Analysis of questions

Literature Review, Solution Proposal

by JaMiMaKa group
Mikołaj Malec, Marceli Korbin, Kacper Grzymkowski, Jakub Fołtyn

Topic introduction

- Not only question answering
- **Clustering based on topic, difficulty**
- Can we measure question complexity?
- Can we classify questions based on Bloom's taxonomy?

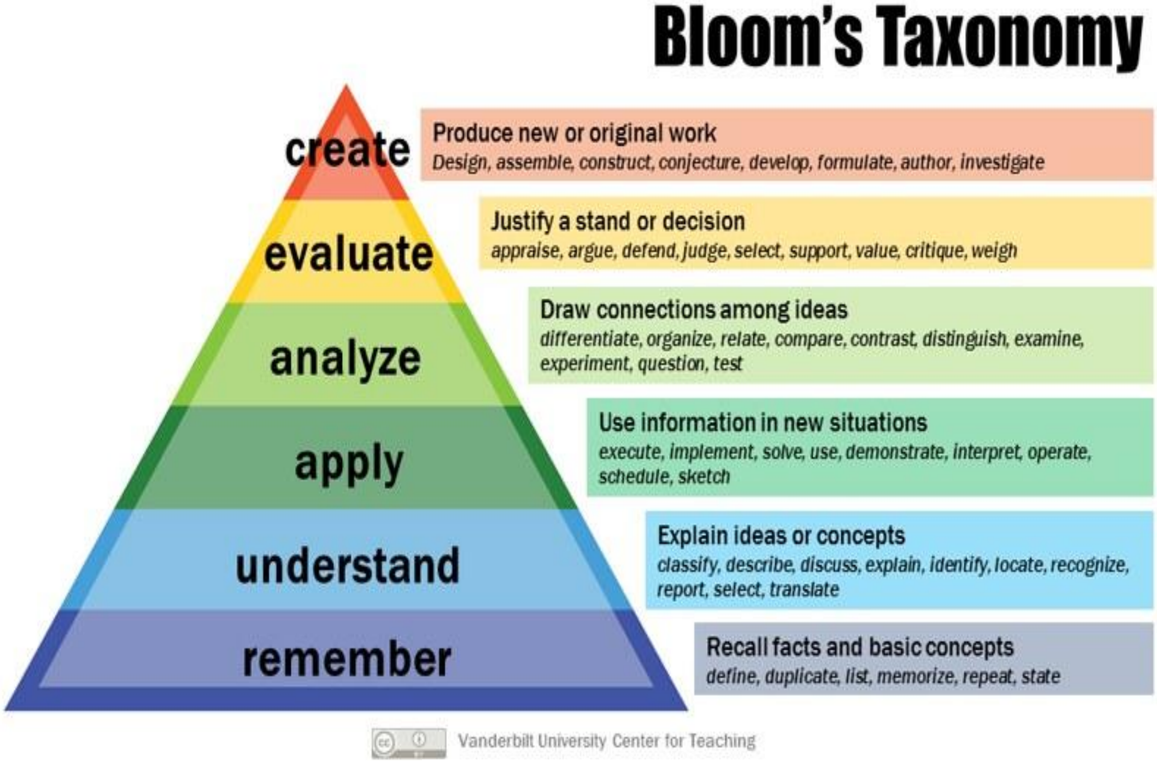
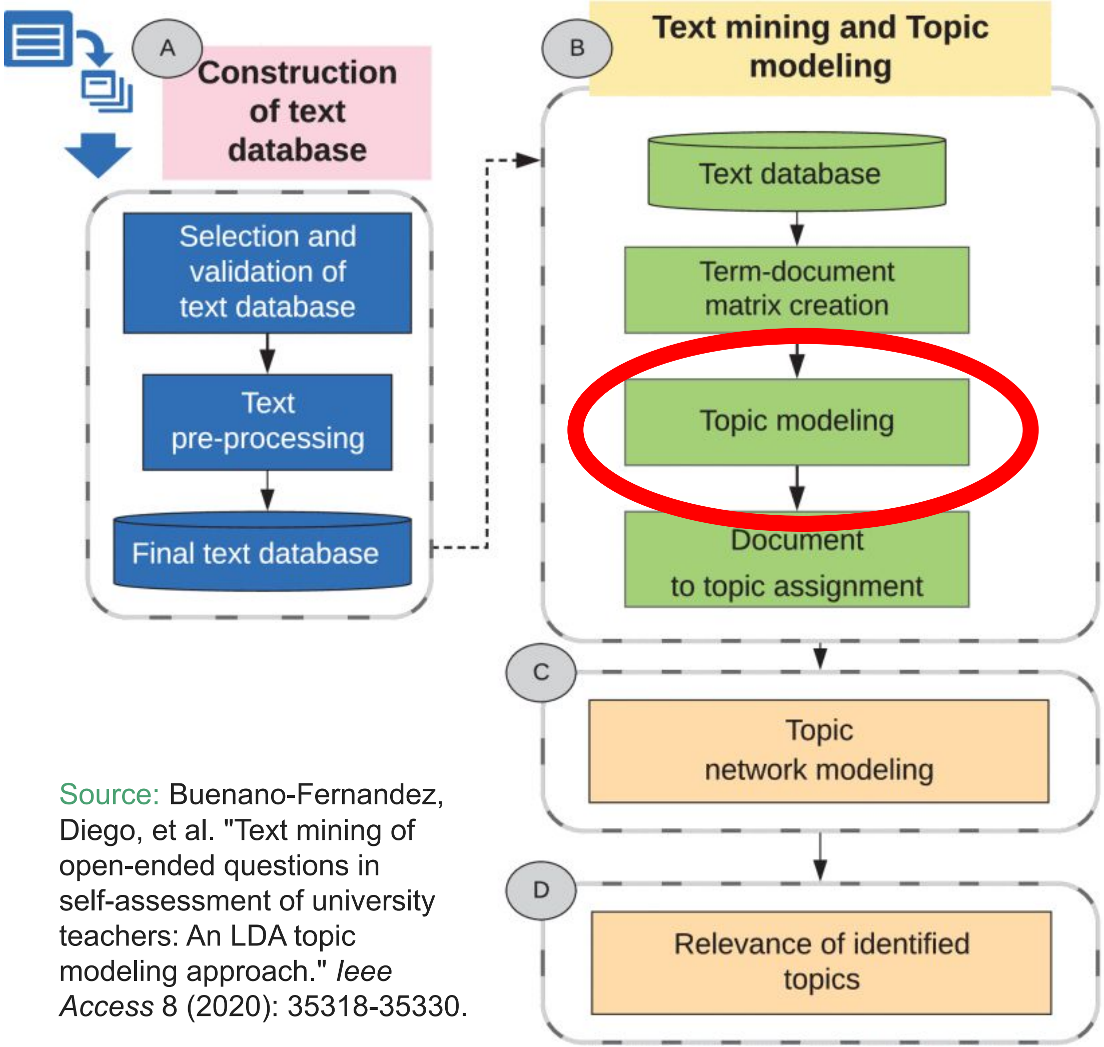

Bloom's Taxonomy
create
Produce new or original work
Design, assemble, construct, conjecture, develop, formulate, author, investigate
evaluate
Justify a stand or decision
appraise, argue, defend, judge, select, support, value, critique, weigh
analyze
Draw connections among ideas
differentiate, organize, relate, compare, contrast, distinguish, examine, experiment, question, test
apply
Use information in new situations
execute, implement, solve, use, demonstrate, interpret, operate, schedule, sketch
understand
Explain ideas or concepts
classify, describe, discuss, explain, identify, locate, recognize, report, select, translate
remember
Recall facts and basic concepts
define, duplicate, list, memorize, repeat, state
Vanderbilt University Center for Teaching

Image source: Vanderbilt University Center for Teaching

Literature Review – topic modeling

- Topic modelling technique: LDA used in clustering open-ended surveys
- Whole pipeline presented in the article
- “Progression” – seed-guided topic modeling


A
Construction of text database
Selection and validation of text database
Text pre-processing
Final text database
B
Text mining and Topic modeling
Text database
Term-document matrix creation
Topic modeling
Document to topic assignment
C
Topic network modeling
D
Relevance of identified topics

Source: Buenano-Fernandez, Diego, et al. "Text mining of open-ended questions in self-assessment of university teachers: An LDA topic modeling approach." *Ieee Access* 8 (2020): 35318-35330.

Literature Review – Latent Dirichlet Allocation

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. Journal of machine Learning research, 3(Jan):993–1022.

- Latent Dirichlet allocation (LDA) is a generative probabilistic model for collections of discrete data.
- LDA is a three-level hierarchical Bayesian model.
- Each article in collection is a mix of different topics, each of these is also thought of as a mix of smaller sub-topics, these are again seen as mixes of even smaller parts.

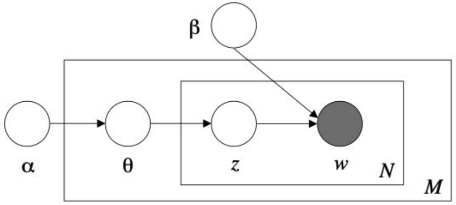
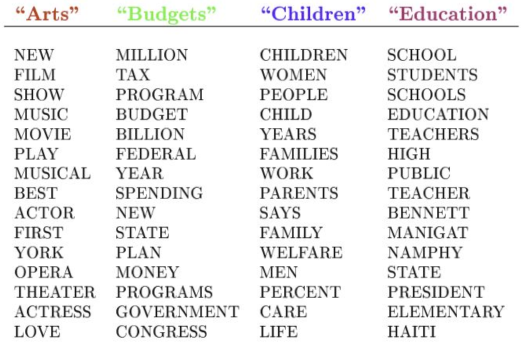
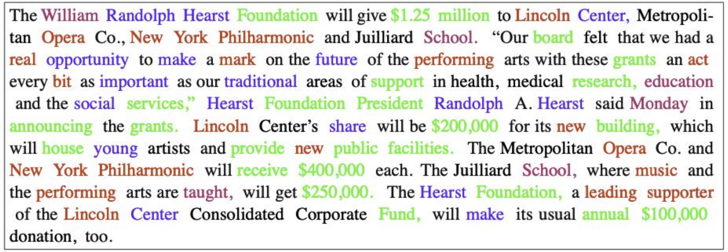


Figure 1: Graphical model representation of LDA. The boxes are “plates” representing replicates. The outer plate represents documents, while the inner plate represents the repeated choice of topics and words within a document.



“Arts”	“Budgets”	“Children”	“Education”
NEW	MILLION	CHILDREN	SCHOOL
FILM	TAX	WOMEN	STUDENTS
SHOW	PROGRAM	PEOPLE	SCHOOLS
MUSIC	BUDGET	CHILD	EDUCATION
MOVIE	BILLION	YEARS	TEACHERS
PLAY	FEDERAL	FAMILIES	HIGH
MUSICAL	YEAR	WORK	PUBLIC
BEST	SPENDING	PARENTS	TEACHER
ACTOR	NEW	SAYS	BENNETT
FIRST	STATE	FAMILY	MANIGAT
YORK	PLAN	WELFARE	NAMPHY
OPERA	MONEY	MEN	STATE
THEATER	PROGRAMS	PERCENT	PRESIDENT
ACTRESS	GOVERNMENT	CARE	ELEMENTARY
LOVE	CONGRESS	LIFE	HAITI



The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. “Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services,” Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center’s share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Figure 8: An example article from the AP corpus. Each color codes a different factor from which the word is putatively generated.

Literature Review – Attention Is All You Need

Ashish Vaswani et al. 2017. Attention is all you need. Advances in neural information processing systems, 30.

- New state of the art model with smaller fraction of the training costs of the best models from the literature.
- Transformer is based solely on attention mechanisms.

commons.wikimedia

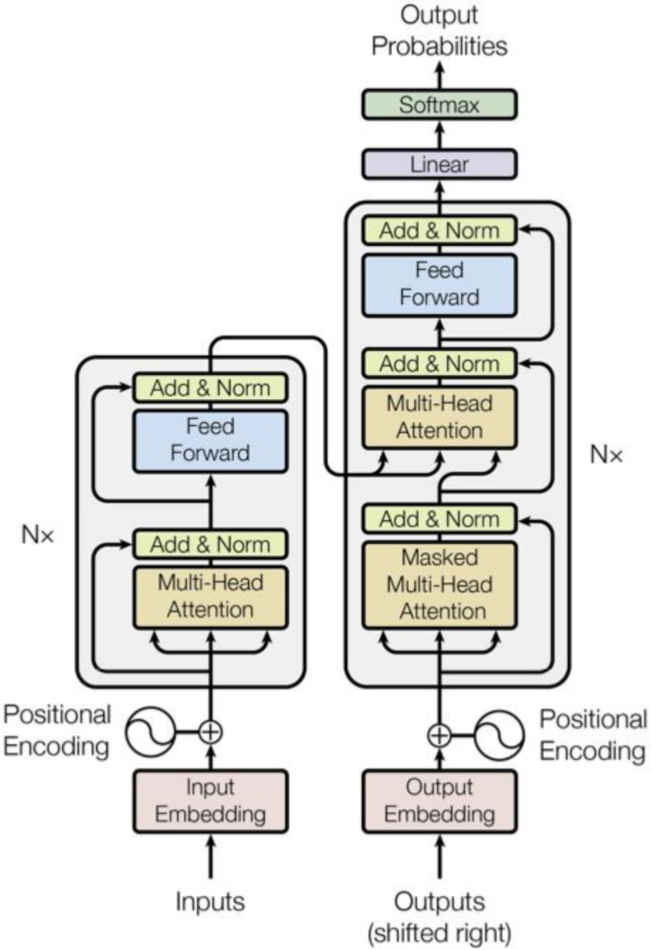


Figure 1: The Transformer - model architecture.

Literature Review – BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

- BERT stands for Bidirectional Encoder Representations from Transformers.
- BERT is designed to pretrain deep bidirectional representations from unlabeled text.
- As a result, the pre-trained BERT model can be fine tuned with just one additional output layer to create state-of-the-art models.

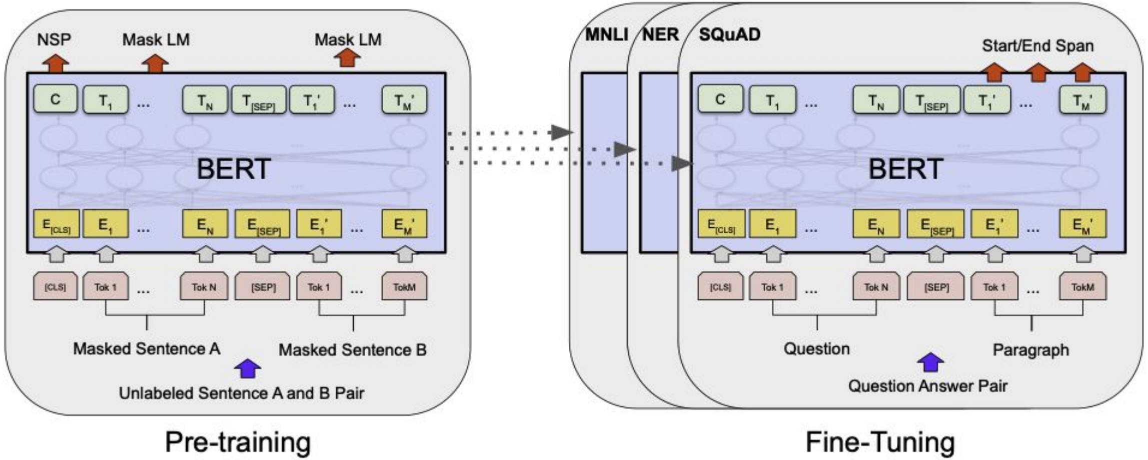
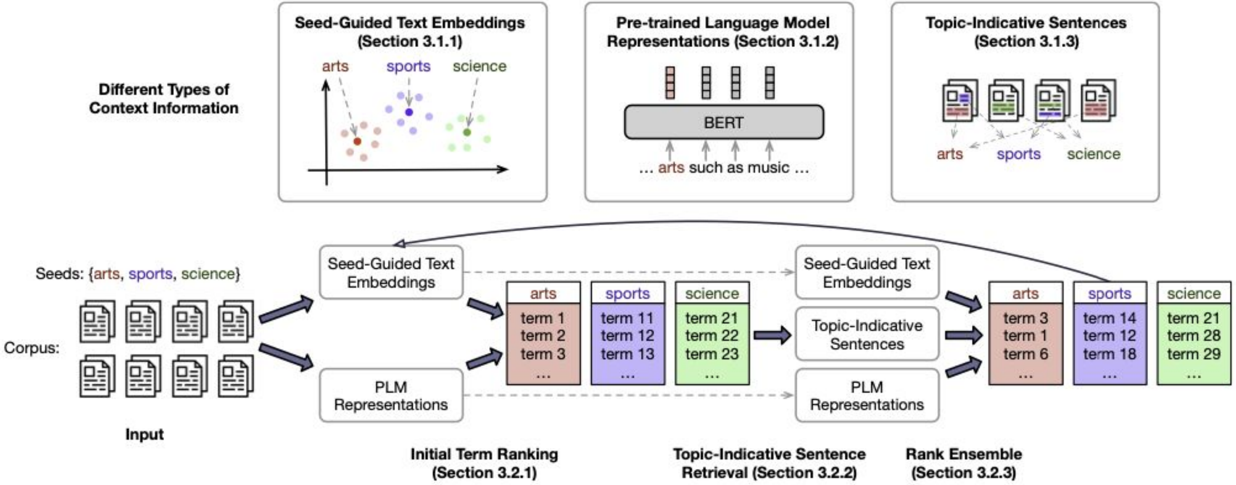


Figure 1: Overall pre-training and fine-tuning procedures for BERT. Apart from output layers, the same architectures are used in both pre-training and fine-tuning. The same pre-trained model parameters are used to initialize models for different down-stream tasks. During fine-tuning, all parameters are fine-tuned. `[CLS]` is a special symbol added in front of every input example, and `[SEP]` is a special separator token (e.g. separating questions/answers).

Literature Review – Effective Seed-Guided Topic Discovery by Integrating Multiple Types of Contexts

Yu Zhang et al. 2023. Effective seed-guided topic discovery by integrating multiple types of contexts. In Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining, pages 429–437.

- "Seed" words are used by model to search for topics we are interested in.
- "SeedTopicMine" algorithm is using three types of information sources.
  - Seed-Guided Text Embeddings (LDA)
  - Pre-trained Language Model Representations (BERT)
  - Topic-Indicative Context (correlation between a given term and a seed)

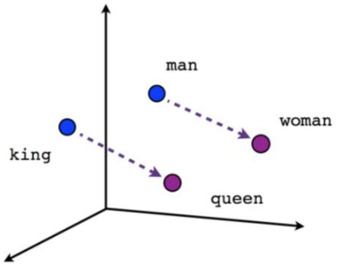


Different Types of Context Information
Seed-Guided Text Embeddings (Section 3.1.1)
arts
sports
science
Pre-trained Language Model Representations (Section 3.1.2)
BERT
... arts such as music ...
Topic-Indicative Sentences (Section 3.1.3)
arts
sports
science
Seeds: {arts, sports, science}
Corpus:
Input
Seed-Guided Text Embeddings
PLM Representations
arts
term 1
term 2
term 3
...
sports
term 11
term 12
term 13
...
science
term 21
term 22
term 23
...
Seed-Guided Text Embeddings
Topic-Indicative Sentences
PLM Representations
arts
term 3
term 1
term 6
...
sports
term 14
term 12
term 18
...
science
term 21
term 28
term 29
...
Initial Term Ranking (Section 3.2.1)
Topic-Indicative Sentence Retrieval (Section 3.2.2)
Rank Ensemble (Section 3.2.3)

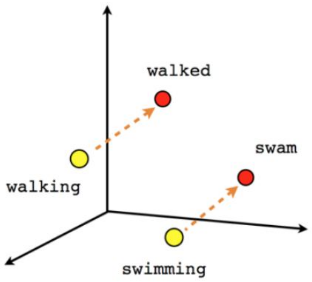
Figure 1: Overview of the SeedTopicMine framework.

Sentence Embedding

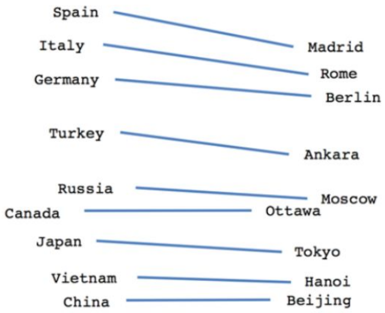
- Sentence embedding is a process of representing a sentence or a piece of text as a fixed-dimensional vector.
- This allows for meaningful comparisons and similarity measurements between sentences.
- Several methods are employed to generate sentence embeddings:
    - Averaging Word Embeddings
    - Pre-trained Models like BERT
    - Neural Network-Based Approaches



Male-Female



Verb tense

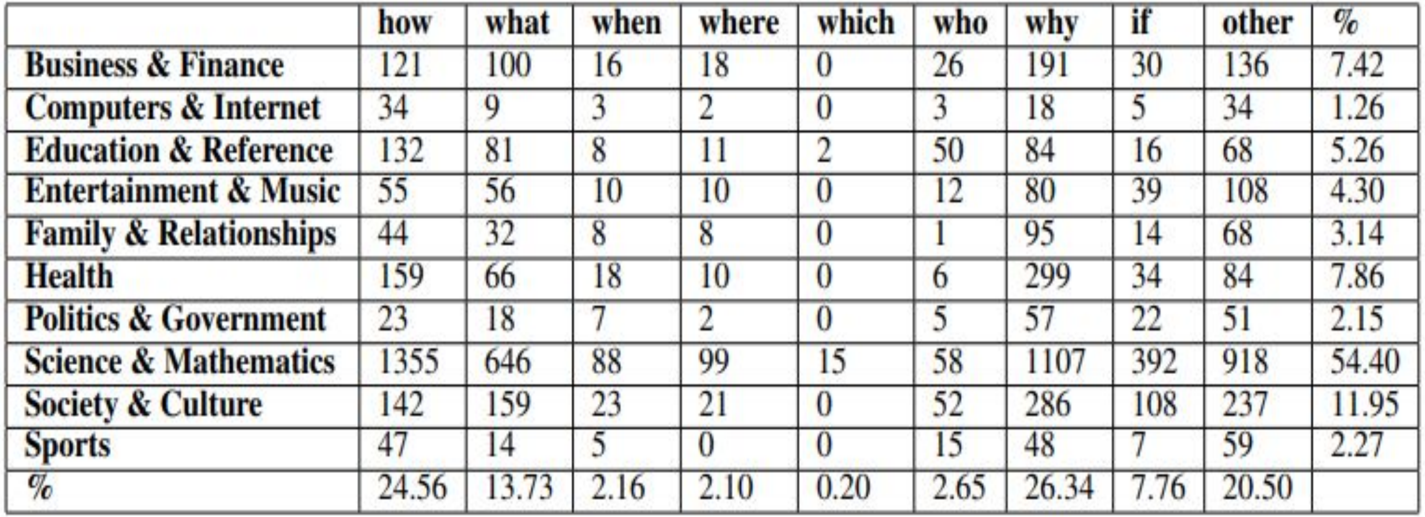


Country-Capital

https://www.analyticsvidhya.com/blog/2020/08/top-4-sentence-embedding-techniques-using-python/

Literature Review – WTC Corpus

- Over 8,000 questions asked by the residents of Bristol
- Cluster and classify the questions
- Try to answer them
- Further research possible



	how	what	when	where	which	who	why	if	other	%
Business & Finance	121	100	16	18	0	26	191	30	136	7.42
Computers & Internet	34	9	3	2	0	3	18	5	34	1.26
Education & Reference	132	81	8	11	2	50	84	16	68	5.26
Entertainment & Music	55	56	10	10	0	12	80	39	108	4.30
Family & Relationships	44	32	8	8	0	1	95	14	68	3.14
Health	159	66	18	10	0	6	299	34	84	7.86
Politics & Government	23	18	7	2	0	5	57	22	51	2.15
Science & Mathematics	1355	646	88	99	15	58	1107	392	918	54.40
Society & Culture	142	159	23	21	0	52	286	108	237	11.95
Sports	47	14	5	0	0	15	48	7	59	2.27
%	24.56	13.73	2.16	2.10	0.20	2.65	26.34	7.76	20.50	

Source: Xu, Zhaozhen, et al. "What makes us curious? Analysis of a corpus of open-domain questions." *arXiv preprint arXiv:2110.15409* (2021)..

Datasets

Most of the datasets we consider contain simple question-answer pairs:

- Large Question Answering Datasets collection
  - https://github.com/ad-freiburg/large-qa-datasets
- R. Tatman's question-answer dataset sourced from Wikipedia (with difficulty assessment)
- Stanford Question Answering Dataset

Datasets

However, we can also use:

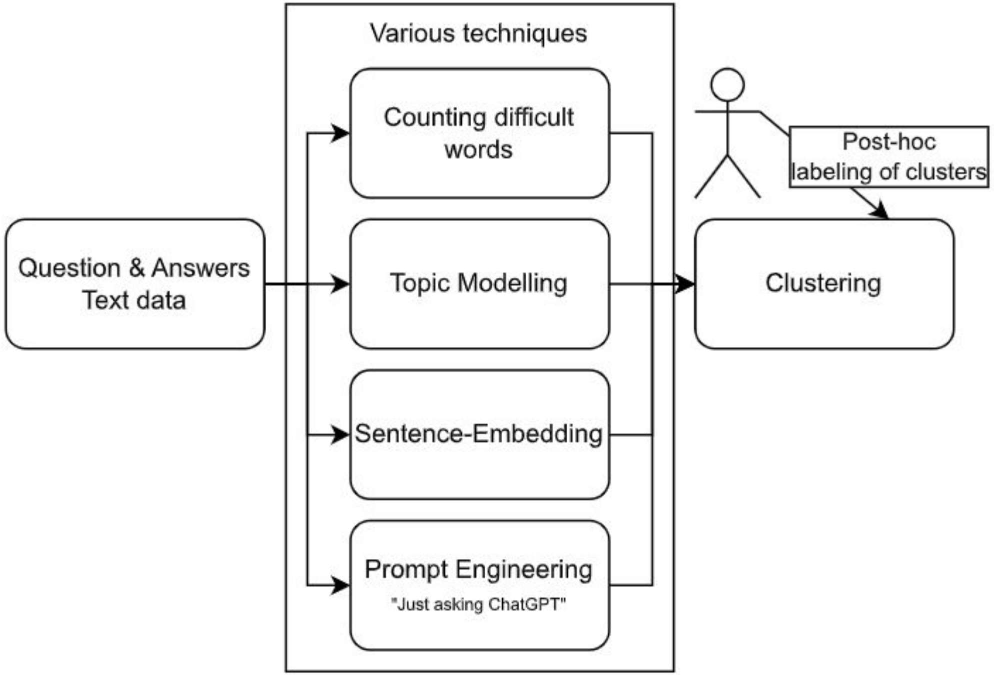
- question-answer **jokes** from r/Jokes subreddit (a part of Reddit)
- Quora’s question pairs (**duplicate questions** detection)
- Yahoo! Answers (grouped into **10 categories**)
- WTC Corpus (if available in time)
- data from STA (Slovene Press Agency)

Solution Proposal

- Take the Q&A data
- Mine them for topics, word complexity
- Use LLM prompt engineered to “ask chatGPT if this question is complex”
- Use sentence embeddings
- Perform clustering on the results from the various mining data
- Post-hoc analysis on the clustering - trying to make sense of the clusters
    - Limited knowledge on psychology
    - Consider simplifications on Bloom’s taxonomy - simple/complex questions
    - Data source considerations - Lectures vs Internet forums

Solution Proposal

Example of what might be created:



Summary

- Interesting topic
- Data labels are very limited
- Possible psychological uses - study of creativity
- LLM - psychology bridge
    - We don’t really know what’s going on in LLMs
    - And we don’t know what’s going on in human brains either

Thank you for your attention!

Any creative questions?