

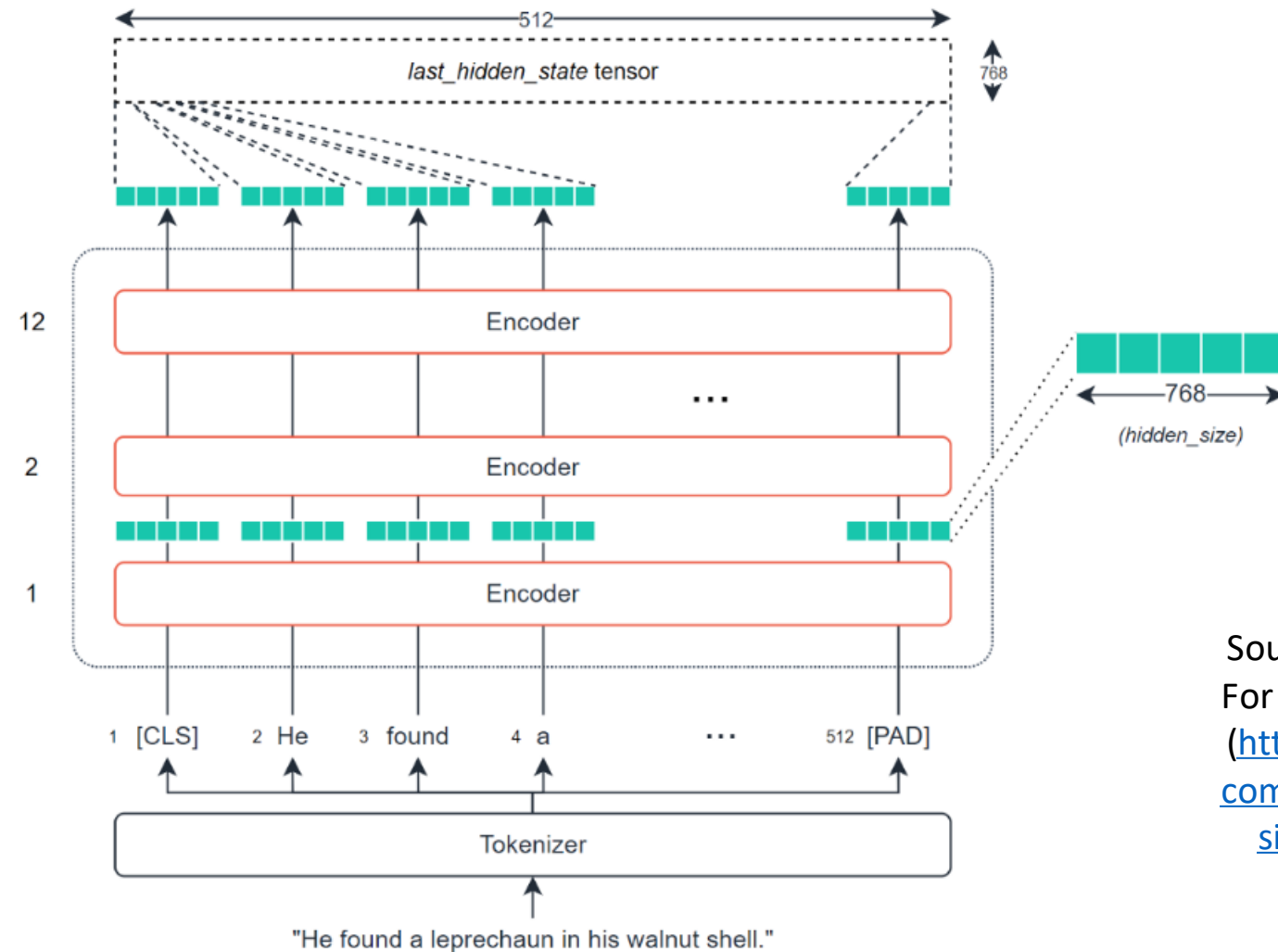


# A New Try: Word Embeddings Based Sentences

# Dataset (Randomized)

index	book_id	review_id	review_rating	sentence
27115	8538058975	R2UDT8IWTB2EFX	5	Muito bom!
133424	8535933395	R1OYCHBFXIGB7	5	Jefferson Tenório, sem dúvidas, é um dos autores mais talentosos que o Brasil.
95619	8578601777	R14B0GDKSSXJKS	5	Demorou um pouco mas chegou perfeito
77519	6555650001	RHIGI19SRUQ4P	4	Ela é simples e direta sem perder a candura.
9265	8550802751	R20SP07M4B2MX4	3	O conteúdo é fenomenal, vindo de JP, não há o que esperar diferente disso!
136839	8565765482	R3FLOAWVI9TSLE	3	Mas deixou muitos buracos na história e fiquei além de triste angustiada por não ter as minhas respostas que quebrei a cabeça o livro todo pra conseguir.
19550	8584391509	R4LQDCZZPZDPW	5	Vou comprar mais livros dessa autora!
109289	8543107202	R22XXN8QVGDUZK	5	Super indico.
169899	8565765695	R3KKLO8FMKDMEI	5	Bem escrito, nos leva passo a passo a evoluir e descobrir a trama junto com a personagem principal.
154399	8545202210	RYPCNV586YPJ9	4	Quem gosta mais da teoria com exemplos práticos sem essa de erga a cabeça!
175988	8544106595	R1Z7V0Q1KA3J4C	5	Sobre a entrega, veio bem embalado, atrasou um pouco, mas tudo bem!

# BERT



BERT base network — with the hidden layer representations highlighted in green.

Source: Extracted from BERT  
For Measuring Text Similarity  
(<https://towardsdatascience.com/bert-for-measuring-text-similarity-eec91c6bf9e1>)

# BERTimbau

- Pre-trained model in portuguese
- Hugging Face
  - <https://huggingface.co/neuralmind/bert-base-portuguese-cased>
- `model = SentenceTransformer('neuralmind/bert-base-portuguese-cased')`

Model	Arch.	#Layers	#Params
neuralmind/bert-base-portuguese-cased	BERT-Base	12	110M
neuralmind/bert-large-portuguese-cased	BERT-Large	24	335M

Source: <https://huggingface.co/neuralmind/bert-base-portuguese-cased>

# SentenceTransformers

## SentenceTransformers Documentation

- <https://www.sbert.net/>

## *mean pooling operation*

- “Each of those 512 tokens has a respective 768 values. This pooling operation will take the mean of all token embeddings and compress them into a single 768 vector space — creating a ‘sentence vector’.” (BERT For Measuring Text Similarity - <https://towardsdatascience.com/bert-for-measuring-text-similarity-eec91c6bf9e1>)

# `sentences_embeddings[0]` (1000 samples)

```
array([  
-3.14204991e-02, -1.72654122e-01, 2.04657361e-01, 2.58500636e-  
01, 5.44240534e-01, 2.48894483e-01, 1.60047114e-01, 6.07270189e-  
02, 4.33031619e-01, -1.59762889e-01, -1.95592731e-01, 5.41238904e-  
01, 3.02740008e-01, -1.30686387e-01, 3.67669351e-02, 2.77642608e-02, ...
```

```
# sentences_embeddings.shape: (1000, 768)
```

# Cosine Similarity

```
cosine_similarity(  
    [sentences_embeddings[0]], sentences_embeddings[1:]  
)
```

# First 20 sentences

```
array([0.6210844, 0.48339698, 0.6691068, 0.5216883, 0.57571626,  
0.6646237, 0.5859783, 0.65288246, 0.41145706, 0.54317814, 0.5807854,  
0.5494274, 0.4787109, 0.54764295, 0.4576561, 0.3734125, 0.5017296,  
0.48248816, 0.5300585, 0.52057725], dtype=float32)
```

# Cosine Similarity - Examples

```
print( f'0: {sentences[0]}')  
print( f'1: {sentences[1]}')  
print( f'3: {sentences[3]}')  
print( f'6: {sentences[6]}')  
print( f'8: {sentences[8]}')
```

0: Eu sei que muito você já ouviu falar desse livro, e confia...

1: Esse sem dúvida é um dos melhores livros que já li.

3: esse primeiro volume de heartstopper é aquele tipo de livro (nesse caso quadrinho) que você lê de novo e de novo e de novo e nunca fica enjoativo.

6: Adorei o livro, li tão vorazmente que vou ter que reler.

8: É um livro q da para ler em um dia, uma coisa muito legal também é o chat entre as duas.





# Results

(clusters\_total = 5)

Total of Sample	1000	5000	10000	30000	50000
Sentences Embedding	4.19 s	8.81 s	13.2 s	33.2 s	49.8 s
K-Means Fit	600 ms	4.83 s	10.9 s	18 s	23.6 s
TSNE Fit	13.4 s	59.3 s	1min 50s	6min 42s	14min 16s
TSNE Graph Generation	101 ms	65.4 ms	90.1 ms	129 ms	186 ms

The notebooks were executed on Google Colaboratory.



# Examples of Labelled Sentences

(**1000** samples)  
(**clusters\_total = 5**)

# Labelled Sentences - Class 0

- **Recomendo demais.**
- **Recomendo ler**
- Transforma a vida!
- Não faz milagre, mas ajuda #
- Show demais
- Atenção!!
- Será muito bem guardado por aqui
- Simplesmente perfeito.
- Ótimo presente também.
- **super recomendo!**

# Labelled Sentences - Class 1

- Que livro maravilhoso.
- Vale a pena **ler!**
- Maravilhosa a **leitura!**
- E bem leve de **ler !**
- Mais que viciante mesmo depois de anos de **leitura!**
- Impossível não amar essa história.
- Livro muito interessante com uma sequencia de estudos de casos.
- **É rápido de ler**, a história é daquelas que prende.
- Enfim, **livro de rápida leitura** e excelente.
- **Ótima leitura!**

# Labelled Sentences - Class 2

- .
- 😊
- !!
- ♥
- 💭
- Gostando
- ☐
- Muito.

# Labelled Sentences - Class 3

- **Resumindo: é aquele livro que você quer marcar todinho (principalmente para reler) e ficar abraçada.**
- **O Pequeno Manual Antirracista já é um dos principais monumentos do ativismo negro no Brasil.**
- **Anthony é um bruto fofo e, realmente, é um Libertino com L maiúsculo.**
- esse livro é tudinho pra mim, tudo é mostrado de uma forma super leve amo tanto as minhas meninas
- **Livro maravilhoso**, como em todos os livros de Julia Quin, não ha uma donzela perfeita e um mocinho heróico, mas sim personagens que com seus defeitos formam uma trama envolvente, com desenvolvimento pessoal e conjugal, uma leitura que literalmente te aconchega, lhe rendendo muitas risadas.
- Pessoas que me recomendaram realmente estão a caminho da promessa de capa, o que me motivou a lê-lo.
- "Esse livro me fez ver a fotografia de forma muito diferente, antes eu via apenas o meu "reflexo" físico nela.
- Se você tem alguém próximo, e sabe que algum dia terá de cuidar desta pessoa, esse livro é INDISPENSÁVEL!
- esse é um problema que ainda persiste nos dias de hoje.
- **Livro muito esclarecedor, vale muito a pena para quem quer mudar de vida e entender os segredos das empresas e de famosos**

# Labelled Sentences - Class 4

- **chegou rápido e bem antes do previsto.**
- **O livro foi entregue antes do prazo estipulado.**
- **Chegou 16 dias antes do prazo e em ótimas condições.**
- O livro veio em ótimas condições, e Chegou bem antes do prazo.
- Chegou perfeitamente embalado e rápido!
- **Entrega da Amazon super rápida.**
- Chegou 10 dias antes do prazo estipulado!
- O livro chegou em 2 dias, é vem bem embalado, ótimo estado e ele é perfeito, eu tô apaixonada.
- Veio tudo certinho e antes do previsto.
- **veio embalado corretamente sem nenhum defeito.**





# Examples of Labelled Sentences

(**50000** samples)  
(**clusters\_total = 5**)

# Labelled Sentences - Class 0

- **Veio antes do previsto da data de entrega**
- **Só dei 4 estrelas pois veio com um defeito**, mas como o livro estava lacrado é culpa da editora e não da amazon.
- **Comprei os três primeiros livros e chegaram em ótimas condições.**
- Comprei ao acaso e foi uma surpresa encantadora.
- Fiquei com medo de não gostar por causa do diálogo em aspas mas foi de boao livro é maravilhoso **e a entrega com prime chegou em 1 dia**
- Comprei outro livro junto e veio um pouco amassado.
- Chegou dentro da data prevista e em ótima qualidade.
- Ainda não li, mas chegou tudo ok a entrega
- Chegou tudo certinho!!
- O produto veio em perfeito estado e chegou em uma semana.

# Labelled Sentences - Class 1

- Na minha opinião o pior livro da CoHo, não gostei nada da história, plot mais sem sentido, como alguém volta pra pessoa que foi responsável pelo seu acidente???
- **Complementa a leitura de outros livros de desenvolvimento pessoal mais famosos.**
- **E mesmo que traga justamente muitas respostas, ainda acrescenta muitas coisas para se pensar...**
- **Esse livro é uma oscilação intensa entre as dificuldades, paixão e esperança.**
- Leitura fácil, nos prende a leitura e vontade de aprender mais ainda sobre o tema e colocar em prática.
- confesso que me cativou bastante, tanto que já comprei o box com todos os livros da série.
- As mensagens nos faz refletir sobre a nossa vida mecanizada, sem propósito cheia de crenças que nos deixam ansiosos e infelizes.
- É assim quando alguém se sente atraído por um pessoa.
- **Gostei das questões de neurociencia sobre o cérebro da criança, mas o que mais me fez compreender as questões da mente foram os exemplos práticos que são colocados no livro, uma leitura complexa em parte mas muito bem explicada.**
- Amei é simplesmente lindo, veio com alguns arranhões mas nada que estrague a leitura

# Labelled Sentences - Class 2

- **Devorei a leitura...**
- Mas esse aqui foi diferente.
- **História maravilhosa**
- Um GPS para uma vida épica.
- **Super recomendo.**
- **Lindo, colorido e durável.**
- Obrigada Samuel Salomé pelo reenvio.
- Excelentes ensinamentos.
- O mundo é dos primos!
- **Interessante**

# Labelled Sentences - Class 3

- livro
- :)
- Top demais
- Amor.
- ♥
- ótimo
- !
- Trás conceitos
- 
- tem ensinamentos

# Labelled Sentences - Class 4


- **Livro** para um Bom Aprendizado !!
- Um **livro** para se reler constantemente.
- O **livro** é ótimo, a escrita fluída, capítulos curtos e a história é interessante.
- O **livro** veio em perfeito estado, aliás darkside nunca decepçiona
- **Livro** muito bom, recomendo!
- **Livro** excelente!
- Vale a pena a leitura.
- Que **livro** e que final.
- Ótimo **livro**.
- Ótimo **livro**, entrega rápida



Trying to find better  
number of clusters...

A large orange circle is positioned on the left side of the slide, partially cut off by the edge.

What I  
could see...

- Possible clusters names:
    - Delivery
    - Sentiment
    - Description
    - Short Sentences
  - Silhouette Coefficient score
    - <https://scikit-learn.org/stable/modules/clustering.html#silhouette-coefficient>
- 
- A yellow dashed line is located in the bottom right corner of the slide, consisting of several short, curved segments.



# Silhouette Coefficient Score

Selected number of clusters: **4**

	Number of Clusters								
	2	3	4	5	6	7	8	9	10
1000	0.12	<b>0.078</b>	<b>0.072</b>	0.058	0.056	0.055	0.052	0.053	0.048
5000	0.11	<b>0.088</b>	<b>0.093</b>	<b>0.09</b>	0.083	0.06	0.065	0.056	0.057
10000	0.12	<b>0.091</b>	<b>0.094</b>	0.062	0.064	0.06	0.057	0.052	0.057
30000	0.12	<b>0.092</b>	<b>0.094</b>	<b>0.081</b>	0.064	0.06	0.058	0.06	0.057
50000	0.12	<b>0.092</b>	<b>0.094</b>	<b>0.081</b>	0.063	0.06	0.057	0.062	0.051



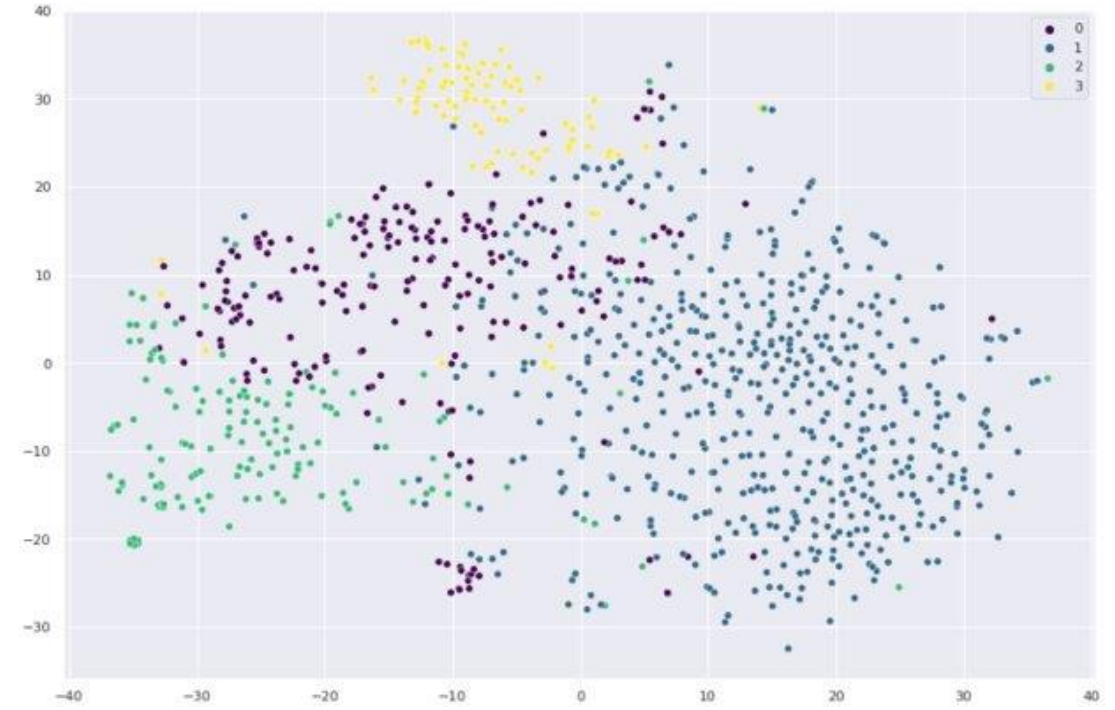
# Results

(clusters\_total = 4)

# Labelled Sentences (**1000** samples)

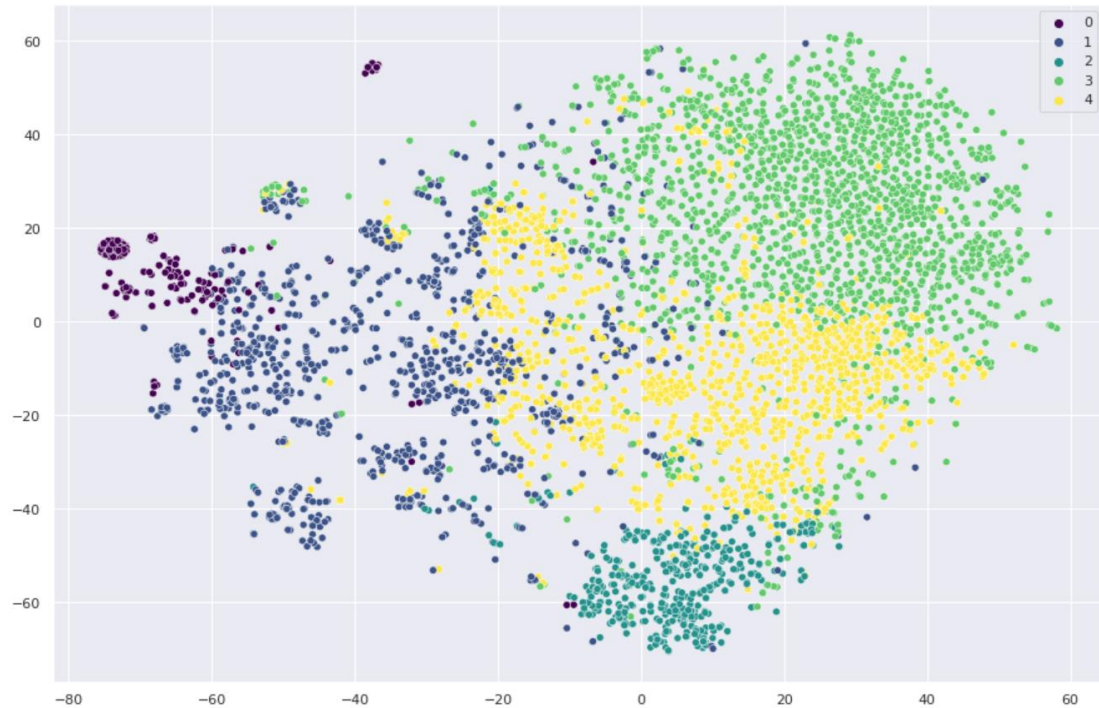


clusters\_total = 5



clusters\_total = 4

# Labelled Sentences (**5000** samples)



clusters\_total = 5



clusters\_total = 4



# Labelled Sentences (**10000** samples)

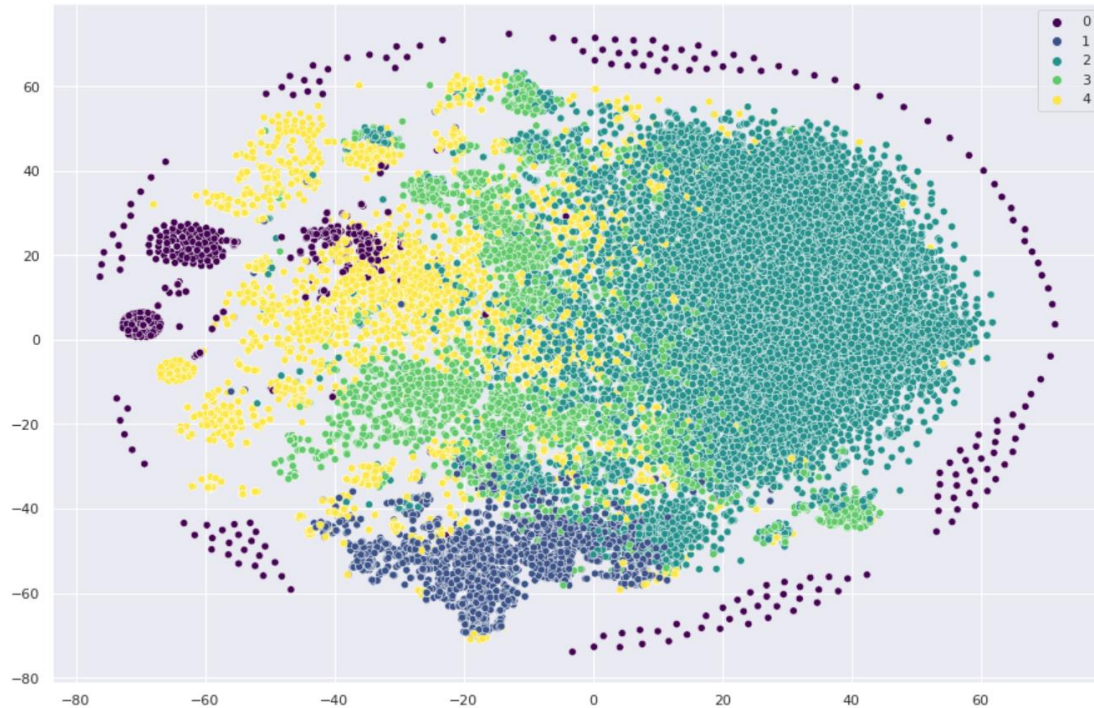


clusters\_total = 5

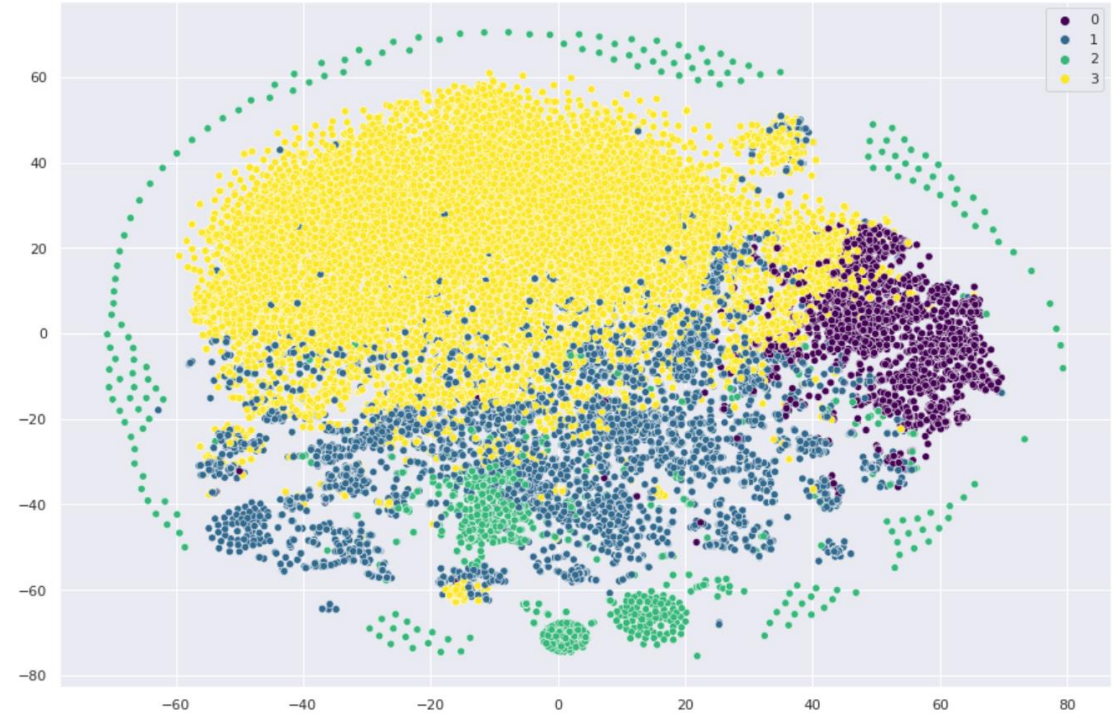


clusters\_total = 4

# Labelled Sentences (**30000** samples)



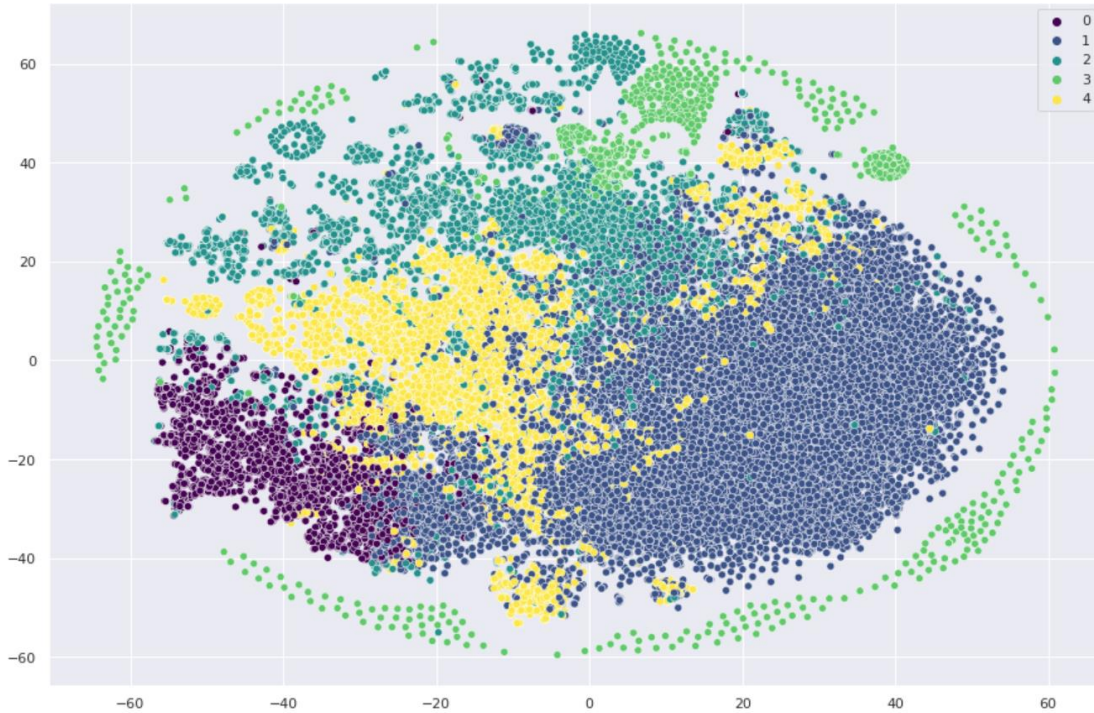
clusters\_total = 5



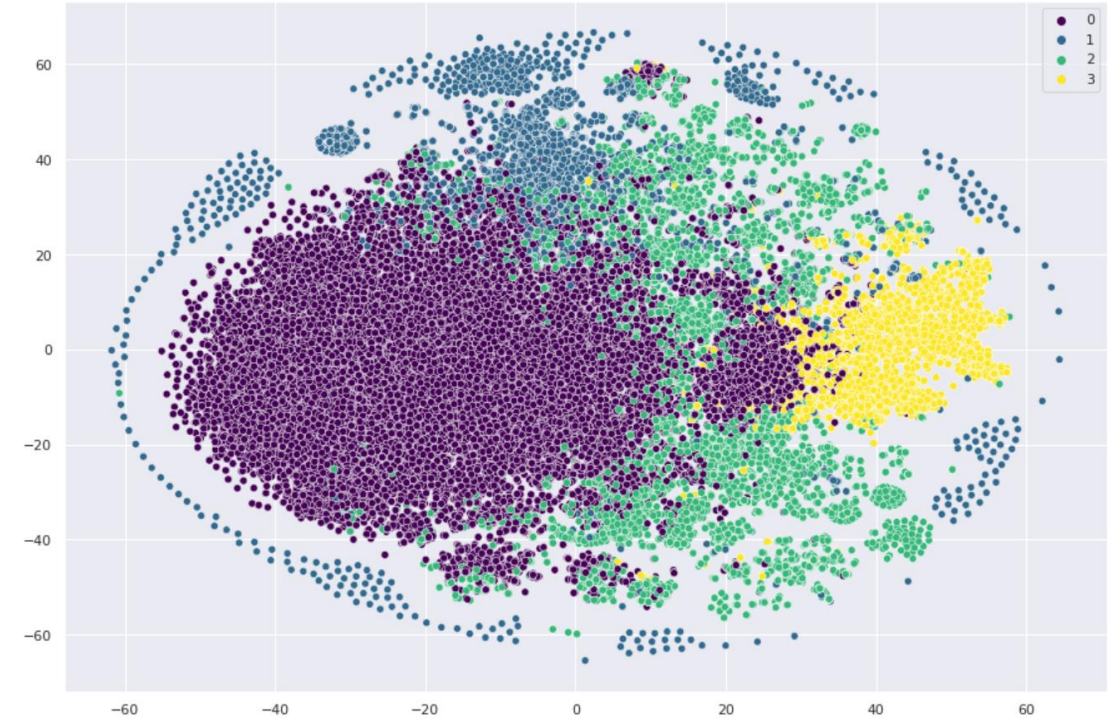
clusters\_total = 4



# Labelled Sentences (**50000** samples)



clusters\_total = 5



clusters\_total = 4



# Examples of Labelled Sentences

(**1000** samples)  
(**clusters\_total = 4**)



# Labelled Sentences - Class 0

- Um livro que deve ser lido e degustado.
- **Muito bom amei**
- **livro lindo e muito bom!**
- Um livro perfeito.
- **Livro excelente!**
- **Agora quero ler todos.**
- Uma grande e incrível jornada.
- A editora nós entrada um produto de autoestima qualidade.
- **Ilustração linda e rica em detalhes, fora o conteúdo que é extremamente precioso.**
- Recomendo a todos.

# Labelled Sentences - Class 1

- Em determinado ponto (37% lido) surge uma mensagem: Algo deu errado, não foi possível exibir a página solicitada.
- **Comprei para dar de presente a um amigo e ele adorou!**
- **Já perdi a conta para quantos amigos já recomendei esse livro.**
- Primeiro, preciso elogiar a forma como a autora escreve: a leitura é prazerosa.
- Só fiquei com raiva do Simon por ter feito a Dafh sofrer desnecessariamente.
- Prende a atenção do começo ao fim.
- Eu simplesmente devorei este livro, que me levou às lágrimas em algumas partes.
- É triste ler esses poemas do início do livro e perceber que a autora passou por tudo isso.
- Os personagens são muito bem construídos e a história é bem desenvolvida, além de tratar de temas importantes
- **Meu sobrinho de dois anos e um mês AMOU, pedia pra ler todo dia e morria de rir, principalmente quando chegava na parte do jacaré (ele ama jacare).**

# Labelled Sentences - Class 2

- <3
- Muito fofo.
- !
- Dentro é mto bonito
- narrado todo o seu conteúdo!
- Divisor de águas.
- Melhor que esperado
- Sai outra pessoa dessa leitura.
- Simples e direto.
- Muito

# Labelled Sentences - Class 3

- porém **a primeira folha veio amassada**
- **A empresa que trouxe foi a loggi (NÃO RECOMENDO)**, era pra presente, paguei entrega expressa e não chegou no dia do aniversário 😞😞
- **Referente ao estado do livro**, chegou em perfeito estado, apenas alguns amassadinhos na parte inferior da lombada , mas acredito que foi questão do manuseio durante a entrega.
- Aqui você encontra os melhores livros e as melhores ofertas do Mercado, entrega rápida e livros bem embalados .
- O livro chegou em perfeito estado
- Chegou antes do prazo,o produto e de ótimo qualidade,além de ter vindo super embalado,amei demais, pretendo comprar mais vezes 😊
- Muito bom o produto , ótima qualidade e entrega repida no prazo ou antes .RECOMENDO
- **A entrega foi rápida, as folhas não são fina e a capa tbm é linda!!**
- Livro de boa qualidade, capa dura com desenhos em alto relevo, folhas amarelas, e letras do tamanho normal.
- O produto chegou em perfeito estado!



# Examples of Labelled Sentences

(**30000** samples)  
(**clusters\_total = 4**)

# Labelled Sentences - Class 0

- Pelo contrário.
- Obs.:
- !
- haha
- tb
- !
- ♥
- ?
- Nossa!

# Labelled Sentences - Class 1

- **O livro é excepcional.**
- **Livro** com conteúdo formidável, e bem escrito.
- **Livro** igual a descrição do site **gostei muito**.
- Tadinho!!!
- **Recomendo a todas as pessoas.**
- Grande autor do nosso tempo.
- Recomendo...
- Boa edição e **bom livro**
- Essencial para um bom desenvolvedor
- Ótimo **livro**.

# Labelled Sentences - Class 2

- **Pelo contrário, ele é cheio de histórias que te prendem.**
- **As crenças de religiões de matriz africana são belamente retratadas dando um toque quase que mágico à obra.**
- Esse foi um livro que eu achei fraco no início, mas no decorrer da leitura comecei a me sentir inspirado a fazer uma análise dos meus hábitos.
- **A autora tem uma escrita muito realista que nós faz até acreditar que tudo que esta é 100% real.**
- Leia, simplesmente leia, e depois entre pro meu grupo de apoio com psicóloga.
- vem também com um fitilho que é muito pratico para usar como marca páginas.
- **Mostra que, com pouco tempo por dia, podemos alcançar grandes feitos, realizar e atingir tudo o que desejamos.**
- Inserir, no começo da história, uma situação que se torna contexto de todo o seu desenvolvimento.
- Tem histórias bem pequenas e outras maiores, tipo de umas 3 ou 4 páginas.
- O livro mostra de maneira bem ilustrativa os meios de se conseguir não ser escravo do dinheiro, mas fazer com que o mesmo trabalhe em seu favor.



# Labelled Sentences - Class 3

- Livro ótimo, como sempre **chegou muito antes do prazo de entrega**, edição original, **sem nenhuma ranhura ou defeito**
- **Chegou muito bem, só veio com a primeira folha com um defeitinho**, mas não importa muito.
- **Chegou muito antes do prazo e em perfeitas condições!**
- Lindo, lindo e lindoooooooo!!!!**Veio super rápido, o livro é lindo.**
- **Adorei o livro, veio em perfeito estado, entrega rápida!**
- Mais uma coisa precisa ser dita, a qualidade dele é INQUESTIONÁVEL, o meu veio perfeito é muito bem embalado.
- Chegou antes da data prevista.
- Chegou bem antes do prazo.
- Entrega padrão
- Amei esse livro, é muito fofoA encomenda veio certinha, nem nenhum amassado ou rasgo

# Labelled Dataset

book_id	review_id	review_rating	sentence	kmeans_label
8578601777	R2MSSL8D6SNN9P	5	Porém, já vi melhores sobre o casamento.	0
<b>8547000240</b>	<b>R2D5L66GQPAJP3</b>	<b>5</b>	<b>Ótimo, chegou suppperr rápido!</b>	<b>3</b>
6555650001	RWDE7J13FQR0L	5	Amei	2
6555871784	R1VIKGOQY56A5X	5	Um livro muito bem escrito, com personagens bem elaborados e uma estória eletrizante, do começo ao fim.	0
8576849941	RCKAYQ6I54EEL	2	só espero que o capítulo 10 me surpreenda.	0
8501117846	RTH2BN3SA1M7I	5	PERFEITO	2
8501112518	R2MFSP46FI8P85	5	Pessoalmente desejo que todos tenham um bom relacionamento afetivo..	0
<b>8595083274</b>	<b>RADTWBO86506V</b>	<b>4</b>	<b>Mas o produto é de ótima qualidade e chegou antes do previsto.</b>	<b>3</b>
8543107202	R2GKA8QZWIPRD7	5	Um livro leve, bonito e tocante apesar do tema tão doloroso.	0
<b>8501304468</b>	<b>RA4RCZW4GVQLD</b>	<b>5</b>	<b>Chegaram todos perfeitos no dia seguinte a compra, bem feliz com a aquisição na minha prateleira</b>	<b>3</b>
6580309318	RNZO7UVGNE9CN	5	É a história do povo brasileiro que fica escondida pela evidência às capitais, principalmente do sudeste.	0
8542209826	R2K01P8K8QWHME	1	muito meloso e o Rune é insuportável!	2
8535909559	R1N2C877M84QYC	5	A história é muito boa e super recomendo.	2

# Future Directions

- Explore the removing of words like "livro" and "leitura".
- 2018 – Deep Representation Learning for Clustering of Health Tweets
  - <https://arxiv.org/abs/1901.00439>
- 2021 – WEClustering: word embeddings based text clustering technique for large datasets
  - <https://link.springer.com/article/10.1007/s40747-021-00512-9>
- Explore the percentage of each cluster in the total of the sentences. (Example: a new book representation.)
- Train a model based on the labels of the clusters and experiment the perception of people.
- Other clustering methods:
  - SpectralClustering
    - <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.SpectralClustering.html>
  - DBSCAN
    - <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html>