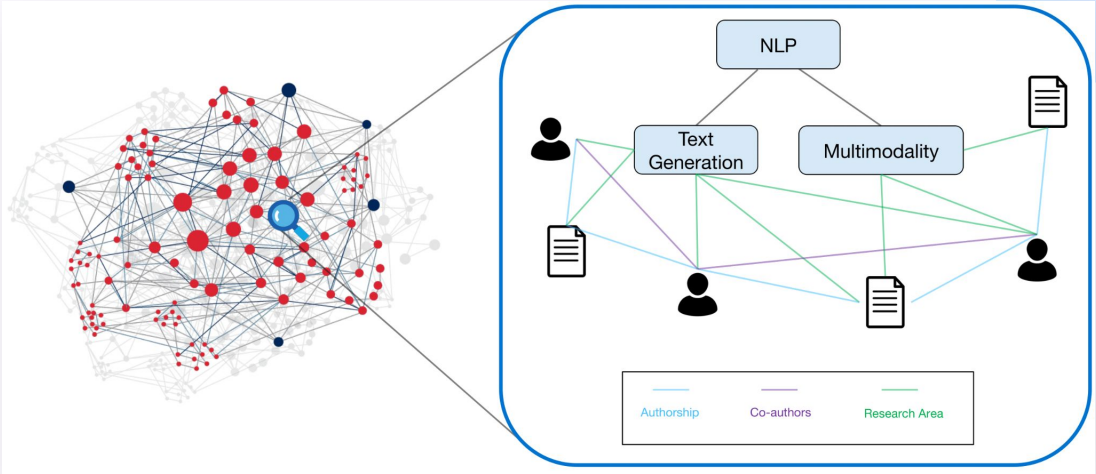
Text Analysis

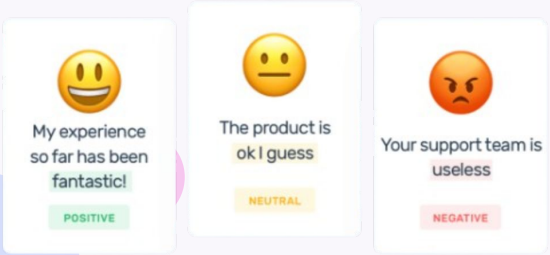
Natural Language Processing by using Graph Theory

Table of Content

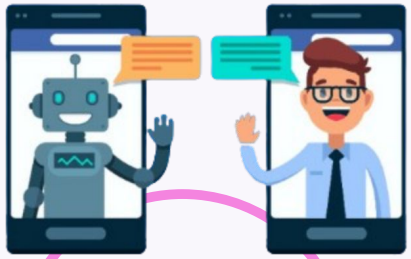
Why NLP with Graph Theory?



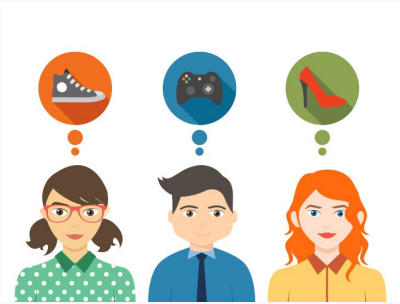
Sentiment Analysis



Chatbot and Virtual Assistants



Penalization



01

Overview about Dataset

Overview about dataset from Reuters

The Reuters-21578 Dataset

The Reuters-21578 dataset is a well-known benchmark in the field of natural language processing, specifically for text categorization and document classification tasks. It contains a collection of **Reuters news articles, and it is often used to evaluate the performance of NLP algorithms**.The original dataset includes a set of 21,578 news articles that were published in the financial Reuters newswire in 1987 from library nltk. It consist of 90 different category. Since, the original dataset of Reuters is quite skew, this chapter use a modified version of its known as **ApteMod from Reuters-21578**.

The Reuters-21578 Dataset

	id	text	label
0	test/14826	ASIAN EXPORTERS FEAR DAMAGE FROM U.S.-JAPAN RI...	[trade]
1	test/14828	CHINA DAILY SAYS VERMIN EAT 7-12 PCT GRAIN STO...	[grain]
2	test/14829	JAPAN TO REVISE LONG-TERM ENERGY DEMAND DOWNWA...	[crude, nat-gas]
3	test/14832	THAI TRADE DEFICIT WIDENS IN FIRST QUARTER Th...	[corn, grain, rice, rubber, sugar, tin, trade]
4	test/14833	INDONESIA SEES CPO PRICE RISING SHARPLY Indon...	[palm-oil, veg-oil]
...	...	...	...
10783	training/999	U.K. MONEY MARKET SHORTAGE FORECAST REVISED DO...	[interest, money-fx]
10784	training/9992	KNIGHT-RIDDER INC <KRN> SETS QUARTERLY Qtl...	[earn]
10785	training/9993	TECHNITROL INC <TNL> SETS QUARTERLY Qtly d...	[earn]
10786	training/9994	NATIONWIDE CELLULAR SERVICE INC <NCEL> 4TH ...	[earn]
10787	training/9995	<A.H.A. AUTOMOTIVE TECHNOLOGIES CORP> YEAR ...	[earn]

Overview on Main Concept of NLP

Then, we will show you some of the main concepts that can be used for dealing with unstructured text data which be able to extract structured information, so that it can be used with ease and detailed of each step.

At the Beginning

Mostly, analytical engines that are used in NLP tasks are trained on documents in a specific language and should only be used for such a language.

```
#Detect Language within Each Article of dataset
from langdetect import detect
import numpy as np
def getLanguage(text: str):
    try:
        return detect(text)
    except:
        return np.nan
corpus["language"] = corpus["text"].apply(detect)
```

```
pd.DataFrame(corpus['language'].value_counts())
```

language	count
en	9906
sv	420
de	374
sw	29
so	25

Text Segmentation and Tokenization

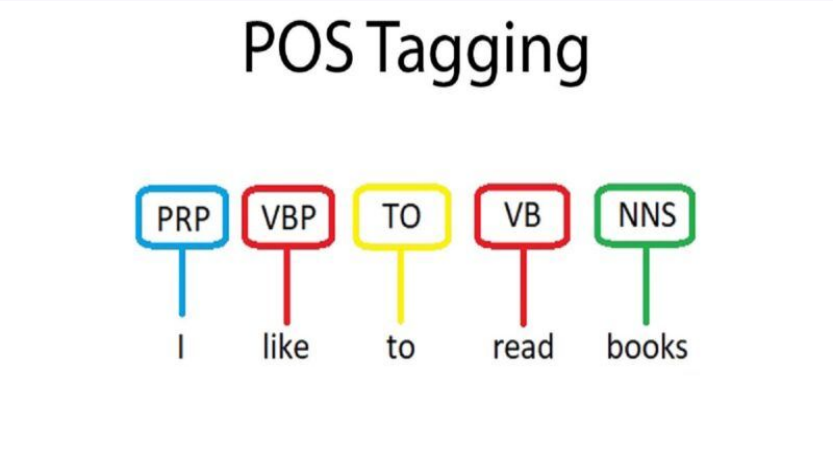
Tokenization is used to break down the text into individual units in order to provide a foundation for further analysis


```
for sent in parsed.sents:
    for token in sent:
        print(token, end=',')
```

```
ASIAN,EXPORTERS,FEAR,DAMAGE,FROM,U.S.-JAPAN,RIFT, ,Mounting,trade,friction,between,the, ,U.S.,And,Japan,has,raised,fears,amo
ng,many,of,Asia,'s,exporting, ,nations,that,the,row,could,inflict,far,-,reaching,economic, ,damage,,,businessmen,and,officia
ls,said,.,      ,They,told,Reuter,correspondents,in,Asian,capitals,a,U.S., ,Move,against,Japan,might,boost,protectionist,sent
iment,in,the, ,U.S.,And,lead,to,curbs,on,American,imports,of,their,products,.,      ,But,some,exporters,said,that,while,the,c
onflict,would,hurt, ,them,in,the,long,-,run,,,in,the,short,-,term,Tokyo,'s,loss,might,be, ,their,gain,.,      ,The,U.S.,Has,s
aid,it,will,impose,300,mln,dlrs,of,tariffs,on, ,imports,of,Japanese,electronics,goods,on,April,17,,,in, ,retaliation,for,Jap
an,'s,alleged,failure,to,stick,to,a,pact,not, ,to,sell,semiconductors,on,world,markets,at,below,cost,.,      ,Unofficial,Japa
nese,estimates,put,the,impact,of,the,tariffs, ,at,10,billion,dlrs,and,spokesmen,for,major,electronics,firms, ,said,they,woul
d,virtually,halt,exports,of,products,hit,by,the, ,new,taxes,.,      ,",We,would,n't,be,able,to,do,business,,,",said,a,spokesm
an,for, ,leading,Japanese,electronics,firm,Matsushita,Electric, ,Industrial,Co,Ltd,&,lt;MC.T,>,.,      ,",If,the,tariffs,rema
in,in,place,for,any,length,of,time, ,beyond,a,few,months,it,will,mean,the,complete,erosion,of, ,exports,(,of,goods,subject,t
```

Part-of-Speech

After the text has been divided into its single words (referred to as token), the next step is to associate each token with a Part-of-Speech (PoS) tag; that is its grammatical type to infer those tags which are usually nouns, verbs, auxiliary verbs, adjective and so on.




```
tokens=[]
posTag=[]
# Display Part-of-Speech tags for each token
for token in parsed:
    tokens.append(token.text)
    posTag.append(token.pos_)
pd.DataFrame(data={
    'Tokens': tokens,
    'PoS Tags': posTag
})
```

	Tokens	PoS Tags
0	ASIAN	ADJ
1	EXPORTERS	PROPN
2	FEAR	VERB
3	DAMAGE	NOUN
4	FROM	ADP
...	...	...
905	end	VERB
906	the	DET
907	dispute	NOUN
908	.	PUNCT
909		SPACE

910 rows × 2 columns

Named Entity Recognition (NER)

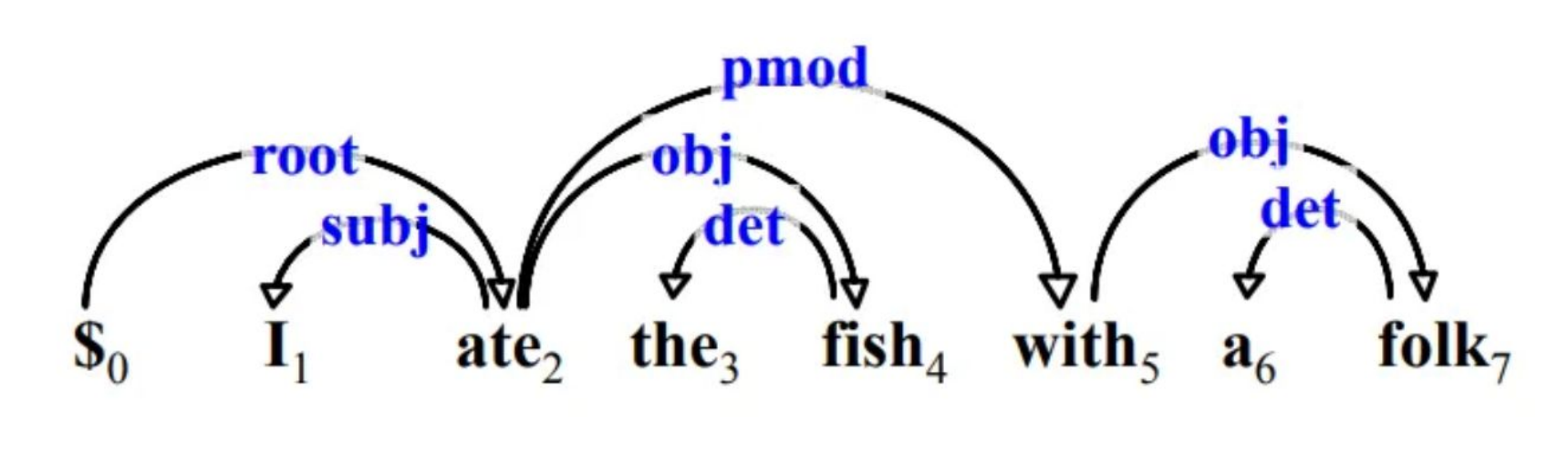
This step is generally a statistical model that is trained to recognize the type of nouns that appear within the text. For instances of entities are **Organization, Person, Geographic Location** and **Addresses, Products, Numbers,** and **Currencies**.

```
# Visualize entities using displacy
spacy.displacy.render(parsed, style="ent", jupyter=True)
```

ASIAN NORP EXPORTERS FEAR DAMAGE FROM U.S.-JAPAN RIFT ORG Mounting trade friction between the U.S. GPE And Japan GPE has raised fears among many of Asia LOC 's exporting nations that the row could inflict far-reaching economic damage, businessmen and officials said. They told Reuter PERSON correspondents in Asian NORP capitals a U.S. GPE Move against Japan GPE might boost protectionist sentiment in the U.S. GPE And lead to curbs on American NORP imports of their products. But some exporters said that while the conflict would hurt them in the long-run, in the short-term Tokyo GPE 's loss might be their gain. The U.S. GPE Has said it will impose 300 CARDINAL mln dlrs of tariffs on imports of Japanese NORP electronics goods on April 17 DATE , in retaliation for Japan GPE 's alleged failure to stick to a pact not to sell semiconductors on world markets at below cost. Unofficial Japanese NORP estimates put the impact of the tariffs at 10 billion CARDINAL dlrs and spokesmen for major electronics firms said they would virtually halt exports of products hit by the new taxes. "We wouldn't be able to do business," said a spokesman for leading Japanese NORP electronics firm Matsushita Electric Industrial Co Ltd <MC.T ORG >. "If the tariffs remain in place for any length of time beyond a few months DATE it will mean the complete erosion of exports (of goods subject to tariffs) to the U.S. GPE ," said

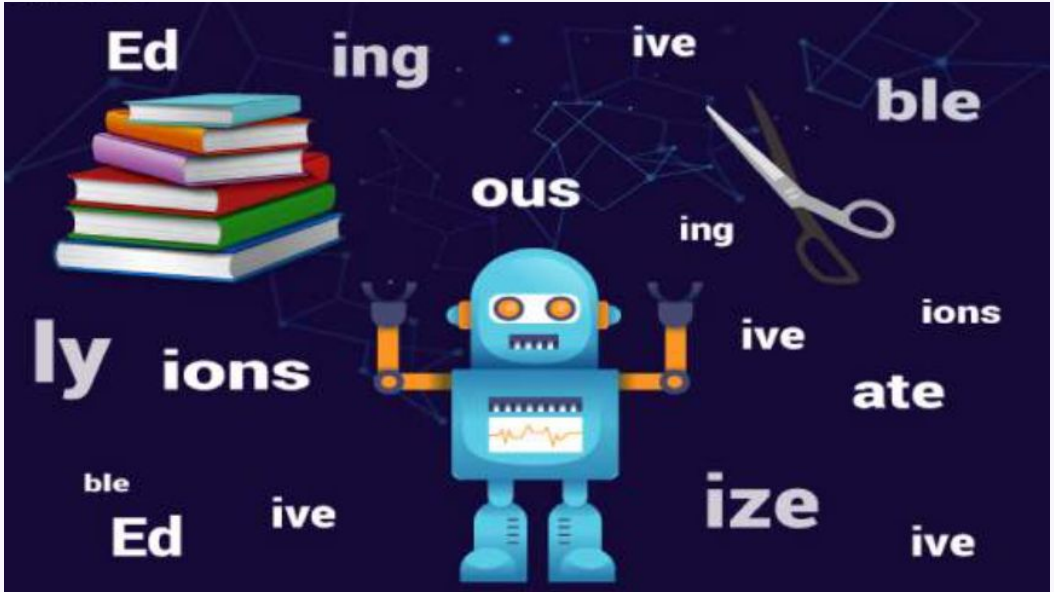
Dependency Parser

A dependency parse tree is **a way of showing the grammatical relationships** between the words in a sentence.



Lemmatizer

Finally, the last step of analytical pipeline is the lemmatizer which allows us **to reduce words to a common root** to provide a cleaner version of it, thus reducing the morphological variation of words. For instance, the verb “to be” has many morphological variations which are “is”, “are”, “was”, all of which are different, valid forms.



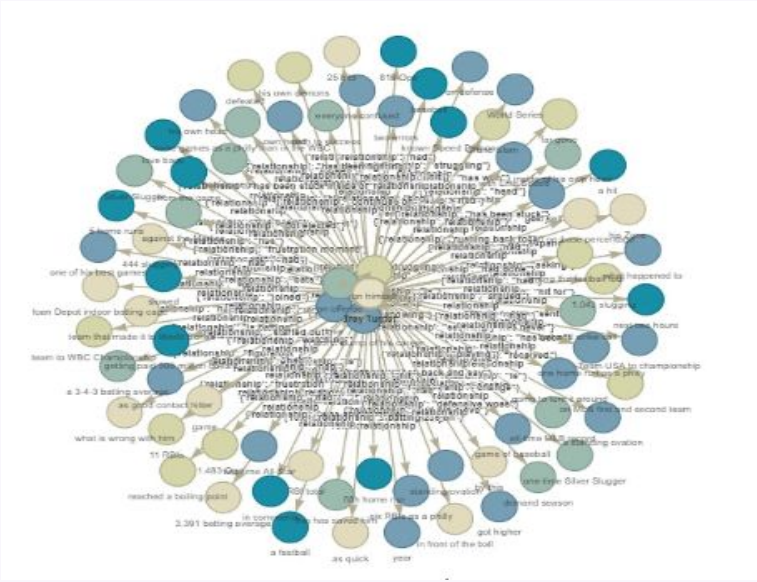
Next Section

In the next section, we will used the information that we have extracted in the previous section using the different text engines to build network that related to the different information and in order ***to capture the meaning of words and how they related to each other in a sentence to infer relationship between different entities***

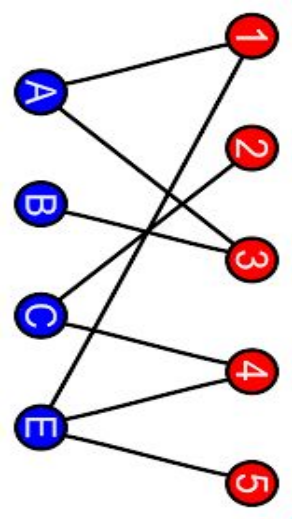
In particular, we will focus on two kinds of graphs, which are ***Knowledge-Based Graph*** and ***Bipartite-Graph***.

Entities: refer to the objects or concepts insight the text. In natural language processing, entities can **be people, organizations, locations, products,** etc. Extracting entities involves identifying and categorizing these elements within the text.

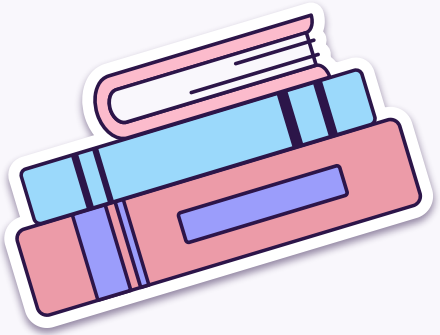
Knowledge Vs Bipartite



Vs



- Role of each one
- Capacity of each



02

Knowledge Graph

Implement knowledge graph to capture the meaning of words and how they related to each other in a sentence to infer relationship between different entities

Knowledge Graph

Knowledge graphs are very interesting as they **not only relate entities but also provide a direction and a meaning to the relationship**. For instance, let's take a look at the following relationship. For instance:

I (->) buy (->) a book

This is substantially different from the following relationship:

I (->) sell (->) a book

Implementation

To build a Knowledg Based Graph - we need a function that identifies the **Subject-Verb-Object (SVO)** for each sentence.

1st step **- Create a Triplet Function of SVO by spacy**
Then, this **triplet SVO function** gonna apply to all sentences in the corpus and it gonna aggregated to generate the graph to see what relationship of each sentence.

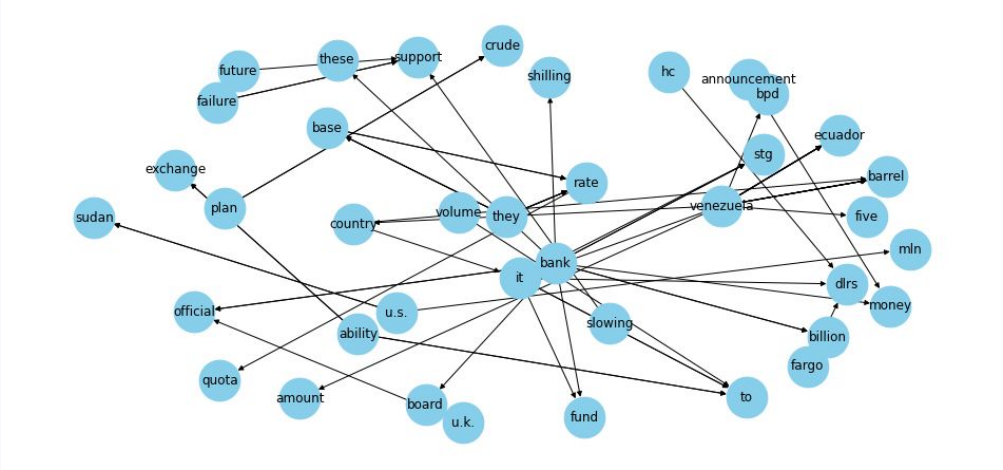
```
from subject_object_extraction import findSVOs

# Assuming corpus is a DataFrame
corpus["triplets"] = corpus["parsed"].apply(lambda x: findSVOs(x))
```

SVO Extractor

The SVO extractor is one of spacy model, which provide through the **dependency tree parser.** By tagging from tree, it helps to separate the main sentence from subordinate and identifies the Subject Verb Object

***2nd step* - Create Knowledge Based Graph**

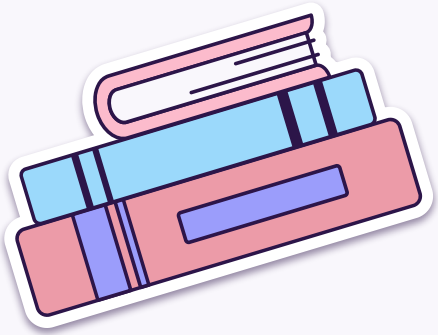
By using the given function in github, we can compute all triplets in the corpus and store them in corpus DataFrame and use netwrokx to plot the Knowledge Based Graph



Conclusion

Knowledge graphs **focus only on capturing direct relationship between entities** which make this graph **unable to capture overall meaning from one context to another** level of relationship within each sentence or across sentence within the text.

To address these limitations, we will encode the information present in the document in the form of a `bipartite graph`.



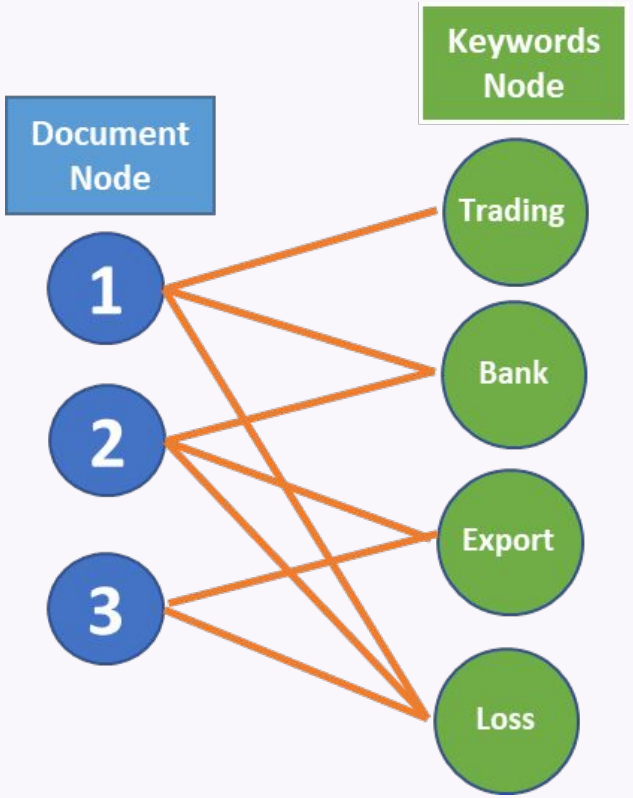
03

Bipartite Graph

To address the limitation of the knowledge graph on clustering each document based on their meaning, we will implement bipartite to cluster document based on their meaning.

Introducing Bipartite Graph for Text Analysis

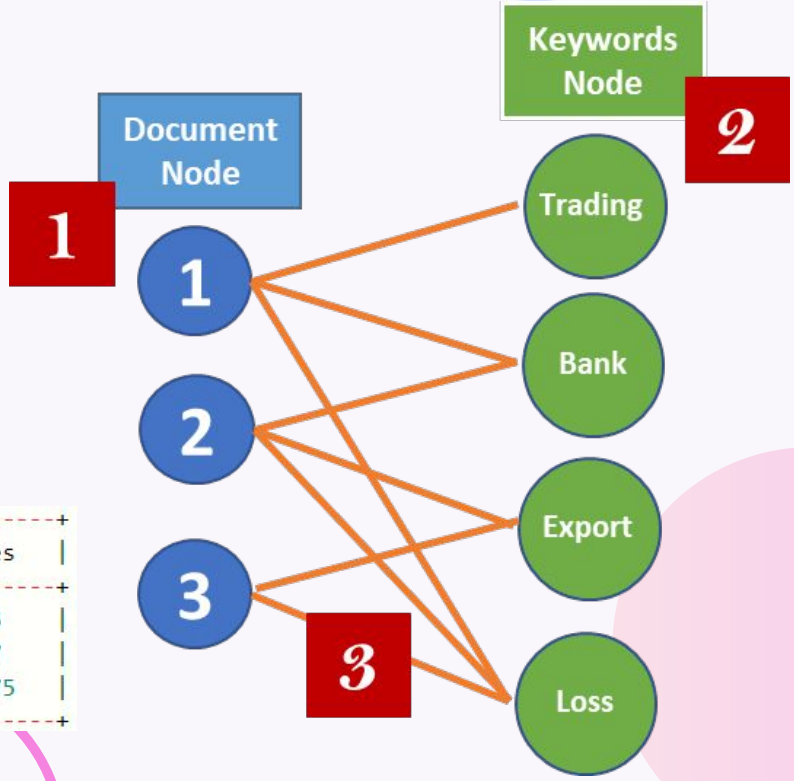
Bipartite graphs connect documents to the entities that appear in the text. The process includes projecting the bipartite graph into a homogeneous graph, consisting of either **document or entity nodes.**



Graph (Document-Entity Graph)

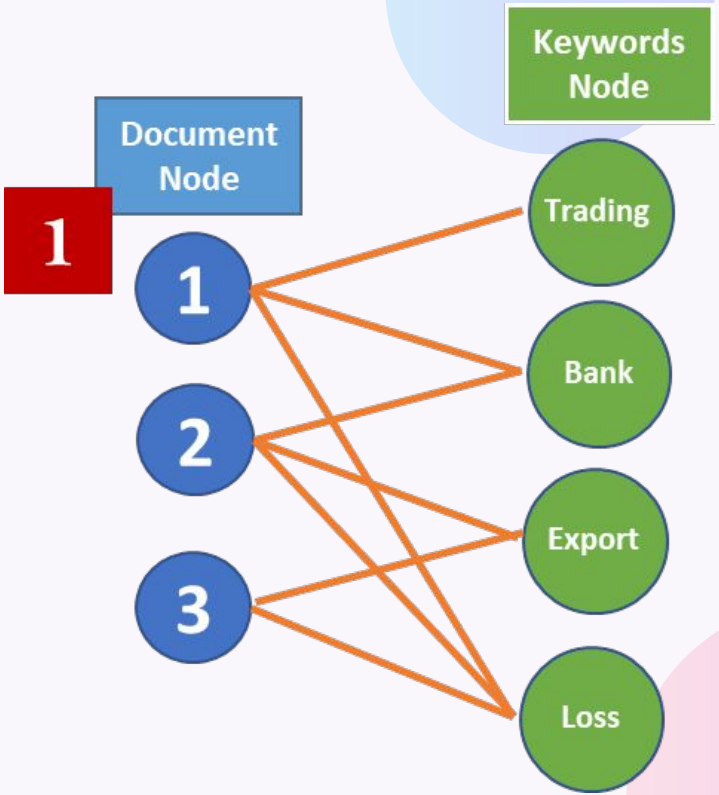
In other to build a bipartite graph we need 3 important elements to do so such as: 2 set of node (where one is document node representation and another is keywords node representation) and the edges between 2 set.

Document-Entity Graph	Values
Number of nodes	22665
Number of edges	71707
Average degree	6.3275



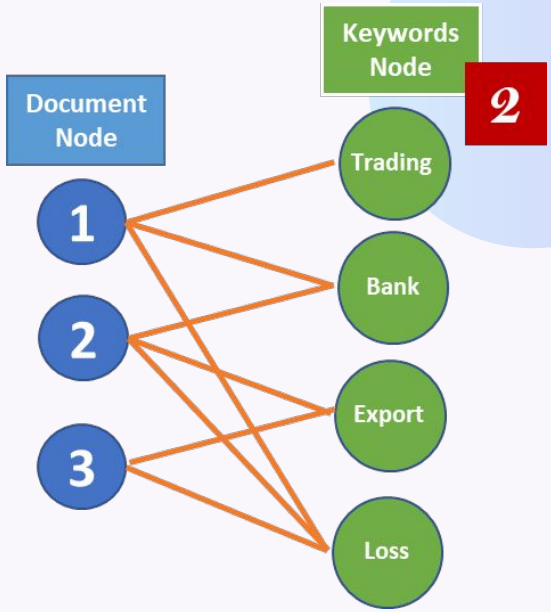
Implementing on Bipartite Graph

1.Document Node: since one row of corpus consist of one document text, so we going to treat document at index 1 as document 1, and so on.



Implementing on Bipartite Graph

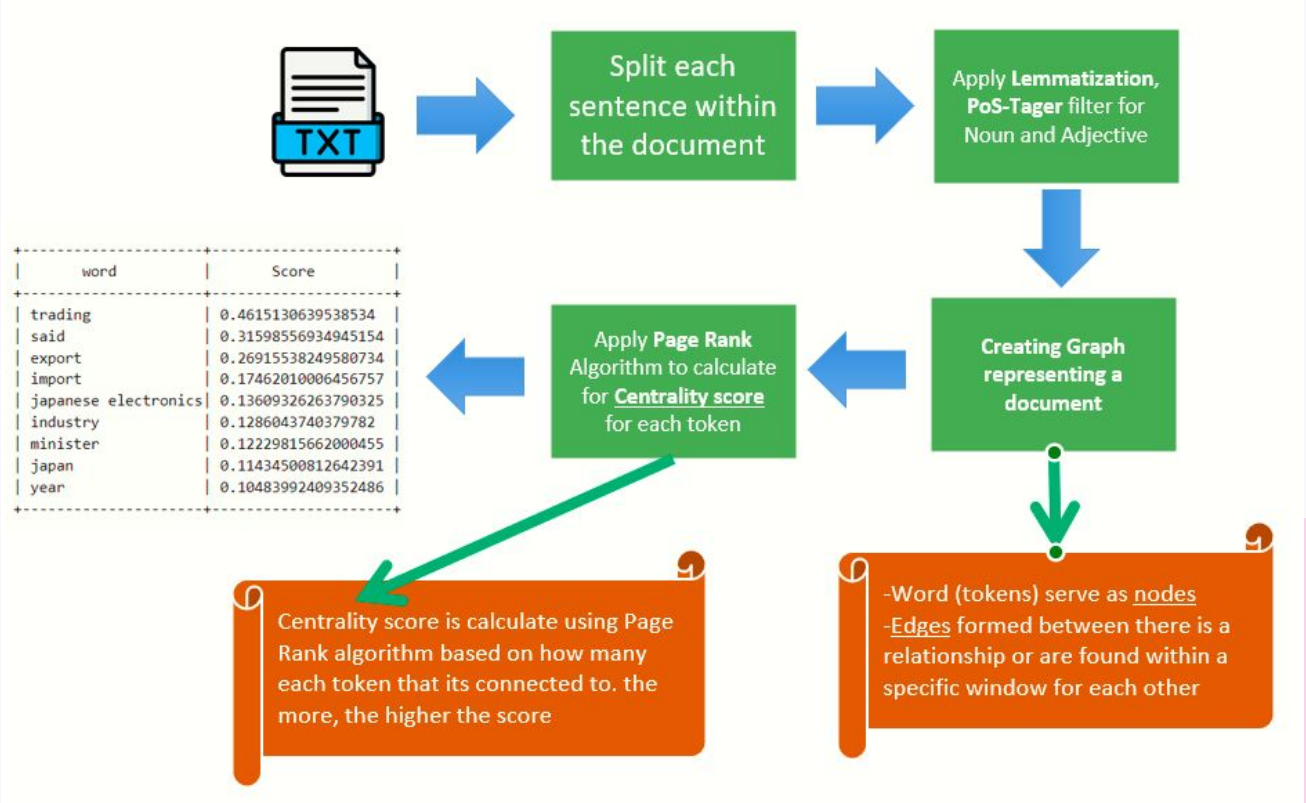
2.Keyword Node: for each document we going to extract keyword by using **Text Rank algorithm** to calculate **centrality score** for each token by using *gensim library* to represent how important the given token is.



word	Score
trading	0.4615130639538534
said	0.31598556934945154
export	0.26915538249580734
import	0.17462010006456757
japanese electronics	0.13609326263790325
industry	0.1286043740379782
minister	0.12229815662000455
japan	0.11434500812642391
year	0.10483992409352486

TextRank Algorithm to extract keyword and calculate score (edges) within each document

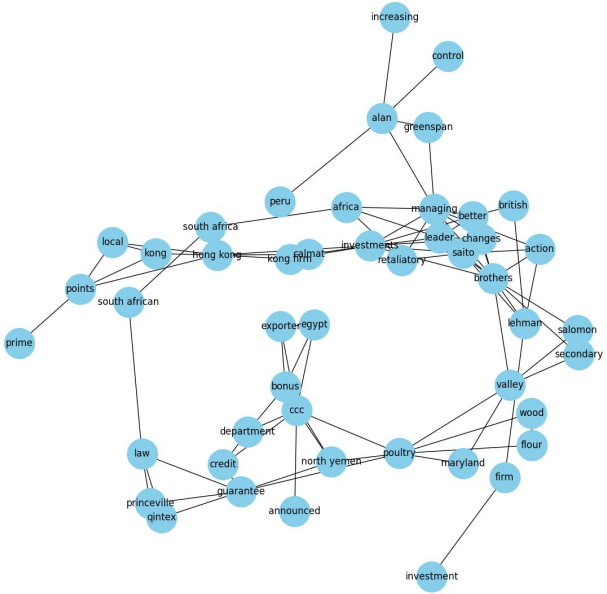
Workflow of how TextRank algorithm use to apply on each document



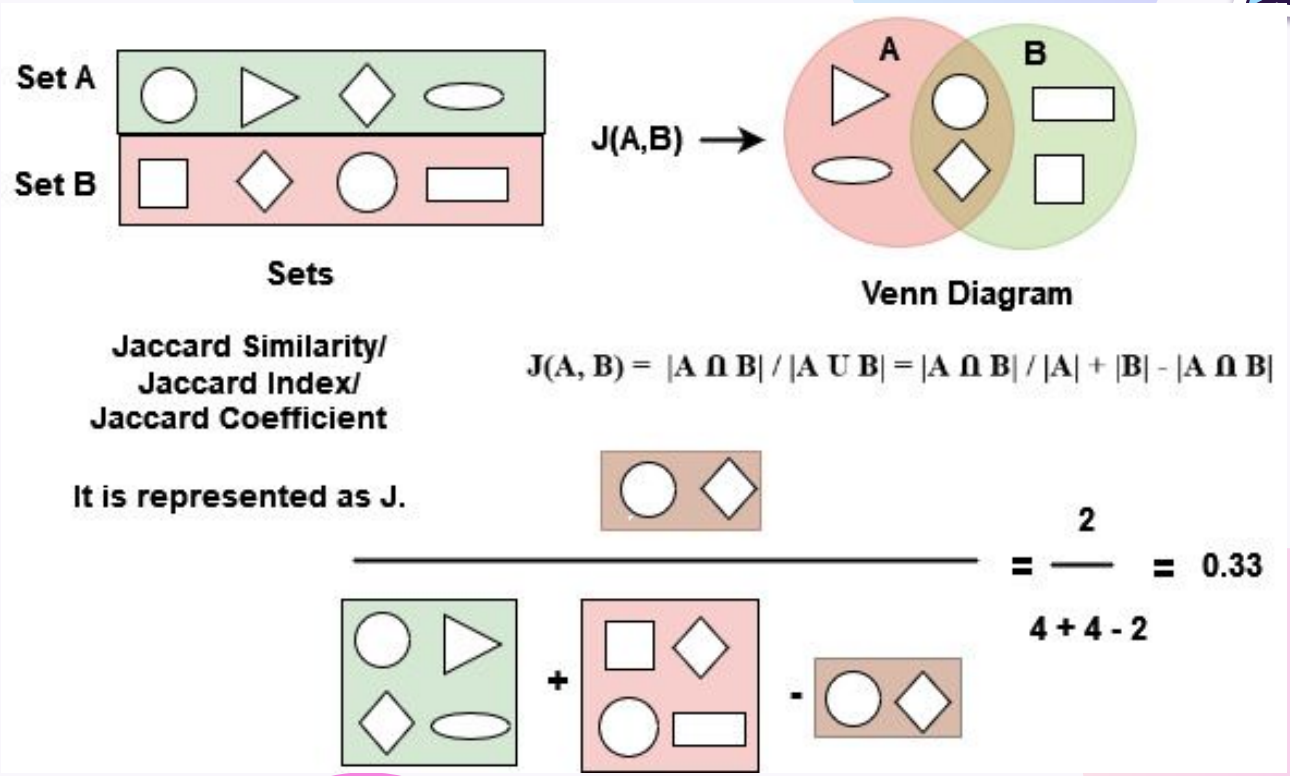
Entity-Entity Graph

- Now we will start project the bipartite graph on a set of entity node to create new graph that show **connection between entities based on their shared relationship within the document.**
- We will use **Jaccard Similarity** to obtain weighted between each node.
- We obtain **`Graph with 12863 nodes and 262365 edges`**

A community within a whole entity-entity graph



Jaccard Similarity between Two Sets

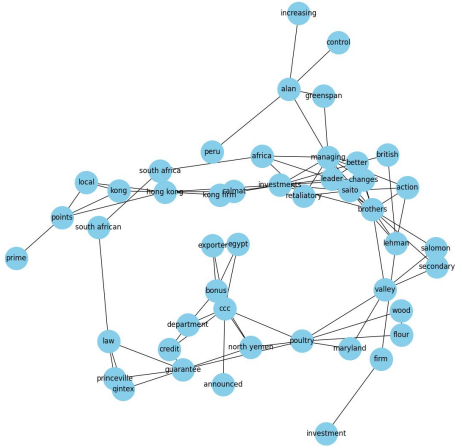


Entity-Entity Graph

It hard to analyst when it has so many node and edges, to reduce this complexity we will consider node that have a certain degree.
We will only focus on strong correlation that support by larger occurrences

Here the average degree of the the graph is too high. Let look at it distribution further

A community within a whole entity-entity graph

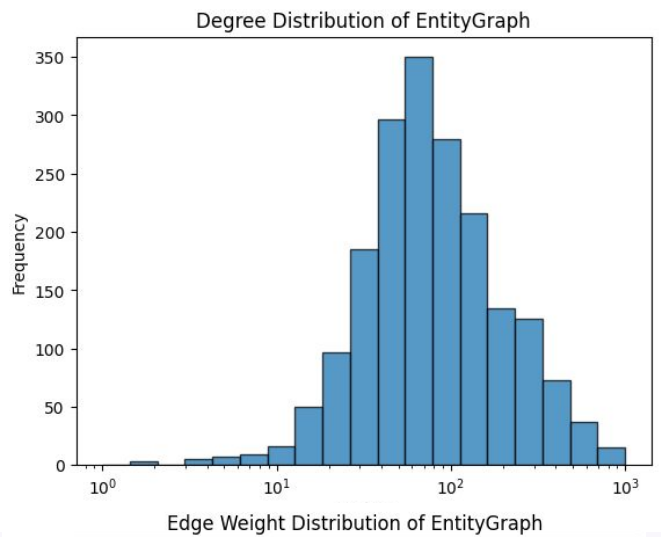


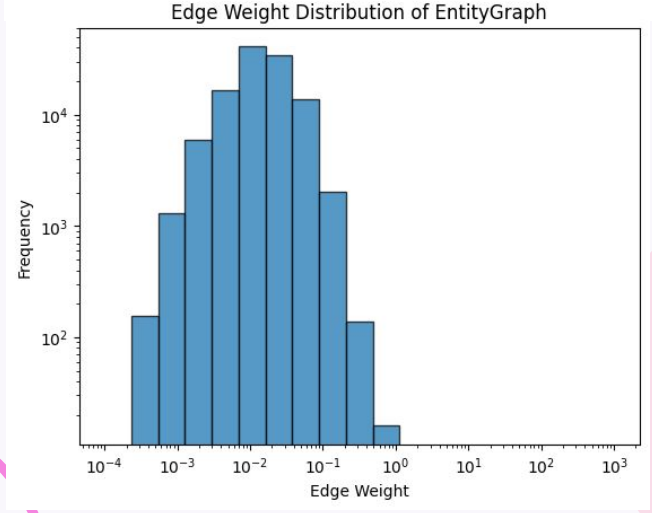
EntityGraph	Values
Number of nodes	1993
Number of edges	131448
Average degree	131.9097

Entity Graph after filter for node with degree more than 5

Entity-Entity Graph - filter the graph

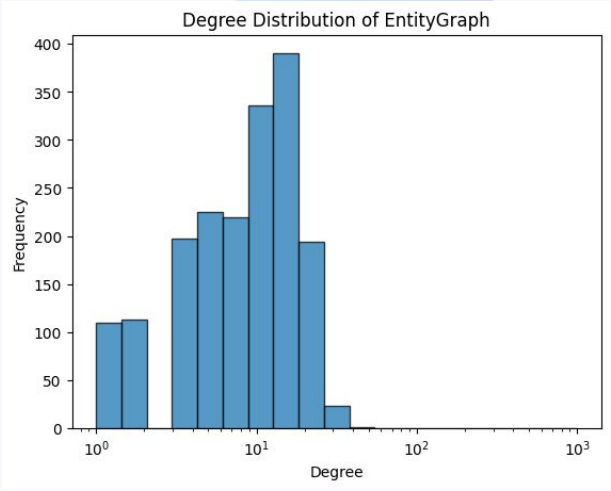
- we can observe one peak in the degree distribution at fairly low values, with a fat tail toward large degree values. Also, the edge weight shows a similar behavior, with a peak at rather low values and fat right tails.
- These distributions suggest the presence of several communities which are connected to each other via some central node
- The edges weight distribution shown in preceding graph suggest that a threshold could be **0.05**

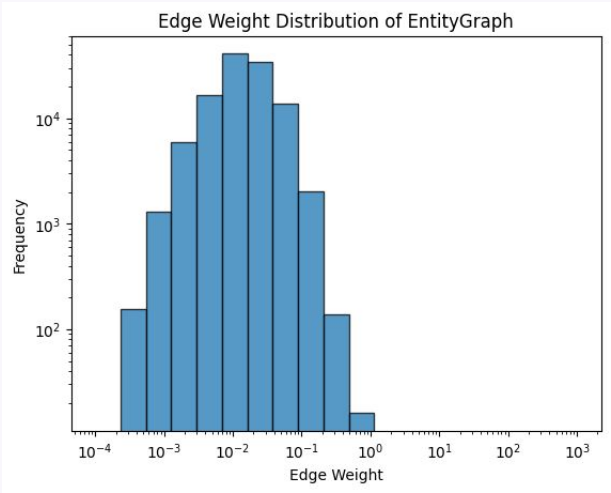




Entity-Entity Graph - filter the graph

FilteredEntityGraph	Values
Number of nodes	1808
Number of edges	9097
Average degree	10.0631





This is less obvious, and it shows the peak for the nodes that have a degree around 10, as opposed to the peak shown in previous slide which show peak at degree around 100. Where the graph consist of **`node 1791 nodes and 9085 edges`**

Entity-Entity Graph - Analyst the graph

Connected Component	Number of Nodes
1	1791
2	3
3	4
4	2
5	2
6	4
7	2

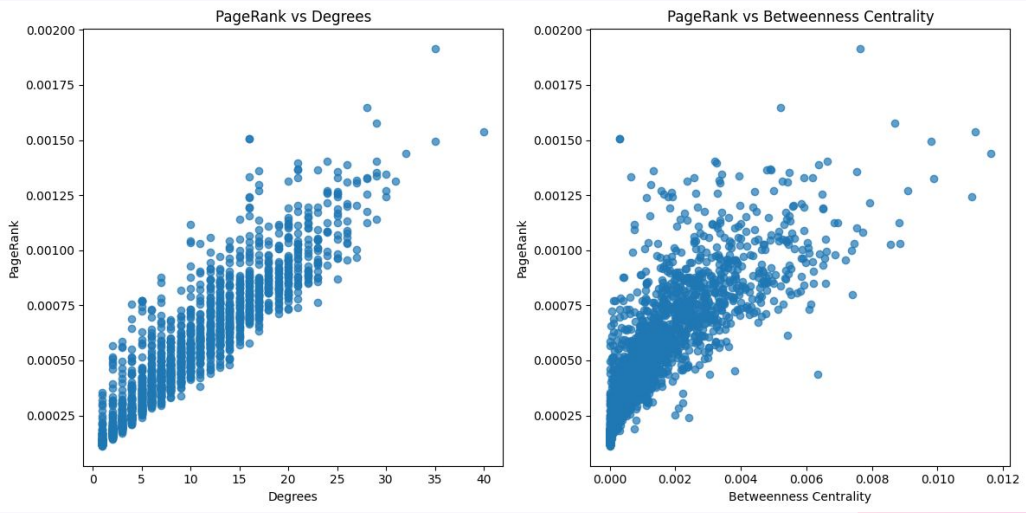
Connected component within the graph after filter

Network Metrics (Component 1)	Values
Shortest Path	3.874338
Clustering Coefficient	0.215627
Global Efficiency	0.278468

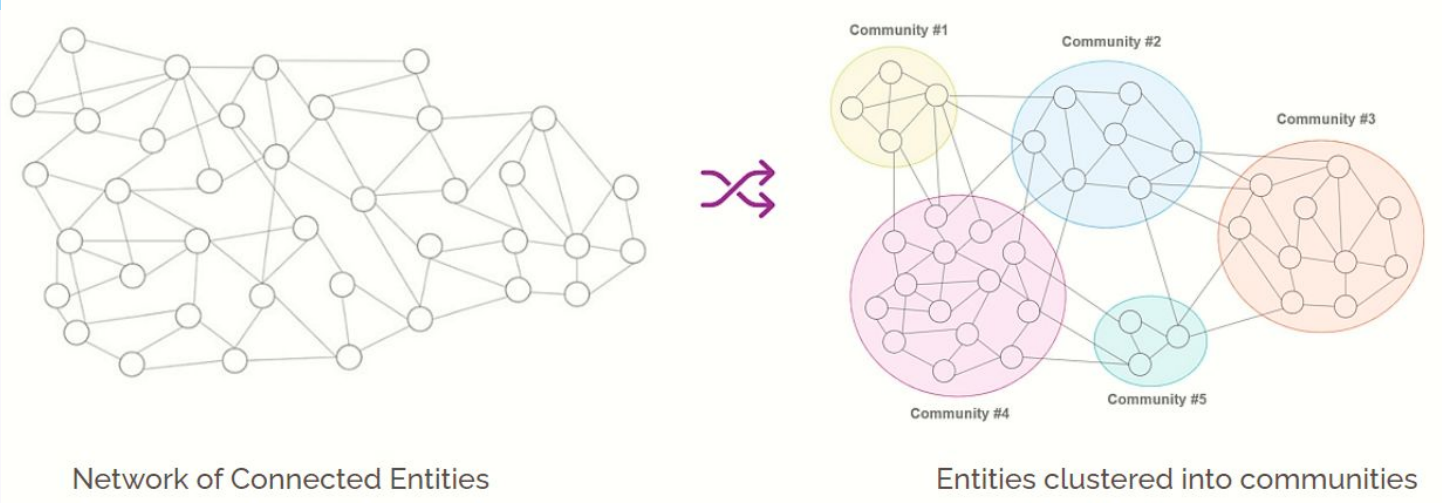
- On average it take about 3.85 or 4 step to go from one node to another.
- By clustering coefficient, it is moderate level that how node are tend to cluster together.
- By global efficiency show how effectively each node connect to each other, and it also moderate level score.

Entity-Entity Graph - Analyst the graph

- There is positive relationship between pagerank and degree of graph which show that the node with high degree tend to have high page rank score or it is well connected to many other node.



Community Detection Algorithm (Louvain)

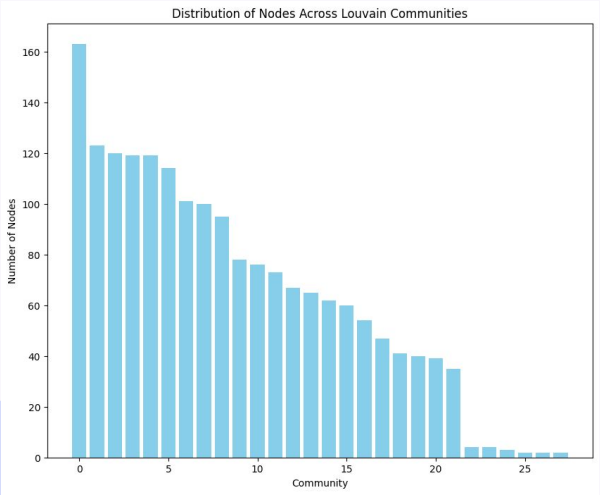


- **The Community Detection algorithm (Louvain)** is used to detect communities (clusters) in networks (interrelated items) by evaluating how much more densely connected the nodes within a community are, compared to how connected they would be in a random network (i.e., nodes more like each other than to the other nodes)
- **Community** here refers to the occurrence of groups of nodes in a network that are more densely connected internally than with the rest of the network.

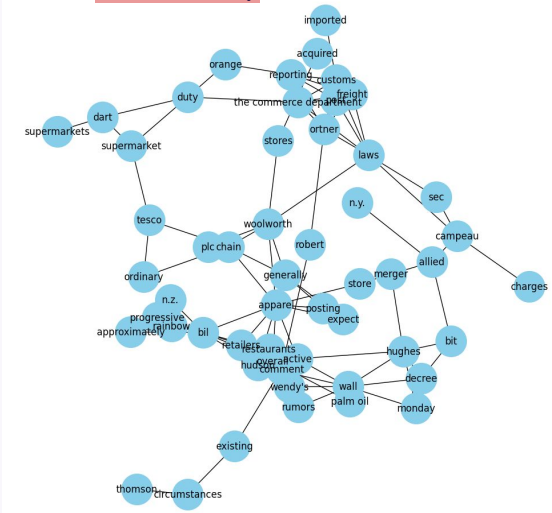
Entity-Entity Graph Analyst the graph

Table of Louvain Communities and Nodes within community

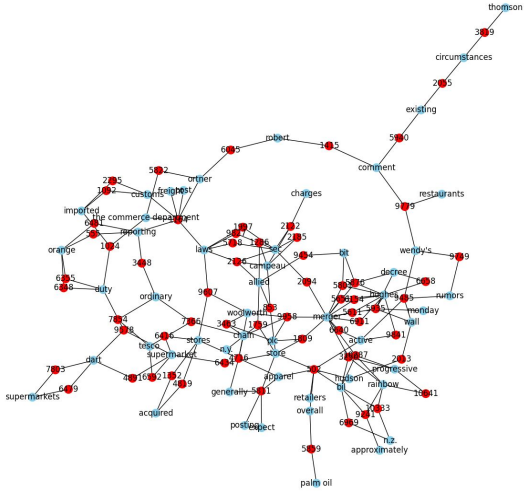
Community	Entities
Community 1	{'previously', 'regular'}
Community 2	{'gelco', 'american express', ..., 'growing'}
Community 3	{'problem', 'affect', 'offered', ...}



Entity-Entity in a community



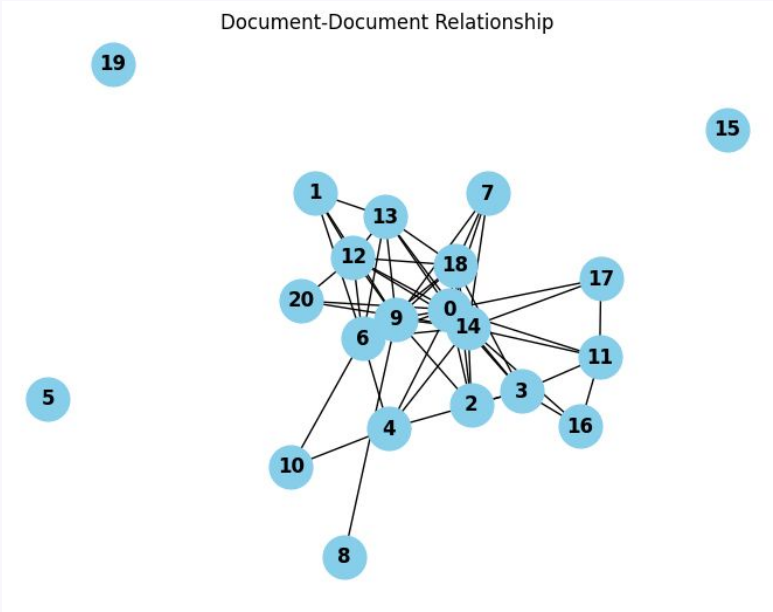
Entity-Document in a community



- We generally observe about 28 communities, with the larger ones containing around 130-160 nodes.
- where some communities also consist of a few nodes within too.

Document-Document Graph

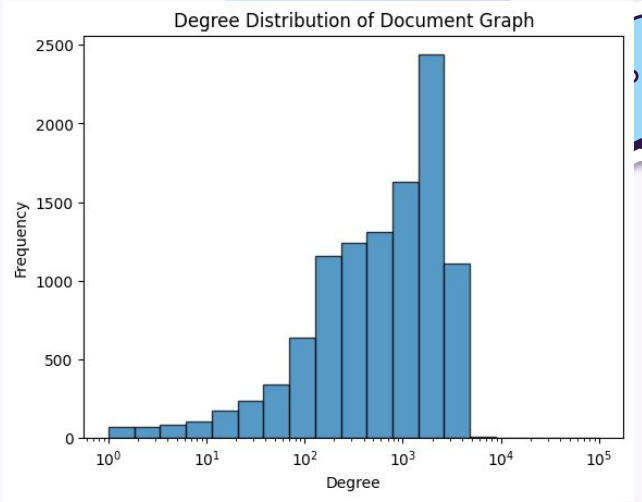
-Same as Entity-Entity relationship, we will project bipartite graph on to a set of document to **create a document-document network.**
-We will use **Jaccard Similarity** to obtain weighted between each node.
-The document-document graph consist of **10694 nodes and 5981285 edges.**

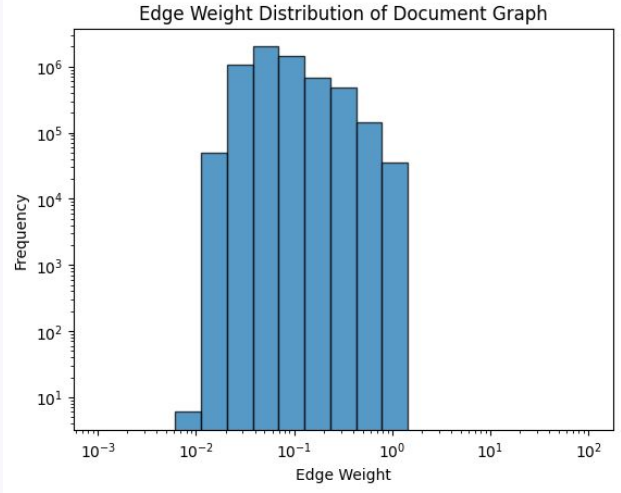


Document-Document Graph - Filter the Graph

-**Distribution of degree** which show that there are many document node that has degree up to 1000 which mean they are highly connected mostly
-**Distribution of edge weight** show that there are even around almost 1000 node has attain weight close to 1 which mean there quiet among number of document that has strong similarity within each other.
-This mean that the presence number of **supernodes**(node with large degrees) are highly connected.

Since we want to filter out edges weight of the graph, so we going to focus on high weight edges, because high weight of edges meaning they have strong similarity within each other. so we going to focus on meaningful relationship here and keeping threshold at **0.6** might be a good one.

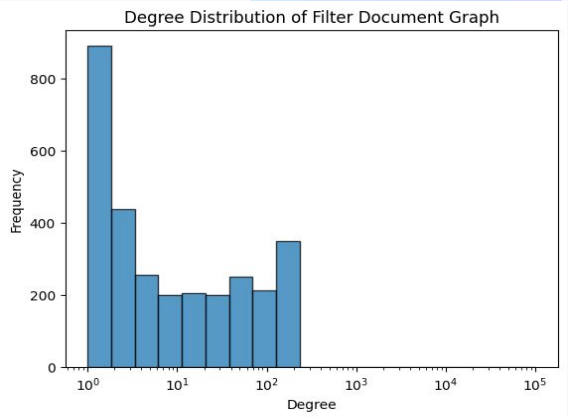


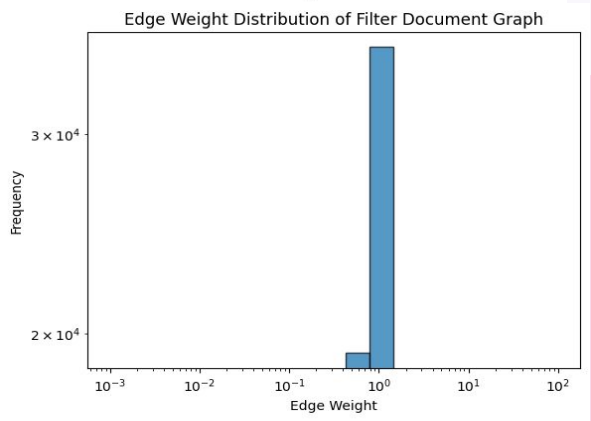


Document-Document Graph - Filter the Graph

The graph after filter for only node with weight bigger than 0.6 we got a graph with **2998 nodes with 55024 edges**. Which is still quite a lot.

These distribution show that there are many node that has few degree and has high relationship within other



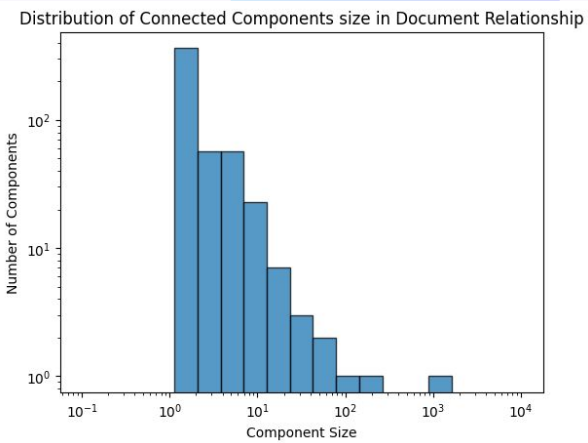


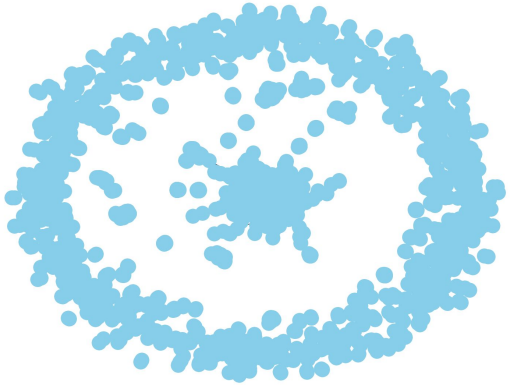
Document-Document Graph - Core Components

There consist of 515 connect component within filter document graph where the number of node within each component is vary from one to another.

we can see the present a large cluster (the core) together with many small sized community (represent the periphery that consist of document around 100)

Component	Document Node
1	{8194, 6149, 4102, 4103, 4109, 8205, 16, 4113,...}
2	{61, 5}
3	{8200, 5846, 5686, 7610, 9082, 5501, 7550, 9503}
4	{2561, 2562, 7169, 3588, 8203, 5136, 8729, 361...}

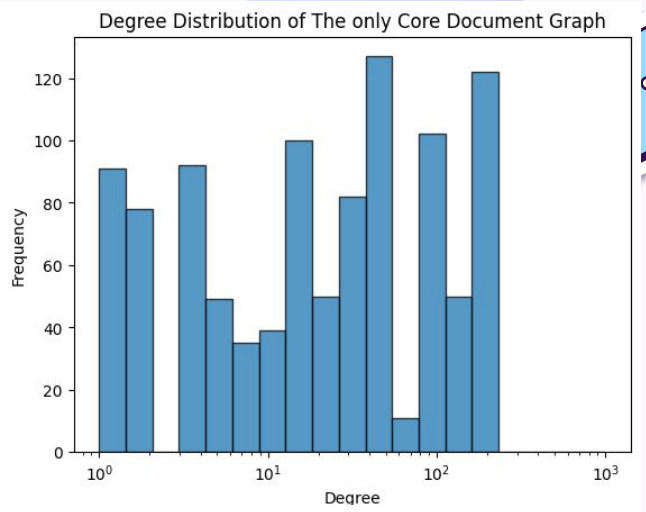


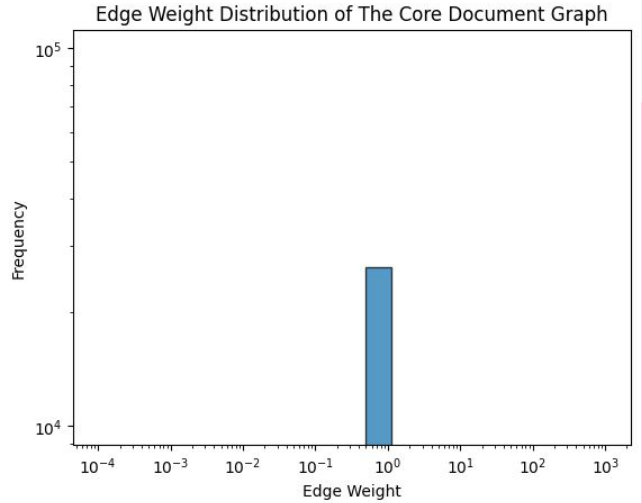


Document-Document Graph - Core Components

-The core document within the filter document graph consist of node with high edge weight where the graph consist of 1028 nodes and 26343 edges.
-The distribution show that the core document component consist of node document with strong similarity and on average, each node in the core document component is connected to approximately 40 other nodes through edges in the graph

Core Document	Value
Number of nodes	1028
Number of edges	26343
Average degree	40.18764302059497





Document-Document Graph - Community Detection And Topic Clustering

Calculate for percentage of every label within a community

Calculate Shanon Entropy of each community to see whether the community that belong to the same class tend to cluster together or not

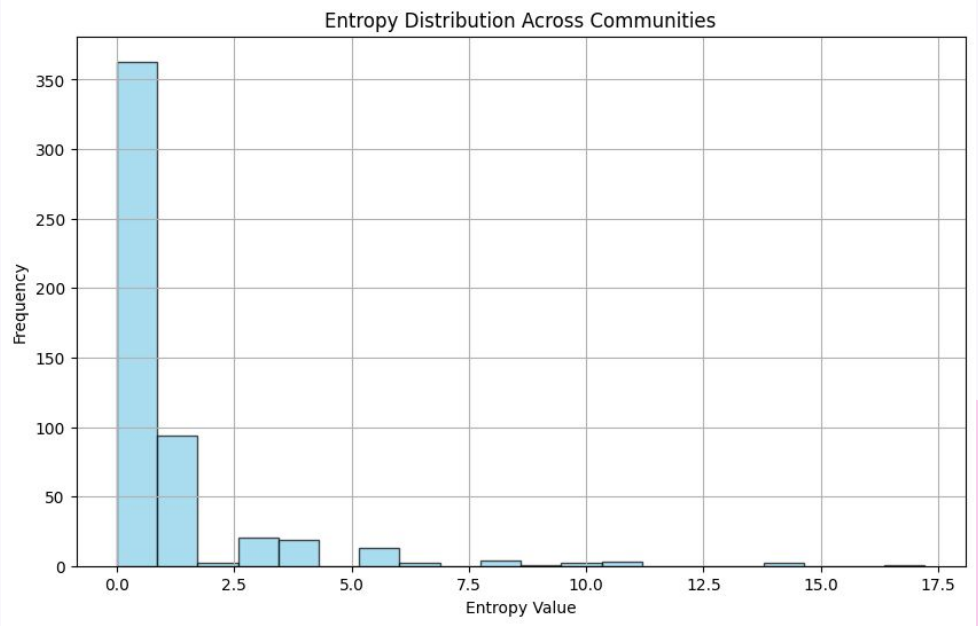
$$I_c = -\sum \log t_{ci}$$

- I_c represents the entropy of the cluster.
- t_{ci} is the percentage of topic i in community c.

We also apply Louvain community detection algorithm, then we will extract topic mixture of each community to see whether all document belong to the same class.

Document-Document Graph - Community Detection And Topic Clustering

We expect the document belong to the same topic to be close and connected, by the distribution of entropy show that **most community have zero or very low entropy mean that document that belong to the same class tend to cluster together**



04

Shallow Learning Multilabel Classification

Next we will focus on using information and connection between entities taken from bipartite entity-document graph to train multi-label classifiers to predict the document topics

Preparing for Topic Classifier - filter dataset

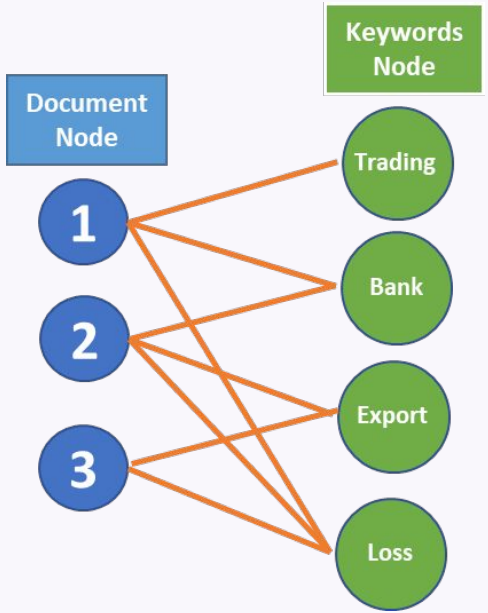
When training topic classifier, we must restrict our focus to only those document that belong to such labels. So, First we will consider the **top 10 common topic across the document**

So we going to take only the dataset that involve within these most common topic to consider and predict for

Top 10 common topic accoross the corpus of document

Topic	Frequency
earn	3964
acq	2369
money-fx	717
grain	582
crude	578
trade	485
interest	478
ship	286
wheat	283
corn	237

Preparing for Topic Classifier - graph embedded



model
node2vec

Table of vector of 10 dimension that represent value of each node within bipartite graph

	0	1	2	3	4	5	6	7	8	9
said	-0.068634	-0.108728	0.279320	0.502636	-0.352387	-0.141525	0.122002	-0.174865	0.270644	-0.536599
mln	0.806816	-0.031284	0.831844	0.338622	-0.450020	-0.301777	0.314351	-0.175108	0.585830	-0.098098
net	0.927049	0.026581	1.356840	0.432589	-0.294489	-0.322550	0.490149	-0.077926	0.527095	0.199232
u.s.	-0.144313	0.180694	0.365246	0.834751	-0.384607	-0.121756	-0.214999	0.088578	0.340159	-0.201328
dlrs	0.167431	-0.229972	0.686198	0.227649	-0.532566	-0.382044	0.405773	0.072278	0.351097	-0.298843
...	...	...	...	...	...	...	...	...	...	...
liedtke	0.161181	-0.653445	0.384399	0.404352	-0.495372	-0.831178	1.201550	-0.669810	-0.097228	-0.398567
minerals properties	-0.126188	-0.666247	1.043137	-0.001206	-1.271500	0.104546	-0.107410	-1.483430	1.339801	0.451293
sand technology	0.562546	-0.375001	0.410943	0.634887	-0.779979	-0.991260	0.834942	-0.043466	0.543723	0.262858
schlecht	-0.109026	0.547412	1.187024	0.542943	0.514080	-0.357270	-0.807087	-0.077646	0.053627	-0.773175
int	0.100956	0.455052	0.339058	0.432554	-0.911695	-0.426637	-0.133107	-1.046802	-0.098659	-0.313944

19836 rows × 10 columns

Preparing for Topic Classifier- train/test split

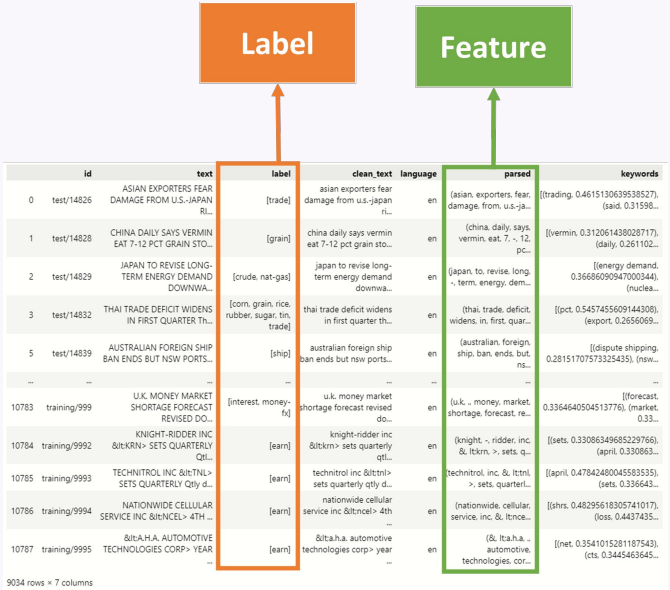
Test Set

Train Set

	id	text	label	clean_text	language	parsed	keywords
0	test/14826	ASIAN EXPORTERS FEAR DAMAGE FROM U.S.-JAPAN RI...	[trade]	asian exporters fear damage from u.s.-japan ri...	en	(asian, exporters, fear, damage, from, u.s.-ja...	[(trading, 0.4615130639538527), (said, 0.31598...
1	test/14828	CHINA DAILY SAYS VERMIN EAT 7-12 PCT GRAIN STO...	[grain]	china daily says vermin eat 7-12 pct grain sto...	en	(china, daily, says, vermin, eat, 7, -, 12, pc...	[(vermin, 0.312061438028717), (daily, 0.261102...
2	test/14829	JAPAN TO REVISE LONG-TERM ENERGY DEMAND DOWNWA...	[crude, nat-gas]	japan to revise long-term energy demand downwa...	en	(japan, to, revise, long, -, term, energy, dem...	[(energy demand, 0.36686090947000344), (nuclea...
3	test/14832	THAI TRADE DEFICIT WIDENS IN FIRST QUARTER Th...	[corn, grain, rice, rubber, sugar, tin, trade]	thai trade deficit widens in first quarter th...	en	(thai, trade, deficit, widens, in, first, quar...	[(pct, 0.5457455609144308), (export, 0.2656069...
5	test/14839	AUSTRALIAN FOREIGN SHIP BAN ENDS BUT NSW PORTS...	[ship]	australian foreign ship ban ends but nsw ports...	en	(australian, foreign, ship, ban, ends, but, ns...	[(dispute shipping, 0.28151707573325435), (nsw...
...	...	...	...	...	...	...	...
10783	training/999	U.K. MONEY MARKET SHORTAGE FORECAST REVISED DO...	[interest, money-fx]	u.k. money market shortage forecast revised do...	en	(u.k., ., money, market, shortage, forecast, re...	[(forecast, 0.3364640504513776), (market, 0.33...
10784	training/9992	KNIGHT-RIDDER INC <KRN> SETS QUARTERLY Qtl...	[earn]	knight-ridder inc <krn> sets quarterly qtl...	en	(knight, -, ridder, inc, &, lt;krn, >, sets, q...	[(sets, 0.33086349685229766), (april, 0.330863...
10785	training/9993	TECHNITROL INC <TNL> SETS QUARTERLY Qtly d...	[earn]	technitrol inc <tnl> sets quarterly qtly d...	en	(technitrol, inc, &, lt;tnl, >, sets, quarterl...	[(april, 0.4784280045583535), (sets, 0.336643...
10786	training/9994	NATIONWIDE CELLULAR SERVICE INC <NCEL> 4TH ...	[earn]	nationwide cellular service inc <ncel> 4th ...	en	(nationwide, cellular, service, inc, &, lt;nce...	[(shrs, 0.48295618305741017), (loss, 0.4437435...
10787	training/9995	<A.H.A. AUTOMOTIVE TECHNOLOGIES CORP> YEAR ...	[earn]	<a.h.a. automotive technologies corp> year ...	en	(&, lt;a.h.a, ., automotive, technologies, cor...	[(net, 0.3541015281187543), (cts, 0.3445463645...

9034 rows × 7 columns

Preparing for Topic Classifier - feature/label extraction



	id	text	label	clean_text	language	parsed	keywords
0	test/14826	ASIAN EXPORTERS FEAR DAMAGE FROM U.S.-JAPAN RI...	[trade]	asian exporters fear damage from u.s.-japan ri...	en	(asian, exporters, fear, damage, from, u.s.-ja...	[(trading, 0.4615130639538527), (said, 0.31598...
1	test/14828	CHINA DAILY SAYS VERMIN EAT 7-12 PCT GRAIN STO...	[grain]	china daily says vermin eat 7-12 pct grain sto...	en	(china, daily, says, vermin, eat, 7, -, 12, pc...	[(vermin, 0.312061438028717), (daily, 0.261102...
2	test/14829	JAPAN TO REVISE LONG-TERM ENERGY DEMAND DOWNWA...	[crude, nat-gas]	japan to revise long-term energy demand downwa...	en	(japan, to, revise, long, -, term, energy, dem...	[(energy demand, 0.3668609094700344), (nuclea...
3	test/14832	THAI TRADE DEFICIT WIDENS IN FIRST QUARTER TH...	[corn, grain, rice, rubber, sugar, tin, trade]	thai trade deficit widens in first quarter th...	en	(thai, trade, deficit, widens, in, first, quar...	[(pct, 0.5457456509144308), (export, 0.2656069...
5	test/14839	AUSTRALIAN FOREIGN SHIP BAN ENDS BUT NSW PORTS...	[ship]	australian foreign ship ban ends but nsw ports...	en	(australian, foreign, ship, ban, ends, but, ns...	[(dispute shipping, 0.28151707573325435), (nsw...
...	...	...	...	...	...	...	...
10783	training/999	U.K. MONEY MARKET SHORTAGE FORECAST REVISED DO...	[interest, money-fx]	u.k. money market shortage forecast revised do...	en	(u.k., money, market, shortage, forecast, re...	[(forecast, 0.3364640504513776), (market, 0.33...
10784	training/9992	KNIGHT-RIDDER INC <KRN> SETS QUARTERLY QTL...	[earn]	knight-ridder inc <krn> sets quarterly qtl...	en	(knight, -, ridder, inc, &, ltkrn, >, sets, q...	[(sets, 0.33086349685229766), (april, 0.330863...
10785	training/9993	TECHNITROL INC <TNL> SETS QUARTERLY QTLY d...	[earn]	technitrol inc <tnl> sets quarterly qtly d...	en	(technitrol, inc, &, lttnl, >, sets, quarterl...	[(april, 0.47842480045583535), (sets, 0.336643...
10786	training/9994	NATIONWIDE CELLULAR SERVICE INC <NCEL> 4TH ...	[earn]	nationwide cellular service inc <ncel> 4th ...	en	(nationwide, cellular, service, inc, &, ltnce...	[(shrs, 0.48295618305741017), (loss, 0.4437435...
10787	training/9995	<A.H.A. AUTOMOTIVE TECHNOLOGIES CORP> YEAR ...	[earn]	<a.h.a. automotive technologies corp> year ...	en	(&, lt;a.h.a., automotive, technologies, cor...	[(net, 0.35410152811875 43), (cts, 0.3445463645...

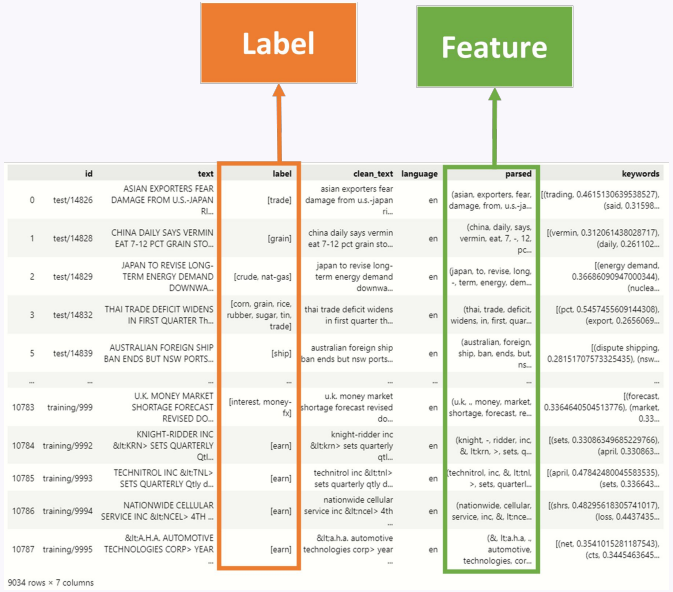
9034 rows × 7 columns

Embeddings
Transformer

Label

	earn	acq	money-fx	grain	crude	trade	interest	ship	wheat	corn
0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0
1	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	1.0	0.0	1.0	0.0	0.0	0.0	1.0
5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0
...	...	...	...	...	...	...	...	...	...	...
3012	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0
3013	0.0	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0
3014	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3015	0.0	0.0	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3016	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0

Building Model Topic Classifier - pipeline



	id	text	label	clean_text	language	parsed	keywords
0	test/14826	ASIAN EXPORTERS FEAR DAMAGE FROM U.S.-JAPAN RI...	[trade]	asian exporters fear damage from u.s.-japan ri...	en	(asian, exporters, fear, damage, from, u.s.-ja...	[(trading, 0.4615130639538527), (said, 0.31598...
1	test/14828	CHINA DAILY SAYS VERMIN EAT 7-12 PCT GRAIN STO...	[grain]	china daily says vermin eat 7-12 pct grain sto...	en	(china, daily, says, vermin, eat, 7, -, 12, pc...	[(vermin, 0.312061438028717), (daily, 0.261102...
2	test/14829	JAPAN TO REVISE LONG-TERM ENERGY DEMAND DOWNWA...	[crude, nat-gas]	japan to revise long-term energy demand downwa...	en	(japan, to, revise, long, -, term, energy, dem...	[(energy demand, 0.3668609094700344), (nuclea...
3	test/14832	THAI TRADE DEFICIT WIDENS IN FIRST QUARTER TH...	[corn, grain, rice, rubber, sugar, tin, trade]	thai trade deficit widens in first quarter th...	en	(thai, trade, deficit, widens, in, first, quar...	[(pct, 0.5457455609144308), (export, 0.2656069...
5	test/14839	AUSTRALIAN FOREIGN SHIP BAN ENDS BUT NSW PORTS...	[ship]	australian foreign ship ban ends but nsw ports...	en	(australian, foreign, ship, ban, ends, but, ns...	[(dispute shipping, 0.28151707573325435), (nsw...
...	...	...	...	...	...	...	...
10783	training/999	U.K. MONEY MARKET SHORTAGE FORECAST REVISED DO...	[interest, money-fx]	u.k. money market shortage forecast revised do...	en	(u.k., money, market, shortage, forecast, re...	[(forecast, 0.3364640504513776), (market, 0.33...
10784	training/9992	KNIGHT-RIDDER INC <KRN> SETS QUARTERLY Qtl...	[earn]	knight-ridder inc <krn> sets quarterly qtl...	en	(knight, -, ridder, inc, &, lt;krn, >, sets, q...	[(sets, 0.33086349685229766), (april, 0.330863...
10785	training/9993	TECHNITROL INC <TNL> SETS QUARTERLY Qtly d...	[earn]	technitrol inc <tnl> sets quarterly qtly d...	en	technitrol, inc, &, lt;tnl, >, sets, quarterl...	[(april, 0.47842480045583535), (sets, 0.336643...
10786	training/9994	NATIONWIDE CELLULAR SERVICE INC <NCEL> 4TH ...	[earn]	nationwide cellular service inc <ncel> 4th ...	en	(nationwide, cellular, service, inc, &, ltnce...	[(shrs, 0.4829561830574101 7), (loss, 0.4437435...
10787	training/9995	<A.H.A. AUTOMOTIVE TECHNOLOGIES CORP> YEAR ...	[earn]	<a.h.a. automotive technologies corp> year ...	en	(&, lt;a.h.a., -, automotive, technologies, cor...	[(net, 0.35410152811875 43), (cts, 0.3445463645...

9034 rows × 7 columns

Train set only

Pipeline

Embeddings Transformer (graphEmbeddings.p)

Multioutput Classifier with Random Forest Classifier (max_feature=0.2, n_estimater=50)

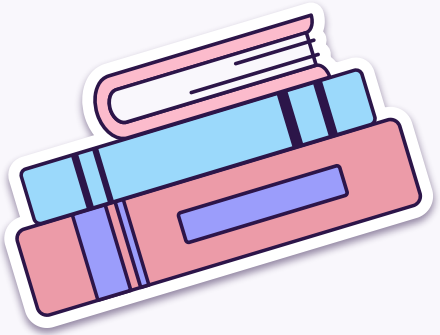
Grid Search Algorithm to find the best combination of hyperparameter

Topic Classifier - Evaluation

Label	Precision	Recall	F1-Score	Support
0	0.97	0.97	0.97	1087
1	0.96	0.90	0.93	719
2	0.76	0.78	0.77	179
3	0.96	0.83	0.89	149
4	0.93	0.84	0.88	189
5	0.85	0.68	0.76	117
6	0.90	0.49	0.63	131
7	0.87	0.52	0.65	89
8	0.69	0.49	0.57	71
9	0.55	0.32	0.40	56
Micro Avg	0.93	0.85	0.89	2787
Macro Avg	0.84	0.68	0.75	2787
Weighted Avg	0.92	0.85	0.88	2787
Samples Avg	0.88	0.87	0.87	2787

Label	Wrong Predictions
earn	66
acq	100
money-fx	85
grain	30
crude	42
trade	51
interest	74
ship	50
wheat	52
corn	53

We obtain accuracy score around 0.783 with …



05

Graph Neural Network

step-by-step guide that will help you train and evaluate a model, based on graph neural networks, for predicting document topic classification

Preparing for creating Graph -Stellar Graph

	source	target	weight	type
0	0	trading	0.471364	keywords
1	0	said	0.305622	keywords
2	0	export market	0.195905	keywords
3	0	import	0.171753	keywords
4	0	japanese electronics	0.123282	keywords
...	...	...	...	...
69767	10785	april	0.478425	keywords
69768	10785	sets	0.336644	keywords
69769	10786	shrs	0.482956	keywords
69770	10786	loss	0.443744	keywords
69771	10787	net shr	0.349324	keywords

Edges are type of each represented keywords

TfidfVectorizer

Nodes of different Type

	0	1	2	3	4	5	6	7	8	9	...	9063	9064	90
3020	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	
3022	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	
3023	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	
3024	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	
3025	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	
3012	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	
3013	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	
3014	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	
3015	0.0	0.0	0.0	0.056513	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	
3016	0.0	0.0	0.0	0.000000	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	

Document Feature

type	keywords	GPE	ORG	PERSON
target				
abandoned	1.000000	0.00	0.0	0.000000
abated	1.000000	0.00	0.0	0.000000
abatement concern	1.000000	0.00	0.0	0.000000
abbett	0.333333	0.00	0.0	0.666667
abbey	1.000000	0.00	0.0	0.000000
...	...	...	...	...
zones	1.000000	0.00	0.0	0.000000
zuckerman	1.000000	0.00	0.0	0.000000
zulia	1.000000	0.00	0.0	0.000000
zurich	0.250000	0.75	0.0	0.000000
zverev	1.000000	0.00	0.0	0.000000

Entity Feature

Preparing for creating Graph -Stellar Graph

After preparing each element we then obtain the information of our stellar graph:

- it have multiple edges with different pair of nodes type, where graph consist of **23585 nodes and 56050 edges**.
- **Two type of node:** entity(14824) and document(9073)
- There are various **edge types** connecting these nodes, such as "entity-GPE->document," "entity-ORG->document," "entity-PERSON->document," and "entity-keywords->document.

```
StellarGraph: Undirected multigraph
 Nodes: 23858, Edges: 56050

 Node types:
  entity: [14824]
    Features: float32 vector, length 4
    Edge types: entity-GPE->document, entity-ORG->document, entity-PERSON->document, entity-keywords->document
  document: [9034]
    Features: float32 vector, length 9073
    Edge types: document-GPE->entity, document-ORG->entity, document-PERSON->entity, document-keywords->entity

 Edge types:
    document-keywords->entity: [50839]
        Weights: range=[0.0503159, 0.854496], mean=0.238046, std=0.109954
        Features: none
    document-GPE->entity: [2309]
        Weights: range=[2, 15], mean=2.93893, std=1.60809
        Features: none
    document-ORG->entity: [2025]
        Weights: range=[2, 18], mean=3.04198, std=1.96378
        Features: none
    document-PERSON->entity: [877]
        Weights: range=[2, 16], mean=2.69897, std=1.34408
        Features: none
```

Preparing for creating Graph -Stellar Graph- (Sub Graph)

To truly test the performance of an inductive approach and avoid information from being linked between the train and test sets, we need to create a subgraph that only contains the data available at training time of the target label.

the following subgraph contain **16528 and edges 41259** compare to the whole graph before

```
StellarGraph: Undirected multigraph
 Nodes: 16528, Edges: 41259

 Node types:
  entity: [10039]
    Features: float32 vector, length 4
    Edge types: entity-GPE->document, entity-ORG->document, entity-PERSON->document, entity-keywords->document
  document: [6489]
    Features: float32 vector, length 9073
    Edge types: document-GPE->entity, document-ORG->entity, document-PERSON->entity, document-keywords->entity

 Edge types:
    document-keywords->entity: [37537]
        Weights: range=[0.0503159, 0.851282], mean=0.234636, std=0.10799
        Features: none
    document-GPE->entity: [1638]
        Weights: range=[2, 15], mean=2.95543, std=1.60895
        Features: none
    document-ORG->entity: [1447]
        Weights: range=[2, 18], mean=2.96752, std=1.76673
        Features: none
    document-PERSON->entity: [637]
        Weights: range=[2, 16], mean=2.69231, std=1.36999
        Features: none
```

Preparing for creating Graph -Train/Test Split

Same as before we will split the data according to the ID label on the filter data

Then we take the train data to split for train and validation for the model

```
from sklearn.model_selection import train_test_split
train, leftOut = train_test_split(
 sampled,
 train_size=0.1,
 test_size=None,
 random_state=42
)
validation, test = train_test_split(
 leftOut, train_size=0.2, test_size=None, random_state=100,
)
```

Building Graph Neural Network

First, we will create a generator able to produce the samples that will feed the neural network. Then we can create our GraphSAGE model. As we did for the generator, we need to use a model that can handle heterogenous graphs, so that why we use HinSAGE model.

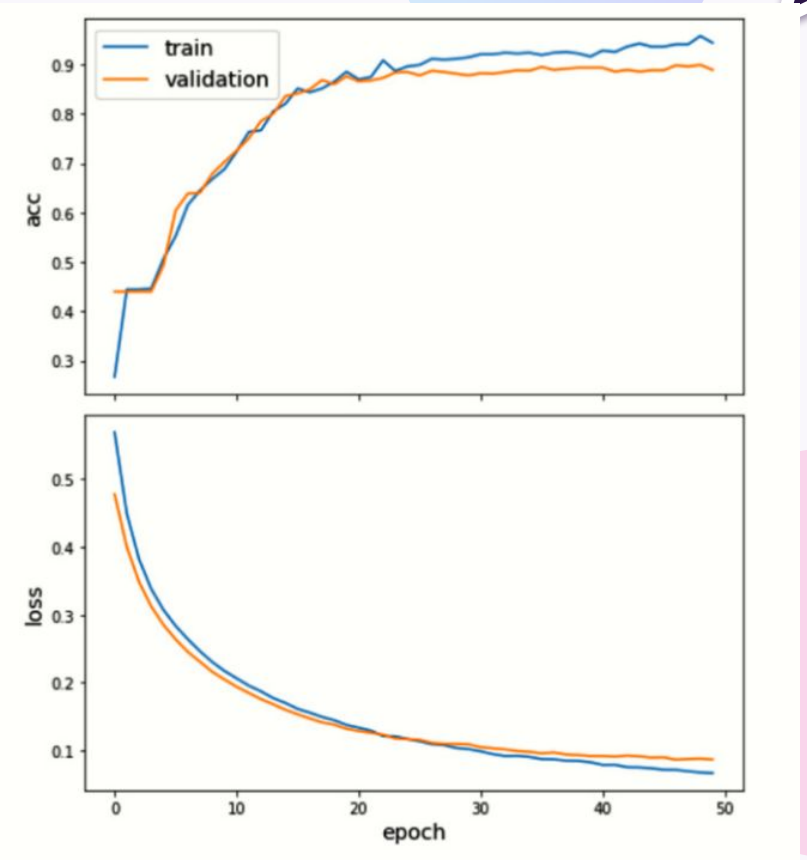
the final dense layer, we use a **sigmoid activation function** instead of a softmax activation function, since the problem at hand is a multi-class, multilabel task. Thus, a document may belong to more than one class, and the sigmoid activation function seems a more sensible choice in this context.

Building Model -Training the model

The preceding graph shows the plots of the evolution of the train and validation losses and accuracy versus the number of epochs.

As we can see, the train and validation accuracy increase consistently, up to around 30 epochs. Here, the accuracy of the validation set settle to a plateau, whereas the training accuracy continues to increase, indicating a tendency for **overfitting.**

Thus, stopping training at **around 50 seems a rather legitimate choice.**



Building Model -Finding best parameter

To identify the best threshold to be used to classify the documents, we will compute the prediction over all the test samples

Then, we will compute the F1-score with a macro average (where the F1-score for the single classes are averaged) for different threshold choices

As shown in the following graph, a threshold value of **0.2 seems to be the best choice as it achieves the best performance**

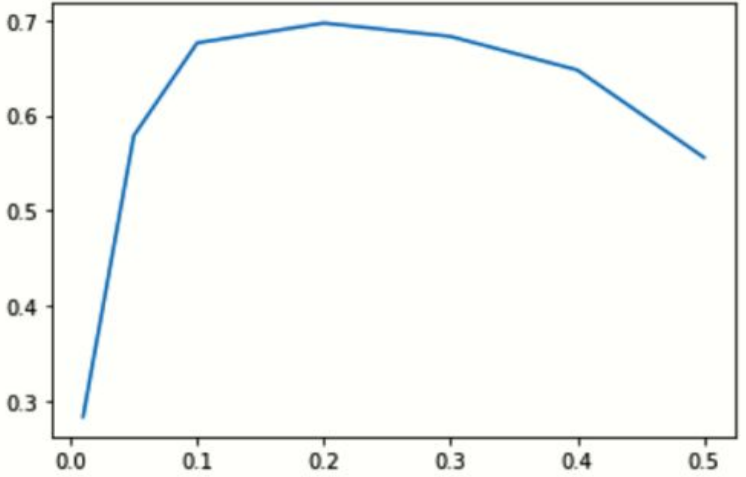


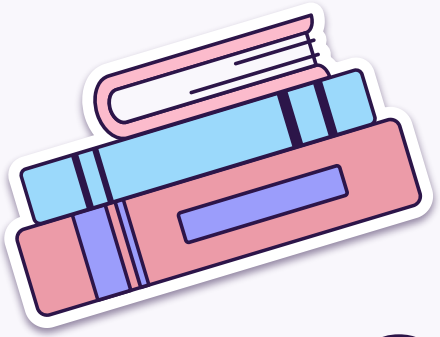
Figure 7.17 – Macro-averaged F1-score versus the threshold used for labeling

```
loss: 0.0933
accuracy: 0.8795
```

Model Evaluation
-Apply threshold value 0.2

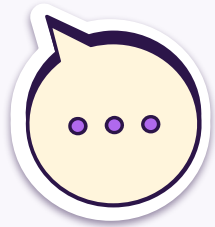
Compared to the shallow learning method, we can see that we have achieved a substantial improvement in performance that's between **5-10%**

	precision	recall	f1-score	support
0	0.93	0.99	0.96	1087
1	0.90	0.97	0.93	719
2	0.64	0.92	0.76	179
3	0.82	0.95	0.88	149
4	0.85	0.62	0.72	189
5	0.74	0.50	0.59	117
6	0.60	0.79	0.68	131
7	0.43	0.03	0.06	89
8	0.50	0.96	0.66	71
9	0.39	0.86	0.54	56
micro avg	0.82	0.89	0.85	2787
macro avg	0.68	0.76	0.68	2787
weighted avg	0.83	0.89	0.84	2787
samples avg	0.84	0.90	0.86	2787



06

Summarization



Thank for your Attention!