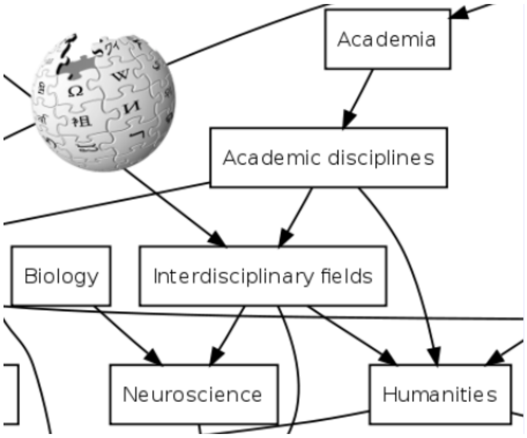


CS 522

ADVANCED DATA MINING

FINAL REPORT

WIKIPEDIA HIERARCHY EXTRACTION

BY: POOJA HEMANTKUMAR PATEL

TEAM MEMBERS:

KRISHNA BHARADWAJ

ANIMESH PATNI

Contents

Abstract

In this report we perform extraction of Wikipedia hierarchies. Wikipedia has many articles with about 300 languages which also allows editing an article and hence the content in the articles are very subjective. Considering only the articles which are in English are more than 5 million. These all the articles on Wikipedia are categorized in hierarchy. Hence, each article has its own hierarchy. Our project extracts the hierarchy of the given topic by making using of dataless approach that is used in a paper 'On Dataless Hierarchical Text Classification' by Yangqiu Song and Dan Roth.

1. Introduction

Wikipedia hierarchy Extraction is extracting the hierarchy of the article which is input by the user. This is very helpful for the categorization of the articles and also can be used in information browsing well as it will decrease the time required to browse the required information. So in order to achieve such goals it is very necessary to understand the existing hierarchy of the Wikipedia and also there is requirement to fetch the Wikipedia hierarchy. In our project, we will take input from the user as an article title and then extract the hierarchy of the same.

Topics we used for hierarchy extraction:

We chose the major topics which are very common in day to day life:

1. Finance
2. Biology
3. Politics

Structure for 'Wikipedia Hierarchy Extraction':

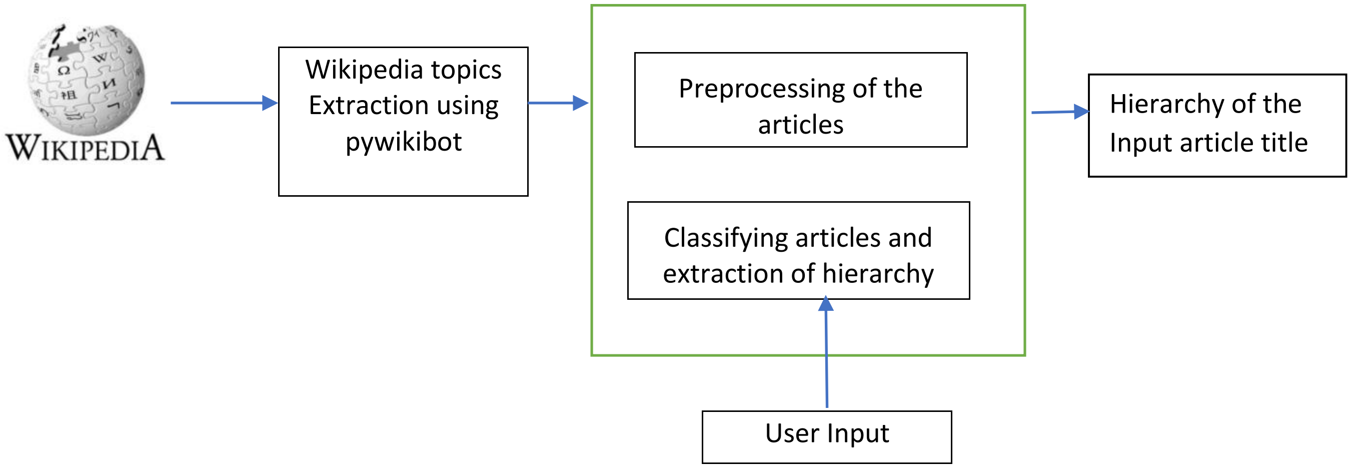

WIKIPEDIA
Wikipedia topics Extraction using pywikibot
Preprocessing of the articles
Classifying articles and extraction of hierarchy
Hierarchy of the Input article title
User Input

Fig. 1. Structure for 'Wikipedia Hierarchy Extraction'

The above diagram shows the major steps that are performed in order to extract the hierarchy of the given input from the user. The vital steps include:

a. The fetching the existing Wikipedia hierarchies of the top level categories like finance, biology, politics etc.
b. Then extracting all the articles related to the above topics thereafter preprocessing all the articles
c. Fetch the hierarchy by extracting the most similar page at each level

2. Dataset

2.1. Data Procurement

The main source of the data which we used for our project is Wikipedia dumps. The following shows the steps for how the data was fetched and we used :

a. We downloaded the xml dumps of Wikipedia from https://dumps.wikimedia.org/enwiki/latest/
b. It has all the information about the hierarchy and the categories which are defined in Wikipedia
c. Category Tree was formed for the categories which we are using which are 'Finance', 'Politics' and 'Biology' and we put these categories at the top level and hence they are at level 0
d. The API in python name 'pywikibot' was used to extract the hierarchy
e. The levels were restricted to two and we did this because as we went down in depth the categories that fetched got irrelevant to the topics at high level
f. We fetched the tree for each top level categories like finance, biology , politics thereafter the we fetched all the articles which are related to tree which we fetched. These all articles which we get here is our dataset which we used for further experiments

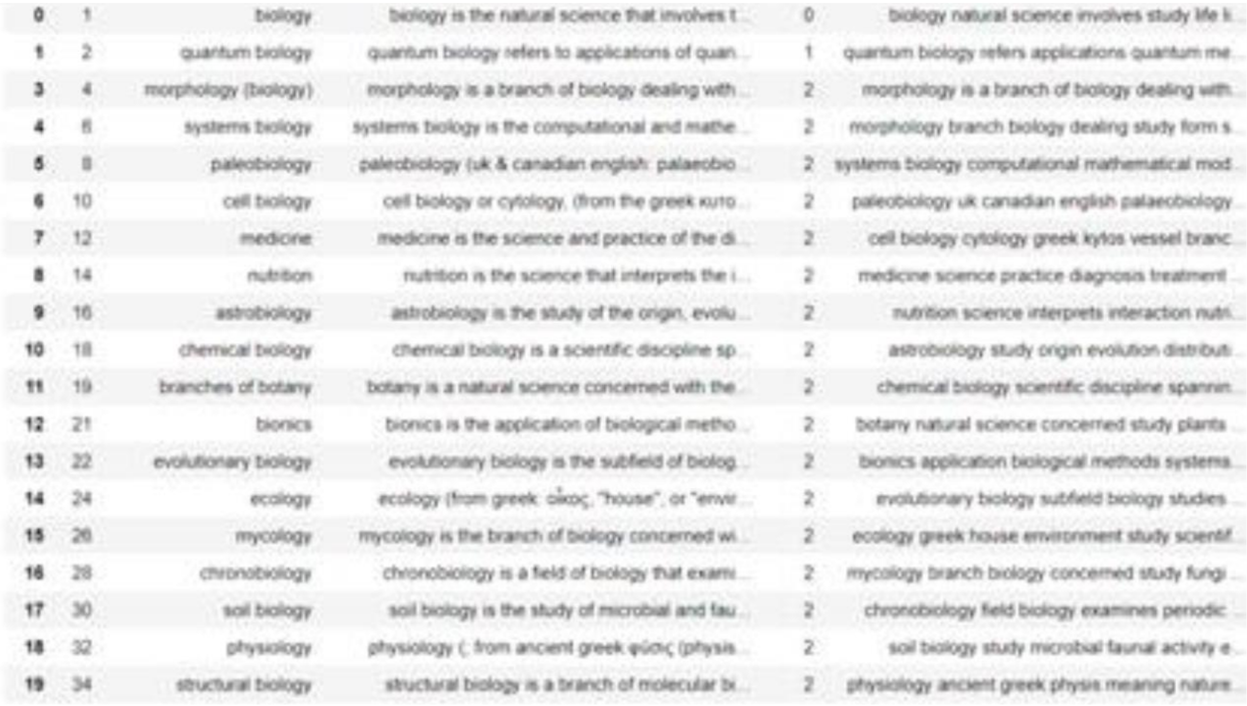
2.2. Data Preprocessing

The following are the steps that we performed to clean the dataset which we are using:

a. Removing the stop words
b. Stemming the data
c. Removed all the lists articles from the corpus ('Lists of Organisms by population')

2.3. Feature Engineering

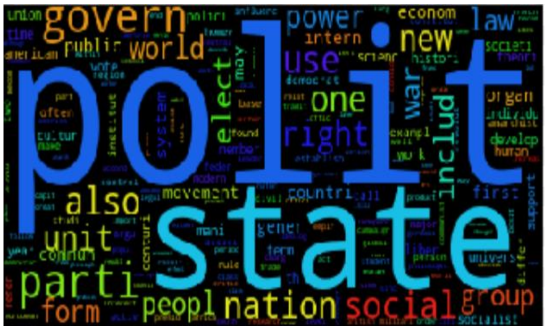
There was not major requirement for feature engineering in our project. But we added a column in our dataframe named 'Level' in which each entry defines the current level of the article.



0	1	biology	biology is the natural science that involves t...	0	biology natural science involves study life k...
1	2	quantum biology	quantum biology refers to applications of quan...	1	quantum biology refers applications quantum me...
3	4	morphology (biology)	morphology is a branch of biology dealing with...	2	morphology is a branch of biology dealing with...
4	6	systems biology	systems biology is the computational and mathe...	2	morphology branch biology dealing study form s...
5	8	paleobiology	paleobiology (uk & canadian english: palaeobio...	2	systems biology computational mathematical mod...
6	10	cell biology	cell biology or cytology, (from the greek κυτο...	2	paleobiology uk canadian english palaeobiology...
7	12	medicine	medicine is the science and practice of the di...	2	cell biology cytology greek kytos vessel branc...
8	14	nutrition	nutrition is the science that interprets the i...	2	medicine science practice diagnosis treatment ...
9	16	astrobiology	astrobiology is the study of the origin, evolu...	2	nutrition science interprets interaction nutri...
10	18	chemical biology	chemical biology is a scientific discipline sp...	2	astrobiology study origin evolution distributi...
11	19	branches of botany	botany is a natural science concerned with the...	2	chemical biology scientific discipline spannin...
12	21	bionics	bionics is the application of biological metho...	2	botany natural science concerned study plants ...
13	22	evolutionary biology	evolutionary biology is the subfield of biolog...	2	bionics application biological methods systems...
14	24	ecology	ecology (from greek: οἶκος, "house", or "envir...	2	evolutionary biology subfield biology studies ...
15	26	mycology	mycology is the branch of biology concerned wi...	2	ecology greek house environment study scientif...
16	28	chronobiology	chronobiology is a field of biology that exami...	2	mycology branch biology concerned study fungi ...
17	30	soil biology	soil biology is the study of microbial and fau...	2	chronobiology field biology examines periodic ...
18	32	physiology	physiology (; from ancient greek φύσις (physis...	2	soil biology study microbial faunal activity e...
19	34	structural biology	structural biology is a branch of molecular bi...	2	physiology ancient greek physis meaning nature...

Fig. 2: Corpus after preprocessing

The following is the word cloud that we made for better visualization of the corpus of the top level hierarchies like finance, biology, politics:



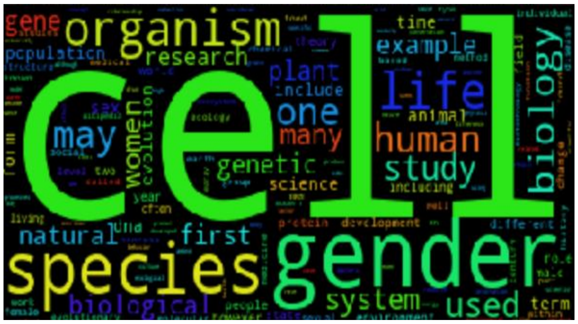


Fig. 3: Politics Wordcloud

Fig. 4: Biology Wordcloud

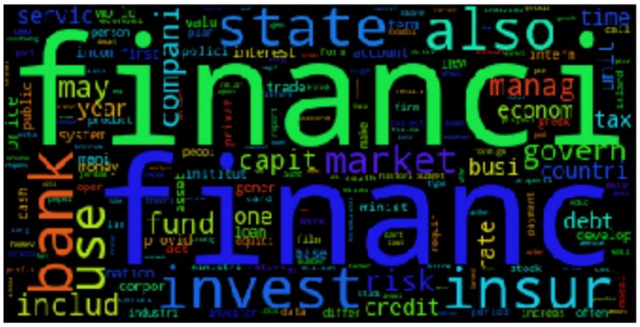


Fig. 5: Finance Wordcloud

3. Experiment

3.1. TF-IDF

It is the value that let us know how important the article is from all the articles in the corpus which we created for Wikipedia and also the value of it increases as the no. of times the words appears in the document and is adjusted by word frequency.

The tf-idf is given by:

$$\text{tf-idf}(t,d) = \text{tf}(t,d) \times \text{idf}(t)$$

The idf is defined as shown below:

$$\text{idf}(t) = log\frac{n_d}{1+\text{df}(d,t)}.$$

Where, nd is total # of documents, and df(d,t) is the number of documents that has or includes terms t

3.2. Bag of Words Representation

The BOW of the data we used was created which means that we vectorized the words into numerical features. Hence one of the below can be done for the same:

- **Tokenizing:** Giving an integer id for each possible token, by tokenizing string using white spaces
- **Counting :** Counting the token occurences in each document
- **Normalizing:** weighing the importance of tokens with respect to articles

To get the BOW for our project we used gensim.corpora.dictionary.Dictionary.doc2bow function of python gensim class thereafter converted documents to vectors.

For instance, vec(Finance) = Vec(all articles that fall under finance category)

3.3. Similarity Calculation

For calculating the similarities, we have worked with the following approaches:

a. Bottom-Up Approach
b. Top-Down Approach

3.3.1 Bottom Up Approach

The following steps were performed for this approach:

1. An article was fetched in order to compare

2. The same was converted to tf-idf representation
3. This was then compared to the summary vectors that were created for the top level categories which gave us the category that was most similar at the top level
4. Once we get the top level hierarchy, fetch the most similar document by comparing the new article with the corpus containing the articles related to the top level hierarchy
5. The article which we obtained was a parent article this is what we assume
6. Intermediate corpus was made by fetching only those articles that are above the parent article found above
7. The article that is most similar to the parent article obtained in step 5 was found
8. Repeat steps 5 to 7 until we reach the top level hierarchy

3.3.2 Top-Down Approach

The following steps were performed for this approach:

1. An article was fetched in order to compare
2. The same was converted to tf-idf representation
3. This was then compared to the summary vectors that were created for the top level categories which gave us the category that was most similar at the top level
4. The children which is immediate one of the top level category was fetched and the most similar article among these was found
5. Repeat step 4 until we reach the leaf node or till we get the article in the tree

3.3.3. Word2Vec

1. We also tried the Word2Vec approach to find the similarity amongst the articles. We converted the corpus as well as the new article into a word2vec representation.

2. Computed the Word2Vec using a predefined function Word2Vec under packages “gensim.models”.

3. Then after that we converted both the word2vec representation into np vectors.

4. Computed the cosine_similarity, using a predefined function in Python under package:

“sklearn.metrics.pairwise”.

5. cosine_similarity(vector1, vector2)

6. We were not able to go ahead with this approach as the outcomes were not satisfactory.

Upon application of both the methods, we found the top down approach to be more efficient since we don’t have to

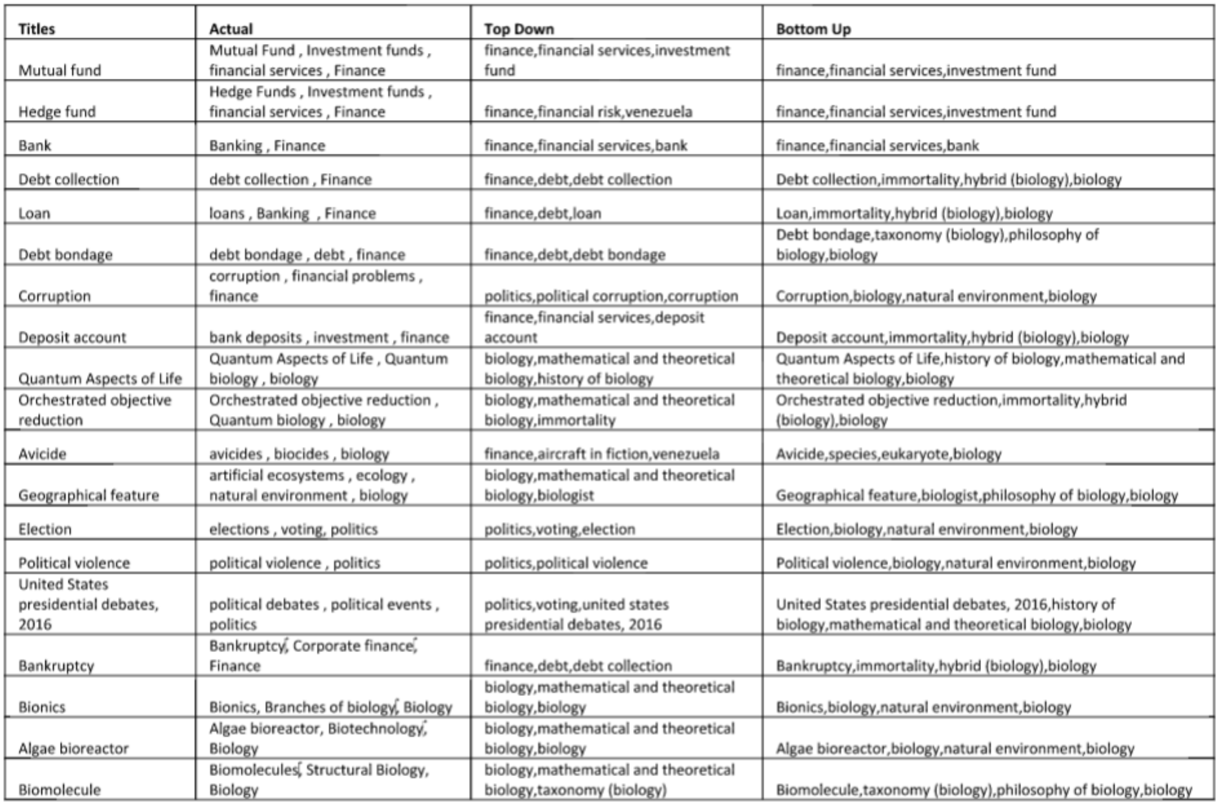
compare with all the articles from the bottom.

4. Results and Analysis

To analyze our result, we took twenty articles related to one of the top level categories from Wikipedia and extracted the hierarchy by implementing both the top down and bottom up approaches.

The results were then compared with the actual hierarchy that was manually extracted for these 20 articles.

The following were the results obtained –



Titles	Actual	Top Down	Bottom Up
Mutual fund	Mutual Fund , Investment funds , financial services , Finance	finance,financial services,investment fund	finance,financial services,investment fund
Hedge fund	Hedge Funds , Investment funds , financial services , Finance	finance,financial risk,venezuela	finance,financial services,investment fund
Bank	Banking , Finance	finance,financial services,bank	finance,financial services,bank
Debt collection	debt collection , Finance	finance,debt,debt collection	Debt collection,immortality,hybrid (biology),biology
Loan	loans , Banking , Finance	finance,debt,loan	Loan,immortality,hybrid (biology),biology
Debt bondage	debt bondage , debt , finance	finance,debt,debt bondage	Debt bondage,taxonomy (biology),philosophy of biology,biology
Corruption	corruption , financial problems , finance	politics,political corruption,corruption	Corruption,biology,natural environment,biology
Deposit account	bank deposits , investment , finance	finance,financial services,deposit account	Deposit account,immortality,hybrid (biology),biology
Quantum Aspects of Life	Quantum Aspects of Life , Quantum biology , biology	biology,mathematical and theoretical biology,history of biology	Quantum Aspects of Life,history of biology,mathematical and theoretical biology,biology
Orchestrated objective reduction	Orchestrated objective reduction , Quantum biology , biology	biology,mathematical and theoretical biology,immortality	Orchestrated objective reduction,immortality,hybrid (biology),biology
Avicide	avicides , biocides , biology	finance,aircraft in fiction,venezuela	Avicide,species,eukaryote,biology
Geographical feature	artificial ecosystems , ecology , natural environment , biology	biology,mathematical and theoretical biology,biologist	Geographical feature,biologist,philosophy of biology,biology
Election	elections , voting, politics	politics,voting,election	Election,biology,natural environment,biology
Political violence	political violence , politics	politics,political violence	Political violence,biology,natural environment,biology
United States presidential debates, 2016	political debates , political events , politics	politics,voting,united states presidential debates, 2016	United States presidential debates, 2016,history of biology,mathematical and theoretical biology,biology
Bankruptcy	Bankruptcy, Corporate finance, Finance	finance,debt,debt collection	Bankruptcy,immortality,hybrid (biology),biology
Bionics	Bionics, Branches of biology, Biology	biology,mathematical and theoretical biology,biology	Bionics,biology,natural environment,biology
Algae bioreactor	Algae bioreactor, Biotechnology, Biology	biology,mathematical and theoretical biology,biology	Algae bioreactor,biology,natural environment,biology
Biomolecule	Biomolecules, Structural Biology, Biology	biology,mathematical and theoretical biology,taxonomy (biology)	Biomolecule,taxonomy (biology),philosophy of biology,biology

Fig. 6: Results

From the above table we observe that the top down is a better approach then bottom up for extracting the hierarchy (The hierarchies of the Wikipedia are subjective). Hence there can be articles that has multiple hierarchies and also vice versa that the parent article can also have multiple children.

For example, take 'Election' which we have taken above has following hierarchy:

Election → Voting → Politics

An alternate hierarchy could be constructed as follows –

Election → Political Events → Politics

If we see both the above hierarchies, both of them makes sense and are correct

5. Conclusion

To conclude, the top down approach gave us the good results for extraction of Wikipedia hierarchy for a given topic. But there are some articles that can have multiple hierarchies for a particular page and this is due to subjective nature of the hierarchies on Wikipedia articles. As we know there can be other ways too for extracting the hierarchy which can be efficient so the future scope will be that we can explore the techniques such as LSA, Word2Vec etc.