be some inaccurate information in this database but we expect the database to be fairly correct.

DBpedia (http://wiki.dbpedia.org/) For this question we need also some background information. As already pointed out, DBpedia is a good source for this.

3 Which characteristics of a politician influence the amount of TV appearances he makes?

Politicians appear on TV quite often, but not all appear equally frequent. A lot of factors can influence this frequency: political party, function, statements made, previous occupations,... the list can go on and on. In this question we will research which of these characteristics actually influence the amount of such TV appearances they make.

First and most importantly, we need to find how often politicians appeared on television within a certain timeframe. This is the *response* measure. Next, we yet again have to build a profile for politicians, similar to that in the first question. Finally, we find out which features of our politician influence the response.

Knowledge bases

In order to build the profile, we need some additional information next to the ToE knowledge base. The most important ones are those for constructing the response, e.g. the television appearances.

BBC Programme (http://www.bbc.co.uk/ontologies/po) contains information about BBC's programmes. Each programme has several `Person`

entities related to it trough the *person* relation, so we can find out which politicians appeared when in BBC television shows. As it is maintained solely by the BBC, we believe it is up-to-date and accurate.

LinkedTV (`http://www.linkedtv.eu/`) attempts to add meta data to television shows, split up into fragments. For example, it attempts to link information on all faces appearing in a fragment to a `Person`. This can also be useful in constructing the response measures for all politicians and even enables us to specify how long they appeared on television. The LinkedTV project aims at automating the process of gathering meta-data [1], which mean it is prone to errors in the software. Some of it might be corrected by hand, but we should still be careful in assuming everything is correct.

Furthermore, we need to gather some background information from for example DBpedia or Freebase. These have already been discussed in detail, so we will not do that again.

References

[1] Daniel Stein et al., *From raw data to semantically enriched hyperlinks: recent advances in the LinkedTV analysis workflow*, NEM Summit 2013, Nantes, France, October 2013.