



# Analysis of questions

## Final presentation

by JaMiMaKa group

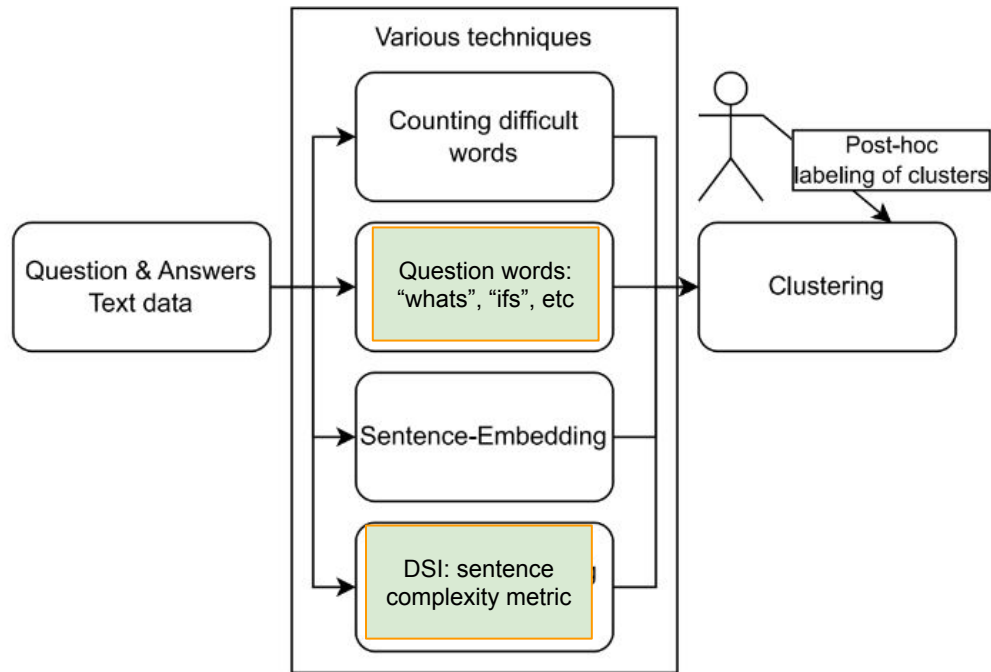
Mikołaj Malec, Marcei Korbin, Kacper Grzymkowski, Jakub Fołtyn



# Topic introduction

- How well can machine learning models understand question complexity?
- What is the influence of the topic being discussed on the complexity of the question?
- What is the relation between the structure of a question and its complexity?
- Can large language models be used to better understand question complexity?
- What is the relation between text complexity and question complexity?

# Proposal





## Data that we used

- **SQuAD** dataset (Stanford Question Answering Dataset)
  - reading comprehension dataset
  - over 100,000 questions asked by crowd workers on 536 articles from Wikipedia, stored with the answers
- due to computational load, we have by now used only around 8,500 questions for analysis



# LLM post-mortem: Beyoncé incident

- We tried different LLM models to evaluate Bloom's taxonomy
- And different prompts
- Most models were very slow
  - Their usefulness was limited anyway
  - The one we used was really biased for BeyoncéFor whatever reason...

Examples:

- which prominent star felt the 2009 female video of the year award should have went to beyoncé instead of taylor swift ? - 6 - evaluation
- hoe did everyone learn that beyonce performed for kaddafi ? 5 - synthesis
- who inspires beyoncé because `` she does it all ? " - 6 - evaluation
- in what year was the slang term from a title of a destiny 's child song that is also used to describe beyoncé put in the dictionary ? 6 - evaluation


## Complexity measures

- How to measure text complexity?
  - There's several ways
- Lexical diversity metrics
  - Longest word
  - % of popular words
  - TTR, CTTR
  - Simpson's D metric (Simpson 1949)
- Text readability metrics
  - Mean sentence length
  - Mean syllables in words
  - Flesch (Flesch 1948)
  - ARI (Senter and Smith 1967)

$$TTR = \frac{V}{N} \quad CTTR = \frac{V}{\sqrt{2N}}$$

$$D = \sum_{i=1}^V f_v(i, N) \frac{i}{N} \frac{i-1}{N-1}$$


$$206.835 - (1.015 \times ASL) - (84.6 \times \frac{n_{sy}}{n_w})$$


$$0.5ASL + 4.71AWL - 21.34$$

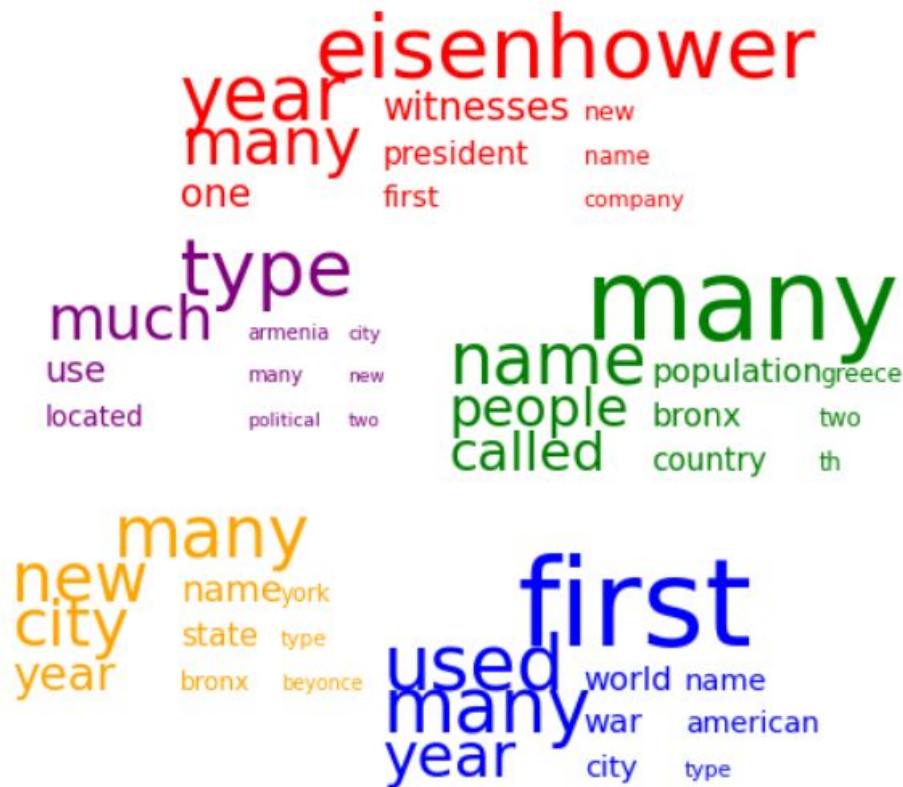


## DSI metric

- DSI — a measure which calculates metric distance of words in a sentence pairwise
  - indicates how diverse the whole sentence is
- Usually cosine distance between embeddings
- **Connected with creativity**
- More info:
  - *Johnson, Dan R., et al. "Divergent semantic integration (DSI): Extracting creativity from narratives with distributional semantic modeling."*

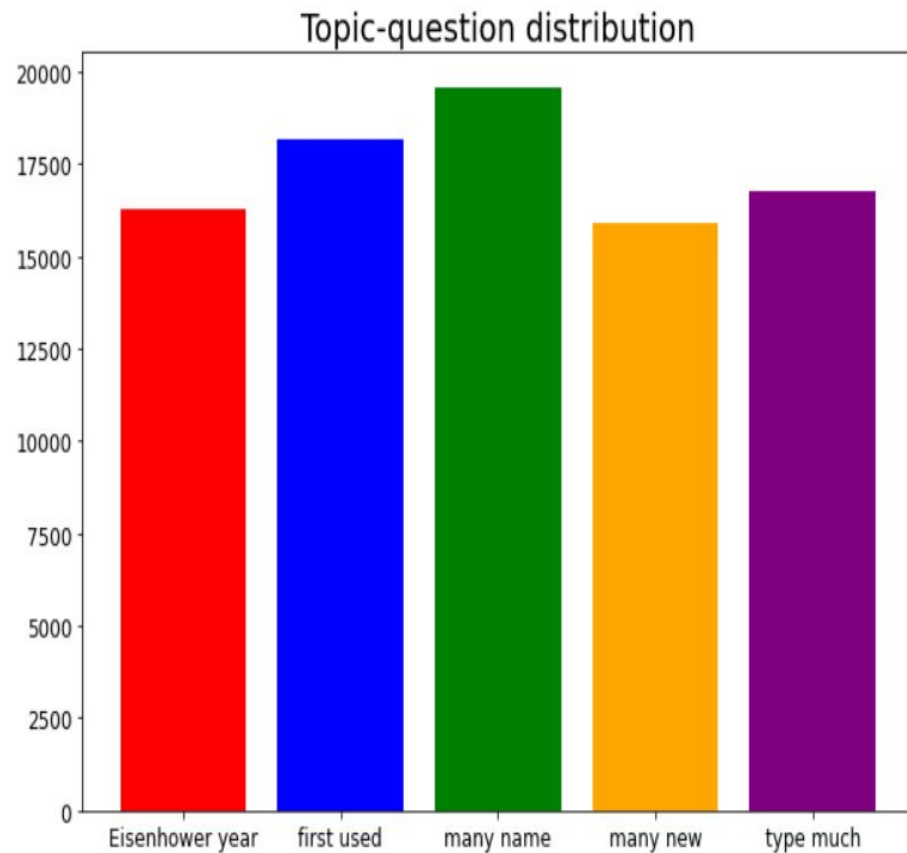
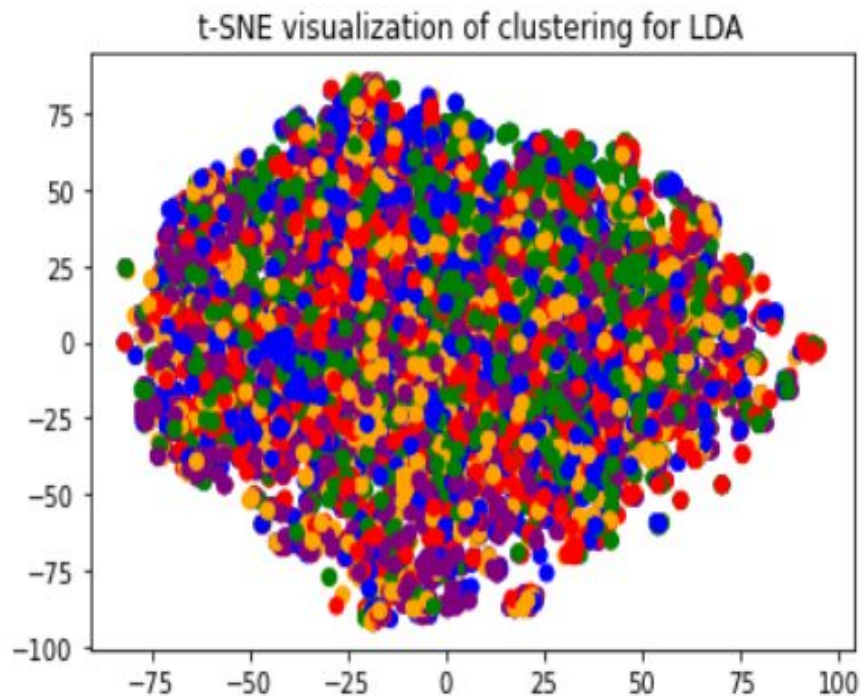
# Latent Dirichlet allocation

- Hard to apply for our use case
- Questions are usually short in nature
  - Relatively small diversity of words in dataset
- Difficult to determine the topic's.. topic





# LDA clustering



# LDA clustering – examples



- adrenalize followed what 1987 def leppard record ?
- what was the term used to describe supervisors of estates ?
- who should be able to challenge administrative orders in court ?
- how did goring expect to regain prestige ?
- what country has failed to ratify sections 87 and 92 of the ilo ?
- who was the last chief minister of norfolk island ?
- which president called kanye west a jackass for his behavior at the 2009 vmas ?
- how long does a wrestler have to leave the ring once they are tagged out ?
- what was the common mosaic theme of iconoclastic churches ?
- what represents a preaspirated bilabial stop ?
- what has a color temperature around 2800 to 3000 kelvin ?
- what entity controlled the prussian army ?
- has new mexico had a majority of spanish speaking natives .
- what was the original stimuli for creating the website ?
- do detritivores have a direct impact on `` donor " organisms ?
- how many feet above sea level is todt hil ?
- how tall is the peak of piz bernina
- what is the largest fishery in the country ?
- on what island did the missionaries live with citizens of the marshall islands ?
- what percentage of the economy is greece 's service sector ?
- in november and december of 1940 what changed to make attacks on civilians a moot point ?
- what do people with a high quality of african descent classify themselves as ?
- who did victoria blame for prince albert 's death ?
- which famous line from kerry 's speech was later featured in one of his later television ad campaigns ?
- what country originally pulled ipods due to higher-than-allowed volume levels ?
- what are some reasons why gas prices may vary in alaska , especially ?
- who was the soviet leader who attended nasser 's funeral ?
- in terms of universities in denmark what is the fee status for citizens of efa states ?

# Bloom's Taxonomy

## Guided LDA

- Another try at LDA
- Guided – we provide seeds as topic “suggestions”
- We tried levels from the bloom’s taxonomy

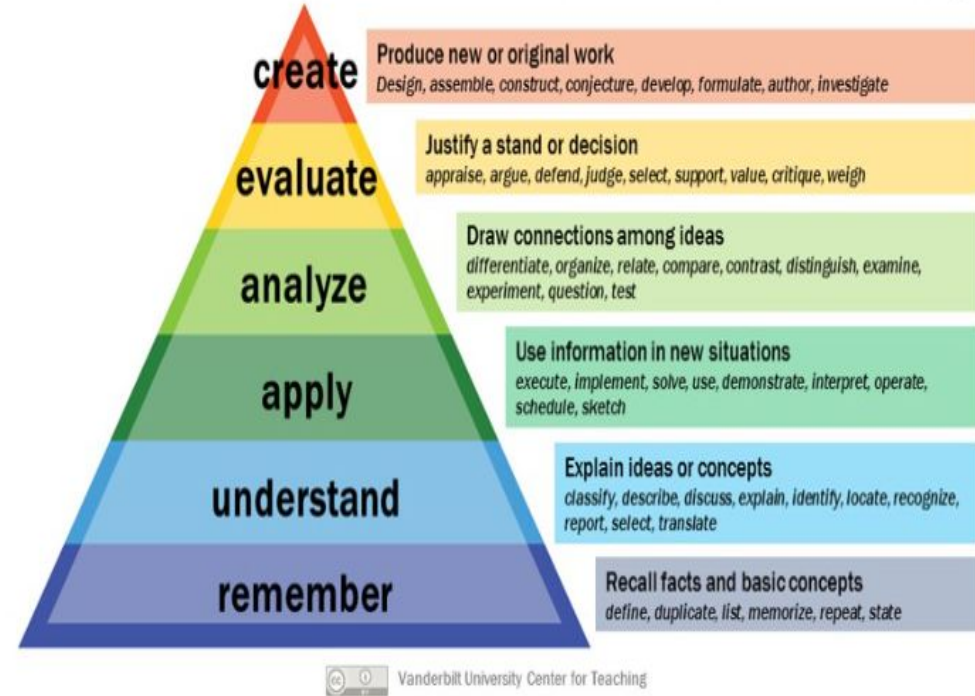
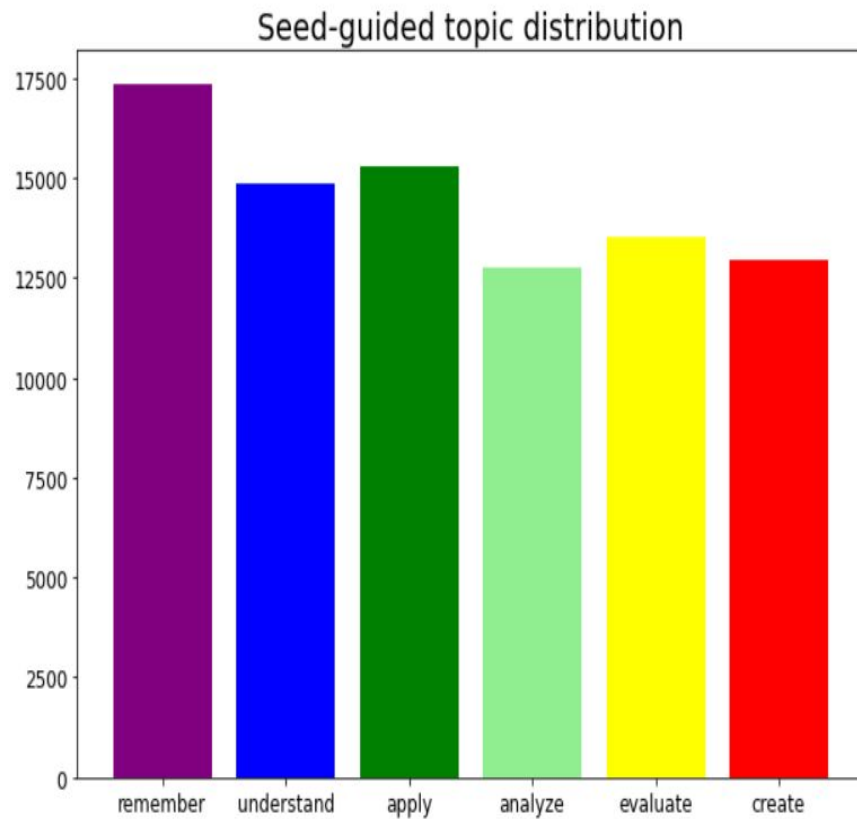
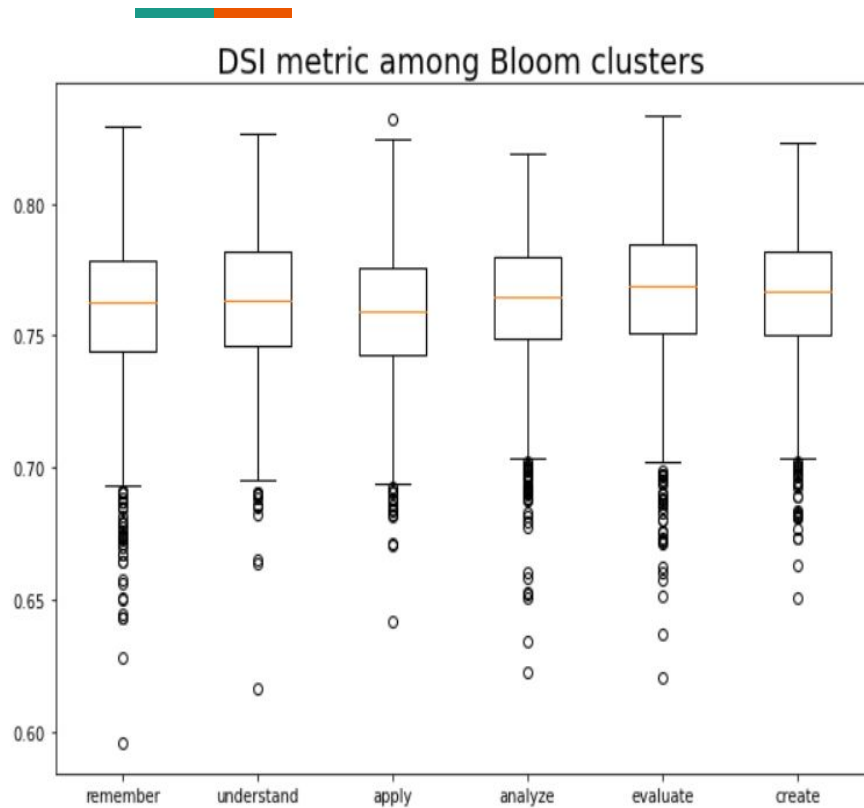


Image source: Vanderbilt University Center for Teaching

# Guided LDA – results

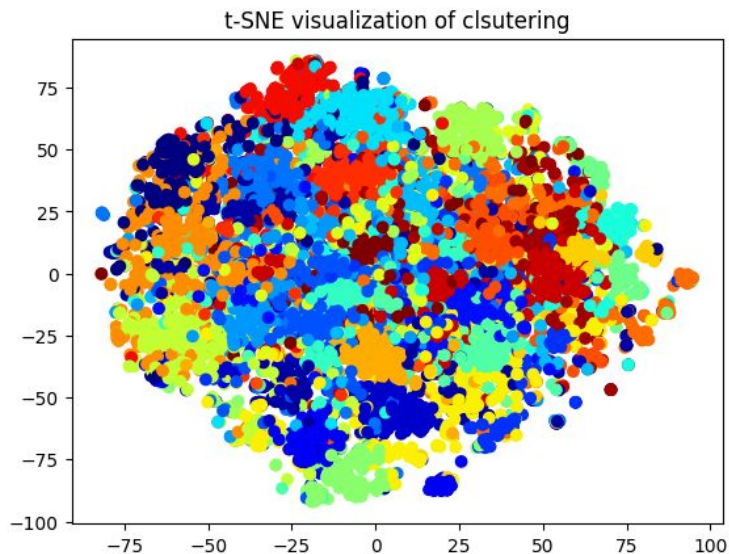




## Guided LDA – results

- Questions in groups did not “seem right”
  - Hard for us to verify
- In reality most questions should fall into “remember” category
- Questions are rather short

## Results and how we have done them



- Clustering on embeddings of all data
- Stratified sampling on clusters
  - 86k -> 8.6k questions
- t-SNE visualization based on just sentence embeddings
- Second clustering on the smaller data combined with other measures
- Looking at clusters for the post-hoc analysis



# Clusters example output

Cluster 7 questions examples:

1. what area of the **hippocampus** plays a role in storing new memories ?
2. what are the six perfections under **mahayana teachings** ?
3. unfpa lists elements that promote what **human right** ?
4. what are some of the findings that support their **argument** ?
5. what is another **nature of dukkha** ?
6. how can patients with **dementia** indicate discomfort exists ?
7. what **narrative technique** does lee use to combine the adult 's perspective with the child's observations ??
8. what is the study of interaction between **living things** ?
9. the oldest **brain** found in a cave was from what gender of human ?
10. what discipline studies the role of **emotions** in neural mechanisms ?

Cluster 23 questions examples:

1. what portion of **india** did the chalukya empire rule ?
2. where was the center of **magadha** ?
3. to what dynasty did **vikramaditya ii** belong ?
4. what **persian** scholar noted the 10th century calculator similar to the antikythera mechanism ?
5. which sura in the **quran** describes the visitation of an angel upon zakariya ?
6. what does "**sarvajna yavanaa**" mean ?
7. when did ali javan , an **iranian scientist** , co-invent the first gas laser ?
8. which **middle east** location was the only area the torch visited ?
9. which people would **mohammad** 's critics have compared him to at the time ?
10. who argued that bonaparte 's admiration for **muhammad** was sincere ?



## Beyoncé cluster?

Cluster 0 questions examples:

Index 759: what are **beyoncé** 's backup singers called ?

Index 282: what was **beyoncé** 's role in destiny 's child ?

Index 348: who did **beyonce** record with for the movie `` the best man ? ''

Index 447: who beat out **beyonce** for best female video ?

Index 295: who is **beyoncé** married to ?

Index 668: who began reporting **beyoncé** 's annual earnings , starting in 2008 ?

Index 699: what theme was **beyonce** 's early music ?

Index 647: when did **beyonce** sign a letter for one campaign ?

Index 736: what song did **beyoncé** sing at a 2006 concert to honor josephine baker ?

Index 717: what number was **beyoncé** on the top 20 hot 100 songwriters list ?

Cluster 4 questions examples:

Index 6808: **buddhist** concept of dependent origination has been compared to what modern thought ?

Index 6719: what two countries after china was the **mahayana** sutras spread ?

Index 6257: most accept that **buddha** lived and taught in what type of order ?

Index 6336: what actions does karma refer to in **buddhism** ?

Index 6405: what is a contributing factor to **dukkha** ?

Index 6792: what **sutras** were transmitted in secret ?

Index 6724: west **buddhism** is often seen as exotic and what ?

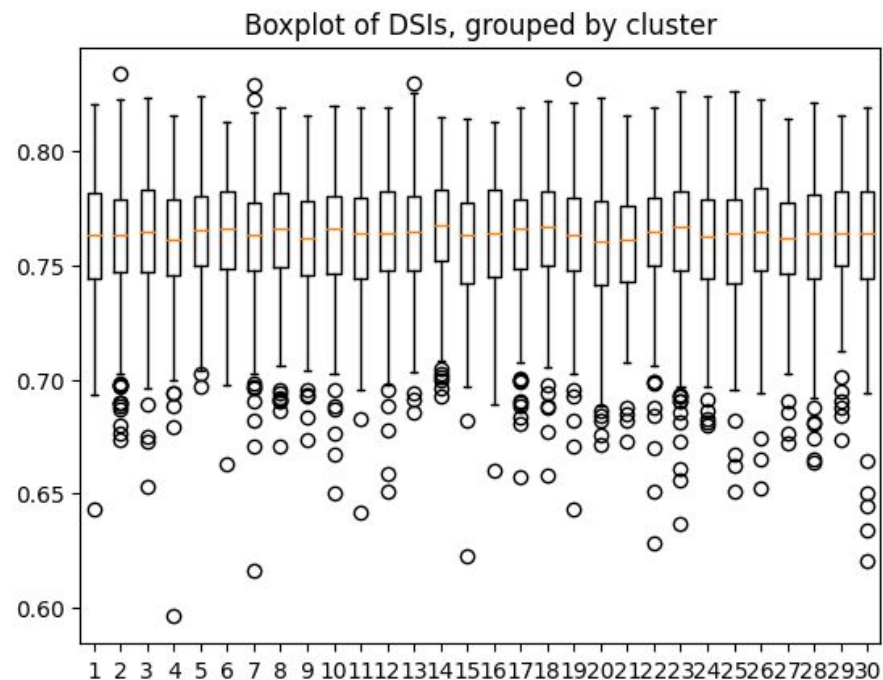
Index 6494: according to what school is **tathgatagarbha** the inseparability of clarity and emptiness of one 's mind ?

Index 6323: is liberation from **samsara** possible ?

Index 6443: what is the goal of the **buddhist** path ?

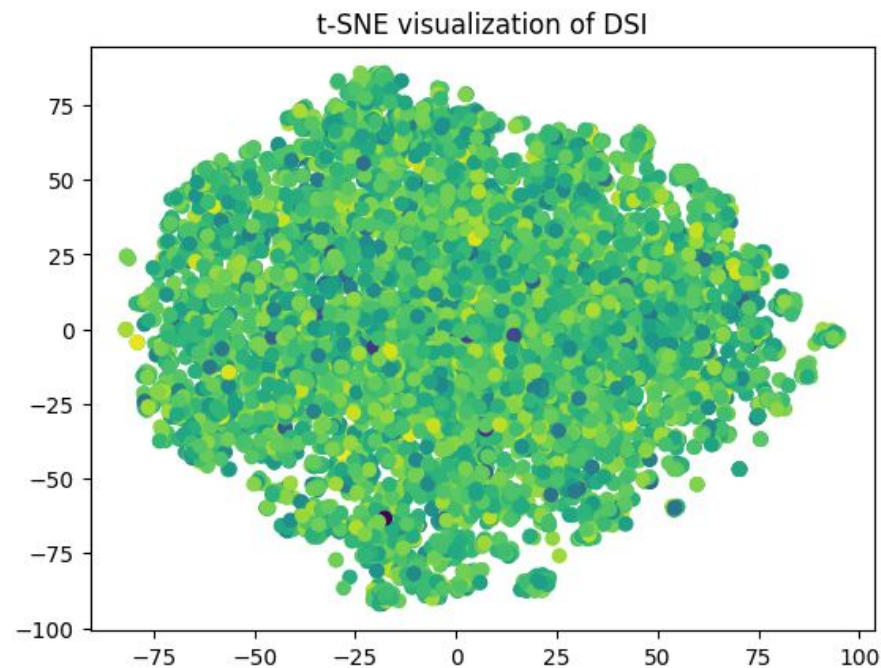


## DSI per cluster



# DSI vs embeddings

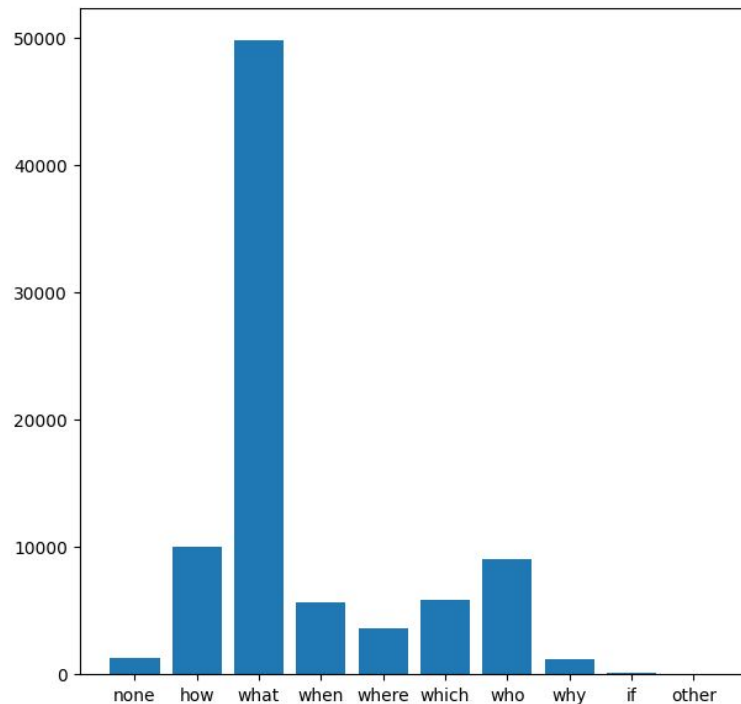
- Confetti plot
- No relationship



## Question words

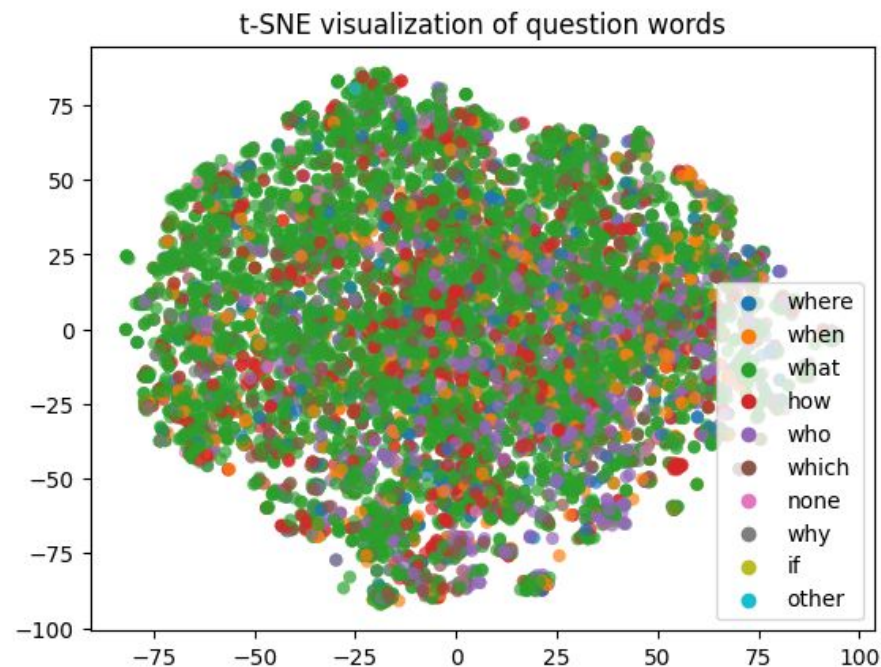
- A lot of “what” questions
- A bunch of “how”, “when”, “where”, etc
- Few “why” questions
- Barely any “if” questions

A flaw of the dataset we chose



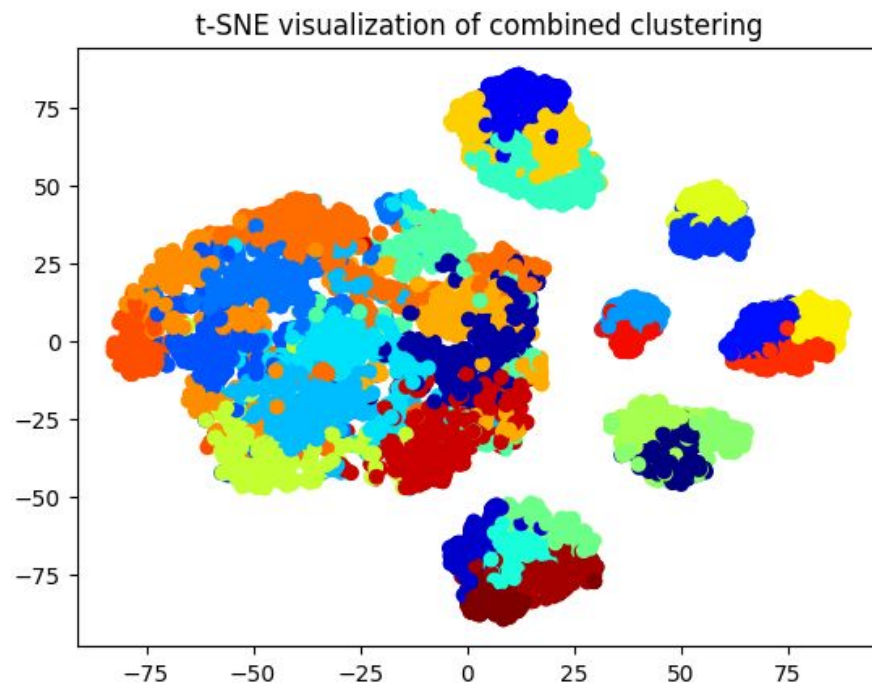
## Question words on t-SNE

- Completely mixed
- We were hoping to maybe get something



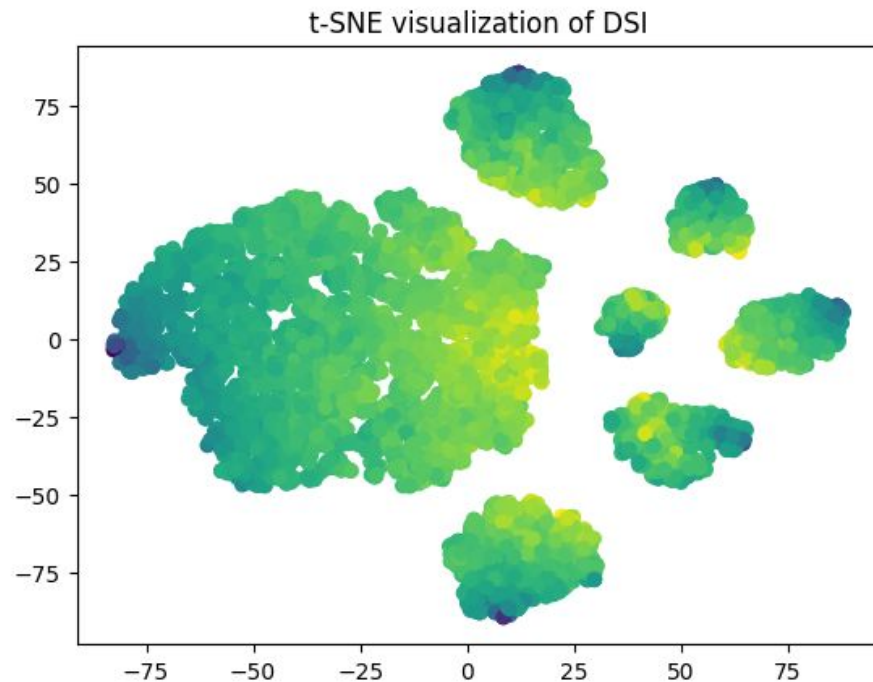
# Using the new metrics for t-SNE and clustering

- Combined DSI and question words with embeddings
- Using categorical data in clustering
  - Not easy to do properly
  - Available libraries to deal with it use 1 CPU core
  - Too much for even the trimmed data
- So we just used k-Means
- Second t-SNE run on combined data



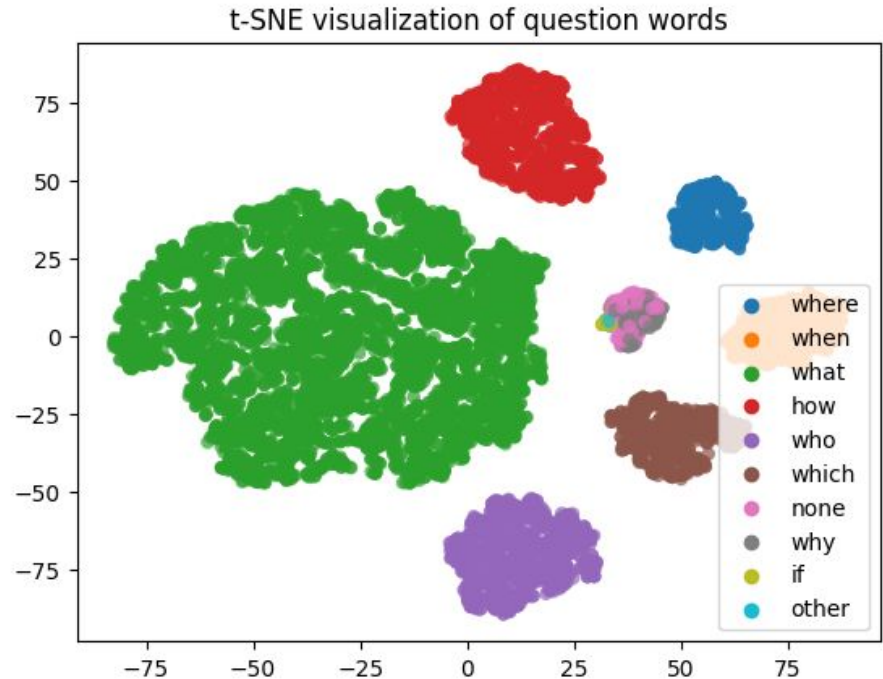
# DSI visualization

- Nice gradients
  - ... but that's about it.
- No hard boundaries
- Clusters we found



# Questions

- Dominance of the question words clear
- The thing we were trying to avoid by using the proper clustering method
- Interesting cluster in the middle
  - Analysed further
  - The thing we were looking for?





# From the interesting cluster

## The good (relatively):

- why was there a large population of algonquian people in bermuda ?
- why is it possible to distinguish utf-8 from other protocols ?
- why are 99 % of pesticide related deaths in underdeveloped countries ?

## The bad:

- did they need parental consent ?
- may welsh clubs enter the competition ?
- are the ewell 's considered rich or poor ?
- name a smaller newspaper ?





# Conclusions

- Sparks of something interesting
- Exploratory nature of the task
- Dataset is very limited in types of question limiting further analysis
  - Most of the questions are queries that could be typed into Google or ChatGPT
  - Useful... but not for us
  - Other datasets share this problem
- Limitations of clustering
- Trying to do structure modelling in a topic modelling world



# What's next?

- Expand to other datasets, preferably with diverse types of questions
- Incorporate context
  - Aiming here at questions like “why is that?”, which are completely meaningless without context, but appear often in e.g. STA data
- Attempt more concrete labelling using Bloom’s taxonomy
  - We have some ideas after this project
  - Avoids the worst parts of clustering