Attribute Template Generation

Daksh, Vasu, Divanshi

Task

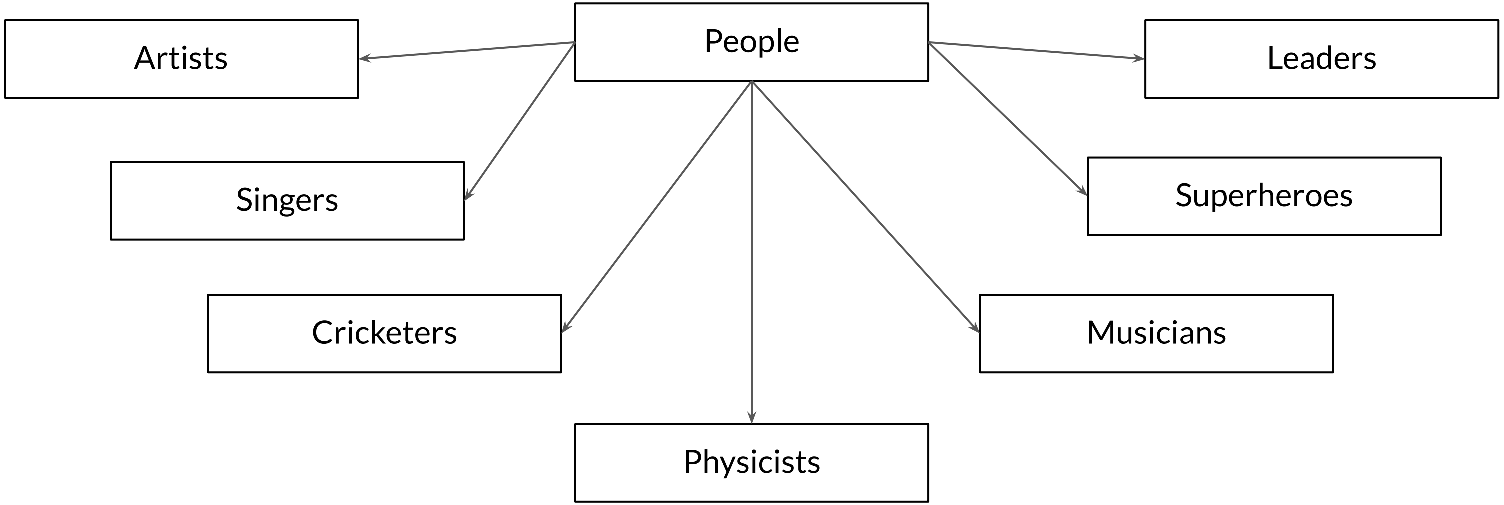
- Given a specific domain, and a set of input documents, the task is to generate an attribute template corresponding to the input domain which can be used in generating new articles for that domain.
- Secondary task - Once the list of attributes for the domain is extracted, we need to divide them in accordance with the section of the wikipedia page a certain group of attributes would belong to.

Scope

- Input
  - Domain name
  - ~1K wikipedia articles
  - Language - English
- Attributes to be extracted from
  - Infobox
  - Wikidata
  - Wikidata page
  - Wikipedia text
- Output
  - Template for that domain
  - Intermediate - List of attributes and their categorization for the domain

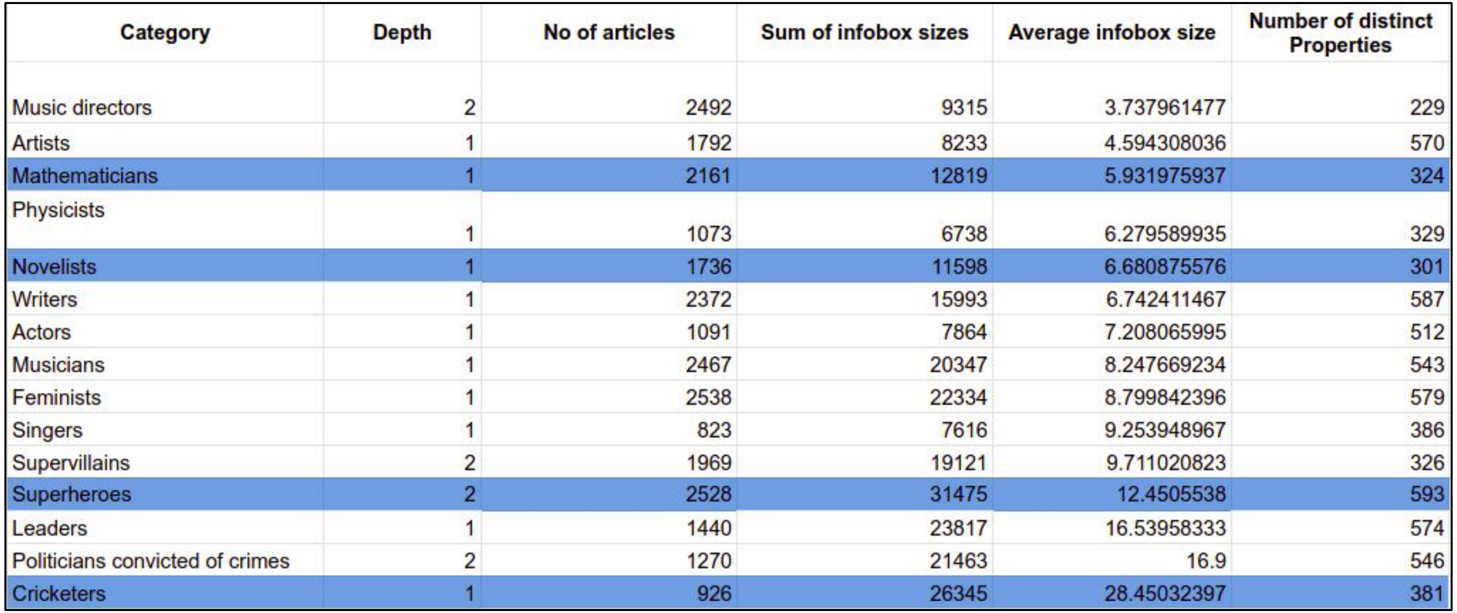
Domain

- For given wikipedia domain, a large number of sub categories exist which may cover very different among themselves. So, instead of focusing on a domain, it will be better to start off by focusing on a sub-domain.



Sub-category stats

The below table shows stats of some sub-categories we manually picked. Out of these, we have decided to work on the 4 highlighted categories for now. We have mainly made the judgement based on number of distinct infobox properties as well as the average size of an infobox for each domain



Category	Depth	No of articles	Sum of infobox sizes	Average infobox size	Number of distinct Properties
Music directors	2	2492	9315	3.737961477	229
Artists	1	1792	8233	4.594308036	570
Mathematicians	1	2161	12819	5.931975937	324
Physicists	1	1073	6738	6.279589935	329
Novelists	1	1736	11598	6.680875576	301
Writers	1	2372	15993	6.742411467	587
Actors	1	1091	7864	7.208065995	512
Musicians	1	2467	20347	8.247669234	543
Feminists	1	2538	22334	8.799842396	579
Singers	1	823	7616	9.253948967	386
Supervillains	2	1969	19121	9.711020823	326
Superheroes	2	2528	31475	12.4505538	593
Leaders	1	1440	23817	16.53958333	574
Politicians convicted of crimes	2	1270	21463	16.9	546
Cricketers	1	926	26345	28.45032397	381

Steps

- First step would be to simply extract the already present attributes in the infobox and wikidata for a given wikipedia page.
- To get attributes from the text
    - Use an OpenIE system to extract relations from the text.
    - We will look at the shortcoming of existing openIE systems and try to come up with heuristics to resolve them ourselves.

Dataset Creation

- To create the dataset, we extracted articles for the highlighted sub-categories in slide 5.
- For each article, we extract their infobox, wikidata and wikipedia text.
- Basic stats for the infoboxes, wikidata and the articles are extracted after processing.

Category	Avg. Article length	#sents	#unique infobox atts	Avg. infobox att. / article	#unique wikidata atts.	Avg. wikidata att. / article
Cricketers	340	17	366	19	278	10
Novelists	1010	57	294	24	1036	42
Mathematicians	571	41	303	18	1055	31
Superheroes	2360	123	501	12	708	13

Cricketers Infobox and Wikidata Attribute Stats

Infobox Attribute	Counts
name	819
source	812
date	790
country	789
birth_date	770

Wikidata Attribute	Counts
Instance of (P31)	926
Sport (P641)	888
Sex or gender (P21)	874
Occupation (P106)	862
Data of birth (P569)	820

Novelist Infobox and Wikidata Attribute Stats

Attribute	**Counts**
Birth Place	1032
Name	1016
Birth data	1005
Occupation	909
Image	761

Attribute	**Counts**
Instance of (P31)	1746
Occupation (P106)	1738
Sex or gender (P21)	1737
Date of birth (P569)	1645
VIAF ID (P214)	1597

Mathematician Infobox and Wikidata Attribute Stats

Attribute	Counts
Name	1071
Birth date	895
Alma mater	864
Birth place	774
Doctoral advisor	712

Attribute	Counts
Instance of (P31)	2151
Sex or gender (P21)	2031
Occupation (P106)	2006
Educated at (P69)	1856
Given name (P735)	1772

Superheroes Infobox and Wikidata Attribute Stats

Infobox Attribute	Counts
caption	1704
image	1699
publisher	1605
debut	1523
creators	1508

Wikidata Attribute	Counts
Instance of (P31)	2436
Freebase ID (P646)	2096
from narrative universe (P1080)	1468
sex or gender (P21)	1172
creator (P170)	951

Observations

- As we can see from the previous slides, some of the attributes which are present in most of the articles are not very useful for template generation.
- So, further processing needs to be done to remove unnecessary attributes from the **most frequent** ones so that we have some possible attributes for the template.
- Example, “*name*” for cricketer is a useful infobox attribute that can be part of the template but “*source*” is not of any value here.