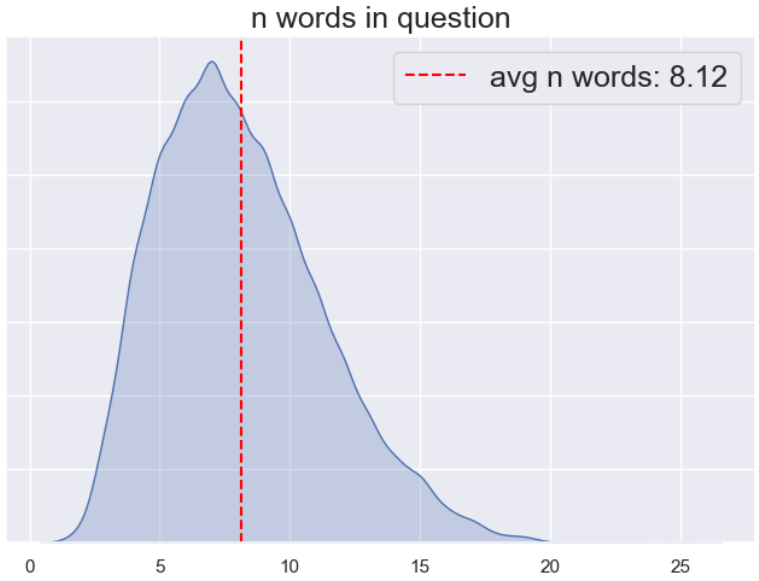
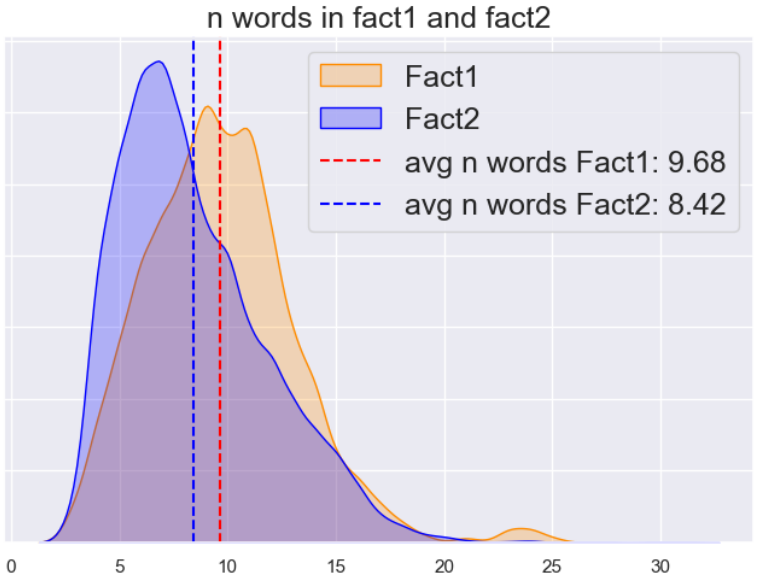
Text Mining and Natural Language Processing 2023-2024

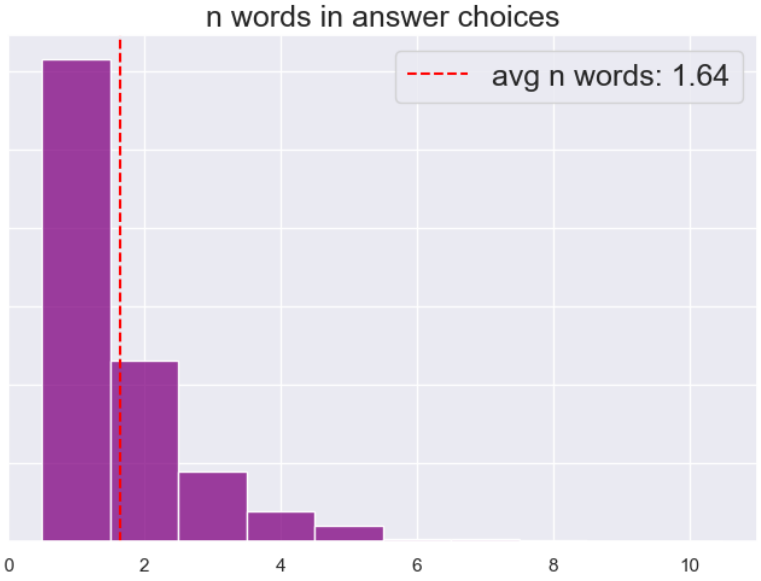
SelectWise

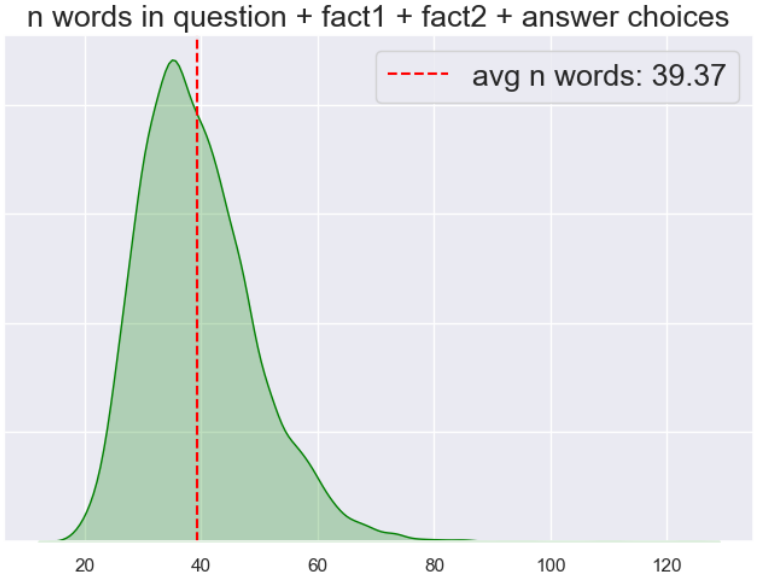
Ghiotto Alessandro 513944
alessandro.ghiotto01@universitadipavia.it

Bachelor of Science in Artificial Intelligence





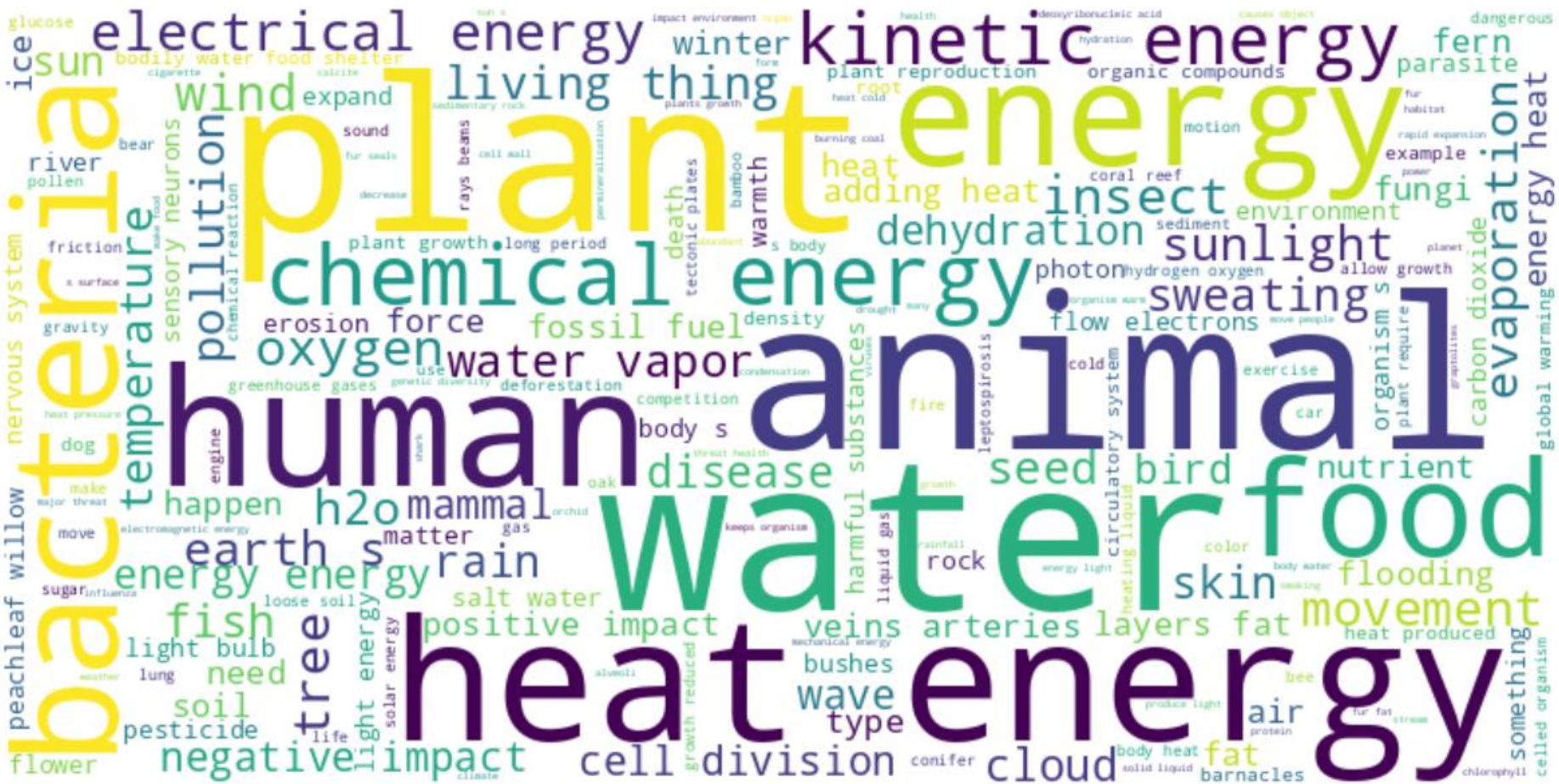




QASC dataset

- **question**: "Climate is generally described in terms of what?
- **fact1**: "Climate is generally described in terms of temperature and moisture."
- **fact2**: "Fire behavior is driven by local weather conditions such as winds, temperature and moisture."
- **8 choices**: {A:"sand",…, H:"city life"]}
- **label**: "F“

	Number of samples
Train	7323
Val	811
Test	926

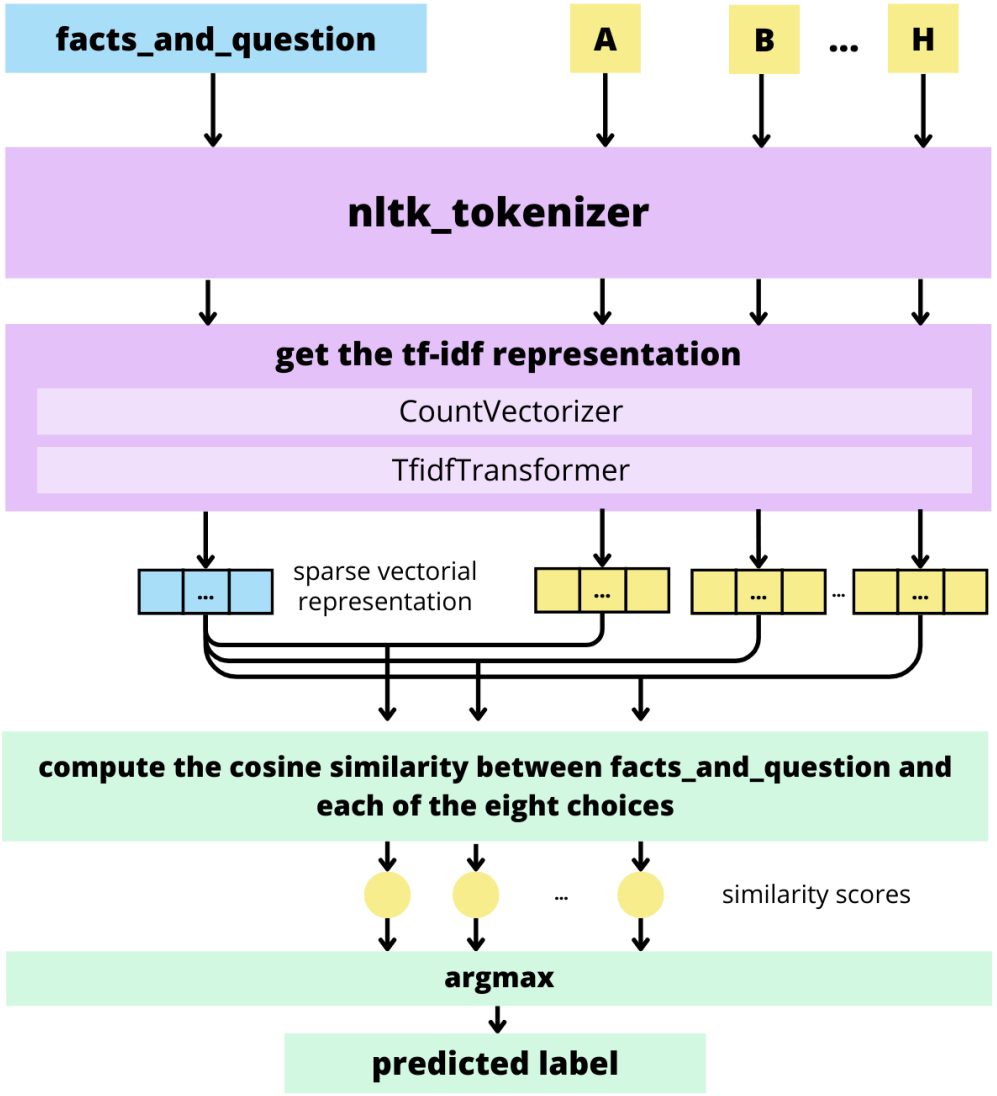


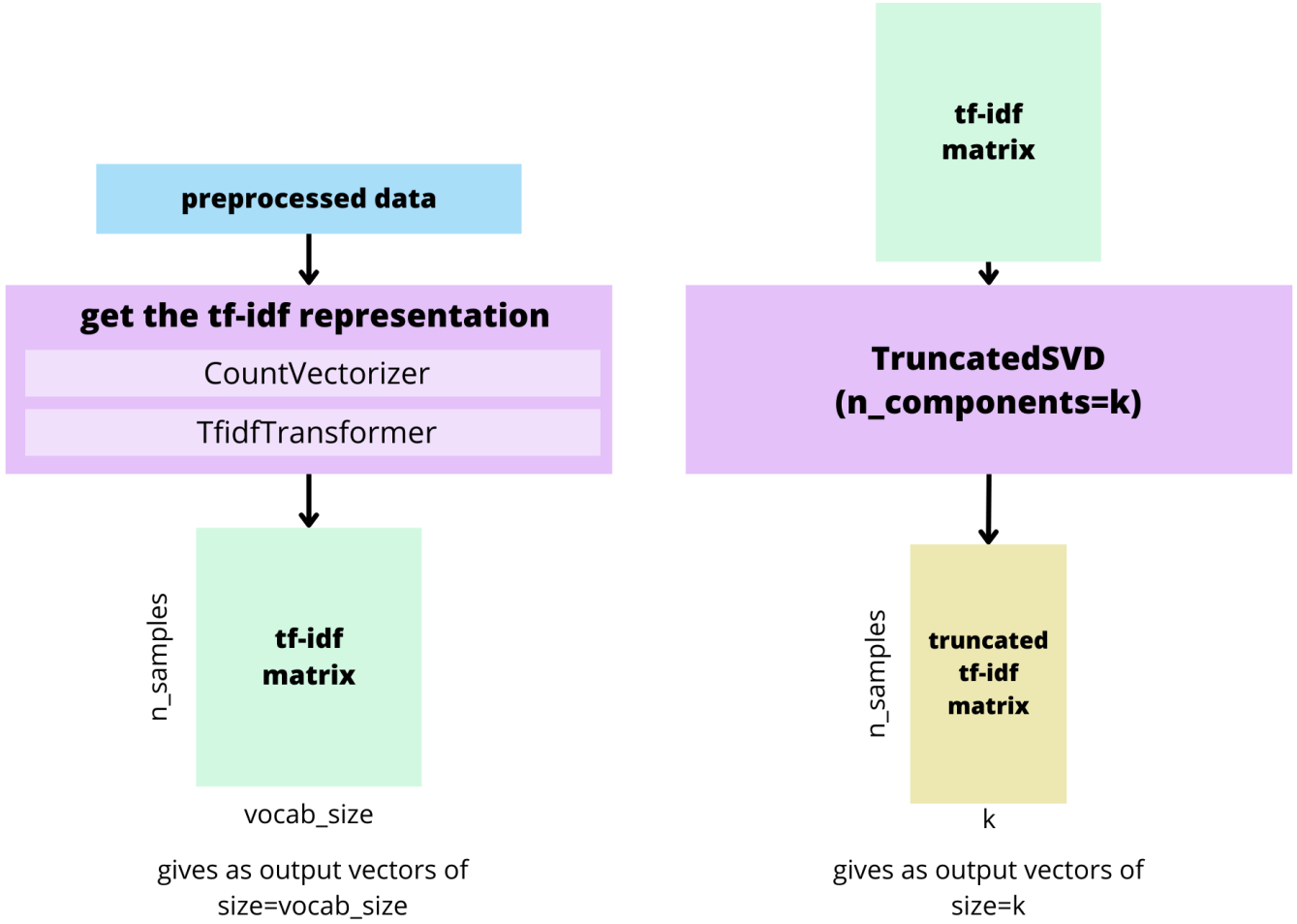
Theme: grade school science

1. Count based methods

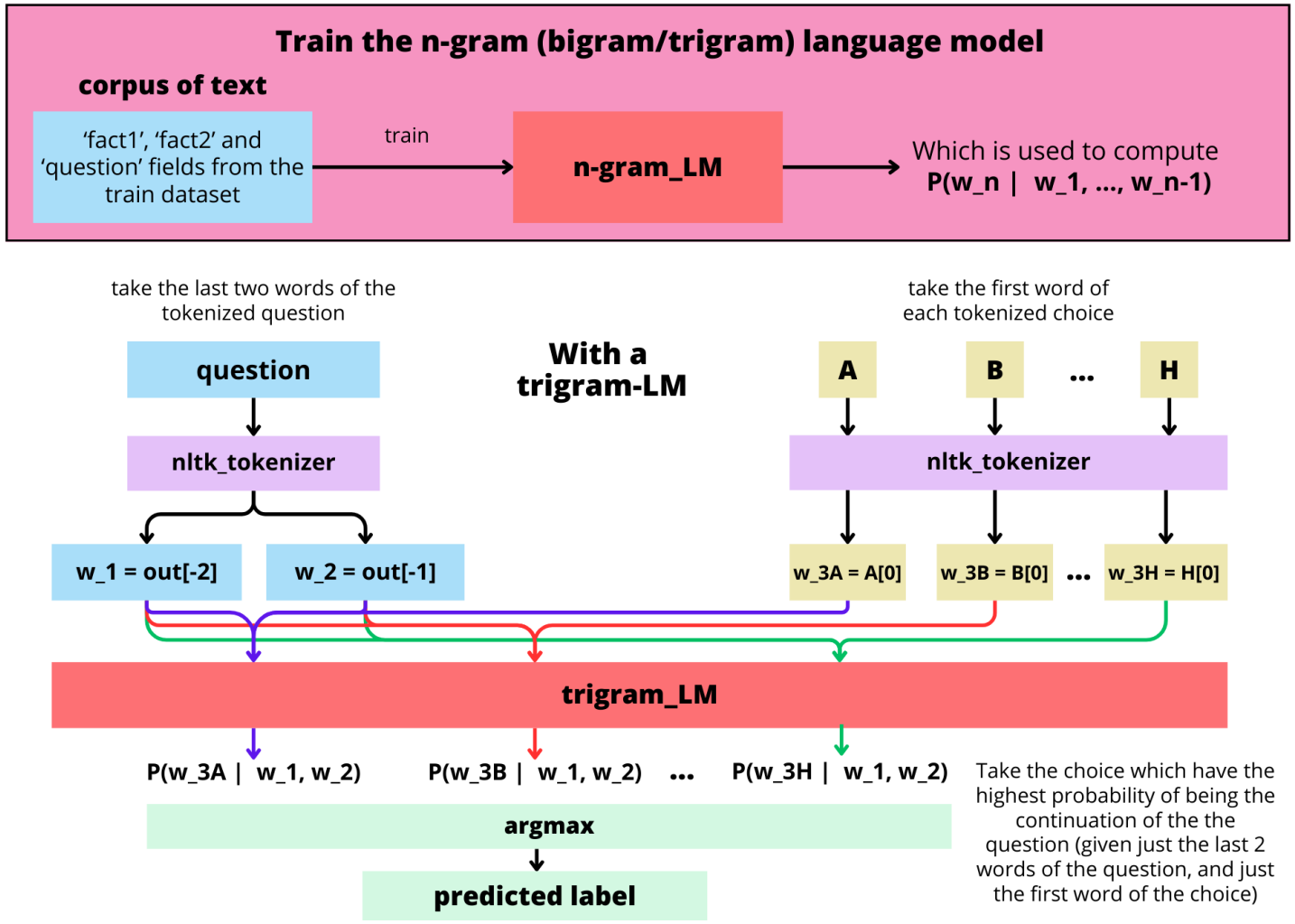
TF-IDF weighting

In the tf-idf representation, which is the choice that is **most similar** to the question?





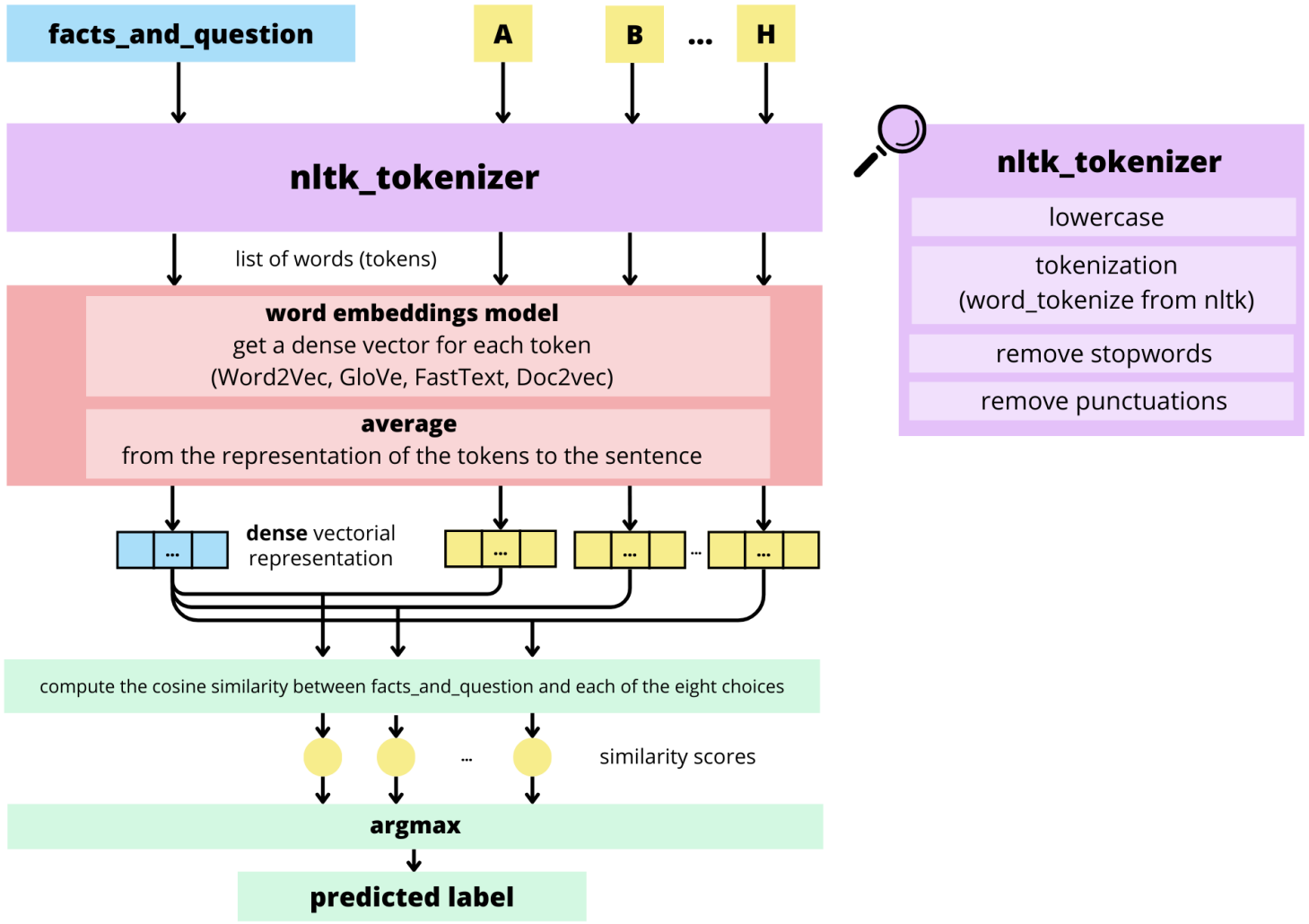
We can truncate the tf-idf matrix with SVD to get a dense representation.



N-gram LM

What is the most likely continuation of the question, between the possible choices?

2. Word embeddings



Static word embedding

We get the representation of a sentence from the **average of the embeddings** of the words in the sentence.

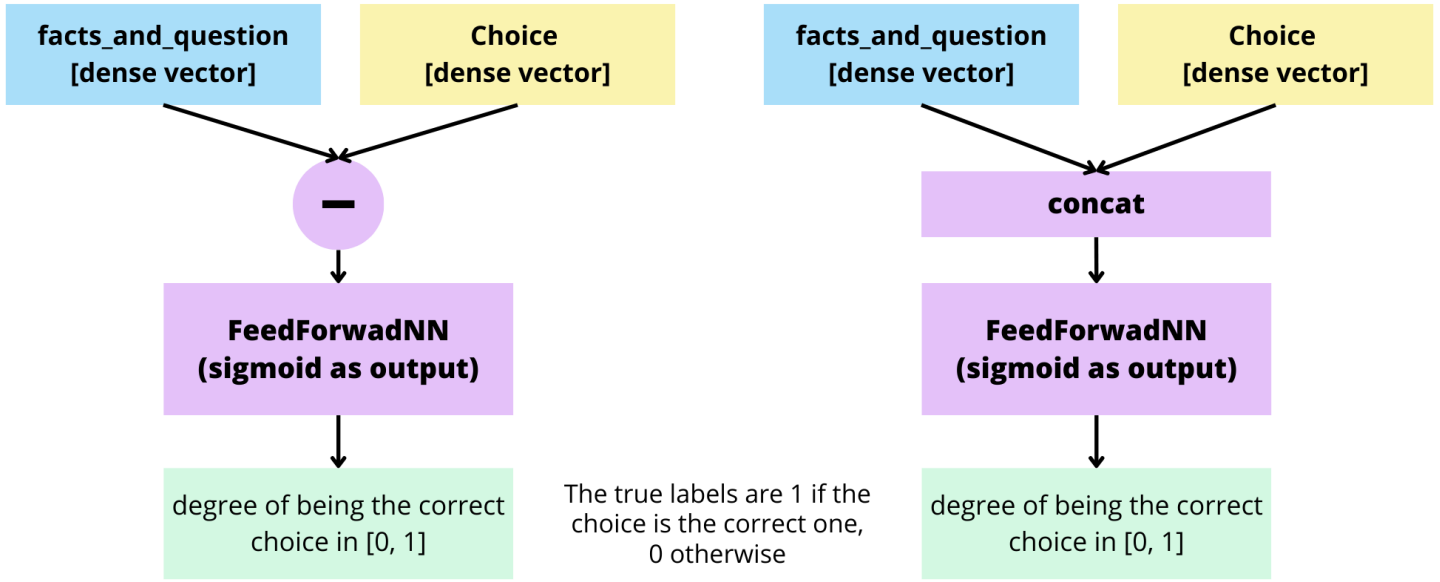
Variants:

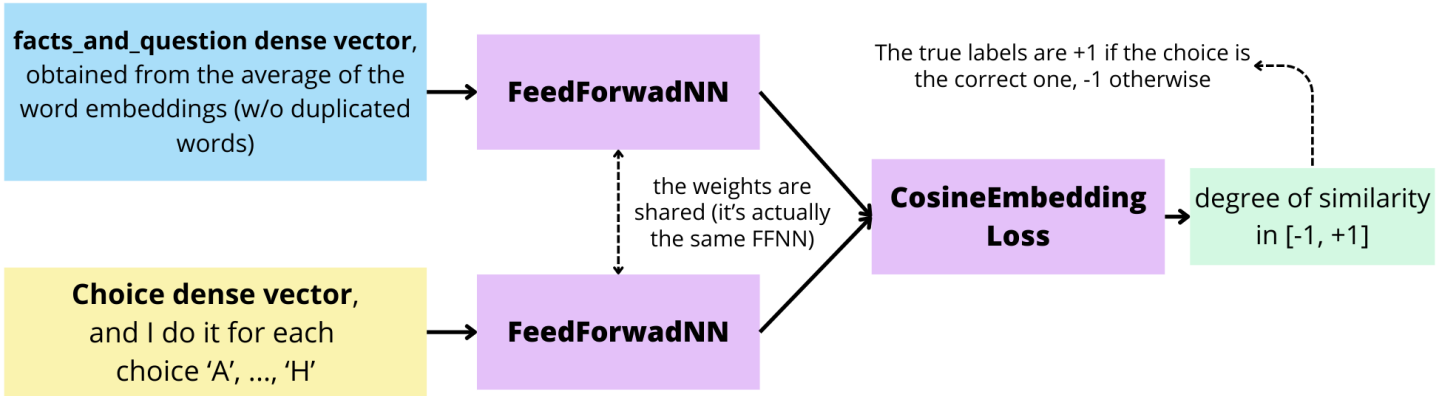
- Take the average after removing the duplicate words.
- Take a weighted average, using the IDF score as the weight for each term.

Alternatives to cosine similarity

I compare the question and the choices with a **neural network**, which is trained to make the question and the correct choice similar.

- **FeedForwardNN** (difference or concatenation)
- **SiameseNN**





3. BERT

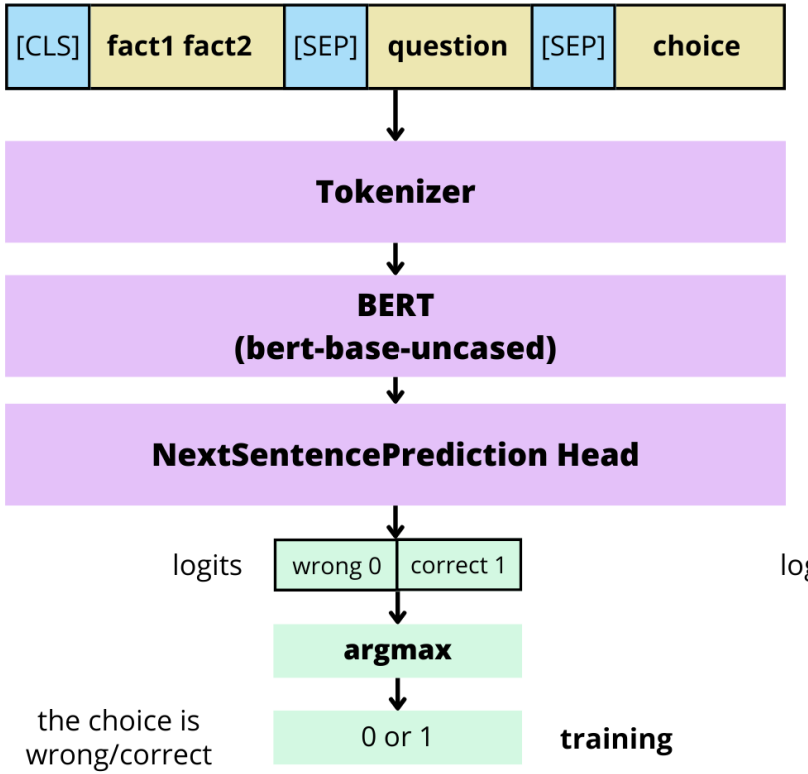
Binary classification

BERT trained for **NextSentencePrediction**.

For each item I have the following 8 inputs:

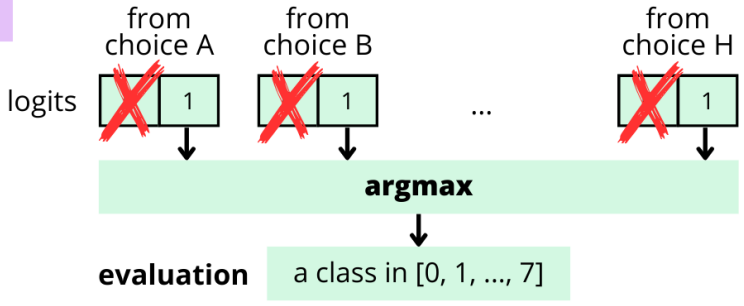
"[CLS] fact_1 fact_2 [SEP] question [SEP] choice_i [SEP]" for each choice i in ['A','B',…,'H’]

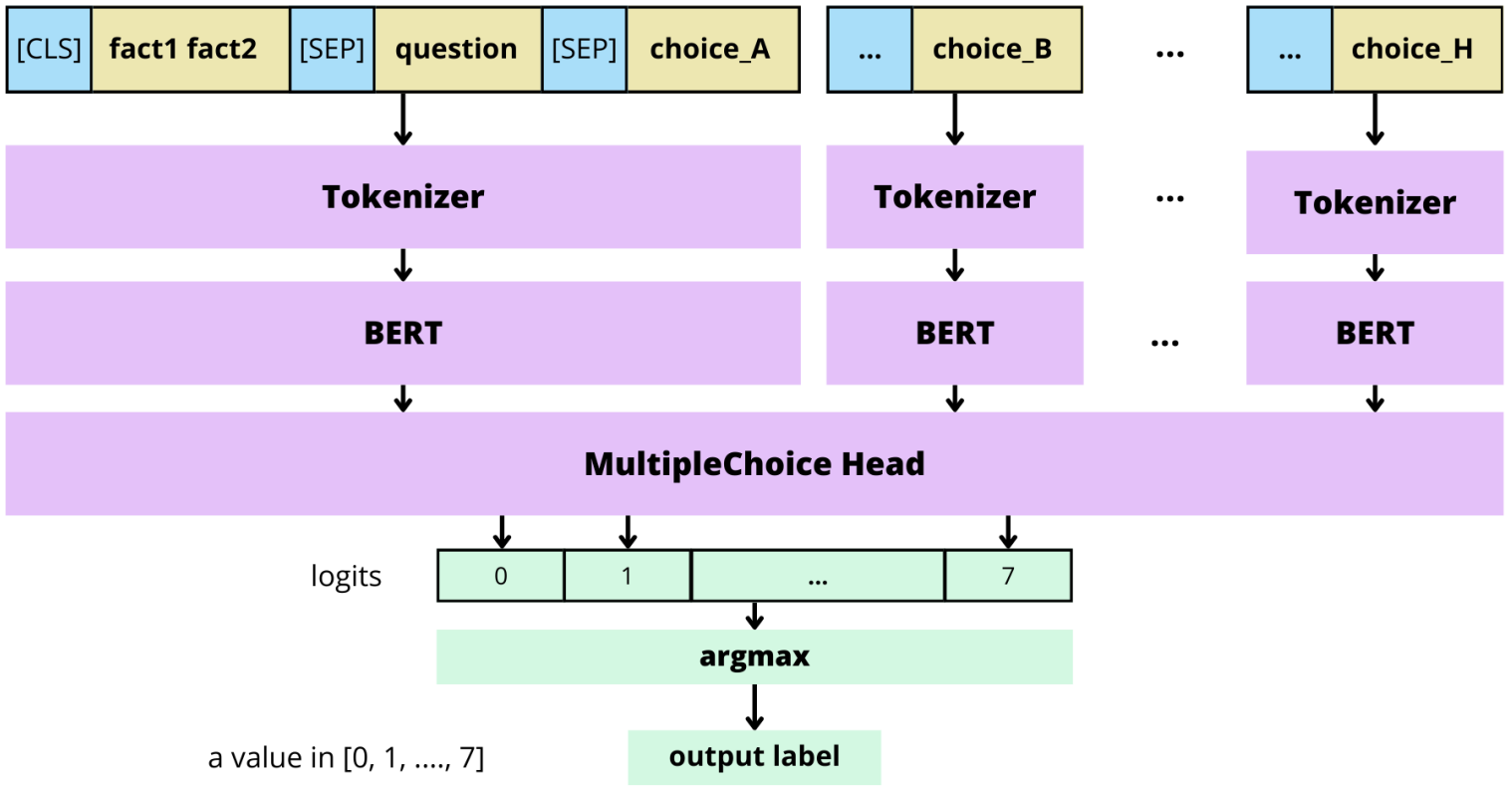
BERT: *'bert-base-uncased'*



TRAINING:
I ask the model to recognize if the choice is correct or not, for each choice. I use the label 1 if it is correct, 0 otherwise.

EVALUTAION:
I don’t want to see if a answer is correct, but I simply want the **most correct** answer between the 8 possible choices. So I feed to the model the 8 choices and take as output label the choice which gives the highest *logits[1],* the choice which belongs more to the correct class





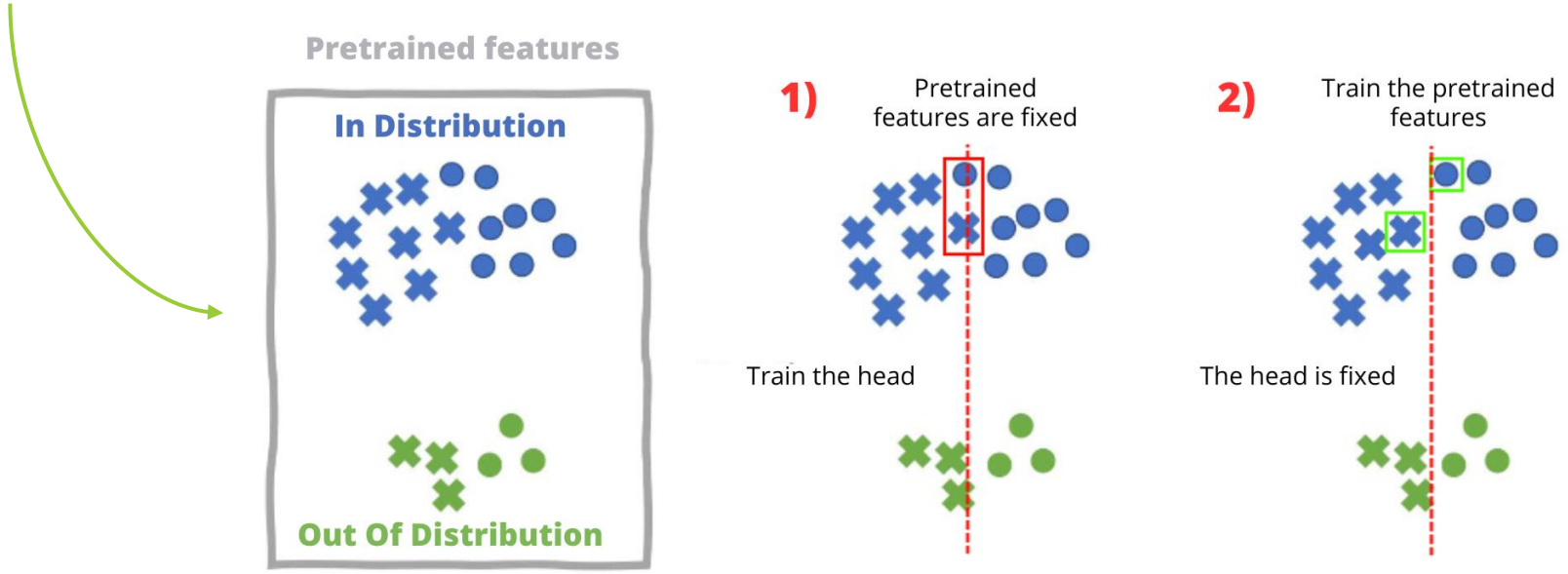
Multiclass classification

BERT trained for **single-label multi-class classification**.

Now I feed all the choices together to the model and I get one output logit for each of them.

Training methods

- **fine-tuning**
- **linear probing**
- **combined method**: after linear probing I train the pretrained features while the classifier is fixed



4. LLM Prompting

zero-shot prompting

```
f"""
fact1: {item['fact1']}
fact2: {item['fact2']}
Question: {item['question']}
A) {item['choices'][0]}
B) {item['choices'][1]}
[...]
H) {item['choices'][7]}
Choose the correct choice.\
Answer with the corresponding letter only.
"""
```

zero-shot chain of thought prompting

```
f"""
Let's think step by step.
fact1: {item['fact1']}
fact2: {item['fact2']}
Question: {item['question']}
A) {item['choices'][0]}
B) {item['choices'][1]}
[...]
H) {item['choices'][7]}
Give the correct choice and a short motivation.\
start the sentence with the letter of the choice.
"""
```

few-shot prompting

```
f"""
fact1: beads of water are formed by\
water vapor condensing
fact2: Clouds are made of water vapor.
Question: What type of water formation\
is formed by clouds?
A) pearls
B) streams
[...]
H) liquid
Answer: F

fact1: {item['fact1']}
fact2: {item['fact2']}
Question: {item['question']}
A) {item['choices'][0]}
B) {item['choices'][1]}
[...]
H) {item['choices'][7]}
Answer:
"""
```

The first example (context) is given by the first sample in the train dataset, **dataset_train[0]**

RAG inspired few-shot prompting

```
f"""
fact1: {best_example['fact1']}
fact2: {best_example['fact2']}
Question: {best_example['question']}
A) {best_example['choices'][0]}
B) {best_example['choices'][1]}
[...]
H) {best_example['choices'][7]}
Answer: {best_example['answerKey']}

fact1: {item['fact1']}
fact2: {item['fact2']}
Question: {item['question']}
A) {item['choices'][0]}
B) {item['choices'][1]}
[...]
H) {item['choices'][7]}
Answer:
"""
```

The **best_example** is the sample that has the highest cosine similarity, computed on the TF-IDF representation of the question, with the question of the current item.

I **just ask** to a Large Language Model which is the correct answer.
In the prompt I give the two facts, the question and the 8 possible choices.

LLM:
'DeciLM-7B-instruct'

5. Evaluation on test set

	Test accuracy (%)	Time elapsed (s) / n_samples
DeciLM – few-shot prompting	**99.03%**	2.23E-01
BERT – multiclass classification – combined method	97.19%	**6.23E-03**

Thanks for your attention!