Decoding Kanji: A Fusion of Word Embeddings, Similarity Measures, and Graph Exploration

Michelle Yi, Graph Geeks, 2024

Agenda

01

Background & goals

How this project came about

02

The anatomy of kanji

Let’s talk data

03

Methodology

The approach for both creating embeddings and the graph

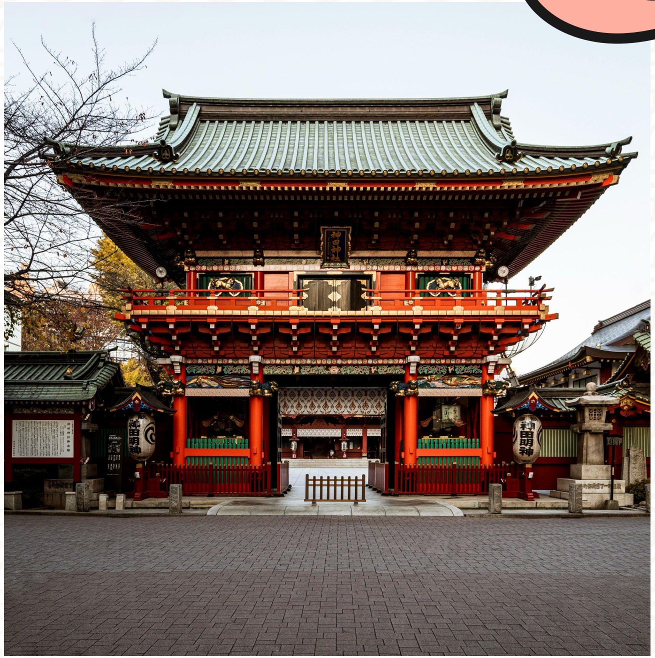
04

Conclusion

Results and next steps

Background

Personally an avid language learner with a background in computer science and deep learning, but I wanted to explore approaches that could integrate with network science in a fun way.



Goals

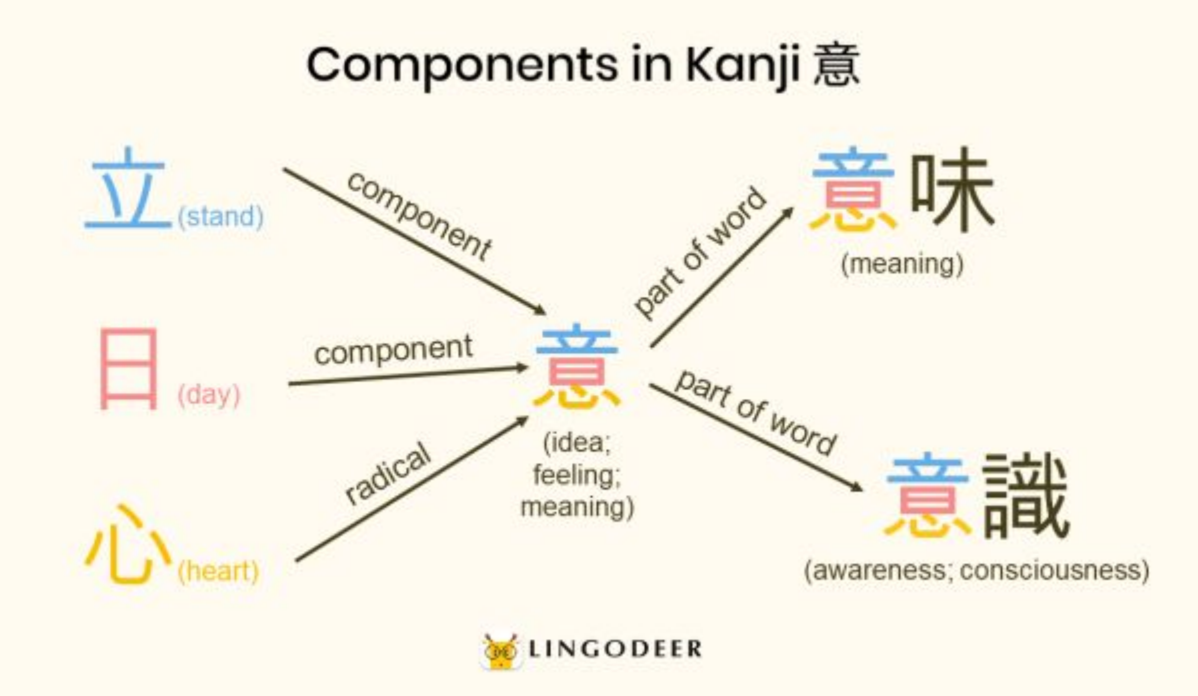
1. Determine whether or not there are new insights into the relationships between characters by taking a look at them by meaning rather than other attributes, such as classification, root, difficulty, etc.
2. Identify any nodes of influence
3. Explore a mechanism for combining deep learning with graph analytics.

02

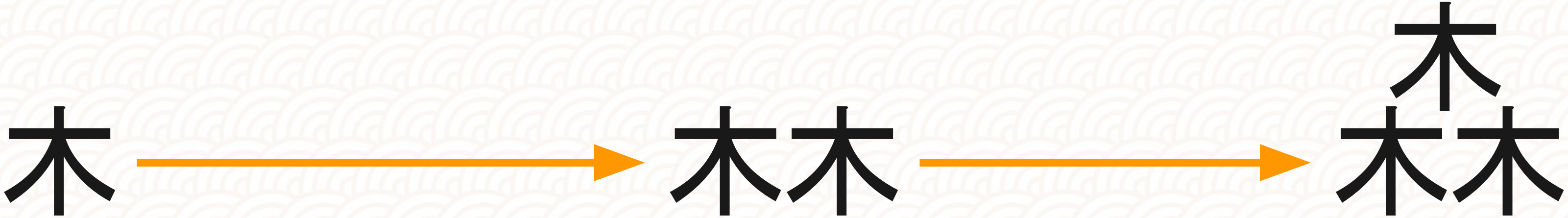
The Anatomy of Kanji

The data used in this project

Components (Simplified)

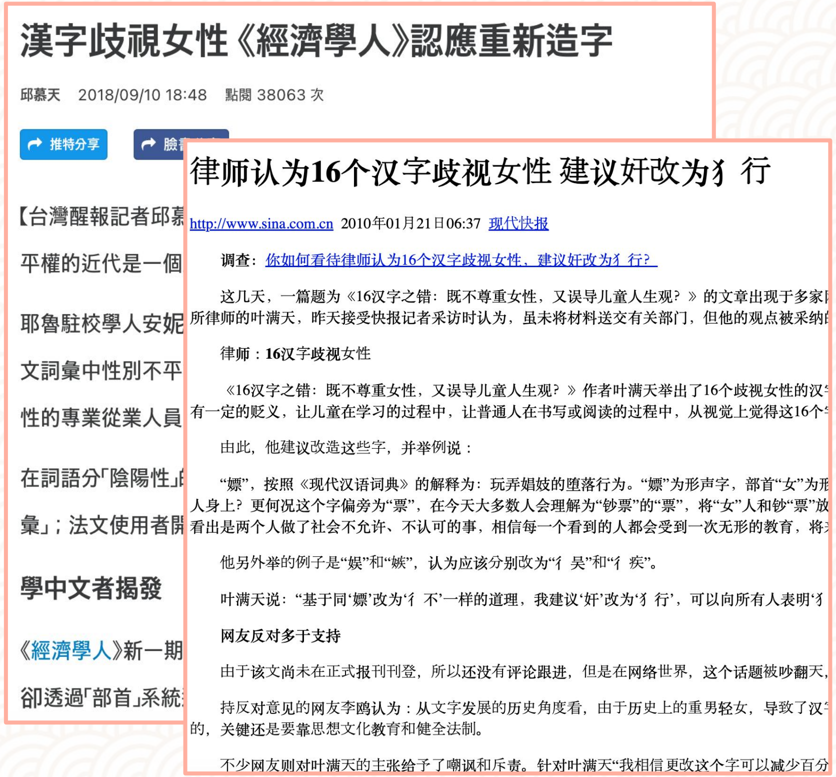


Source: https://blog.lingodeer.com/kanji-radicals/



Fun Example 1

女



漢字歧視女性《經濟學人》認應重新造字

邱慕天 2018/09/10 18:48 點閱 38063 次

【台灣醒報記者邱慕

平權的近代是一個

耶魯駐校學人安妮

文詞彙中性別不平

性的專業從業人員

在詞語分「陰陽性」的

彙」；法文使用者開

學中文者揭發

《經濟學人》新一期

卻透過「部首」系統

律师认为16个汉字歧视女性 建议奸改为犭行

http://www.sina.com.cn 2010年01月21日06:37 现代快报

调查： 你如何看待律师认为16个汉字歧视女性，建议奸改为犭行？

这几天，一篇题为《16汉字之错：既不尊重女性，又误导儿童人生观？》的文章出现于多家
所律师的叶满天，昨天接受快报记者采访时认为，虽未将材料送交有关部门，但他的观点被采纳

律师：**16汉字歧视女性**

《16汉字之错：既不尊重女性，又误导儿童人生观？》作者叶满天举出了16个歧视女性的汉
有一定的贬义，让儿童在学习的过程中，让普通人在书写或阅读的过程中，从视觉上觉得这16个

由此，他建议改造这些字，并举例说：

"嫖"，按照《现代汉语词典》的解释为：玩弄娼妓的堕落行为。"嫖"为形声字，部首"女"为
人身上？更何况这个字偏旁为"票"，在今天大多数人会理解为"钞票"的"票"，将"女"人和钞"票"放
看出是两个人做了社会不允许、不认可的事，相信每一个看到的人都会受到一次无形的教育，将

他另外举的例子是"娱"和"嫉"，认为应该分别改为"彳吴"和"彳疾"。

叶满天说："基于同'嫖'改为'彳不'一样的道理，我建议'奸'改为'犭行'，可以向所有人表明'犭

网友反对多于支持

由于该文尚未在正式报刊刊登，所以还没有评论跟进，但是在网络世界，这个话题被吵翻天，

持反对意见的网友李鸥认为：从文字发展的历史角度看，由于历史上的重男轻女，导致了汉
的，关键还是要靠思想文化教育和健全法制。

不少网友则对叶满天的主张给予了嘲讽和斥责。针对叶满天"我相信更改这个字可以减少百分

Don't put three together!

One of the Most Controversial

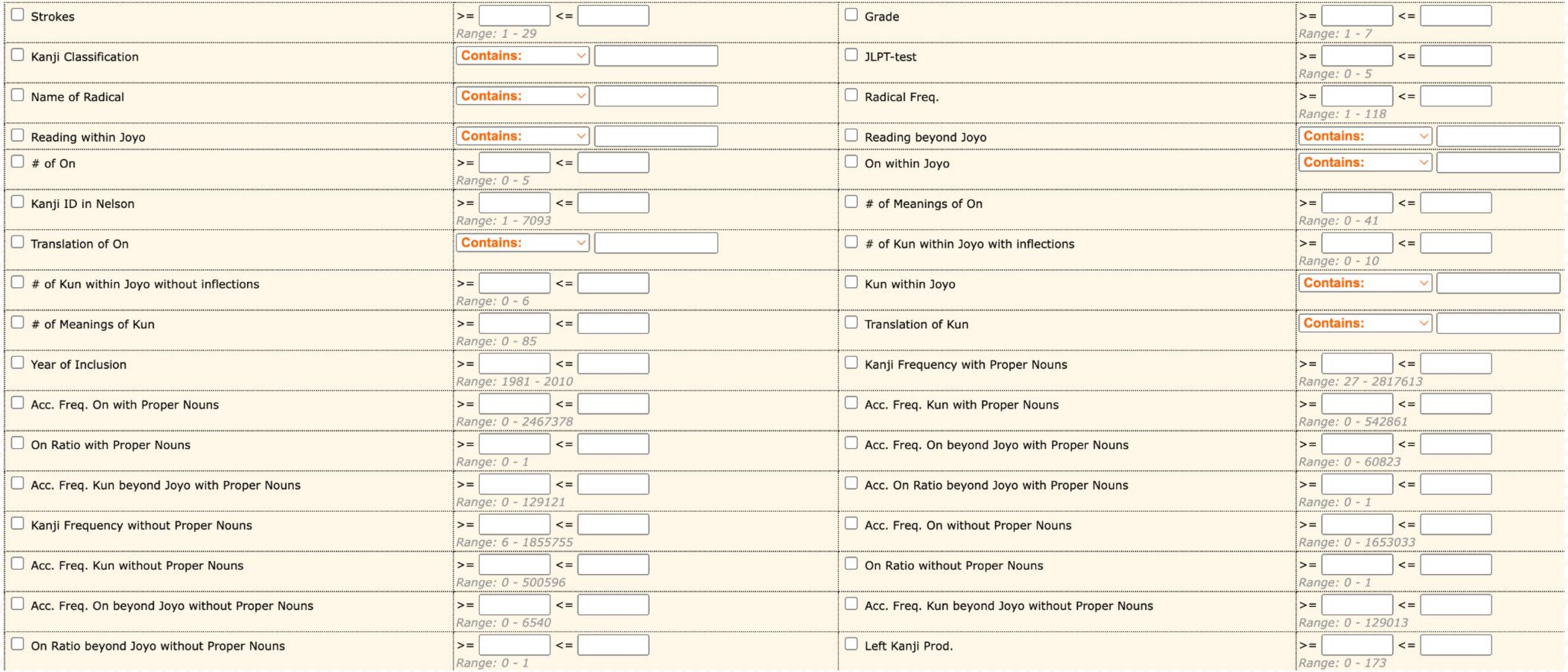
Core Concepts

01

Classification

02

Meaning



□ Strokes	>= <= Range: 1 - 29	□ Grade	>= <= Range: 1 - 7
□ Kanji Classification	Contains:	□ JLPT-test	>= <= Range: 0 - 5
□ Name of Radical	Contains:	□ Radical Freq.	>= <= Range: 1 - 118
□ Reading within Joyo	Contains:	□ Reading beyond Joyo	Contains:
□ # of On	>= <= Range: 0 - 5	□ On within Joyo	Contains:
□ Kanji ID in Nelson	>= <= Range: 1 - 7093	□ # of Meanings of On	>= <= Range: 0 - 41
□ Translation of On	Contains:	□ # of Kun within Joyo with inflections	>= <= Range: 0 - 10
□ # of Kun within Joyo without inflections	>= <= Range: 0 - 6	□ Kun within Joyo	Contains:
□ # of Meanings of Kun	>= <= Range: 0 - 85	□ Translation of Kun	Contains:
□ Year of Inclusion	>= <= Range: 1981 - 2010	□ Kanji Frequency with Proper Nouns	>= <= Range: 27 - 2817613
□ Acc. Freq. On with Proper Nouns	>= <= Range: 0 - 2467378	□ Acc. Freq. Kun with Proper Nouns	>= <= Range: 0 - 542861
□ On Ratio with Proper Nouns	>= <= Range: 0 - 1	□ Acc. Freq. On beyond Joyo with Proper Nouns	>= <= Range: 0 - 60823
□ Acc. Freq. Kun beyond Joyo with Proper Nouns	>= <= Range: 0 - 129121	□ Acc. On Ratio beyond Joyo with Proper Nouns	>= <= Range: 0 - 1
□ Kanji Frequency without Proper Nouns	>= <= Range: 6 - 1855755	□ Acc. Freq. On without Proper Nouns	>= <= Range: 0 - 1653033
□ Acc. Freq. Kun without Proper Nouns	>= <= Range: 0 - 500596	□ On Ratio without Proper Nouns	>= <= Range: 0 - 1
□ Acc. Freq. On beyond Joyo without Proper Nouns	>= <= Range: 0 - 6540	□ Acc. Freq. Kun beyond Joyo without Proper Nouns	>= <= Range: 0 - 129013
□ On Ratio beyond Joyo without Proper Nouns	>= <= Range: 0 - 1	□ Left Kanji Prod.	>= <= Range: 0 - 173

The data: https://www.kanjidatabase.com/

Data Scope

- 2,136 characters
- Single Kanji (not combined e.g. Jukugo)
- All columns from the database
- Key on meaning (each character can have many meanings)

03

Methodology

Embeddings and Graph

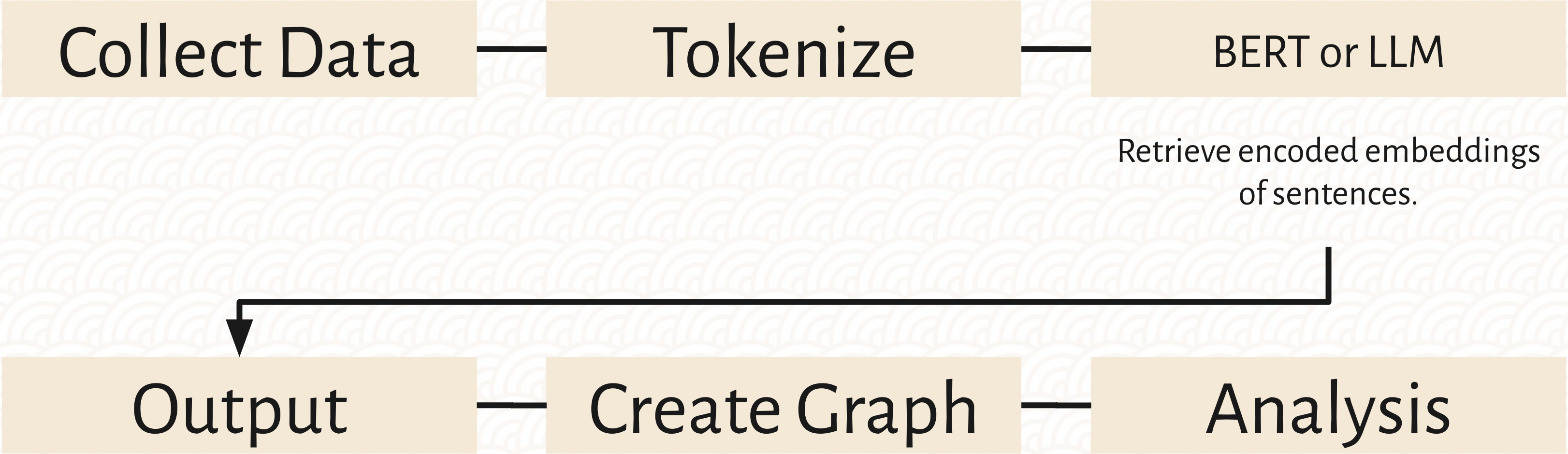
Approach

- Similar to density based clustering on top of word embeddings and cosine similarity as a way to look at the relationship between concepts, but different in that it also allows for graph analysis.
- How it's similar - if you look at DBscan (density based clustering), this approach is actually very similar because cosine similarity is an angle (theta) and the angle that we accept is a vector different than the one we are evaluating.
- So this approach allows for a circle around this vector parameterized by theta, making it a cone. Density based clustering is just a circle.
- There are various ways to do density based clustering on embeddings, ours is interesting to get graphs out of it.
- To map embeddings directly into 2D space, T-SNE is also a valid option.

Research Alert!!

1. Is Cosine-Similarity of Embeddings Really About Similarity?
2. A Survey of Large Language Models on Generative Graph Analytics: Query, Learning, and Applications

Process



Collect Data

Tokenize

BERT or LLM

Retrieve encoded embeddings of sentences.

Output

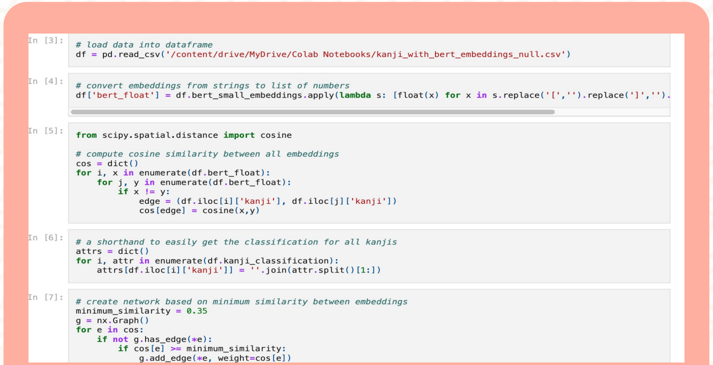
Create Graph

Analysis

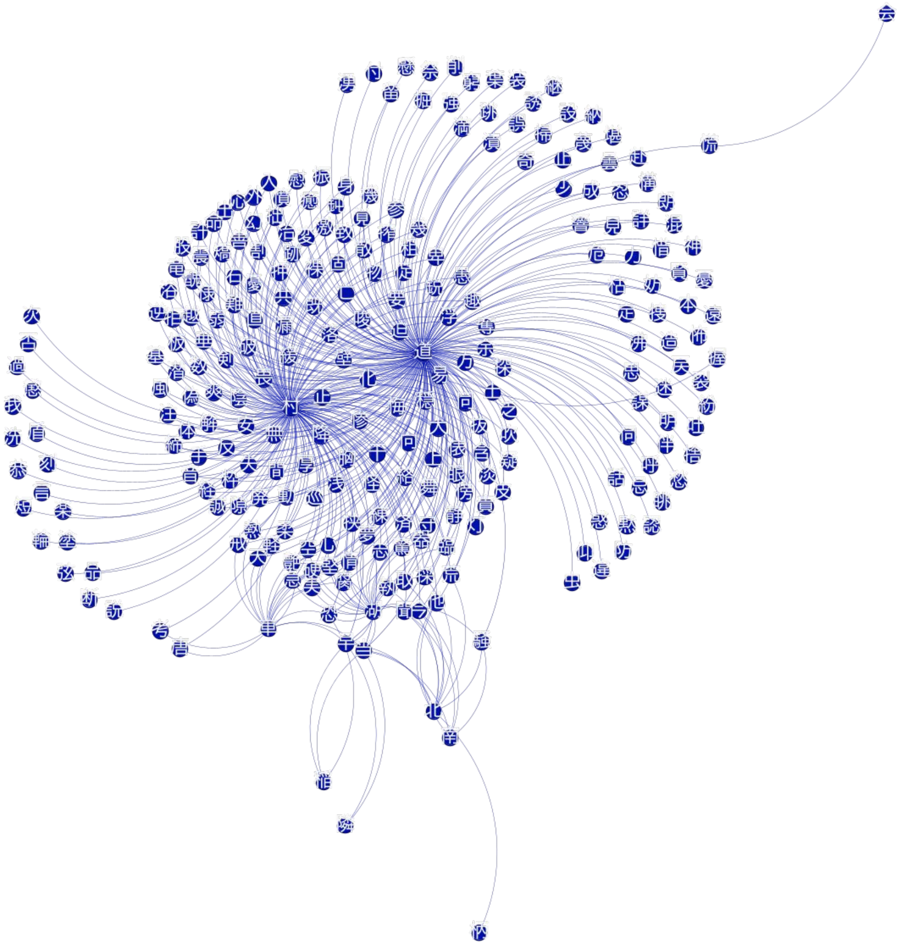
Structure of embeddings as vectors with original attributes

Notebook Walkthrough

https://github.com/yulleyi/bert-kanji-graph/blob/main/graph/Analysis.ipynb



```
In [3]:  # load data into dataframe
         df = pd.read_csv('/content/drive/MyDrive/Colab Notebooks/kanji_with_bert_embeddings_null.csv')

In [4]:  # convert embeddings from strings to list of numbers
         df['bert_float'] = df.bert_small_embeddings.apply(lambda s: [float(x) for x in s.replace('[','').replace(']','').

In [5]:  from scipy.spatial.distance import cosine

         # compute cosine similarity between all embeddings
         cos = dict()
         for i, x in enumerate(df.bert_float):
             for j, y in enumerate(df.bert_float):
                 if x != y:
                     edge = (df.iloc[i]['kanji'], df.iloc[j]['kanji'])
                     cos[edge] = cosine(x,y)

In [6]:  # a shorthand to easily get the classification for all kanjis
         attrs = dict()
         for i, attr in enumerate(df.kanji_classification):
             attrs[df.iloc[i]['kanji']] = ''.join(attr.split()[1:])

In [7]:  # create network based on minimum similarity between embeddings
         minimum_similarity = 0.35
         g = nx.Graph()
         for e in cos:
             if not g.has_edge(*e):
                 if cos[e] >= minimum_similarity:
                     g.add_edge(*e, weight=cos[e])
```

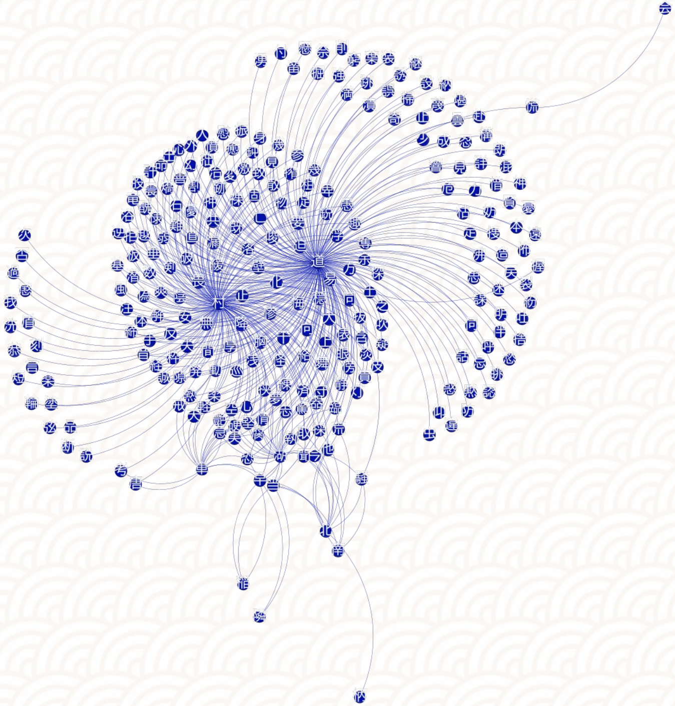

云
道
易
学
要
古
洛
之
力
专
止
女
反
大
上

04

Conclusion

Results and next steps

Summary of Results



1. Looking at the variations of meaning (sentence tokens) at the Kanji level, we identified two influential nodes: Community (Village), and Path/Way/Journey

THANKS!

Questions, comments, or feedback?

Michelle Yi
https://www.linkedin.com/in/michelleyulleyi/
michelle@generationship.ai

GitHub: https://github.com/yulleyi/bert-kanji-graph

CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon**, and infographics & images by **Freepik**