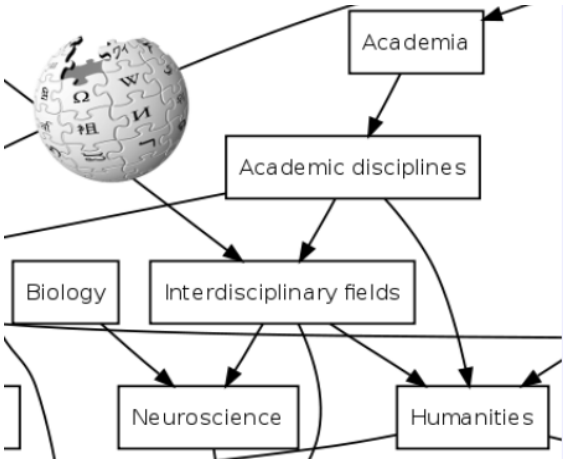


Wikipedia Hierarchy Extraction.

Final Report

CS-522: Advance Data Mining

Submitted By:

Animesh Patni (A20403240)

Pooja Patel (A20396099)

Krishna Bharadwaj (A20398222)

Content Table:

Abstract:

In this project work, we are working on finding the correct Wikipedia hierarchy level, for any new incoming article. Wikipedia is very unstructured and a huge dataset to work on. It is a huge pool of information and English Wikipedia contains as many as 5,525,674 articles. These articles are in a dense hierarchy, with a few top level and bubbles down to the leaf nodes. The structure of these hierarchy is very dense and well defined, this is mainly due to the fact that the articles are submitted by the users. This could mean there's a possibility that many people can name the same hierarchy model for two or three different pages. In this work our is towards more data centric approach towards the categorization of wiki articles. We attempt to make use of the dataless approach that has been described in Yangqiu Song et al in their paper 'On Dataless Hierarchical Text Classification'.

Introduction:

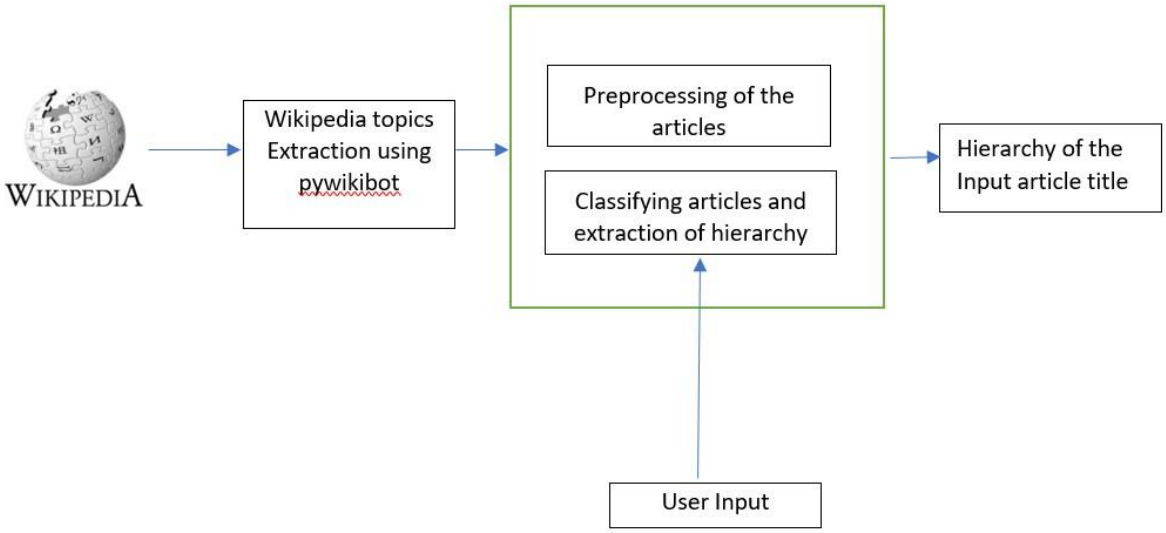
Wikipedia Hierarchy Extraction extracts the hierarchy of the new incoming article entered by the user. This technique helps in the categorizing the data and can work as a search engine for information browsing as it considerably reduces the time to browse information. For this we should have a good understanding about the data we are working on, that is Wikipedia Hierarchy and we should be able to extract the Hierarchy. We have extracted three categories, we will take the input from the user as an article title and then extract the hierarchy.

Topics we used are:

We have taken some of the most common topics:

- Finance
- Biology
- Politics

The Figure below shows the Data Flow and Architecture of Wikipedia Hierarchy Extraction:



THE ARCHITECTURE

The above figure shows the flow of our model:

- Fetching the top-level categories like Finance, Biology, Politics etc.
- Then extracting the articles in the above hierarchy and pre-processing them.
- Fetch the hierarchy by getting the most similar page at each level.

Dataset:

Data Procurement:

We need to get the data and Wikipedia dumps were our primary source of data. The process to get the data is as follow:

- Read the latest category xml dumps file available on https://dumps.wikimedia.org/enwiki/latest/.
- The file downloaded contains information about the categories defined in Wikipedia.
- We extracted the category tree. These categories include 'Finance', 'Biology' and 'Biology'. For our experiments, we consider these categories as the top level categories and hence they are called Level 0.
- We used an API "pywikibot" for extracting the hierarchy which is written in Python.
- We have only extracted data upto two levels cause as we increased the depth, the categories extracted got irrelevant.

- After getting the categories, we crawl Wikipedia site for articles. This is our dataset for further evaluations.

Data Pre-processing:

- We removed the stop words.
- Then we went on to do Data Stemming.
- Removing the list articles from the corpus, for example: “Lists of organisms by population”.

Feature Engineering:

- In this we added one column which tells us the level, which pointed to the current level of the article in the hierarchy as a whole.

FIGURE SHOWS THE CORPUS AFTER PRE-PROCESSING:

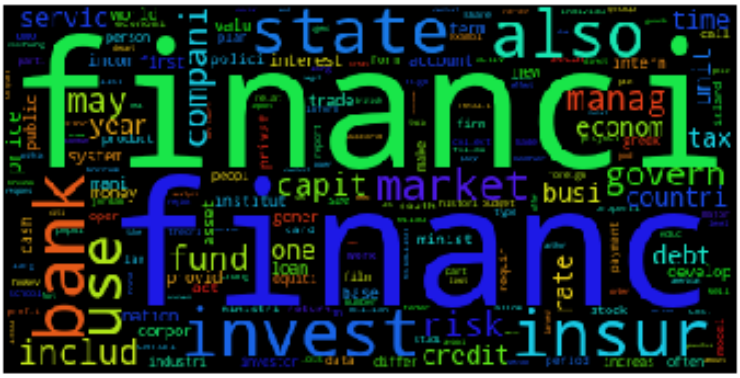
	ID	Title	Text	Level	new_column
0	1	biology	biology is the natural science that involves t...	0	biology natural science involves study life li...
1	2	quantum biology	quantum biology refers to applications of quan...	1	quantum biology refers applications quantum me...
3	4	morphology (biology)	morphology 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 kyto...	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...

For better Visualization, we have generated a word clouds:

POLITICS:

BIOLOGY:

FINANCE:



Experiments:

TF-IDF:

Term Frequency- Inverse Document Frequency gives us the numerical value, indicating the importance of each Wikipedia article in the corpus of wiki article. The TF-IDF value increases exponentially according to the number of times a word occurs in document and is adjusted by the frequency of it.

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

The inverse document Frequency is given as:

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

Bag of Words Representation:

Converting the corpus into Bag of Words Representation. This involves vectorizing the text into numerical features. This can be implemented as:

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

We have used gensim.corpora.dictionary.Dictionary.doc2bow function of python genism class to get the Bag of Words representation.

Then we convert this into vectors.

For example:

Vec(Biology): Vec(all articles under biology)

Calculating Similarity:

Bottom Up Approach:

- Get an article.
- Convert this article into TF_IDF representation.
- Compare this representation with the summary vectors that were created for the top level categories. This would eventually give us the most similar top level category.

- Once we get the hierarchy, fetch the most similar document by comparing the new article with the corpus containing the articles related to the top level hierarchy.
- At this point we assume that the article obtained above is the parent article.
- Make the corpus of the articles that are just above the parent article found above.
- Find the article that are most similar to the parent article obtained in step above.
- Repeat the last 3 steps.

Top Down Approach:

- Get an article.
- Transform an article into its TF_IDF representation.
- Compare this with the summary vector that were created for the top-level categories. This will give us the most similar document.
- Fetch the immediate children from the top and find the most similar article amongst them.
- Repeat step 4 until we get the hierarchy.

Word2Vec Approach:

- 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.
- Computed the Word2Vec using a predefined function Word2Vec under packages "gensim.models".
- Then after that we converted both the word2vec representation into np vectors.
- Computed the cosine_similarity, using a predefined function in Python under package: "sklearn.metrics.pairwise".
    - cosine_similarity(vector1, vector2)
- We were not able to go ahead with this approach as the outcomes were not satisfactory.

Upon implementing all the approaches, we found that the Top Down approach to be more efficient and less time consuming as we don't have to compare all the leaf node articles to go up the hierarchy.

Result and Analysis:

For the analysis part, we picked 20 articles related to one of the three selected categories from Wikipedia and found result using both top down and bottom up approaches.

The evaluation was done with the actual hierarchy, extracted for these 20 categories.

The result obtained are:

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

As we can see above, the top down approach seems to do better compared to the bottom up approach. However, while making comparisons, it has to be kept in mind that the actual hierarchy that is available in Wikipedia is subjective. That is, it is the view of the user submitting the article may be different for every user. As a result, it cannot be a hard and fast rule that the actual hierarchy that exists on Wikipedia is the only hierarchy that exists. Hence, for a given hierarchy, there is always a possibility of having multiple hierarchies for a given article. This is illustrated by taking the example of **Election** which we have taken above.

Election →Voting → Politics

An alternate hierarchy could be constructed as follows –

Election →Political Events → Politics

Conclusion:

As mentioned above the Top down approach yields a better outcome and a more efficient outcome for extracting the hierarchy of the Wikipedia hierarchies given a particular topic. But there are chances that there can be multiple hierarchies for a particular topic.

For future work we would like to implement this using different techniques like Word2Vec and Semantic Analysis, as there might be more suitable techniques to find the Wikipedia Topic Hierarchy.