THE DEVELOPER'S CONFERENCE

Trilha – Java

Carla Vieira

Coordenadora do Perifacode

Mineração de textos com Java

Carla Vieira
@carlaprvieira

Quem sou eu?

Carla Vieira

Graduanda e Aluna Especial de SI – USP

Coordenadora do Perifacode

Evangelista de Inteligência Artificial e Ética

@carlaprvieira

carlavieira.dev

{ perifacode(); }

WOMAKERS CODE

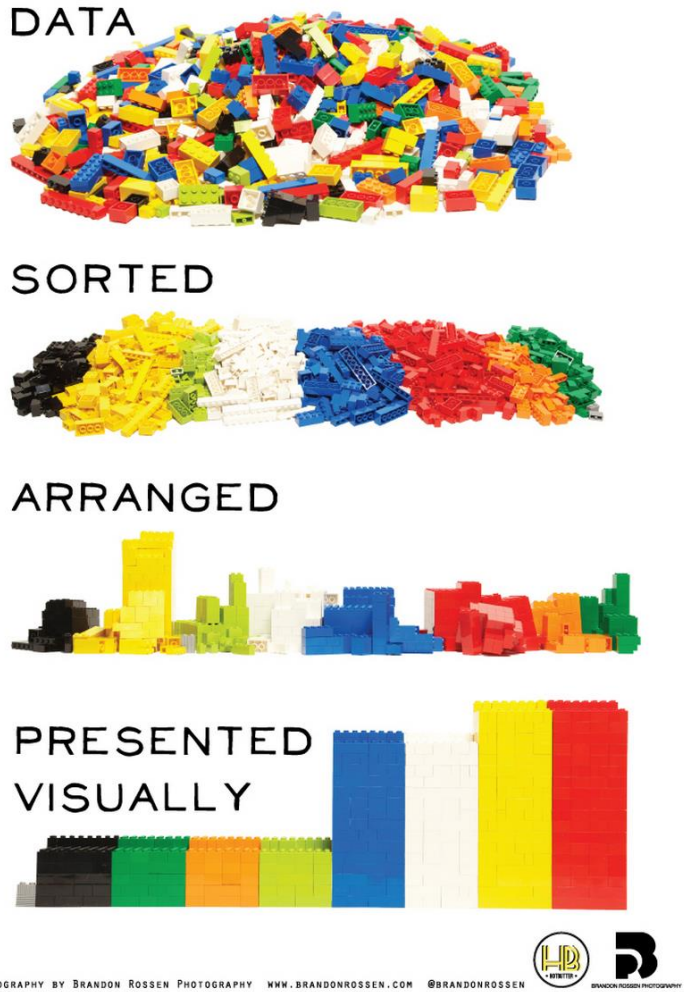
DEVS JAVAGIRL

Agenda

- Conceitos básicos de Machine Learning
- Workflow da mineração de textos
- Apresentação de um case de mineração de textos
- Conclusões

O que é Mineração de Dados?

Data Mining define o processo automatizado de captura e análise de grandes conjuntos de dados para extrair um significado, sendo usado tanto para ***descrever características do passado*** *como para* ***predizer tendências para o futuro****.*



"Um **dado** não vira informação se você não souber o que ele significa; uma **informação** não vira **conhecimento** se você não enxergar **relevância** nela e conhecimento não serve pra nada se não aplicá-lo de maneira apropriada."

INTELIGÊNCIA ARTIFICIAL

Programas com habilidade de agir como humanos

1950

MACHINE LEARNING

Algoritmos com habilidade de aprender sem programação expressa

1980

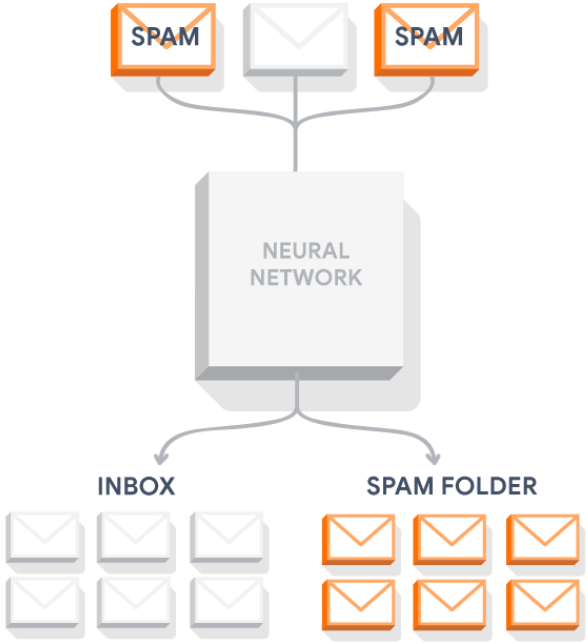
DEEP LEARNING

Redes neurais artificiais que aprendem através de um grande volume de dados

2010

Programação Tradicional x Machine Learning

Regras



Rede Neural

“Um programa de computador aprende se ele é capaz de melhorar seu **desempenho** em determinada **tarefa**, sob alguma medida de **avaliação**, a partir de **experiências** passadas.”

(Tom Mitchell)

PLN x Mineração de Textos

Processamento de Linguagem Natural	Mineração de textos
Reconhecimento de Fala	Agrupamento
Sistemas de diálogo (chatbots)	Classificação
Tradução	Descoberta de Padrões

Mineração de textos

github.com/carlaprv/k-means-clustering

Etapas da mineração de textos

Pré-processamento

- Limpeza;
- Formatação;
- Redução complexidade e dimensões dos dados

Algoritmo

- Execução do algoritmo

Pós-processamento

- Análise dos resultados;

Qual problema queremos resolver?

Para onde vai a gordura que queimamos quando perdemos peso?
BBC Brasil - 9 horas atrás
Nas aulas de física e química aprendemos que energia não se cria nem se destrói. Ela, na verdade, se transforma. Com base na chamada lei da conservação ...

PT deveria realizar 'comissão da verdade' para examinar seus erros ...
BBC Brasil - 14 horas atrás
BBC News Brasil - Lula nomeou Fernando Haddad como seu sucessor. Se ele vencer, terá que lidar com um forte sentimento anti-PT no país, ...

Eleições 2018: Haddad e Bolsonaro avançam, mas sombra da ...
BBC Brasil - 3 horas atrás
... Marina 7%, branco/nulo/nenhum 12%. A BBC News Brasil analisou os detalhes das pesquisas e destaca abaixo alguns dos aspectos mais interessantes.

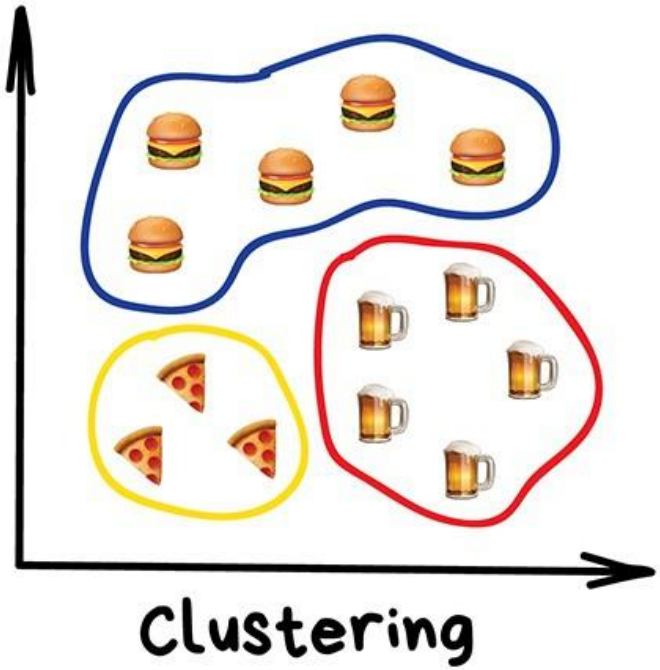
O país que conseguiu recuperar um 'mar' que havia sido extinto
BBC Brasil - 7 horas atrás
Para Madi Zhasekenov, o verão na costa do
nostálgicas. Quando ainda estava na escola,

Canadense e executiva: a verdadeira h
BBC Brasil - 19 de set de 2018
A advogada Karina Kufa disse à BBC News
no vídeo do ponto de vista eleitoral" e que "n

BBC NEWS

Categorização automática de notícias de acordo com o conteúdo

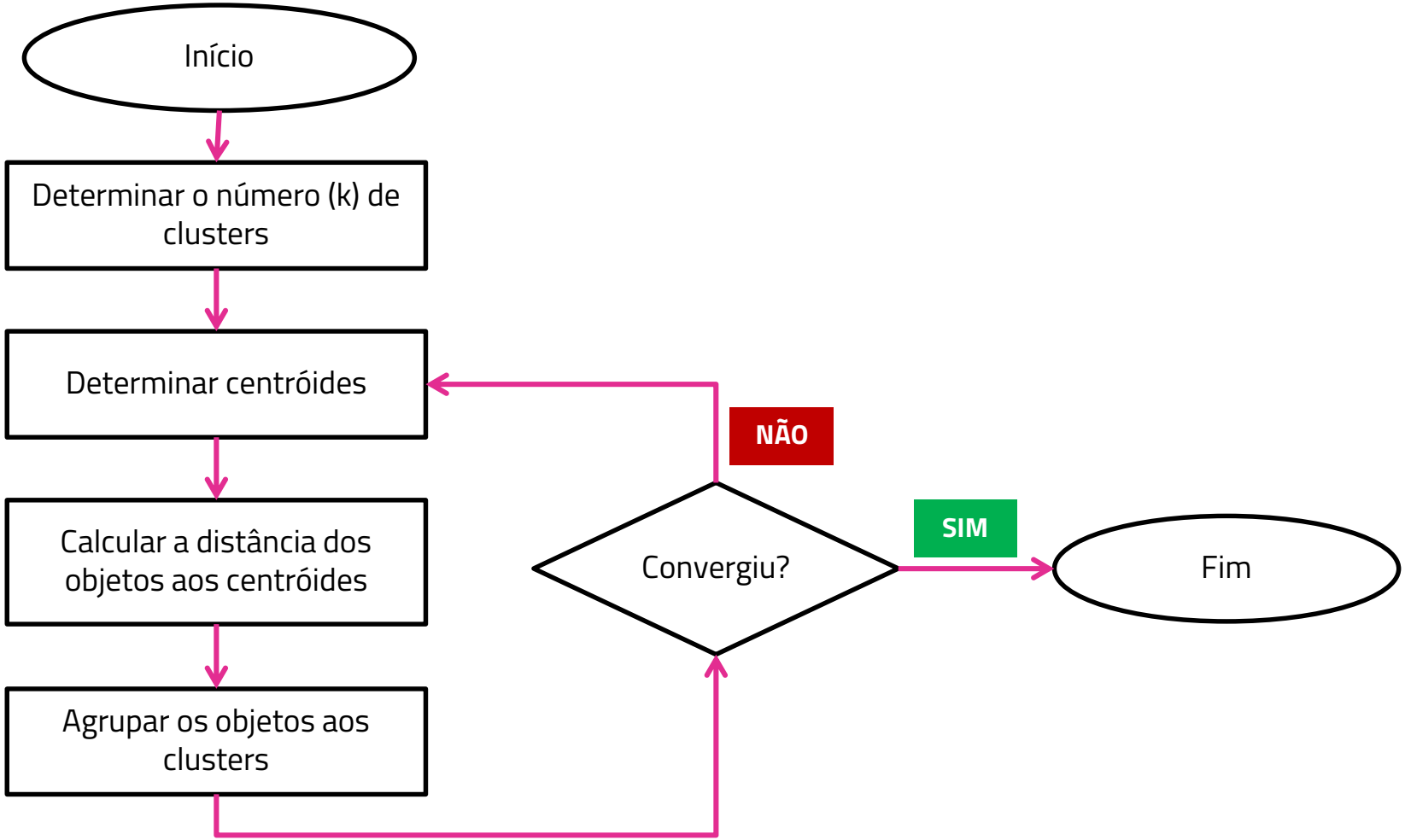
Classificação do problema

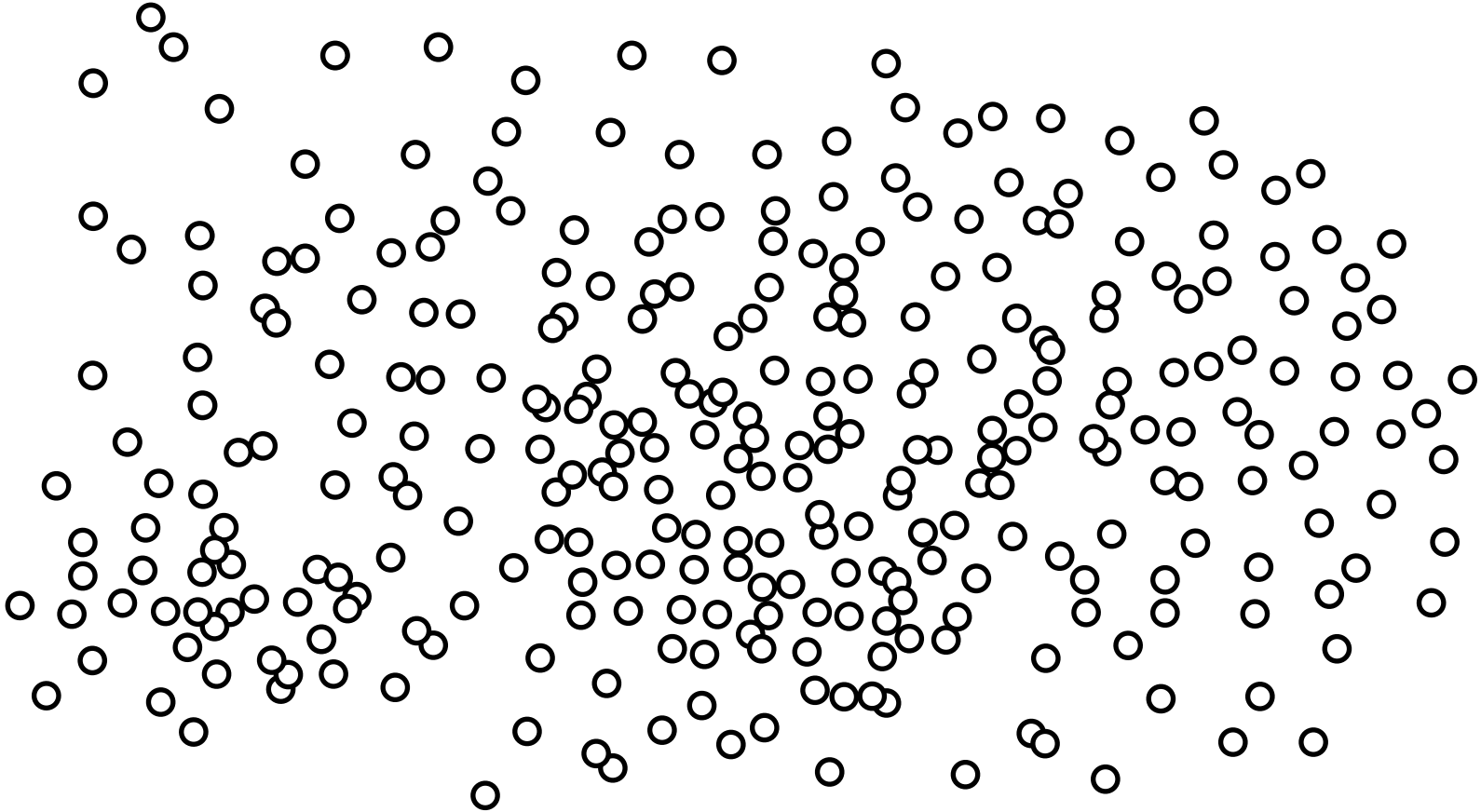


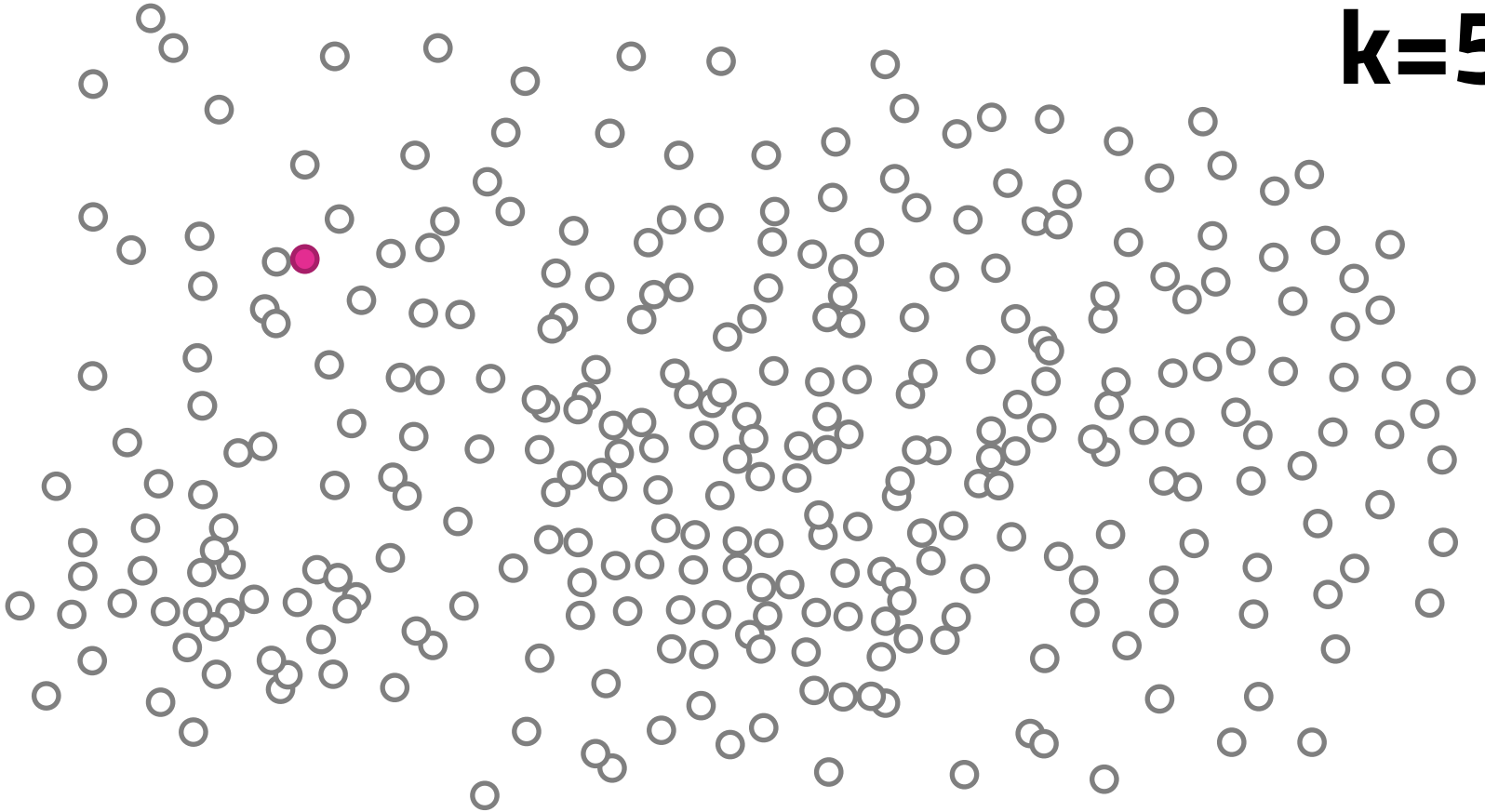
Fonte: https://vas3k.com/blog/machine_learning/

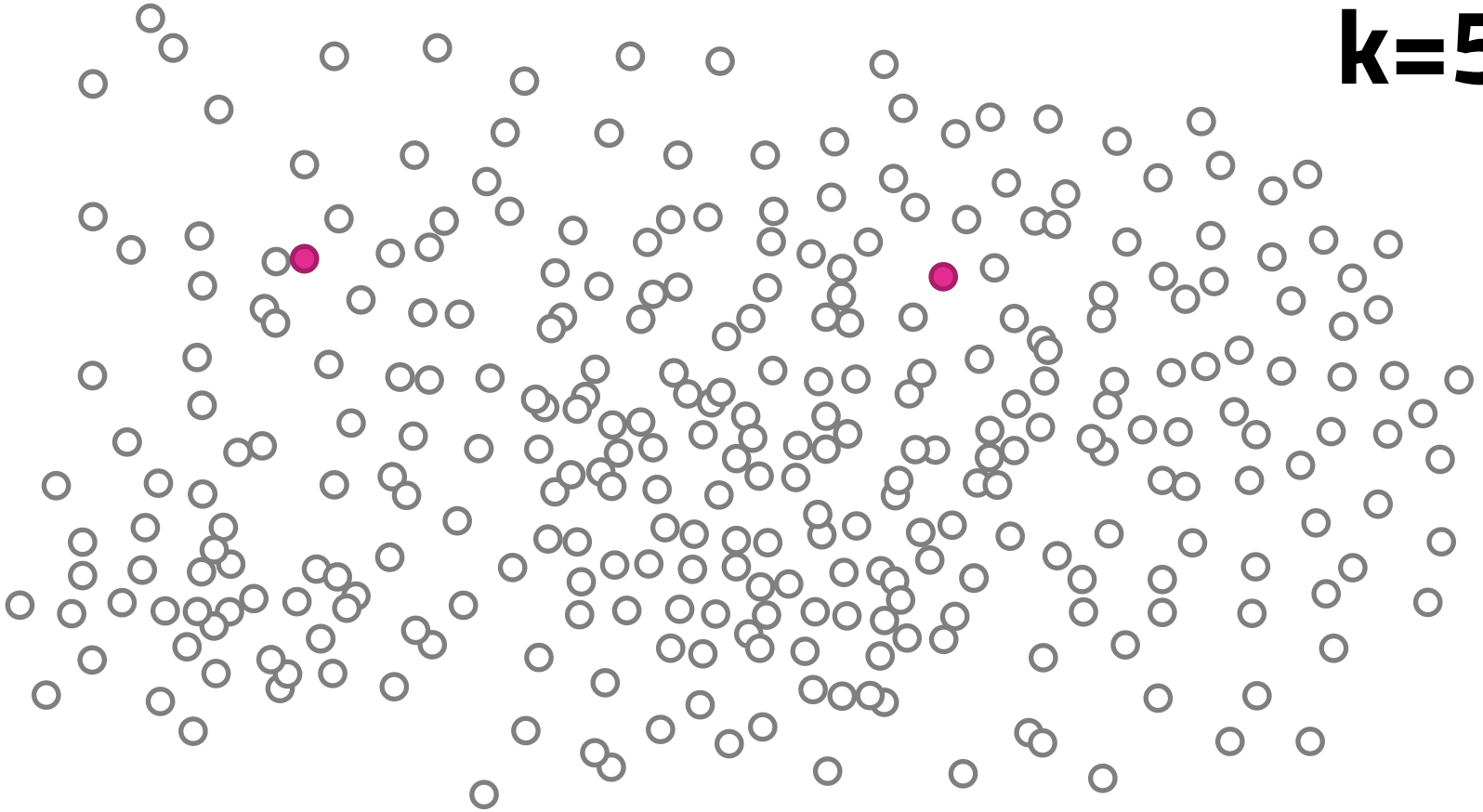
Clusterização

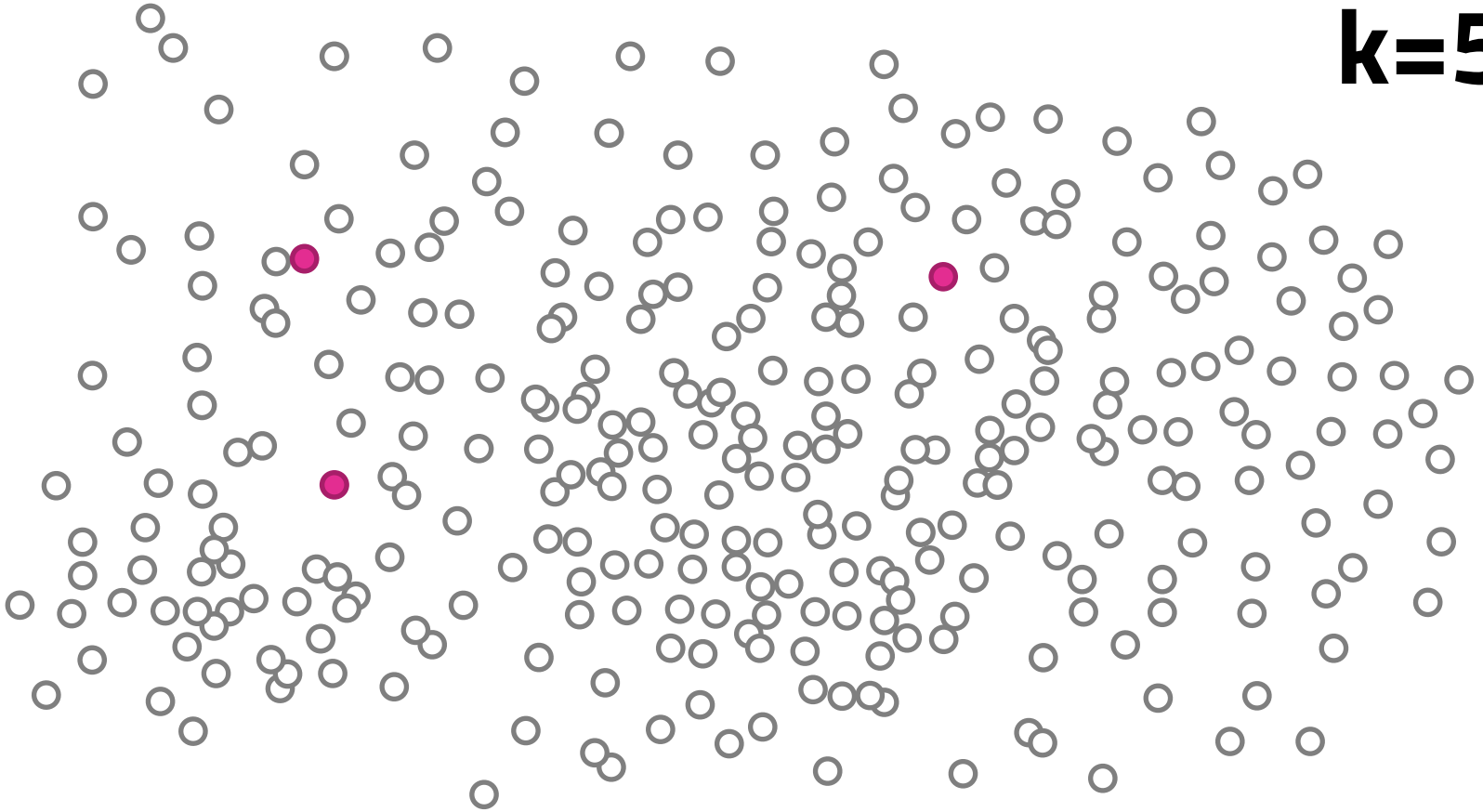
"Divides objects based on unknown features. Machine chooses the best way"

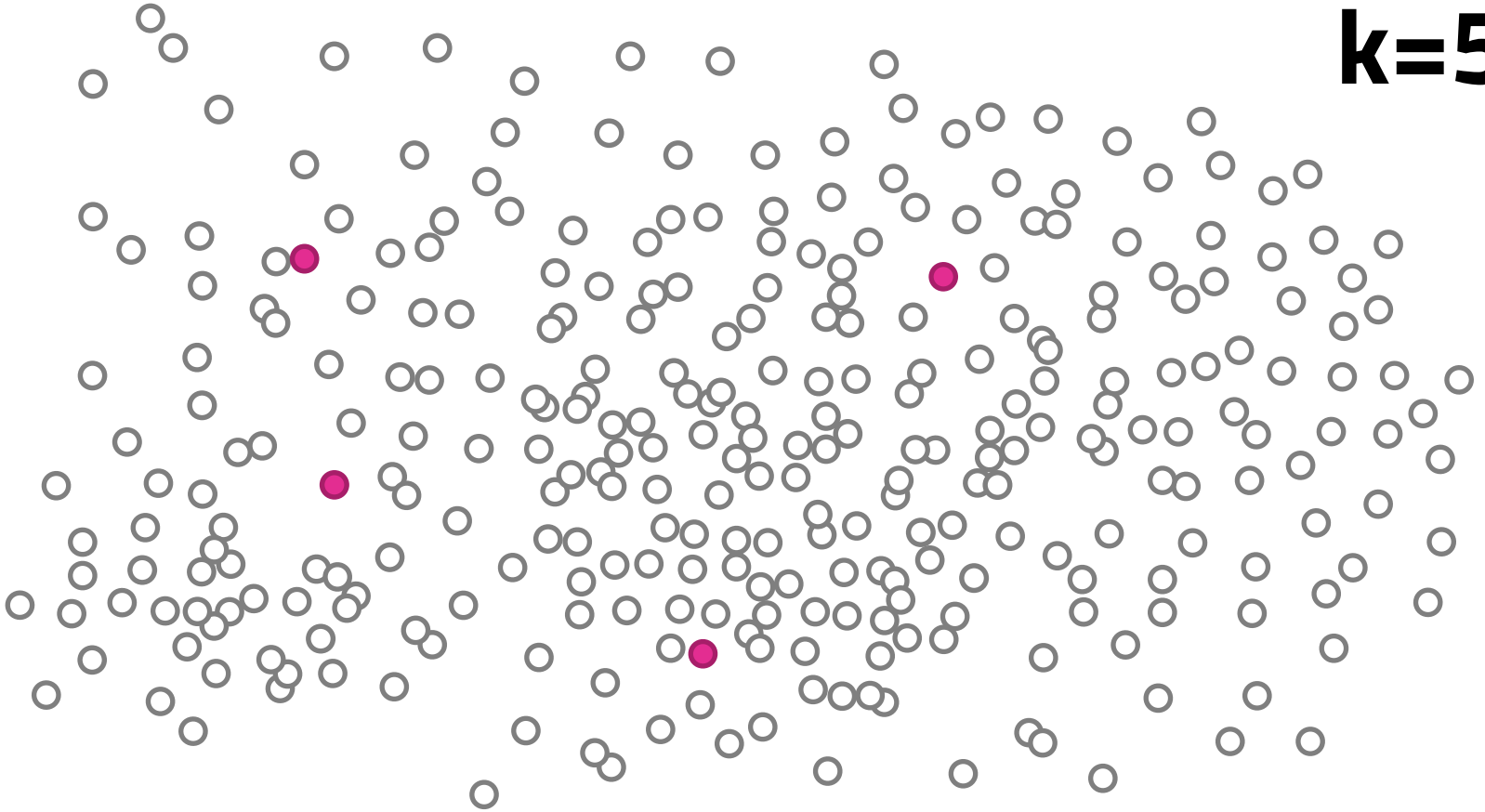

Início
Determinar o número (k) de clusters
Determinar centróides
Calcular a distância dos objetos aos centróides
Agrupar os objetos aos clusters
Convergiu?
NÃO
SIM
Fim

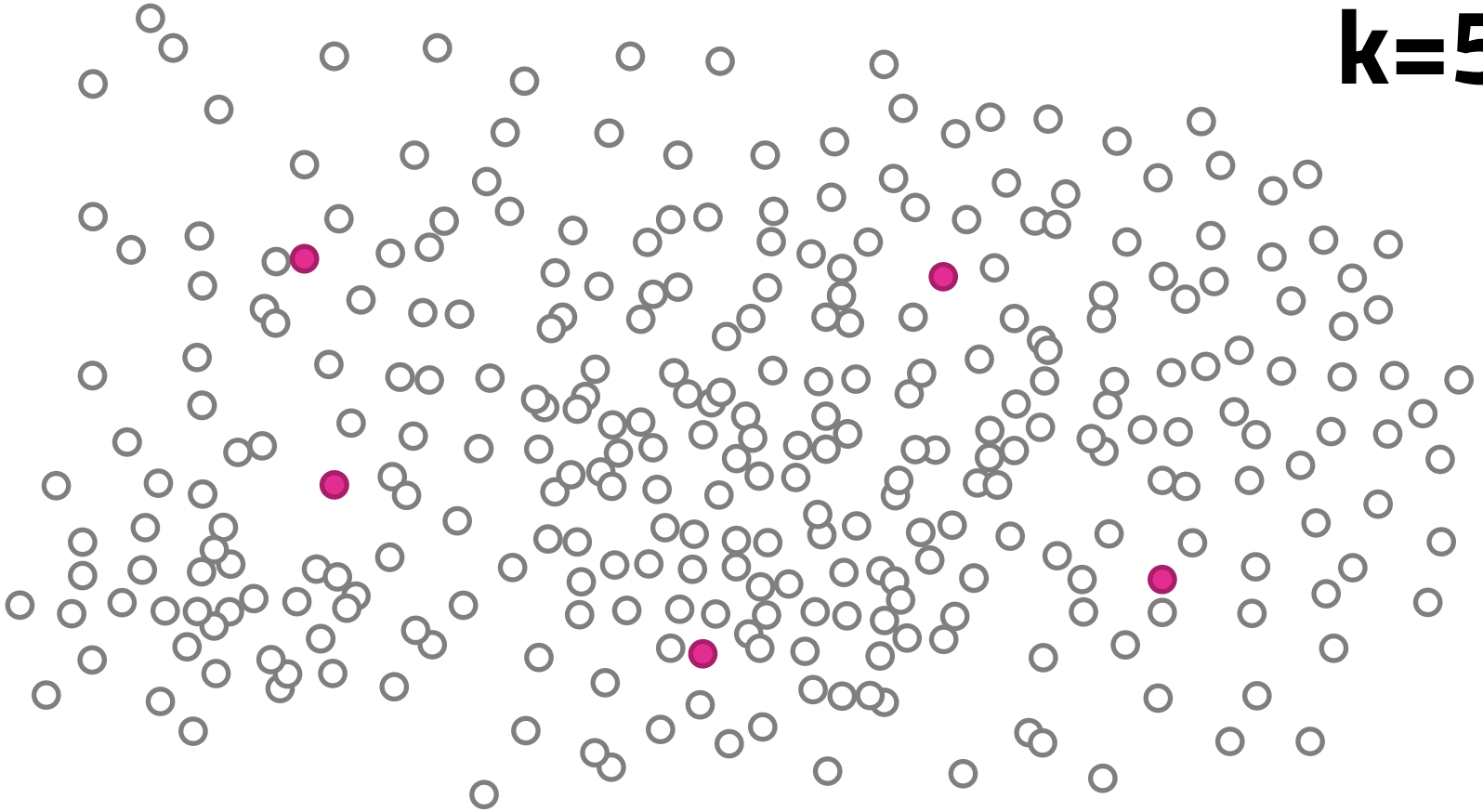


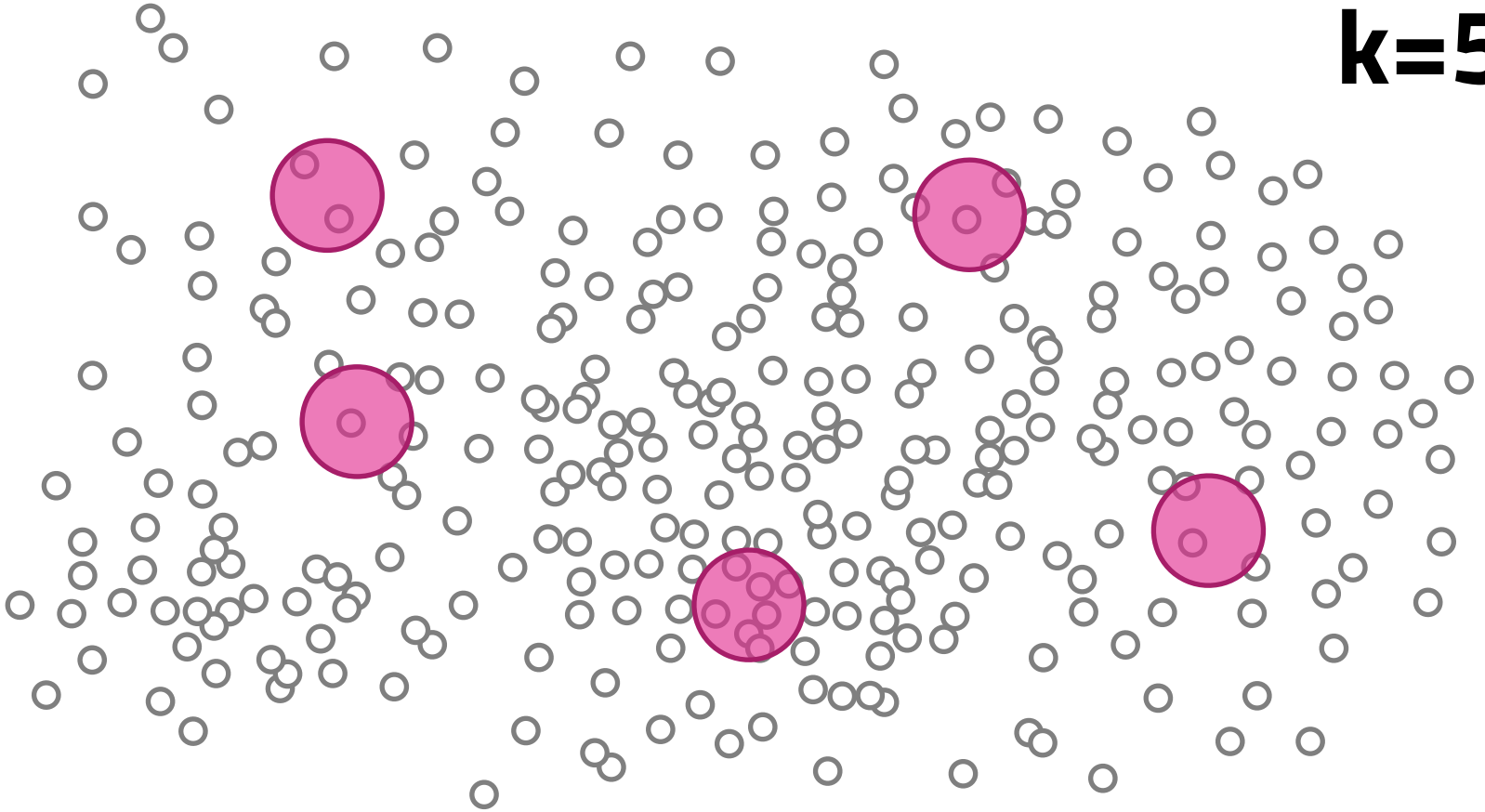

k=5

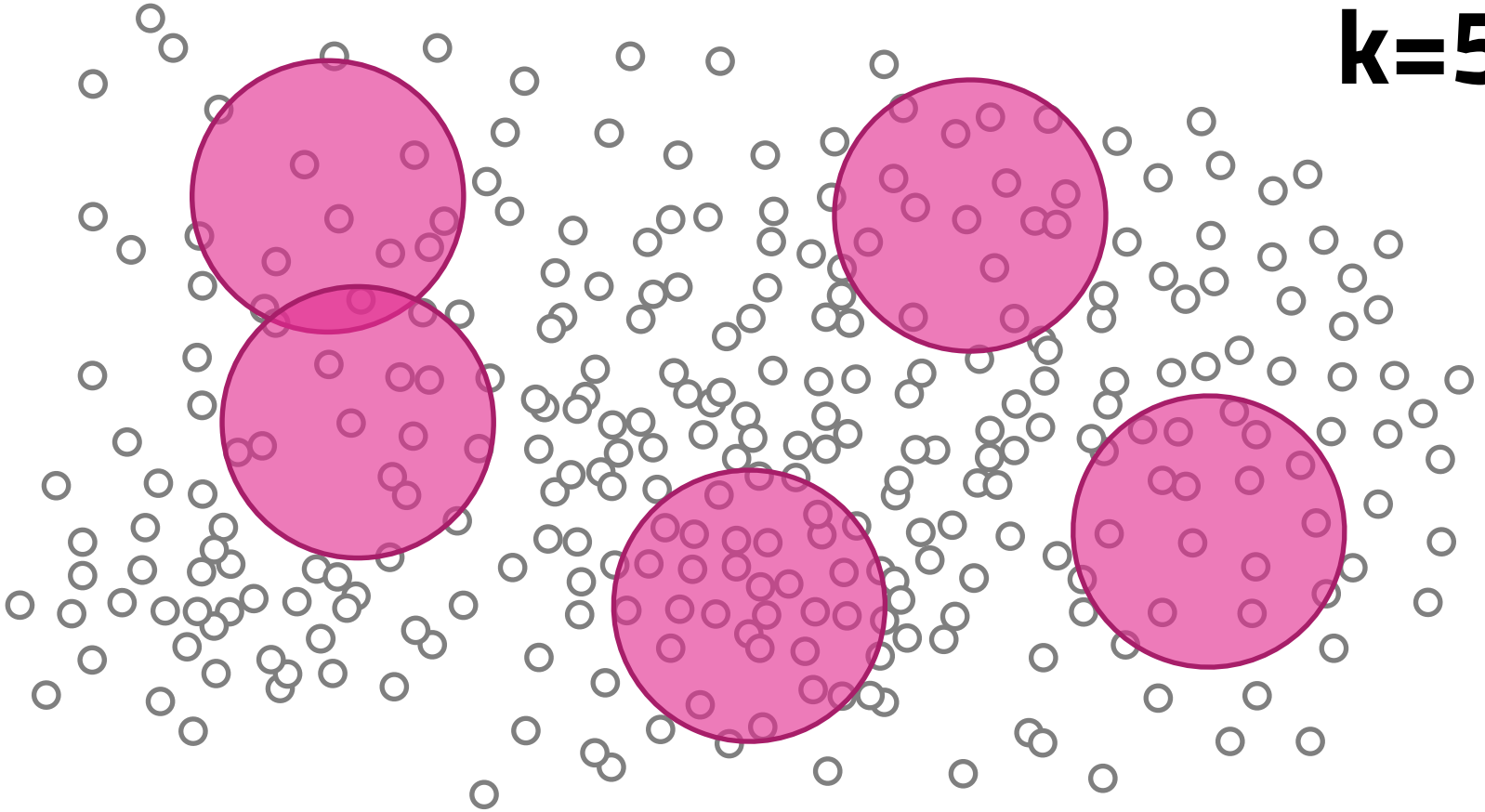

k=5

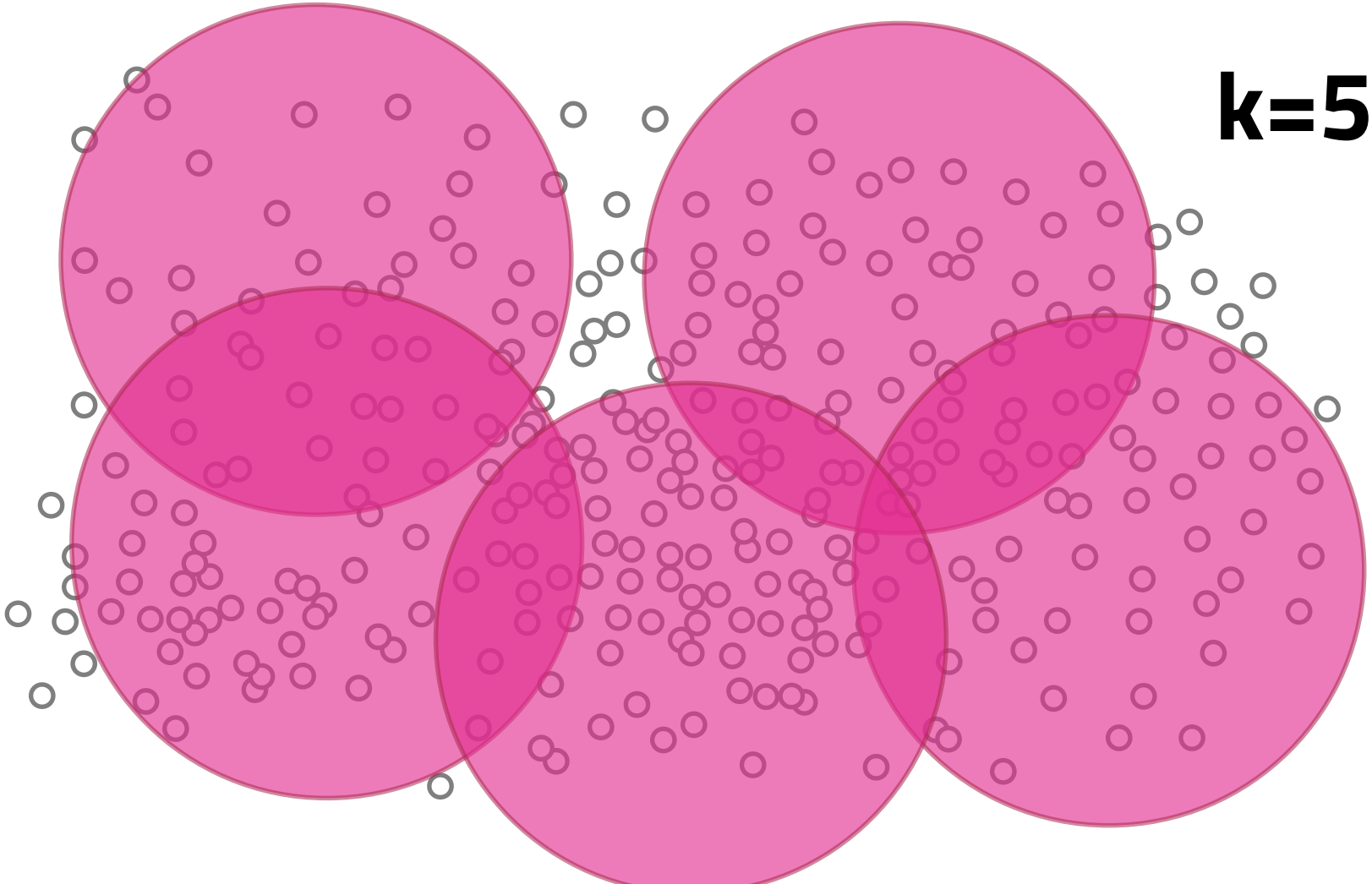

k=5

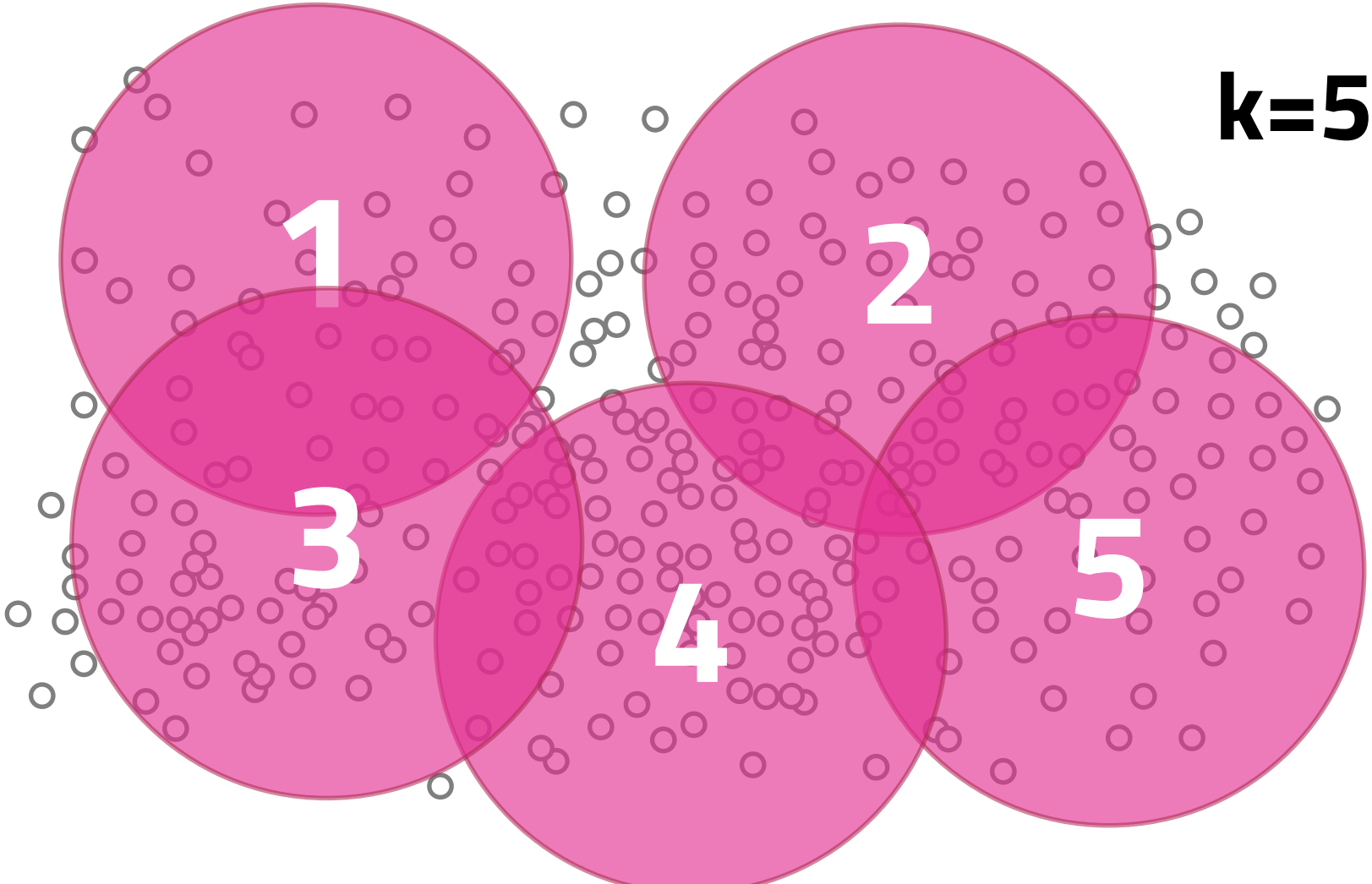

k=5

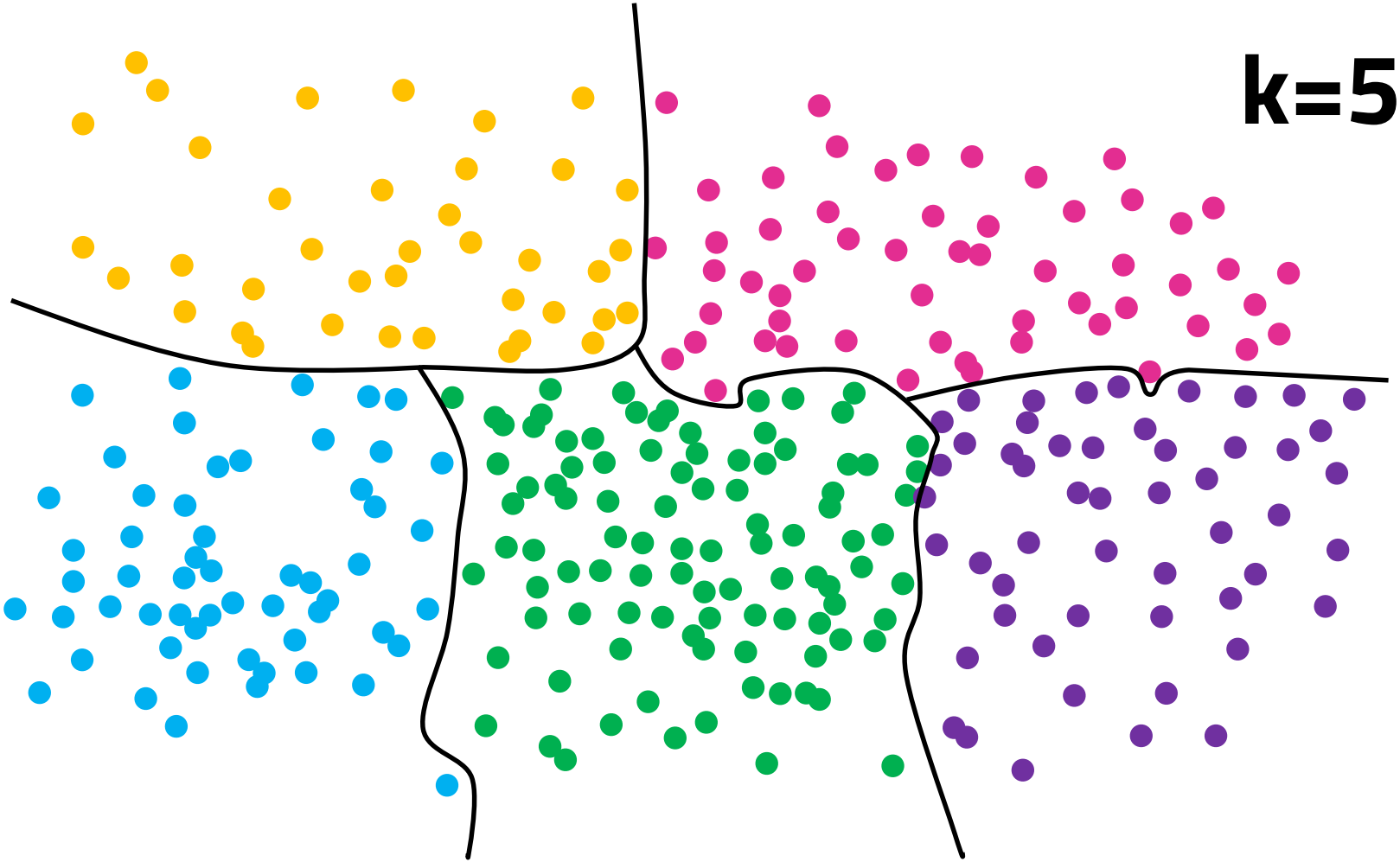

k=5

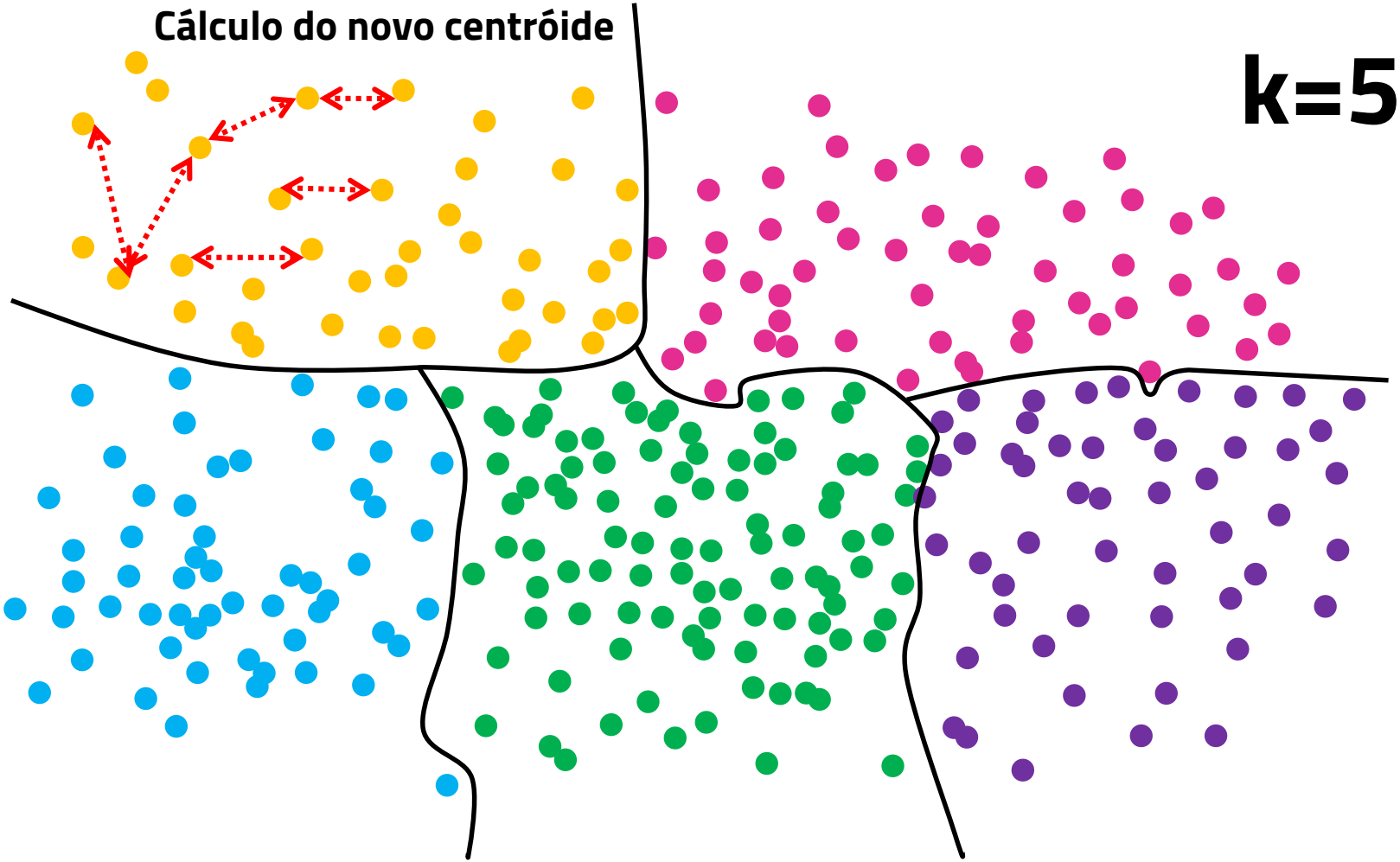

k=5

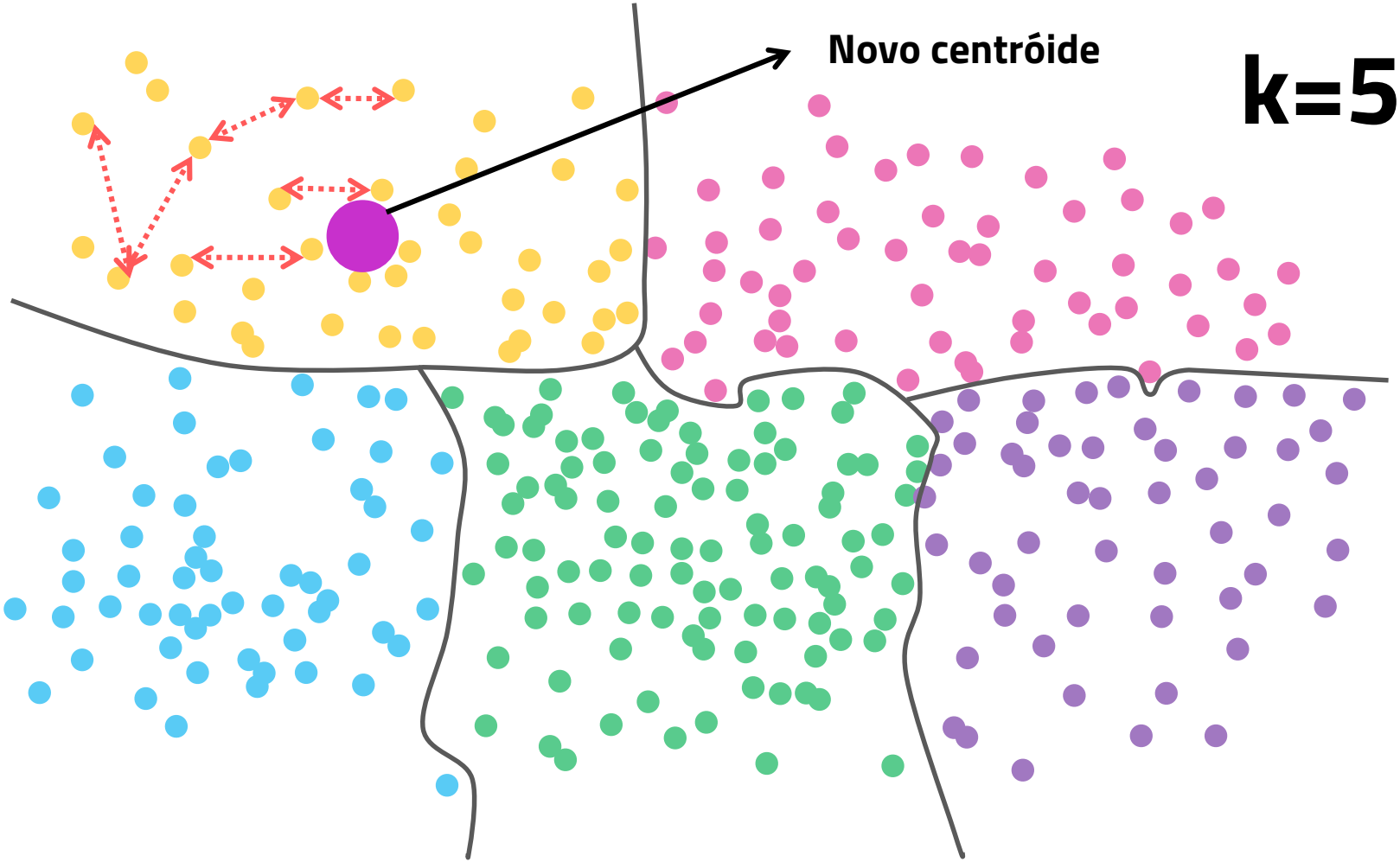

k=5

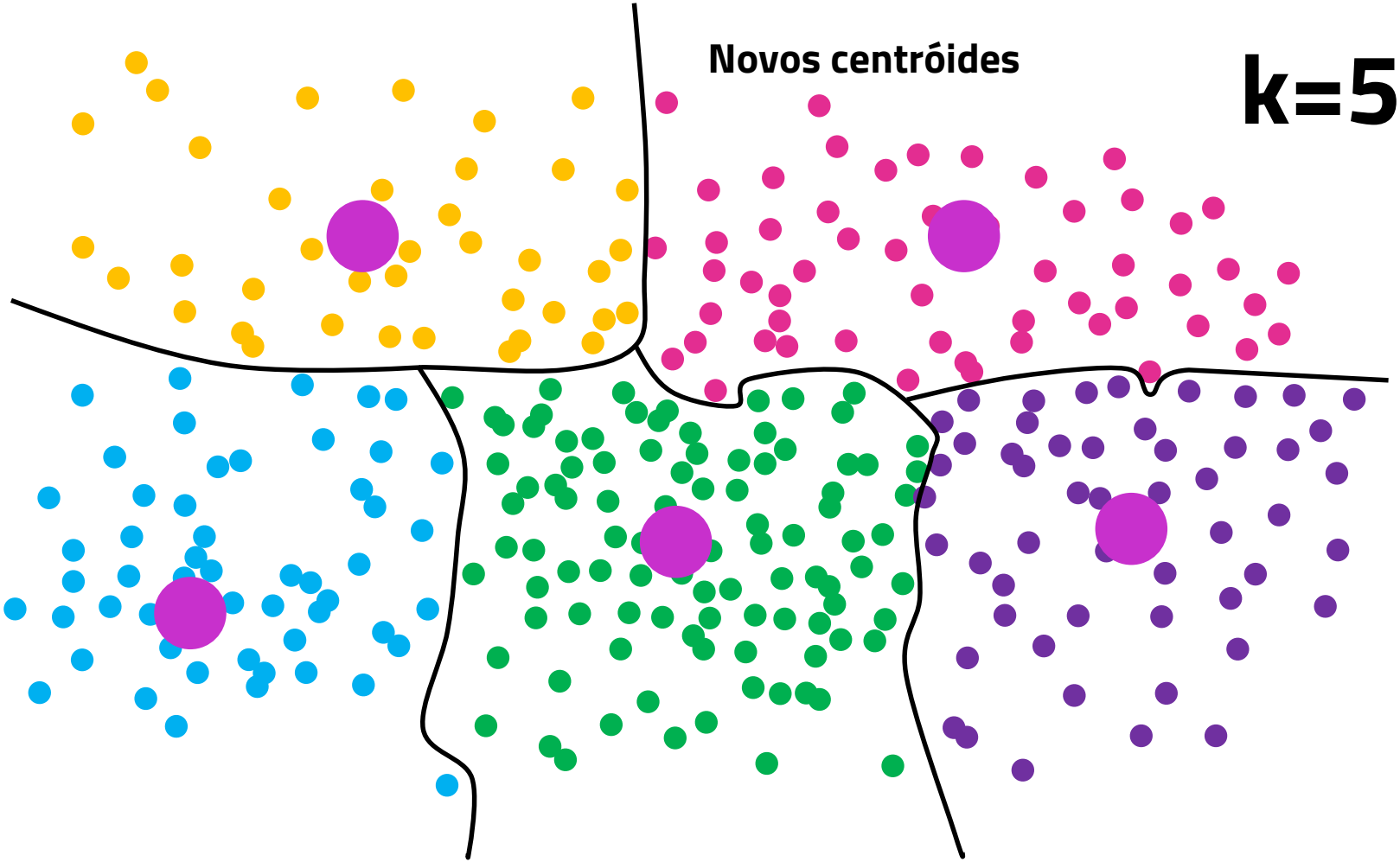

k=5

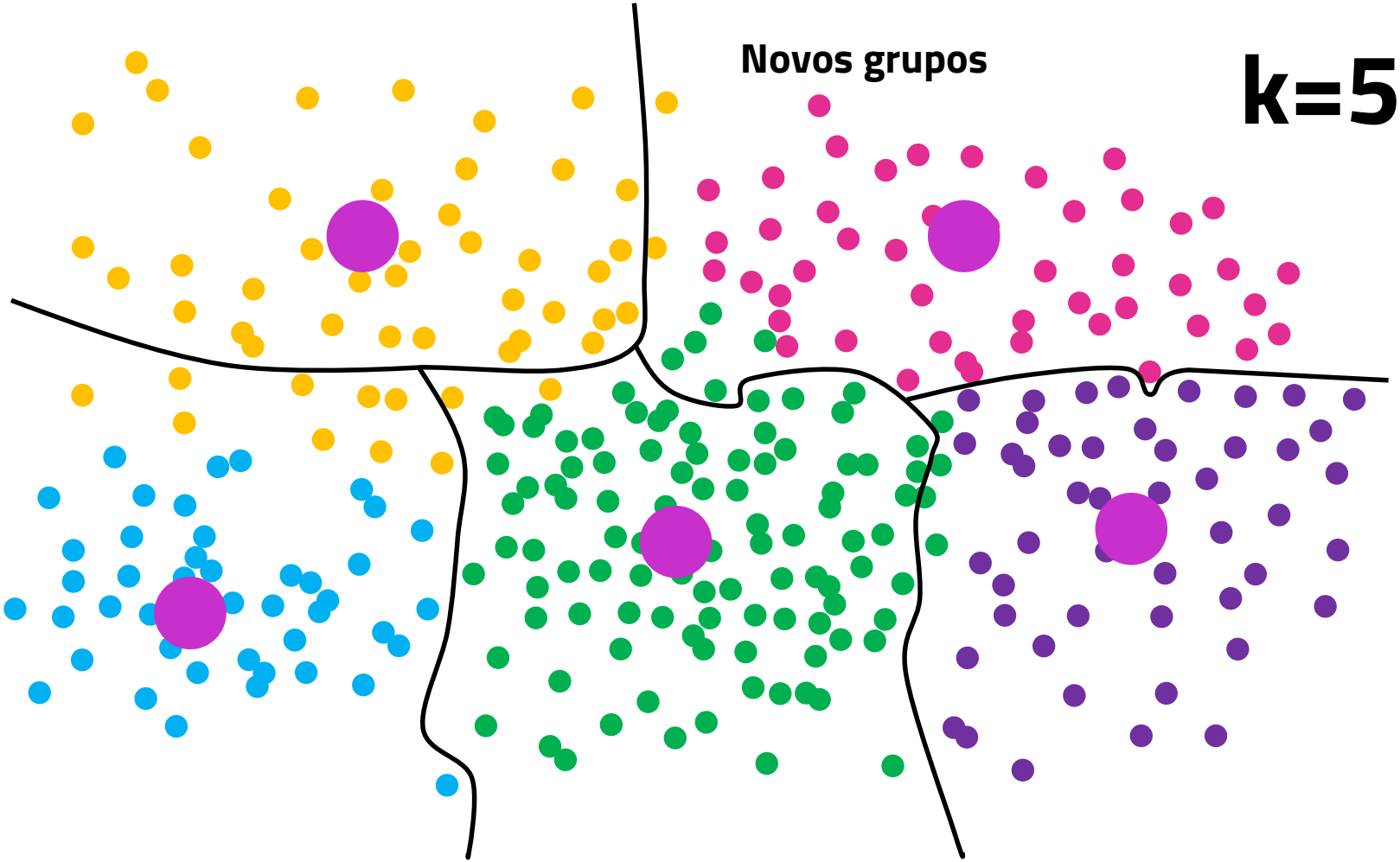

k=5
1
2
3
4
5


k=5


Cálculo do novo centróide
k=5

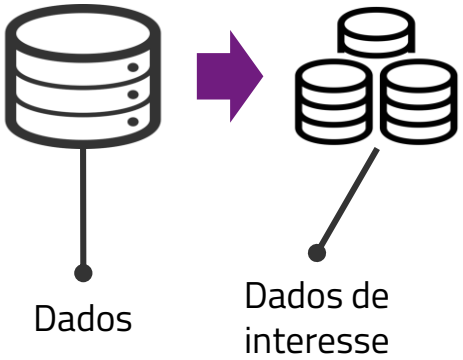

Novo centróide
k=5


Novos centróides
k=5


Novos grupos
k=5

Seleção de dados

Seleção



Corpora BBC (2225 notícias em inglês)

20.246 termos distintos

http://mlg.ucd.ie/howmanytopics/index.html

Objetivo do pré-processamento

Redução de dimensão dos dados

20.246

termos distintos

?

termos distintos

Etapas do pré-processamento

1. **Case Folding**
2. **Tokenização**
3. **Remoção de stop words**
4. **Redução ao radical**
5. **Corte de termos**

1) Case Folding

Converter todas as palavras para minúsculas ou maiúsculas

Por quê?

A mesma palavra seria contada muitas vezes por diferentes escritas

Exemplo: *Amor e amor (seriam consideradas palavras diferentes)*

2) Tokenização

```
from nltk import regexp_tokenize
```

1) u.s foi transformado em usa;

2) palavras compostas com hífen foram unidas;

3) expressões que continham subtraço (_) foram separadas

Tokens com menos de 3 caracteres foram retirados da lista

3) Remoção de stop words

```
from nltk.corpus import stopwords
```

O que são stop words (ou palavras vazias)?

Lista de stop words da NLTK

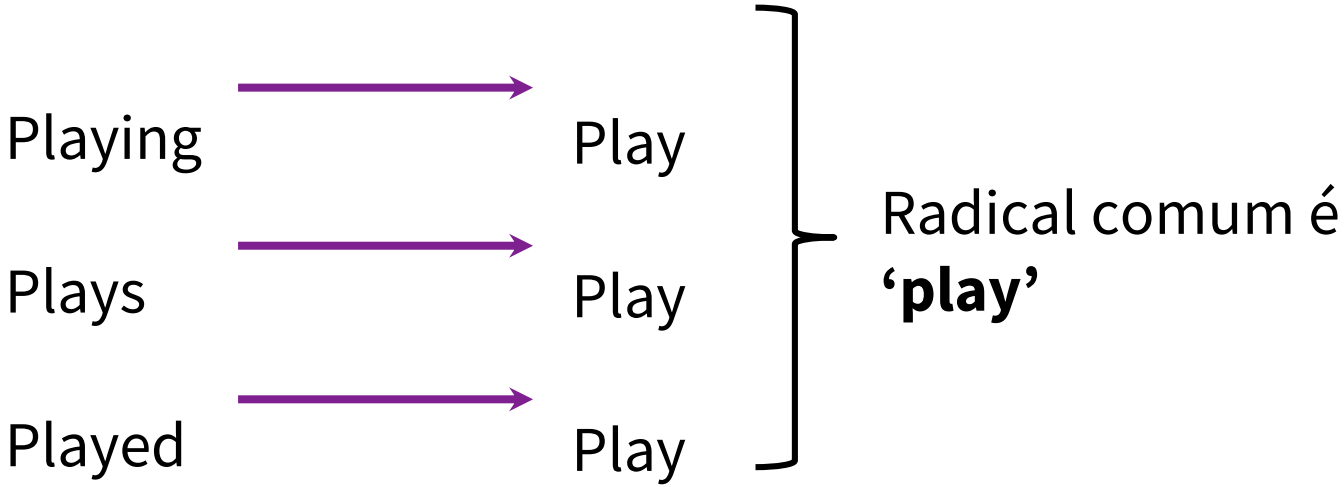
Natural Language
Tool Kit (NLTK)
Basic Text Analytics

a, agora, ainda, alguém, algum, alguma, meus, minha, minhas, muita, muitas, muito, muitos, na, não, nas, nem, nenhum,

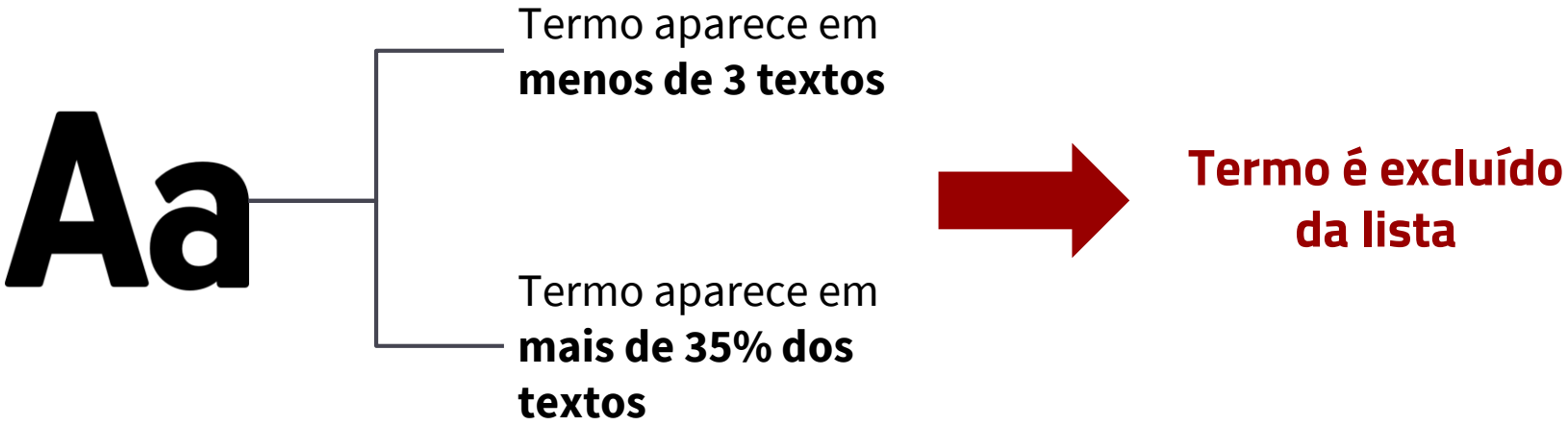
a, able, about, across, after, all, almost, also, am, among, an, and, any, are, as, at, be, because, been, but, by, can, cannot,

4) Redução ao radical (Stemming)

```
from nltk import SnowballStemmer
```



5) Corte de termos



Objetivo do pré-processamento

Redução de dimensão dos dados

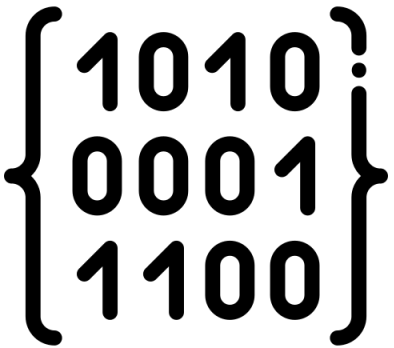
20.246

termos distintos

6997

termos distintos

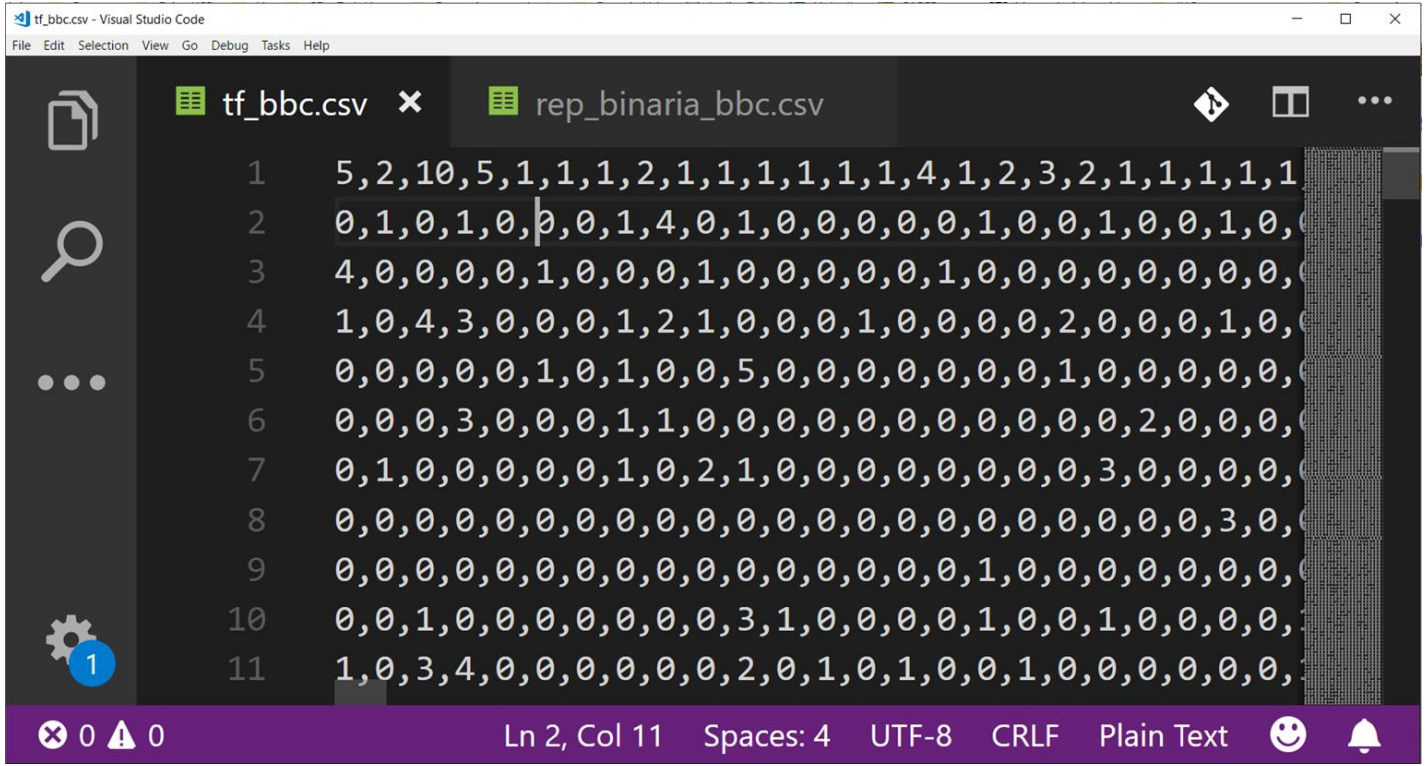
Mas, como iremos representar os textos para rodar o algoritmo?



Representação TF

	termo 0	termo 1	termo 2	.....	termo n
Texto 0	2	0	5	0	15
Texto 1	7	0	1	1	0
Texto 2	1	1	0	1	0
Texto	0	0	1	0	0
Texto 2225	1	1	1	0	1

Representação TF



Algoritmo k-means

Algoritmo de clusterização implementado em java (~700 linhas)

Entradas: 2225 textos

Saída: k grupos

```
// iniciacao aleatoria dos centroides
iniciarCentroides(centroides, numClusters, tamCorpus);

do {
   // limpar as variaveis que armazenam os grupos
   limparGrupos(grupos);

   //determinar o centroide de cada elemento
   adicionaNoCluster(centroides, pontos, grupos);
   novosCentroides = new ArrayList <Coordinate>();

   //calculo dos novos centroides
   novosCentroides = novosCentros (centroides, grupos, novosCentroides);
   epocas --;
} while(condicaoDeParada (centroides,novosCentroides) && epocas > 0);
```

```
//inicializando o vetor de centroides aleatorios
public void iniciarCentroides(ArrayList <Coordinate> centroides, int numClusters,
int tamCorpus) {

  for(int i = 0; i < numClusters; i++) {
     Coordinate centro = new Coordinate();
     double [] coordenadas = new double[dimensao];

     //definindo aleatoriamente as coordenadas de cada centro
     for(int j = 0; j < dimensao; j++){
       coordenadas[j] = random.nextInt((int)(maiorValor[j]+1));
     }
     centro.setCoordenadas(coordenadas);

     centroides.add(i, centro);

  }
}
```

```java
public void adicionaNoCluster(ArrayList <Coordinate> centroides, ArrayList
<Coordinate> pontos, Coordinate [][] grupos) {
    double menorDist; int cluster;
    for(int p = 0; p < pontos.size(); p++) {
        menorDist = Double.MAX_VALUE; cluster = 0;

        //para cada ponto encontrar o centroide mais proximo
        for(int c = 0; c < centroides.size(); c++) {
            double distAtual = distancias(pontos.get(p), centroides.get(c));
            if(distAtual < menorDist) {
                menorDist = distAtual; cluster = c;
                pontos.get(p).setClusters(cluster);
            }
        }
    //adiciona esse ponto no cluster mais proximo
    for(int i = 0; i < pontos.size(); i++) {
        if(grupos[cluster][i] == null) {
           grupos[cluster][i] = pontos.get(p);
           break;
        }
    }
}
```

```
public ArrayList<Coordinate> novosCentros (ArrayList <Coordinate> centroides, Coordinate [][]
grupos,
ArrayList <Coordinate> novosCentroides){
    int c;
    for(int k = 0; k < centroides.size(); k++) {
    double [] novaCoordenada = new double[dimensao];

        for(int j = 0; j < dimensao; j ++){
           Coordinate centro = centroides.get(k);
           novaCoordenada[j] = centro.getCoordenadas()[j];
        }
        for(int i = 0; i < dimensao; i ++){
            for(c = 0; c < tamCorpus && grupos[k][c] != null; c++) {
                novaCoordenada[i] = Math.abs(novaCoordenada[i] + grupos[k][c].getCoordenadas()[i]);
            }
            novaCoordenada[i] = novaCoordenada[i]/(c+1);
        }
        Coordinate novoCentro = new Coordinate(novaCoordenada);
        novosCentroides.add(k, novoCentro);
    }
    return novosCentroides;
}
```

Kmeans.java		
Leitura dos dados	**Execução do algoritmo**	**Arquivos para pós-processamento**
lercsv( ) : void lerPalavrasCorpus( ) : void	executarKmeans( ) : void iniciarCentroides( ) : void adicionaNoCluster( ) : void distanciaEuclidiana( ) : double similaridadeCoseno( ) : double distancias( ) : double novosCentros : ArrayList<Coordinate> condicaoDeParada( ) : boolean limparGrupos( ) : void	gerarArquivoCsvGrafico( ) : void gerarArquivoCsv( ) : void gerarArquivoNuvemPalavra( ) : void gerarLogFinal( ) : void

Após a execução do algoritmo, como saber se obtivemos um bom resultado?

Saídas do algoritmo

Log de execução

Para cada iteração do algoritmo, como os textos foram agrupados

```
grupo 0 - filhos: 520
grupo 1 - filhos: 350
grupo 2 - filhos: 577
grupo 3 - filhos: 427
grupo 4 - filhos: 351

grupo 0 - filhos: 520
grupo 1 - filhos: 350
grupo 2 - filhos: 577
grupo 3 - filhos: 427
grupo 4 - filhos: 351
```

Saídas do algoritmo

$$s(i) = \frac{b(i) - a(i)}{max(a(i), b(i))}$$

$a(i)$ - distância do dado i a todos os demais dados do seu grupo

$b(i)$ - distância do dado i a todos os demais dados que não pertencem ao seu grupo

Silhouette - cálculo do quão bem representado aquele texto está no grupo

Saídas do algoritmo

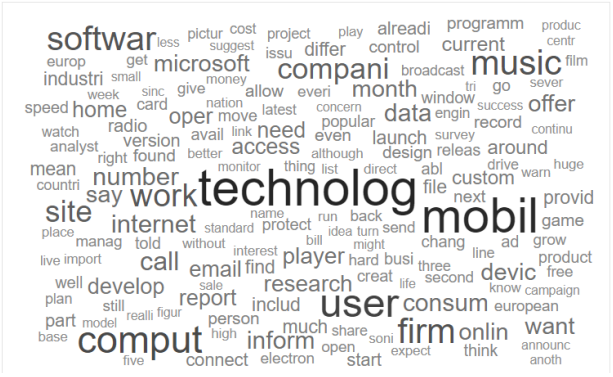
Arquivos para pós-processamento

Nuvem de palavras para cada grupo

https://worditout.com/word-cloud/create

```
problem 85
work    119
sell    86
largest 83
statement   99
deutsch 107
fraud   92
gazprom 73
stock   190
propos  95
list    90
glazer  74
```

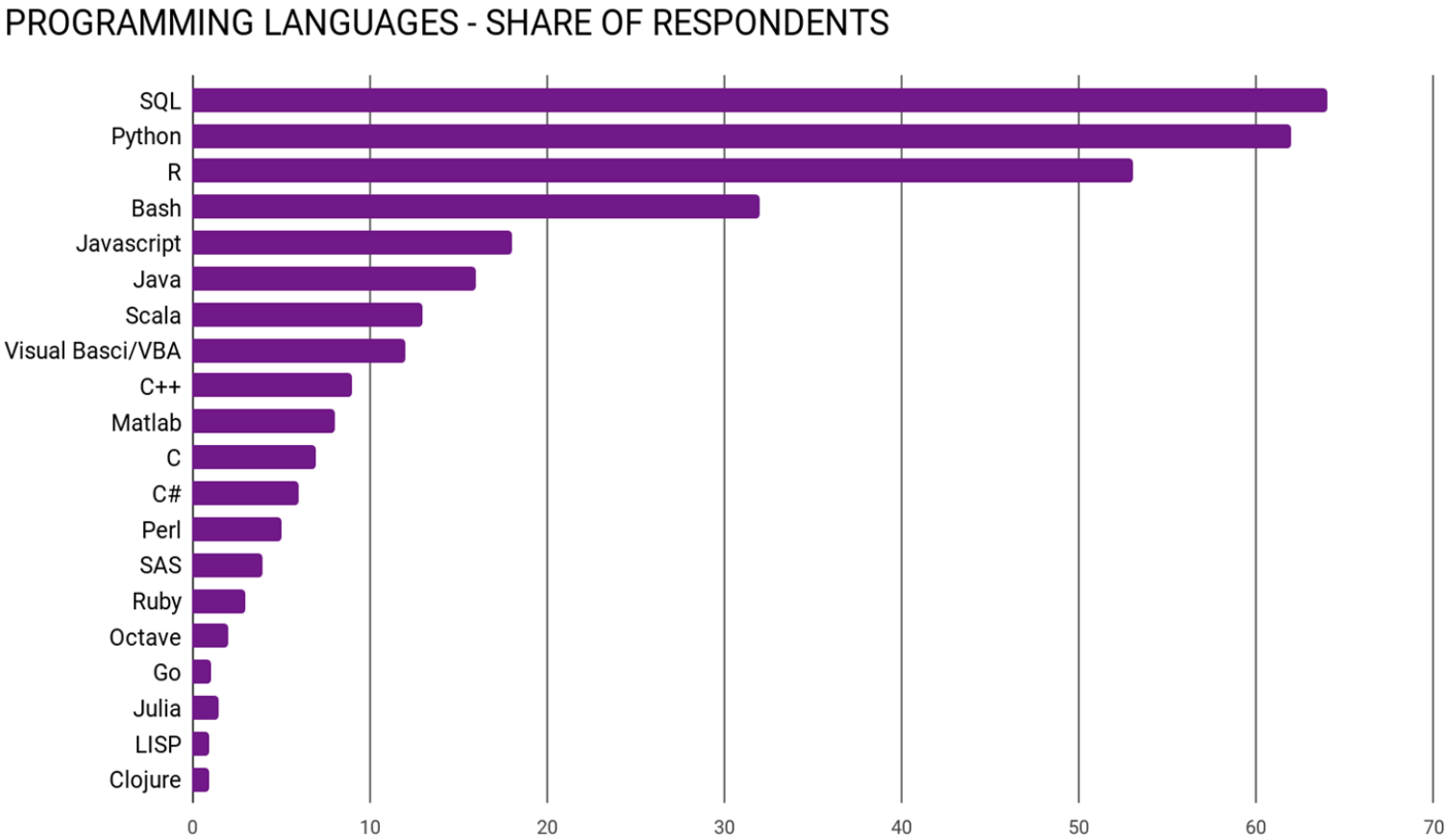
Resultado do agrupamento



Categorias corpus BBC

Qtd. textos	Categorias
510	Negócios
386	Entretenimento
417	Política
511	Esportes
401	Tecnologia

Qual a melhor linguagem de programação para Machine Learning?



Ferramentas mais utilizadas de Data science (Fonte: O'Reily Data science Survey 2017)

Python x Java

1. Syntax
2. Community Support
3. Documentation
4. Performance

   Java: ~3 minutos de execução | **Python: ~**10 minutos

Obrigada!

Carla Vieira

carlaprv@hotmail.com

@carlaprvieira

Referências

Introdução aos fundamentos de Machine Learning

- https://www.freecodecamp.org/news/a-history-of-machine-translation-from-the-cold-war-to-deep-learning-f1d335ce8b5/
- https://vas3k.com/blog/machine_learning/
- https://sebastianraschka.com/blog/2015/why-python.html
- https://towardsdatascience.com/do-you-know-how-to-choose-the-right-machine-learning-algorithm-among-7-different-types-295d0b0c7f60

Referências

Python

- https://www.cursoemvideo.com/course/curso-python-3/
- https://paulovasconcellos.com.br/10-bibliotecas-de-data-science-para-python-que-ningu%C3%A9m-te-conta-706ec3c4fcef
- https://cappra.com.br/2018/12/27/data-thinking-2019/
- https://discuss.analyticsvidhya.com/t/difference-between-nlp-and-text-mining/2977/2
- https://www.infoq.com/br/presentations/processamento-de-linguagem-natural-com-deep-learning/
- https://github.com/fernandojvdasilva/nlp-python-lectures

Referências

Machine Learning

- https://stanford.edu/~shervine/l/pt/teaching/cs-229/dicas-truques-aprendizado-maquina
- https://stanford.edu/~shervine/l/pt/teaching/cs-229/dicas-aprendizado-supervisionado
- https://stanford.edu/~shervine/l/pt/teaching/cs-229/dicas-aprendizado-nao-supervisionado
- https://hackernoon.com/best-machine-learning-libraries-for-java-development-4eccb88e1348
- https://medium.com/nexo-ai/machine-learning-x-deep-learning-qual-a-diferen%C3%A7a-entre-eles-665c0739f78a

Referências

Java x Python for NLP

- https://stackoverflow.com/questions/22904025/java-or-python-for-natural-language-processing