Análise
Descrições Empresas

Bruno Leme

Teste Técnico Cientista de Dados

- Análise 360º com descrições de empresas.

- Objetivo:
  - Geração de Insights
  - Segmentação de Empresas
  - Geração de Oportunidades
  - Extração de Informação
  - Modelagem de Linguagem

Análise Exploratória de Dados

- Para realização da EDA, seguimos os seguintes passos:

  - Pre-processamento
  - Vetorização de Descrições
  - Análise Global de Ocorrência de Palavras.
  - Análise Semântica Latente
  - Clusterização de Descrições

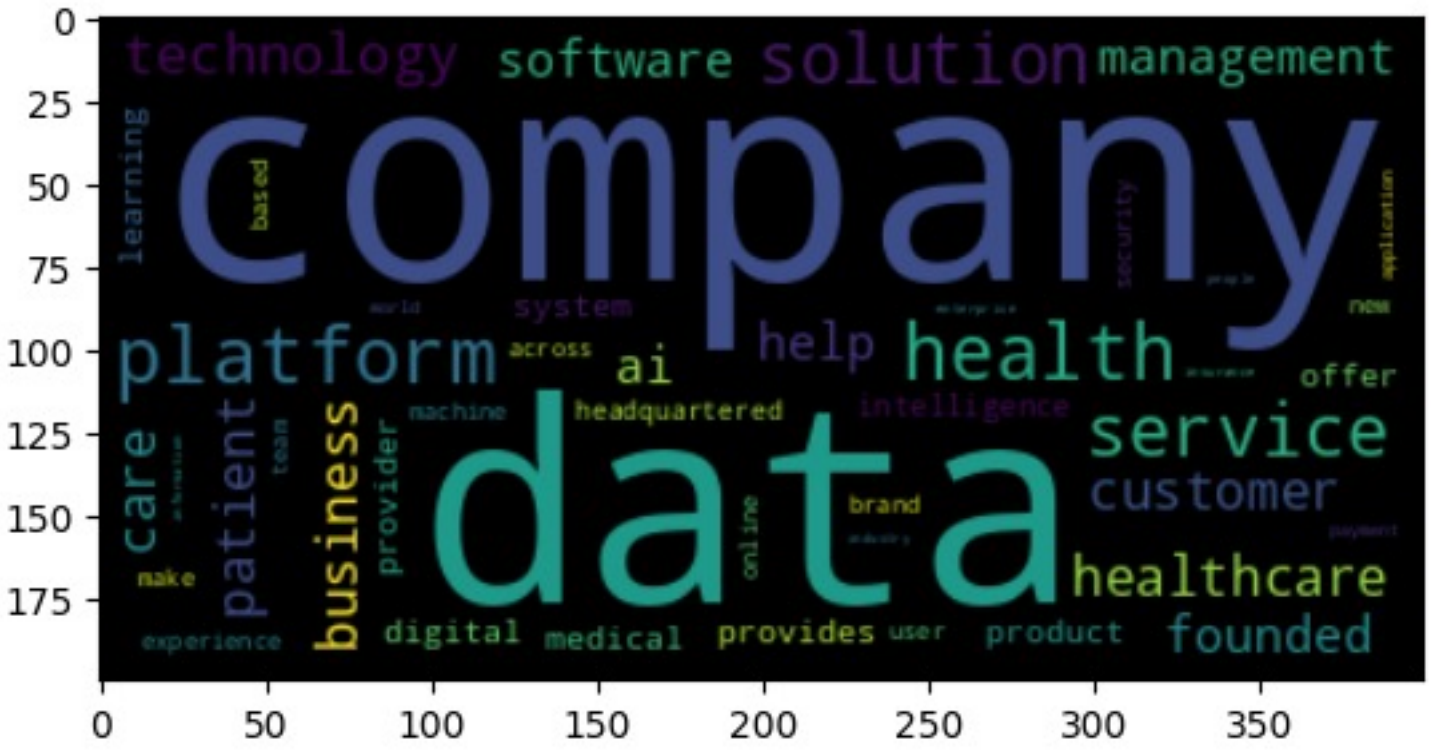
Pre-Processamento

- Limpeza
- Remoção de Caracteres Especiais
- Remoção de Stopwords
- Lematização

Vetorização

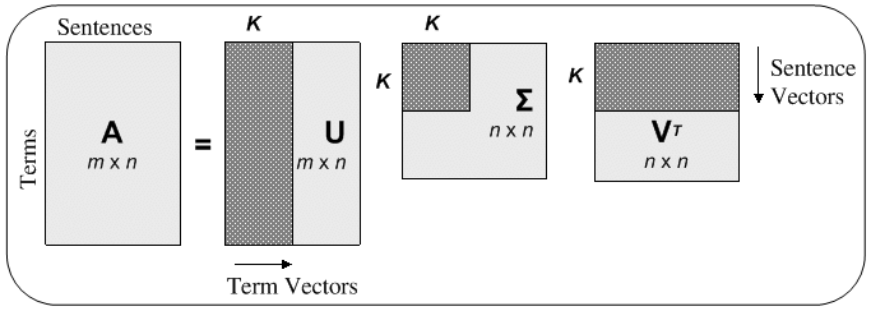
- Normalização Tfidf
- Max Features = 2.000

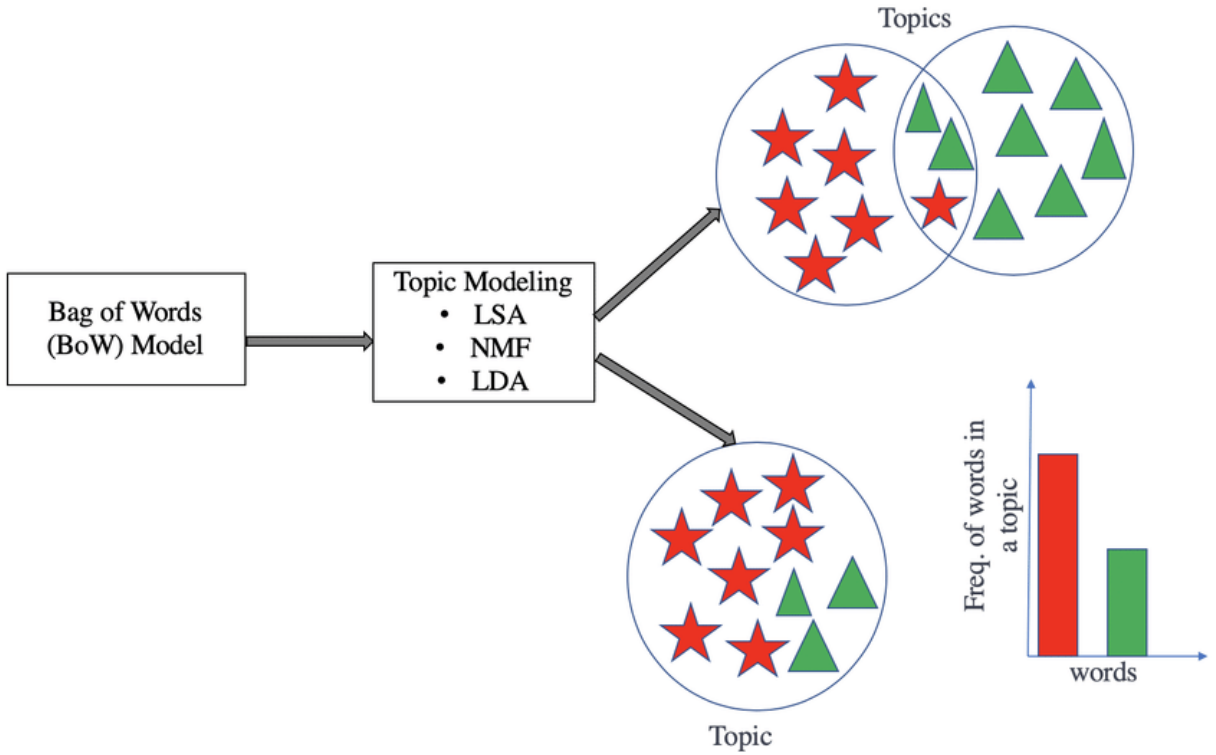
Análise Global de Ocorrência de Palavras



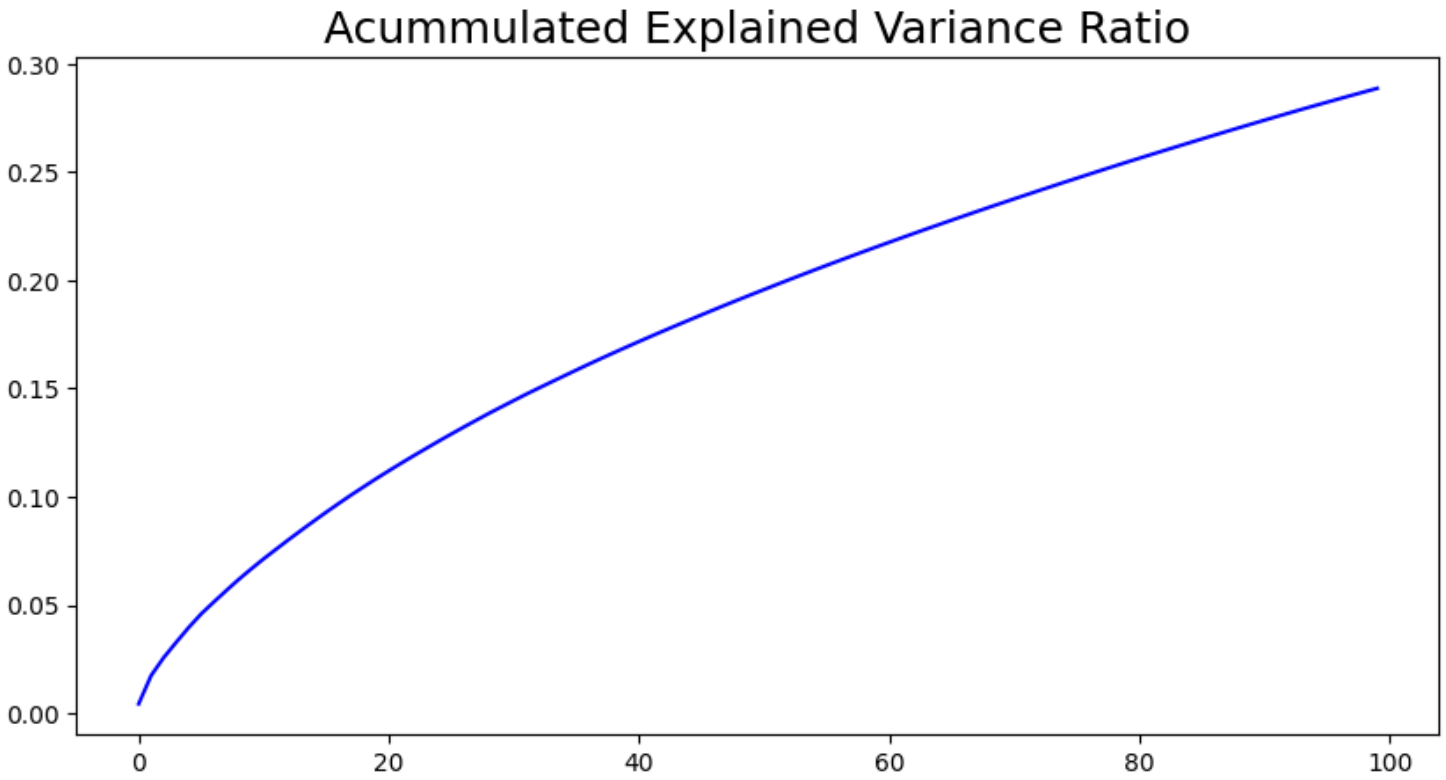
Análise Semântica Latente

- Decomposição de Valores Singulares

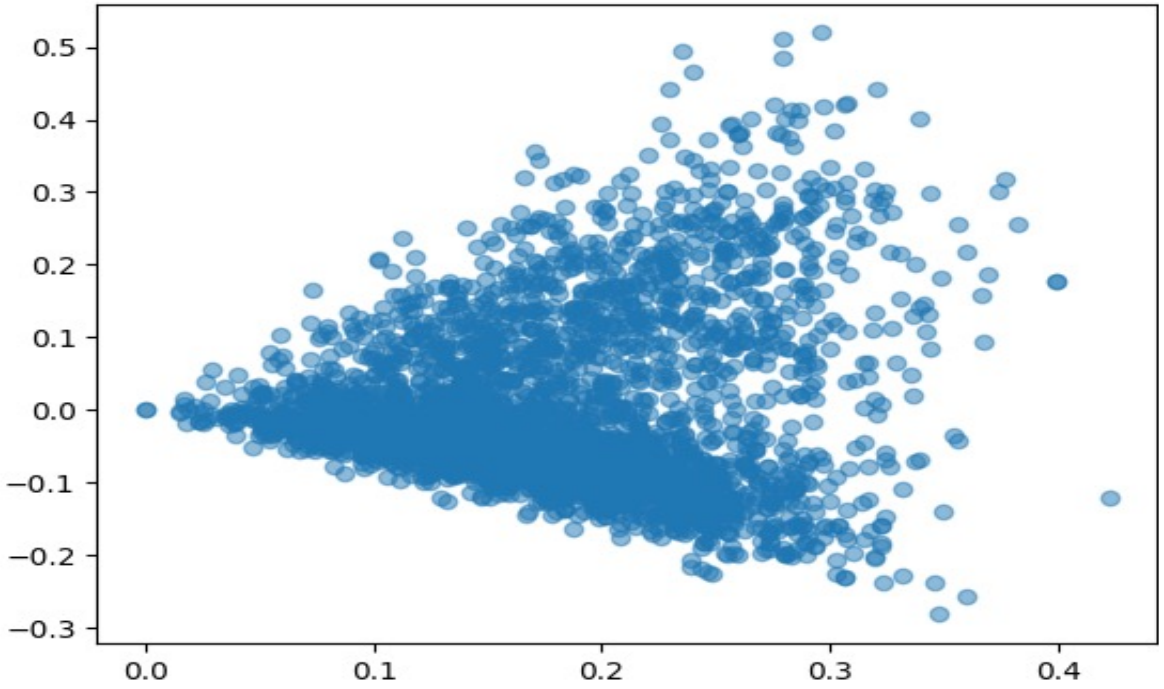




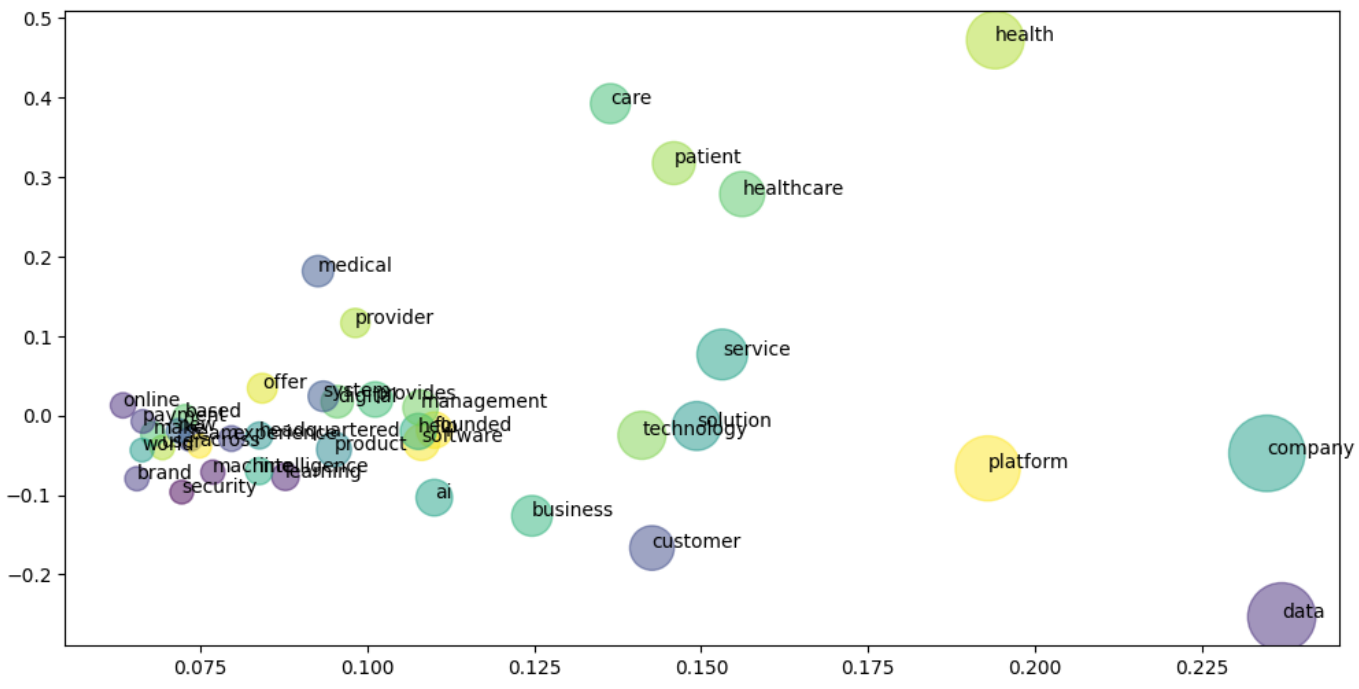
Análise Semântica Latente



Análise Semântica Latente



Descrições dispostas em 2 dimenções



Top 40 tokens dispostos em 2 dimenções

Análise Semântica Latente

`top_tokens_on_dims[0]`

	token	weight
0	data	0.236927
1	company	0.234708
2	health	0.194038
3	platform	0.192911
4	healthcare	0.156139

`top_tokens_on_dims[1]`

	token	weight
0	health	0.472303
1	care	0.392219
2	patient	0.317365
3	healthcare	0.278442
4	medical	0.181515

`top_tokens_on_dims[2]`

	token	weight
0	data	0.599208
1	learning	0.172951
2	ai	0.167069
3	machine	0.158412
4	patient	0.136679

`top_tokens_on_dims[3]`

	token	weight
0	healthcare	0.338967
1	service	0.248632
2	payment	0.244145
3	solution	0.211205
4	management	0.185214

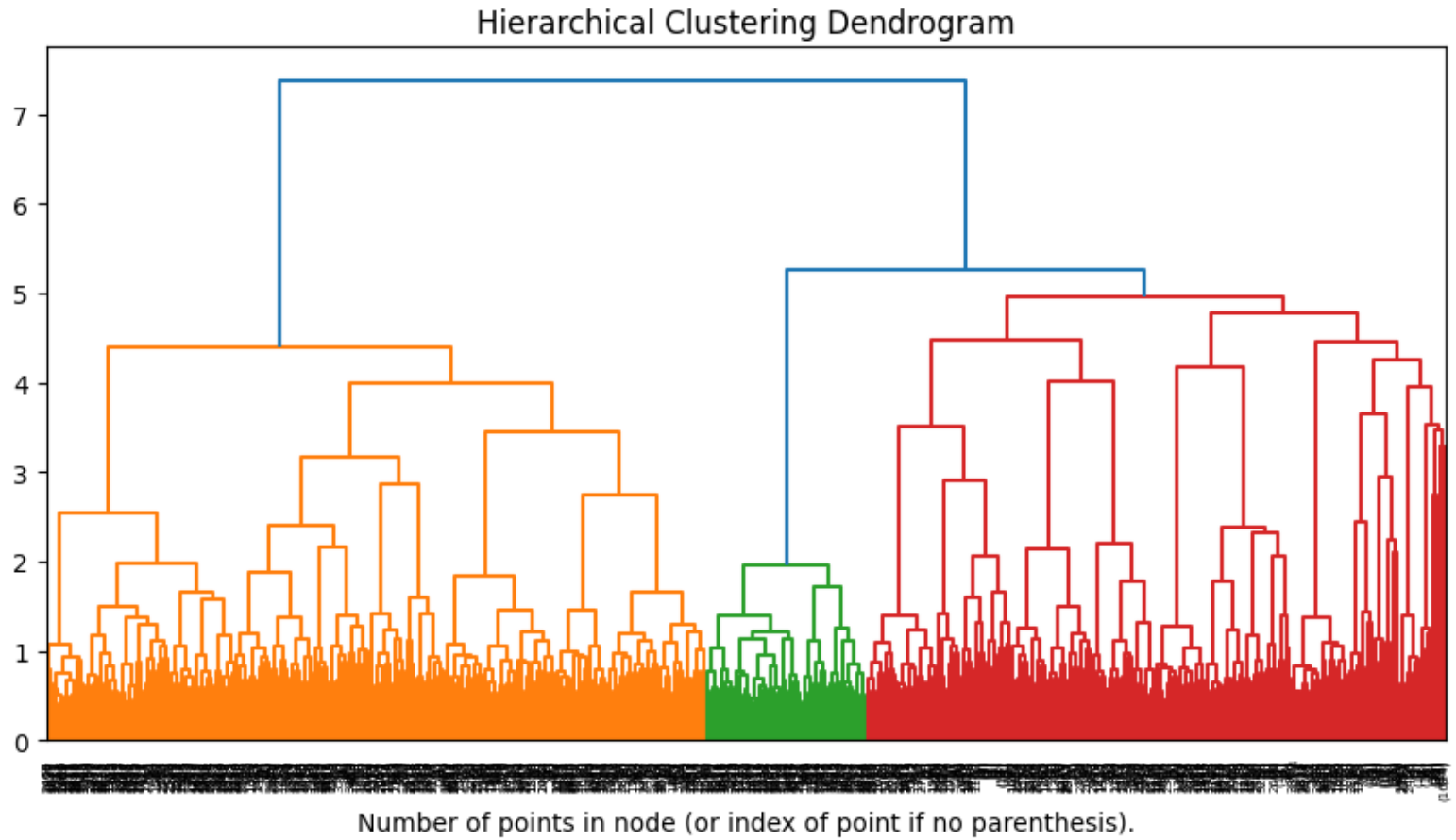
`top_tokens_on_dims[4]`

	token	weight
0	data	0.404468
1	health	0.248170
2	security	0.172029
3	care	0.147860
4	customer	0.139286

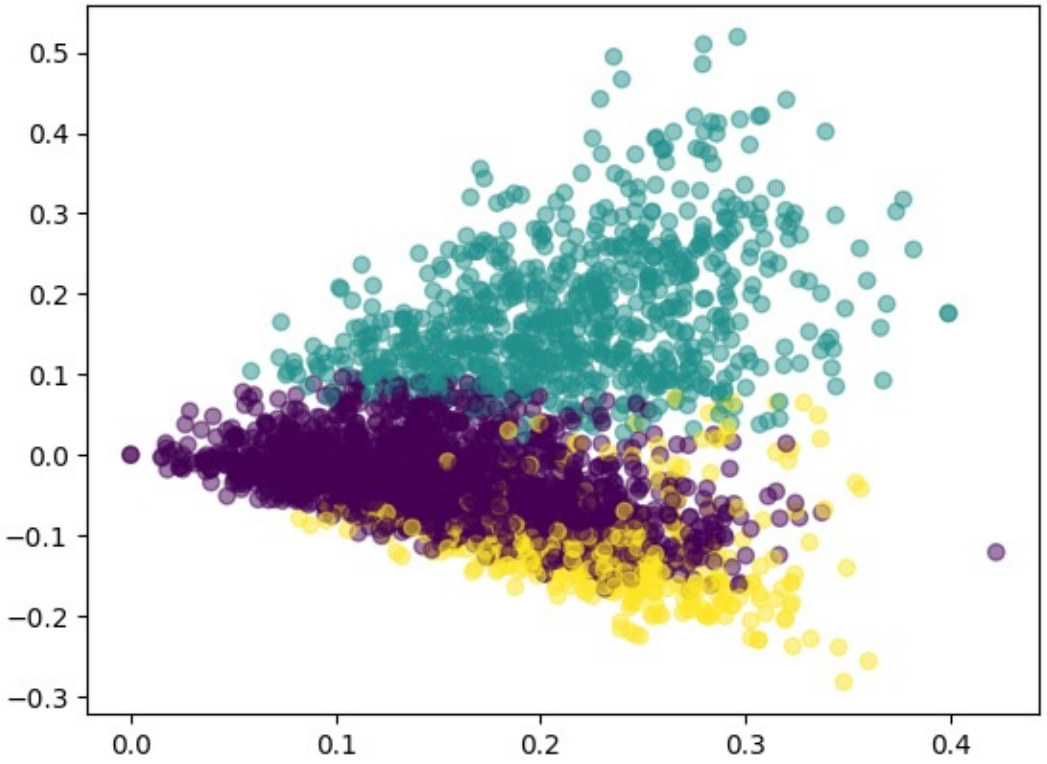
`top_tokens_on_dims[5]`

	token	weight
0	patient	0.457729
1	healthcare	0.237910
2	customer	0.176873
3	experience	0.148743
4	brand	0.131781

Análise de Cluster

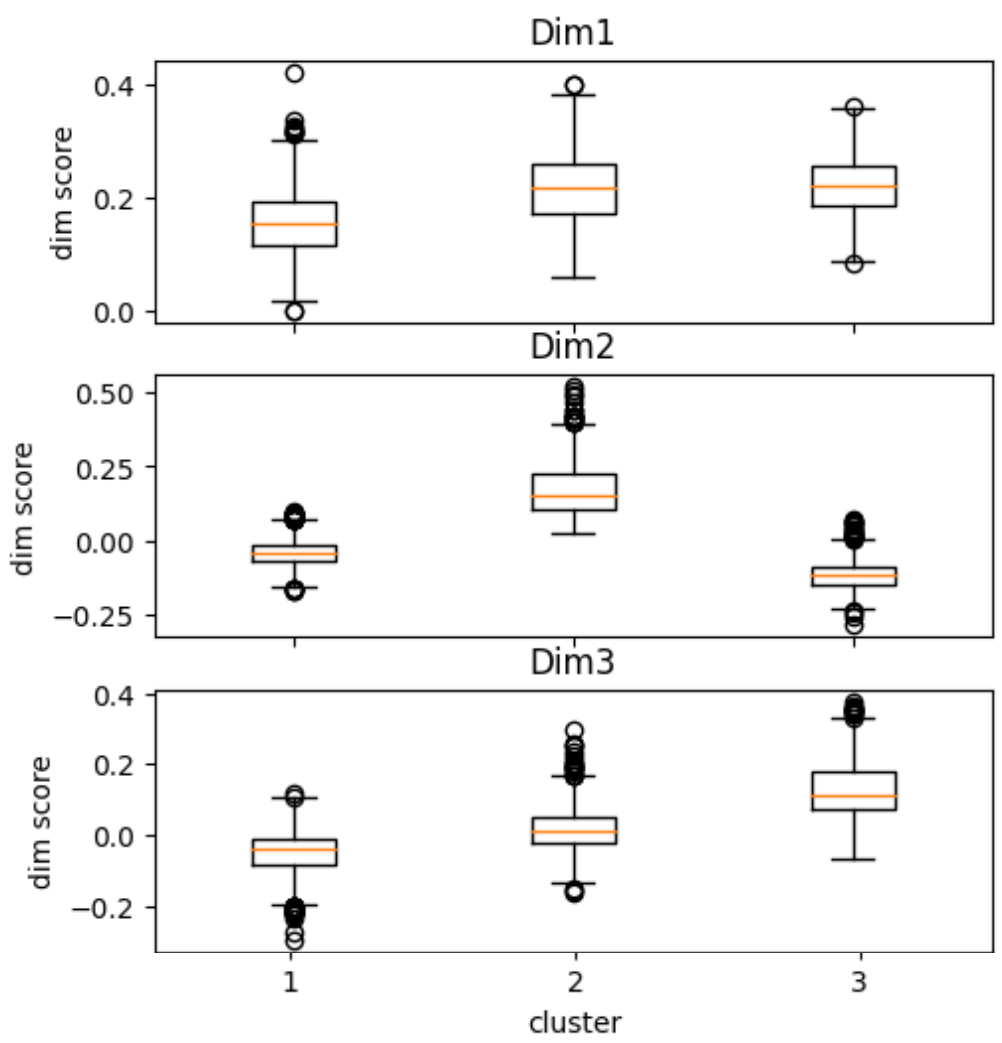


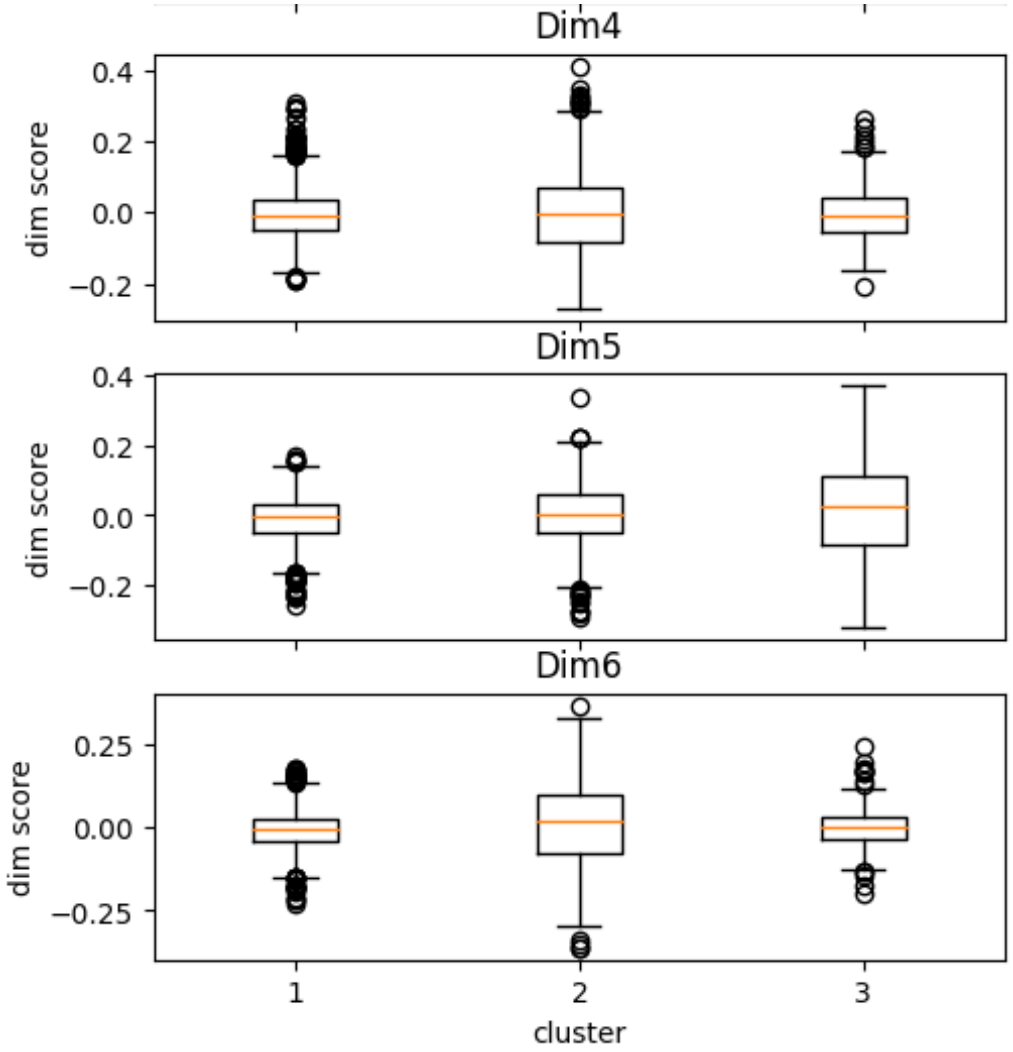
Análise de Cluster



cluster	
0	2275
1	764
2	514

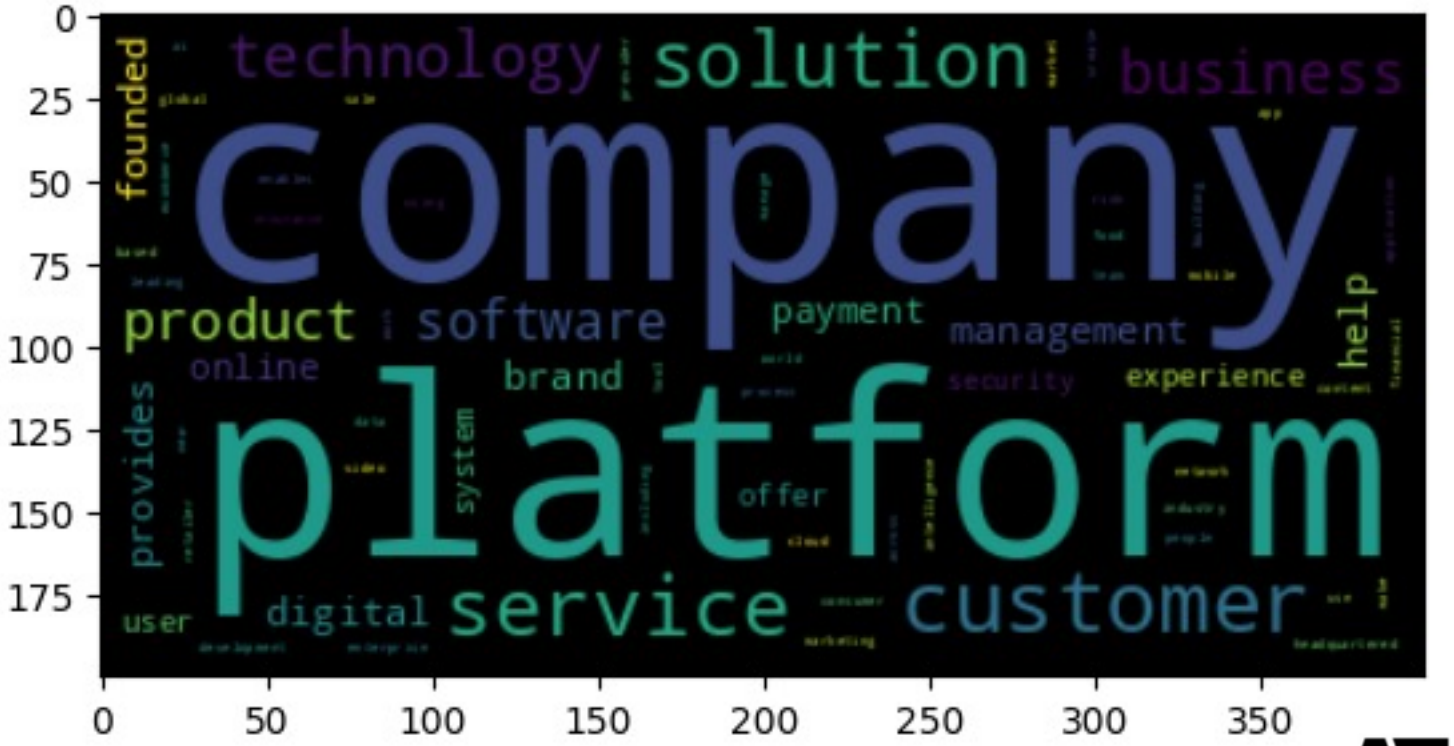
Análise de Cluster





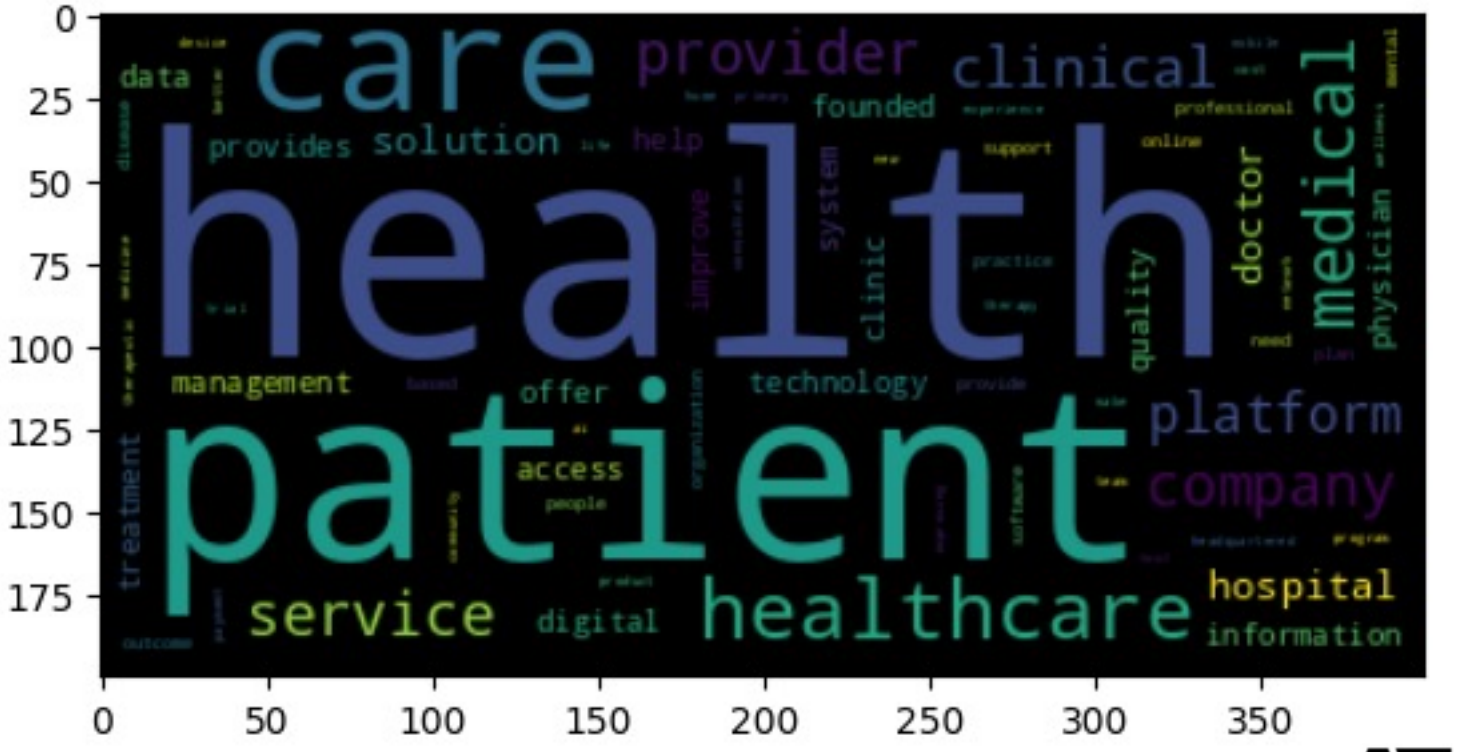
Análise de Cluster

- Cluster 0



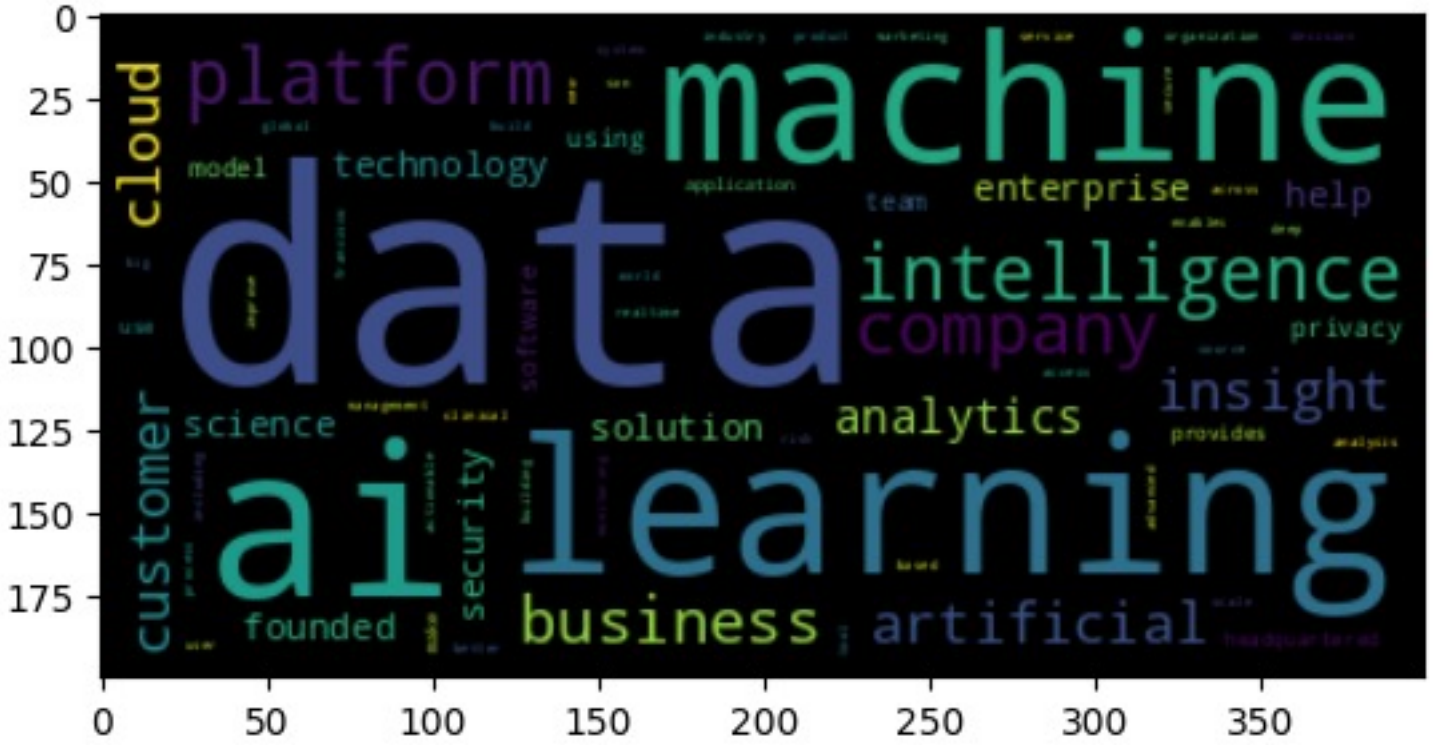
Análise de Cluster

- Cluster 1



Análise de Cluster

- Cluster 2



Natural Language Understanding and Language Models

- Suposição:
  - Problema: o cadastro manual das empresas gera um esforço operacional de tamanho significativo.
  - Solução 1: treinar modelos de extração de informação para identificar e extrair a empresa (NER), bem como extrair o que é a empresa (Answering Extraction) no conjunto de descrições das empresas.
  - Solução 2: treinar modelos de linguagem para disponibilizar um autocomplete de escrita.

Data Labeling

```
def sample_id_generator(data_sample_ids):
```

```
def insert_new_labels(sample_id, skip_ner = False, skip_answer = False):
```

```
sample_id = next(data_sample_ids)
print(sample_id)
encoded_input = tokenizer(dict_data[sample_id]['desc'])
dict_data[sample_id]['desc']
```

```
2094
'Amelia renders banking, HR, banking, insurance, healthcare, telecommunication and IT services.'
```

```
" ".join([tokenizer.decode(token) + get_super(str(i)) for i, token in enumerate(encoded_input["input_ids"])])
```

```
'[CLS]⁰ Amelia¹ render² ##s³ banking⁴ ,⁵ H⁶ ##R⁷ ,⁸ banking⁹ ,¹⁰ insurance¹¹ ,¹² healthcare¹³ ,¹⁴ te¹⁵ ##le¹⁶ ##
```

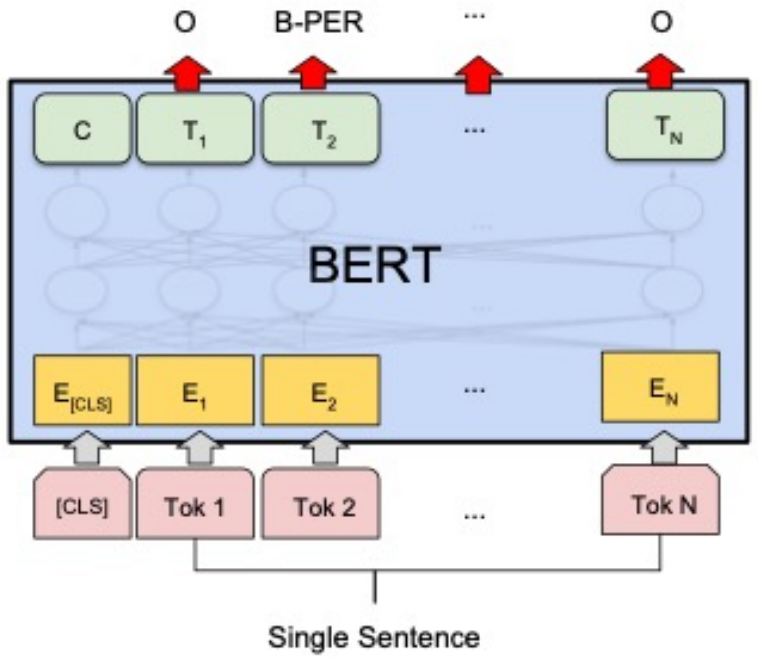
```
start_ner_company = 1
end_ner_company = 1 + 1

#Q1
start_answer = 1
end_answer = 22 + 1

print(tokenizer.decode(encoded_input["input_ids"][start_ner_company:end_ner_company]))
print(tokenizer.decode(encoded_input["input_ids"][start_answer:end_answer]))
```

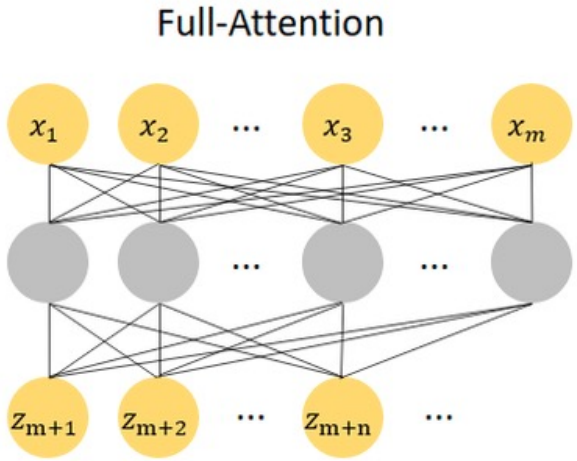
```
Amelia
Amelia renders banking, HR, banking, insurance, healthcare, telecommunication and IT services.
```

```
insert_new_labels(sample_id)
```

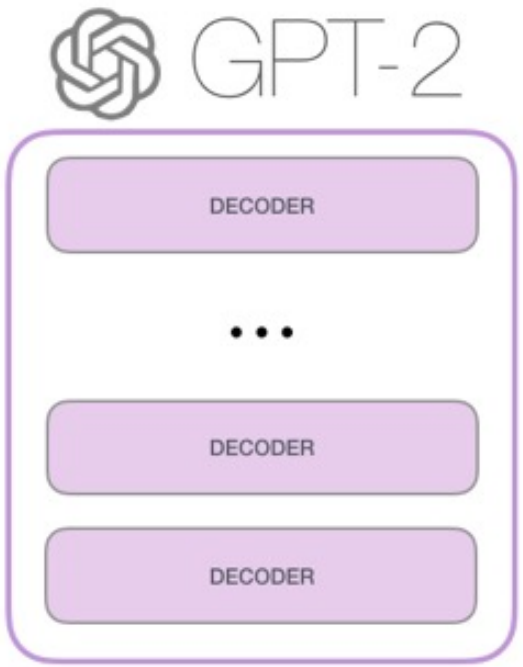
```
NER: {'start': 1, 'end': 2}
Answer: {'start': 1, 'end': 23}
Total Labeling Count: 7
```

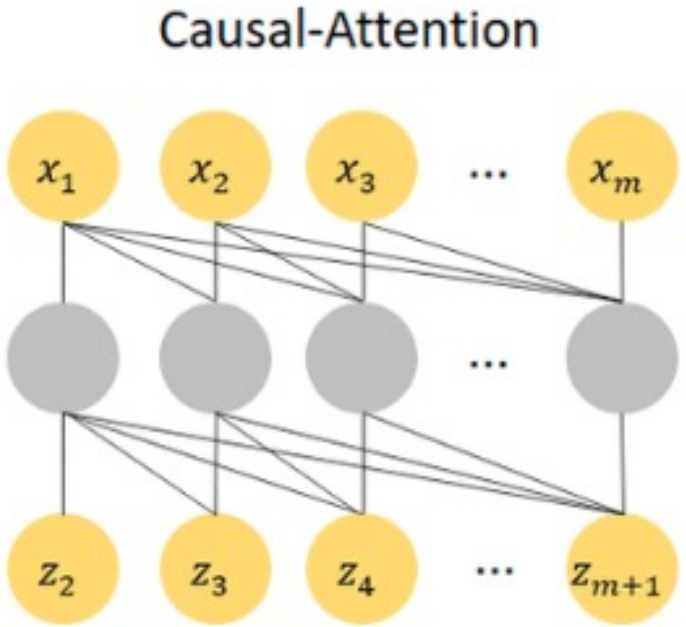
NER and Answering Extraction (unfinished)





Language Model (unfinished)





Conclusões

- NLP é muito útil para análise e automatização de problemas de texto.
- Observamos associações relevantes entre palavras, bem como tópicos que permitem uma análise semântica das descrições
- Através da Análise de Cluster, segmentamos as descrições em 3 grupos: (1) tecnologia, (2) saúde e medicina e (3) AI, machine learning e analytics.
- Observamos a possibilidade de realizar tarefas adicionais como:
  - NER, Answering Extraction, Language Models, entre outras.