Dataset Search Engine

Ahmad Shayaan IMT2014004
H. Vijaya Sharvani IMT2014022
Indu Ilanchezian IMT2014024

December 10, 2017

1 Introduction

With the emerging trends in data analysis, it has increasingly become important for the data analysts and data scientists to find the right dataset. However, datasets are spread across multiple websites like kaggle, data.gov, uci repository and so on.It often becomes a tiresome and time consuming task to find the correct dataset. This motivates us to design a common platform where we can search for datasets, based on the domain, the type of the dataset, the size of the dataset and miscellaneous features like the popularity, number of downloads and number of views. We build an ontology which describes the various components and features that may be necessary to describe a dataset. Using this ontology, we label the datasets and then use a probabilistic classifier to determine the domain of the dataset. Our dataset search engine then consults the database of datasets and the probabilistic classifier to search for the appropriate datasets, based on the keyword searched by the user.

2 Building the ontology

In order to search for datasets efficiently, we have to first build a representation of a dataset. By representation, we mean a way to depict the properties, features and components of datasets. For representing the datasets, we build an ontology of the datasets. The ontology for the datasets is shown in Figure 1. Figure 2 shows the sub-ontology of the dataset domains. The domain ontologies will be used for classification of datasets into different domains.

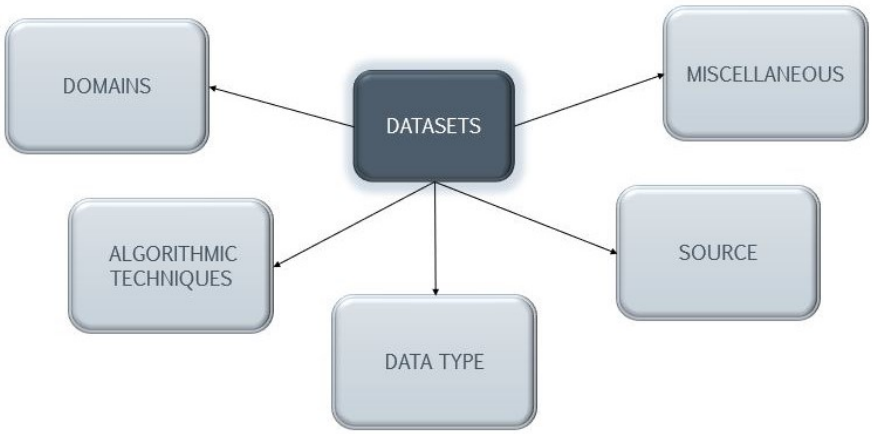


Figure 1: **Ontology of datasets:** The dataset ontology captures the attributes and properties of the datasets. The datasets can belong to different domains. Datasets can be classified based on the learning algorithms or statistical methods that can be applied on them. Datasets have a type and the columns in the datasets also have datatypes which capture important aspects of the datasets. The size of the dataset is also an important attribute. The source of the dataset can be private or public. There may be miscellaneous attributes like number of views, number of downloads, popularity which may be important to characterise the dataset.

3 Data collection

Ironically, even a dataset search engine requires a dataset of datasets. Hence, we scrape data about these datasets from three different websites: Kaggle, data.gov and UCI repository. We used python BeautifulSoup to scrape the webpages.
A total of 1964 datasets were collected from these webpages. The information scraped includes the title of the dataset, the url of the dataset, the description, the author/publisher, keywords/tags, column names, the column datatypes and the date of publishing.

4 Labelling the datasets

4.1 Labels

The labels of the dataset correspond to the leaf nodes of the domain ontology. We will use a probabilistic classifier because the datasets can often fall in multiple categories. The probability of a dataset belonging to the class represented by a leaf node will be equal to probabililty of the dataset belonging to that

class predicted by the classifier.The probability of a dataset belonging to a class represented by a non-leaf node will be the sum of probabilities of the dataset belonging to the class represented by each of its child nodes. The number of classes will be equal to the number of leaf nodes. The total number of leaf nodes in the ontology designed is 71.
Let x be the vector representing the features of a dataset. Let y be the label. Then,

$p(y = l|x) = p(class = l|x, M)$ where l is a leaf node

$p(y = p|x) = \sum_c p(class = c|x, M)$ where each p is a non-leaf node and $c \in \{children(p)\}$

Here, M is the model learnt by the classifier.

4.2 Label Propagation

Annotating over a 1000 datsets to be trained by a classifier is a challenging task. Around 450 datasets were annotated manually and for the remaining we used the label propagation method. The label propagation method works as follows:

Problem Setup[1]:

Let $(x_1, y_1)...(x_l, y_l)$ be labelled data, where $Y_L = y_1...y_l$ are the class labels. Let $(x_{l+1}, y_{l+1})...(x_{l+u}, y_{l+u})$ be unlabelled data, where $Y_U = y_{l+1}...y_{l+u}$ are unobserved, usually $l << u$.
The probability transition matrix is then defined as:
$T_{ij} = P(j \rightarrow i) = \frac{w_{ij}}{\sum_{k=1}^{l+u} w_{kj}}$
Where:

- T_{ij} is the probability to jump from node j to i
- $w_{ij} = exp(-\frac{\sum_{d=1}^{D}(x_i^d - x_j^d)^2}{\sigma^2}$

Also consider a $(l+u) \times C$ label matrix Y, whose i^{th} row representing the label probability distribution of node x_i.

Algorithm[1]:

1. Propagate $Y \leftarrow TY$
2. Row-normalize Y
3. Clamp the labelled data. Repeat from step 1 until Y converges

It can also be shown that the algorithm converges to a simple solution [1].

5 Feature extraction

5.1 Data pre-processing

The scraped data was saved in a database with utf encoding. The fields title, description, column names, tags/keywords combined together form documents. The following steps are taken to preprocess the text in these documents.

1. Punctuation marks are stripped from the documents
2. Non-ascii characters are removed from the text. Some of the descriptions included characters from other languages such as Chinese, Spanish and so on. These were removed in this step.
3. The words in the document are stemmed to obtain the root form. This will help reduce the number of redundant features.

5.2 TFIDF features

After initial pre-processing of the text documents, the documents are converted to TFIDF features. TF-IDF features are real valued features calculated for each word in a document through an inverse proportion of the frequency of the word in a particular document to the percentage of documents the word appears in. Words with high TF-IDF numbers imply a strong relationship with the document they appear in. Thus, the TF-IDF features have a high correlation with the labels, in this case, the domain categories [2].

6 Probabilistic classification models

We train the following probabilistic models with the labelled datasets and the TFIDF features:

- Logistic Regression
- SGD Classifier
- Bernoulli Naive Bayes
- Multinomial Naive Bayes

7 Evaluation

Figure 3 shows the accuracies, the training and test time for the different probabilistic classifiers. The logistic regression classifier with L2 penalty and the SGD classifier with elastic net penalty give the best accuracies.

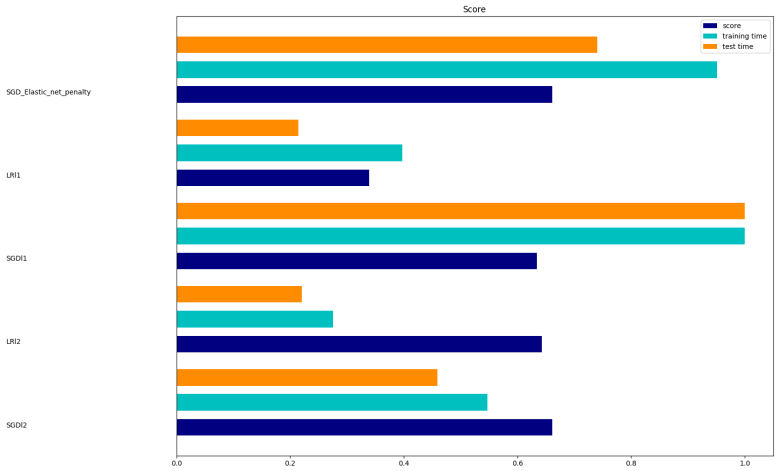


Figure 3: The figure shows the accuracies, training and testing time for the logistic regression and SGD classifer. The SGD classifier is trained with L1 penalty, L2 penalty and elastic net penalty. The logistic regression classifier is trained with L1 and L2 penalty. The orange bars indicate the test time, the cyan bars represent training time and the dark blue bars indicate the accuracies. The accuracy is highest for SGD model with L2 or elastic net penalty and the logistic regression model with L2 penalty. The logistic regression method requires the minimum time for training and testing. And the resuls are comparable to that of the SGD model. Hence, we choose the logistic regression model as the best probabilistic classifier for categorising datasets.

Figure 4 shows the ROC curves for the logistic regression and the SGD classifier. The logistic regression model with L2 penalty has the highest area under ROC.

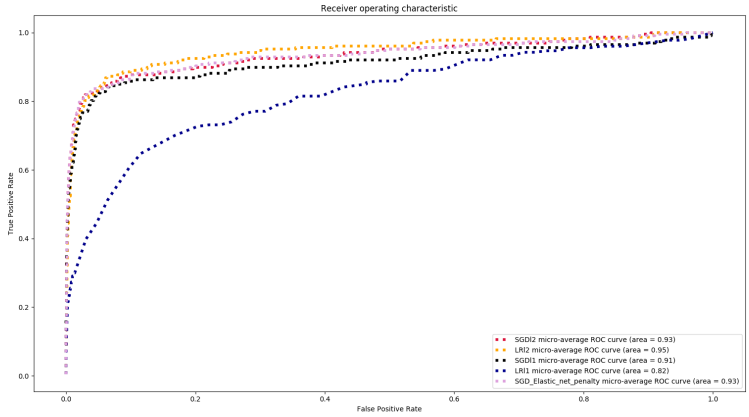


Figure 4: ROC curves for the logistic regression and the SGD classifier. We have a multiclass classification. The ROC curves plot the microaverage for each one vs rest classification. All the models, except the logistic regression model with L1 penalty, show good performance. The area under ROC are tabulated in Table 1.

Table 1: Area Under Receiver Operating Characteristic for the different models

Model	Area under ROC
Logistic Regression (L2 penalty)	0.95
SGD Classifier (Elastic Net Penalty)	0.93
SGD Classifier (L2 Penalty)	0.93
SGD Classifier (L1 Penalty)	0.91
Logistic Regression (L1 penalty)	0.82

Thus, we choose the logistic regression classifier for the classification of datasets.

8 Web Portal

In order to make the search more interactive and user friendly, we have built a web portal for the application. The application is a rudimentary one which accepts one of the nodes in the domain ontology as input and displays all the datasets which have a high probability corresponding to that node. The server for the web application is built using the flask library for python.

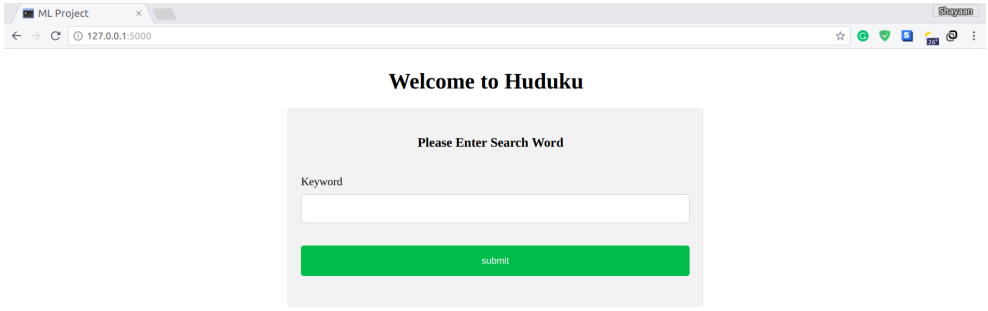
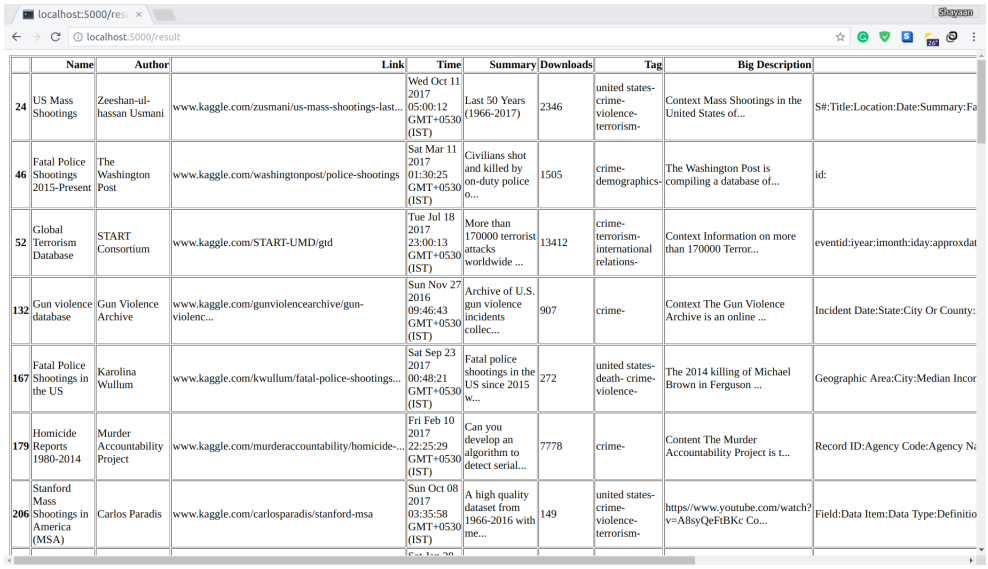


Figure 5: Home page of the dataset search web application



	Name	Author	Link	Time	Summary	Downloads	Tag	Big Description	
24	US Mass Shootings	Zeeshan-ul-hassan Usmani	www.kaggle.com/zusmani/us-mass-shootings-last...	Wed Oct 11 2017 05:00:12 GMT+0530 (IST)	Last 50 Years (1966-2017)	2346	united states- crime- violence- terrorism-	Context Mass Shootings in the United States of...	S#:Title:Location:Date:Summary:Fa
46	Fatal Police Shootings 2015-Present	The Washington Post	www.kaggle.com/washingtonpost/police-shootings	Sat Mar 11 2017 01:30:25 GMT+0530 (IST)	Civilians shot and killed by on-duty police o...	1505	crime- demographics-	The Washington Post is compiling a database of...	id:
52	Global Terrorism Database	START Consortium	www.kaggle.com/START-UMD/gtd	Tue Jul 18 2017 23:00:13 GMT+0530 (IST)	More than 170000 terrorist attacks worldwide ...	13412	crime- terrorism- international relations-	Context Information on more than 170000 Terror...	eventid:iyear:imonth:iday:approxdat
132	Gun violence database	Gun Violence Archive	www.kaggle.com/gunviolencearchive/gun-violenc...	Sun Nov 27 2016 09:46:43 GMT+0530 (IST)	Archive of U.S. gun violence incidents collec...	907	crime-	Context The Gun Violence Archive is an online ...	Incident Date:State:City Or County:
167	Fatal Police Shootings in the US	Karolina Wullum	www.kaggle.com/kwullum/fatal-police-shootings...	Sat Sep 23 2017 00:48:21 GMT+0530 (IST)	Fatal police shootings in the US since 2015 w...	272	united states- death- crime- violence-	The 2014 killing of Michael Brown in Ferguson ...	Geographic Area:City:Median Incor
179	Homicide Reports 1980-2014	Murder Accountability Project	www.kaggle.com/murderaccountability/homicide-...	Fri Feb 10 2017 22:25:29 GMT+0530 (IST)	Can you develop an algorithm to detect serial...	7778	crime-	Content The Murder Accountability Project is t...	Record ID:Agency Code:Agency Na
206	Stanford Mass Shootings in America (MSA)	Carlos Paradis	www.kaggle.com/carlosparadis/stanford-msa	Sun Oct 08 2017 03:35:58 GMT+0530 (IST)	A high quality dataset from 1966-2016 with me...	149	united states- crime- violence- terrorism-	https://www.youtube.com/watch?v=A8syQeFtBKc Co...	Field:Data Item:Data Type:Definitio

Figure 6: The figure shows the results retrieved for the keyword crime by our application. The results are the datsets with highest probabilities corresponding to the keyword crime.

9 Results

Figure 5 and 6 show the web portal for the dataset search and the results retrived for the keyword crime.

10 Future Work

In this project, we have built an extensive ontology of the domains of the dataset. The search engine can search for datasets belonging to each node of the ontology. The following extensions are possible for the search engine:

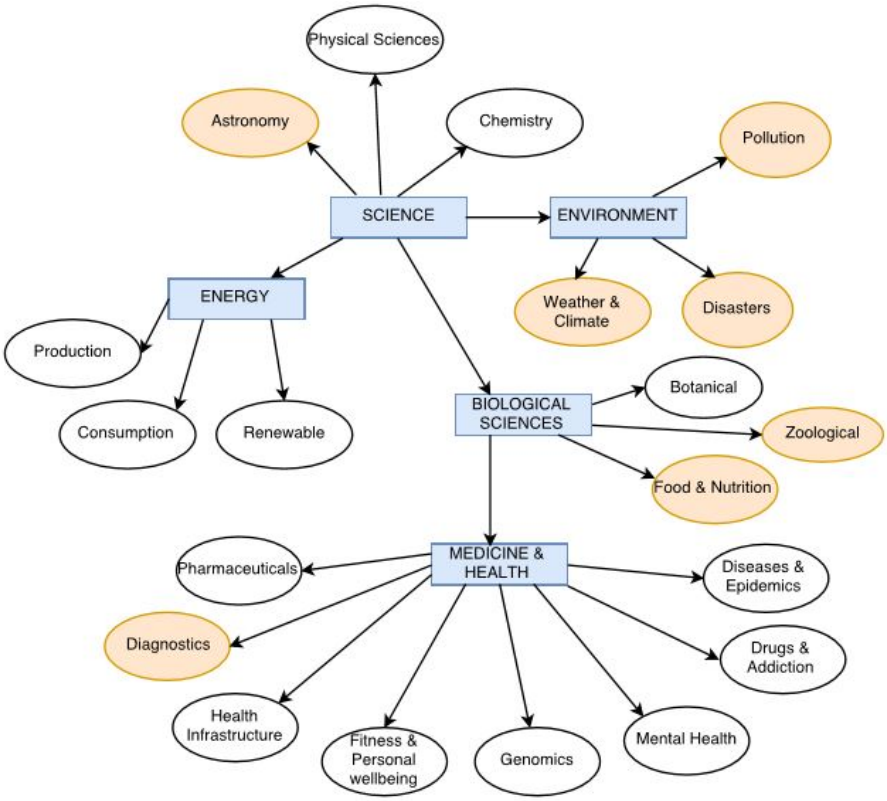
- Build an ontology for the learning algorithms. And learn a classifier to categorise datasets based on the algorithms which could be possibly applied on them.
- Retreive datasets based on multiple search factors, for example, based on popularity, size requirements, dataset types and so on
- Accept synonyms of the nodes as input to the keyword field
- Extend support for addition of nodes to the ontology

There may be several other extensions possible, in the context of user interface, ease of use but we limit ourselves to the possible extensions which can make use of machine learning.

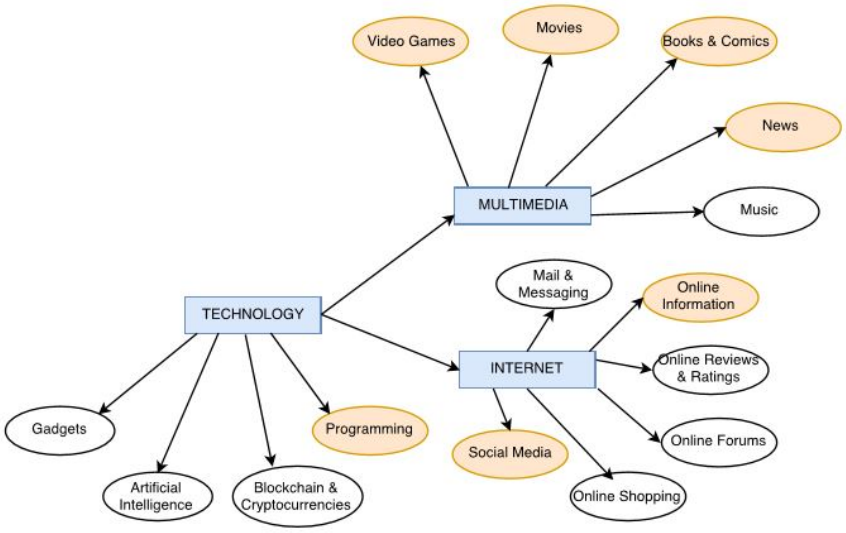
References

[1] Zhu Xiaojin, Ghahramani Zoubin.*Learning from Labeled and Unlabeled Data with Label Propagation.*(2003). Retrieved from: http://pages.cs.wisc.edu/ jerryzhu/pub/CMU-CALD-02-107.pdf

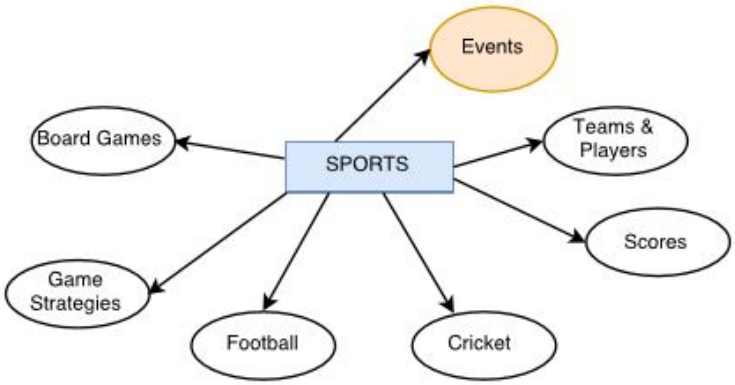
[2] Ramos Juan. *Using TF-IDF to Determine Word Relevance in Document Queries.* (2003).



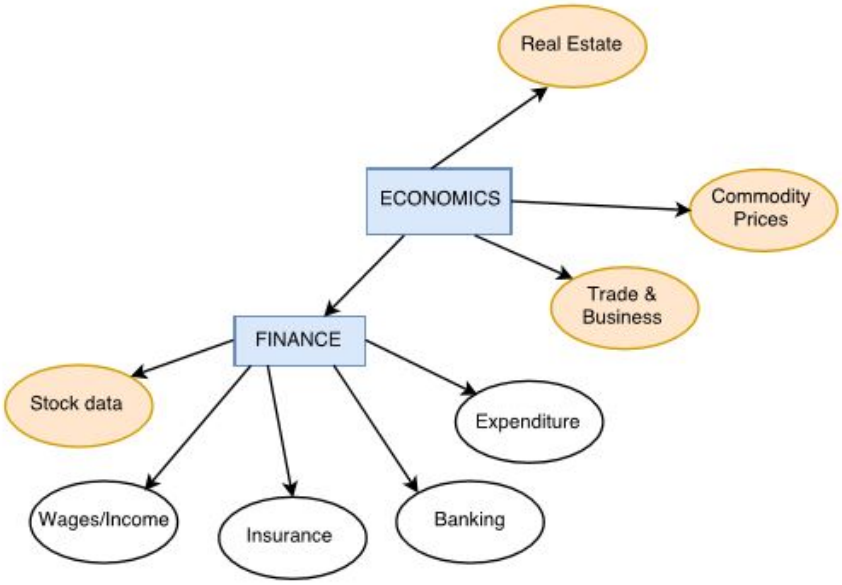
(a) Ontology for science domain



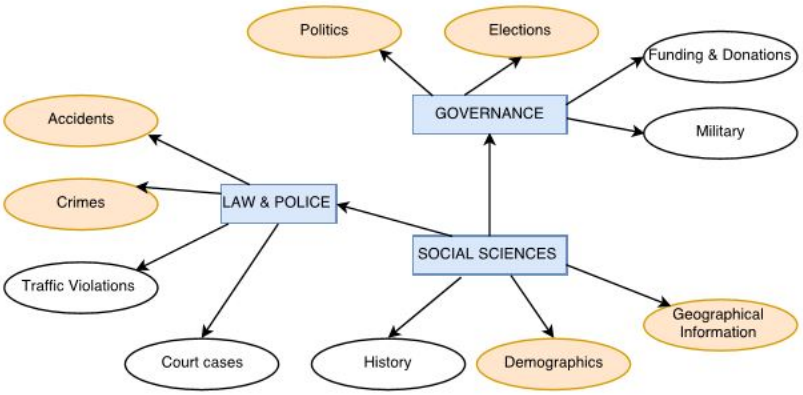
(b) Ontology for technology domain



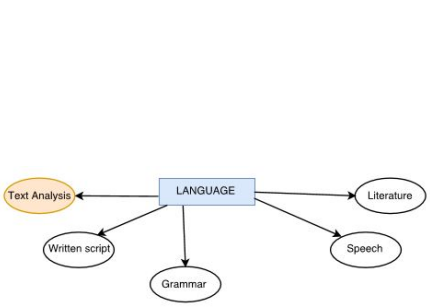
(a) Ontology for sports domain



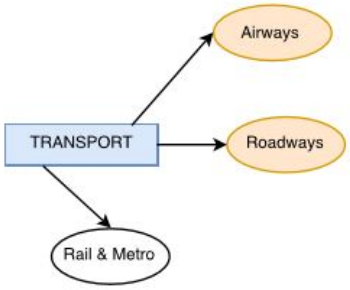
(b) Ontology for economics domain



(a) Ontology for social sciences domain



(b) Ontology for language domain



(c) Ontology for transport domain

Figure 2: The domains are divided into seven broad categories: Science, Economics, Technology, Education, Transport, Social Sciences, Language and Sports. The colored nodes indicate the most frequently occuring labels in the dataset. The datasets collected so far, exhibit class imbalance.