Regressão Linear

Prática 08: Clusterização

Profª Deborah Magalhães
Monitor: Davi Luis de Oliveira

Universidade
Federal do Piauí

Olá!

Curso: Bacharelado em Sistema de Informação

Disciplina: Sistemas Inteligentes

- **K-médias e medida de similaridade**
- **Estudo de caso: Recuperação de documentos**

Você pode me encontrar em **deborah.vm@gmail.com**
(Dúvidas e sugestões serão bem-vindas =D)

Estudo de caso: recuperação de documentos

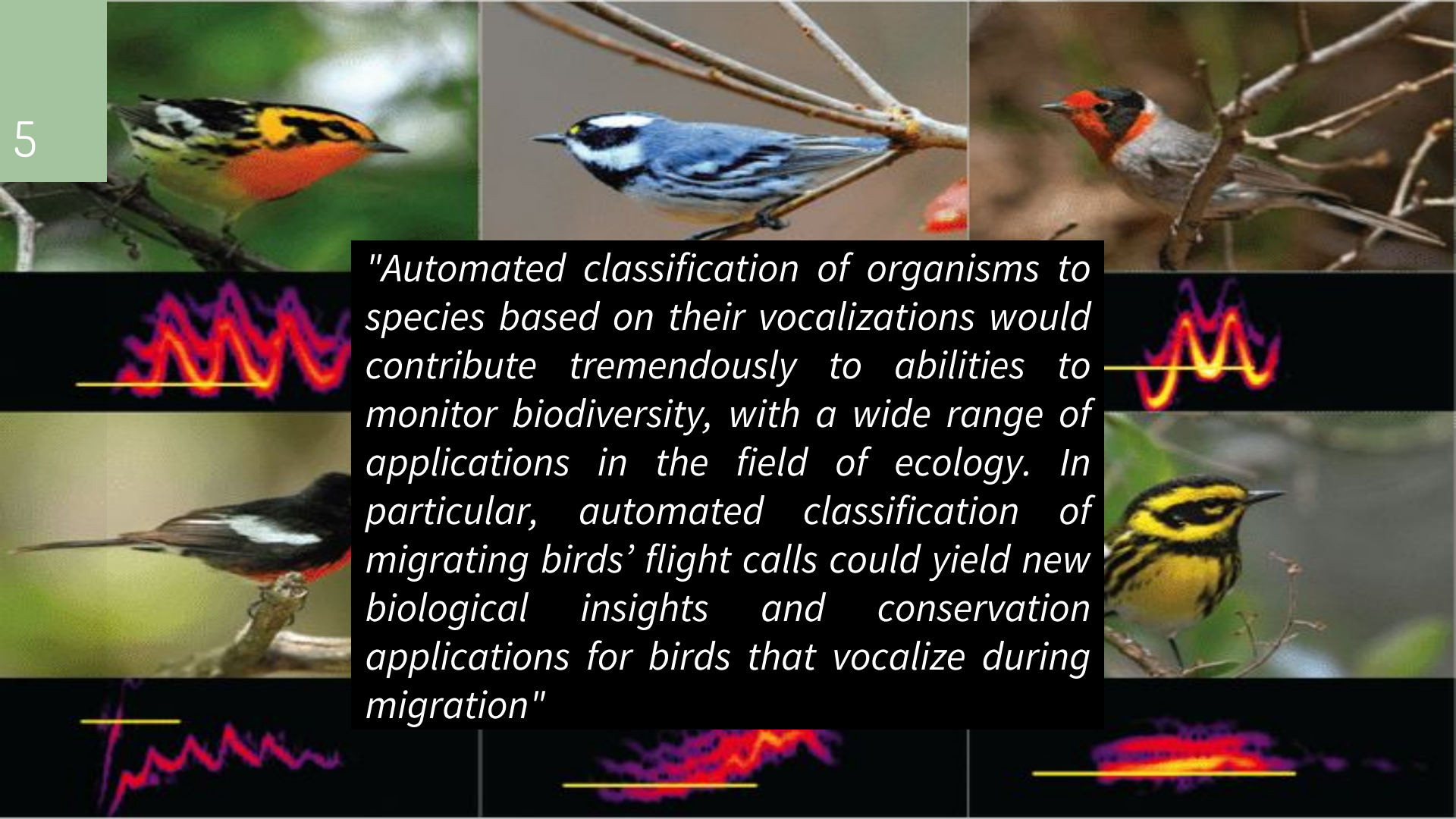
Atualmente, você está lendo um artigo para seu TCC e precisa encontrar um artigo similar para escrever seus trabalhos relacionados

”

Desafios

Como vamos medir a similaridade entre os artigos?

Entre todos os artigos existentes, como selecionar quais serão lidos?

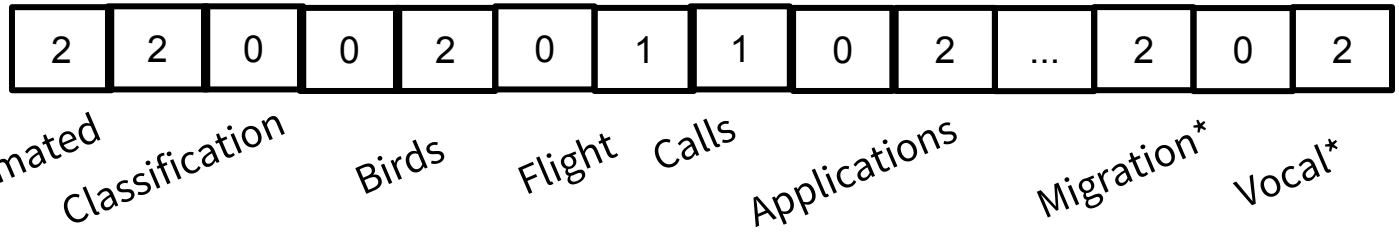


"Automated classification of organisms to species based on their vocalizations would contribute tremendously to abilities to monitor biodiversity, with a wide range of applications in the field of ecology. In particular, automated classification of migrating birds' flight calls could yield new biological insights and conservation applications for birds that vocalize during migration"

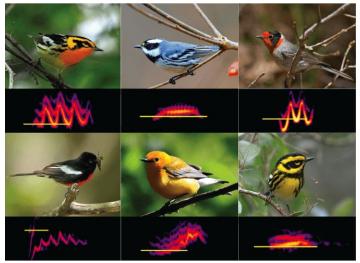
Modelo "Saco de Palavras"

- Ignora a ordem das palavras
- Conta o número de ocorrência das palavras

*"**Automated classification** of organisms to **species** based on their **vocalizations** would contribute tremendously to abilities to **monitor biodiversity**, with a wide range of **applications** in the field of **ecology**. In particular, **automated classification** of **migrating birds' flight calls** could yield new **biological** insights and conservation **applications** for **birds** that **vocalize** during **migration**"*



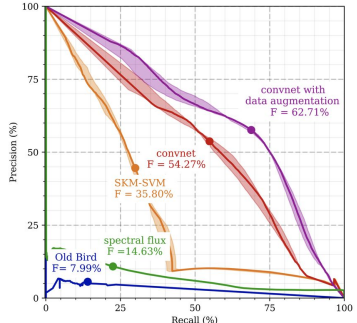
Medindo Similaridade



*"**Automated classification** of organisms to **species** based on their **vocalizations** would contribute tremendously to abilities to **monitor biodiversity**, with a wide range of **applications** in the field of **ecology**. In particular, **automated classification** of **migrating birds' flight calls** could yield new **biological** insights and conservation **applications** for **birds** that **vocalize** during **migration**"*

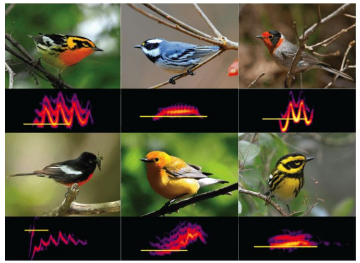
2	2	0	0	2	0	1	1	0	2	...	2	0	2
0	0	0	0	1	0	1	1	0	0	...	1	0	1

2*1 + 1*1 + 1*1 + 2*1 + 2*1 = 8



*"This article addresses the automatic detection of **vocal**, nocturnally **migrating birds** from a network of acoustic sensors. We correlate recall with the density of **flight calls** over time and frequency and identify the main causes of false alarm."*

Medindo Similaridade



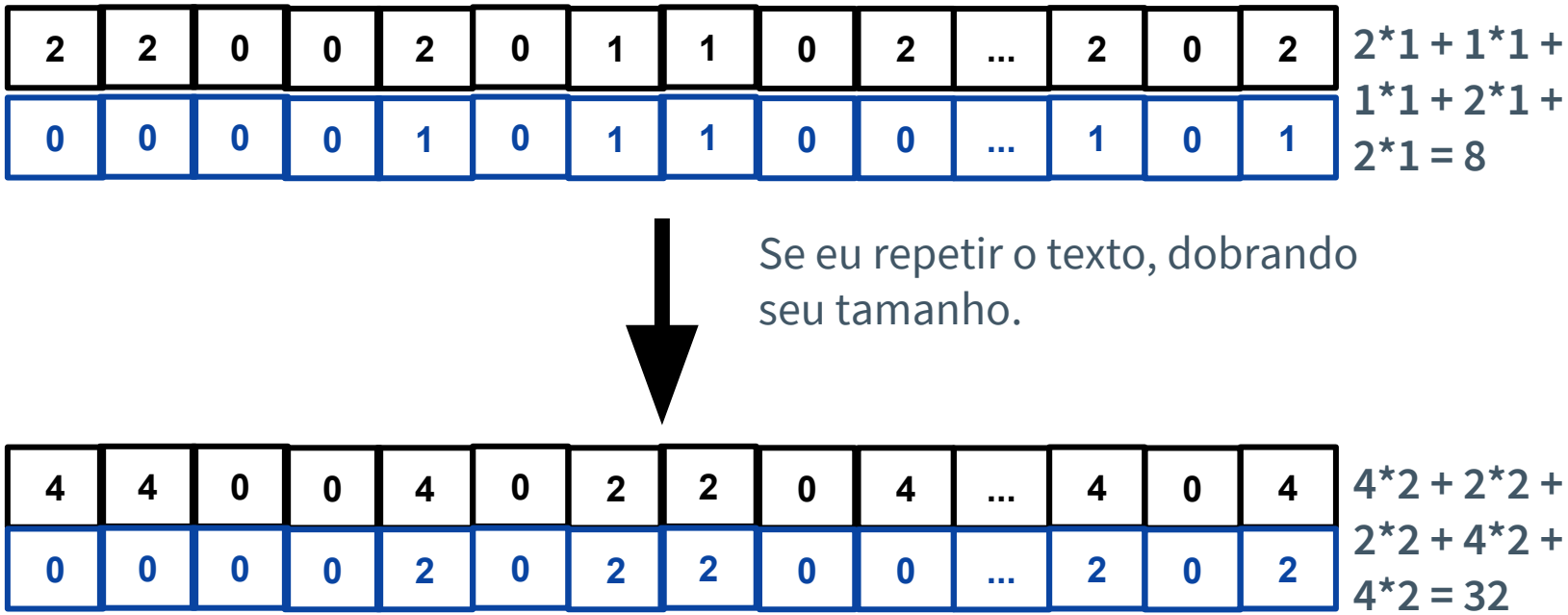
*"**Automated classification** of organisms to **species** based on their **vocalizations** would contribute tremendously to abilities to **monitor biodiversity**, with a wide range of **applications** in the field of **ecology**. In particular, **automated classification** of **migrating birds**' **flight calls** could yield new **biological** insights and conservation **applications** for **birds** that **vocalize** during **migration**"*

2	2	0	0	2	0	1	1	0	2	...	2	0	2
0	0	0	0	0	0	0	0	0	0	...	0	0	0

0

"Brazil plays its final friendly in preparation for the World Cup when it visits Austria on Sunday. It's another opportunity for this team to gel, further implement Neymar since his return from a foot injury and build some more positive vibes ahead of Russia 2018"

Problema: tamanho do documento



Solução: normalizar

2	2	0	0	2	0	1	1	0	2	...	2	0	2

$\sqrt{(2)^2 + (2)^2 + (2)^2 + (1)^2 + (1)^2 + (2)^2 + (2)^2 + (2)^2} = \sqrt{4+4+4+1+1+4+4+4} = \sqrt{26} = 5.09$

4	4	0	0	4	0	2	2	0	4	...	4	0	4

$\sqrt{(4)^2 + (4)^2 + (4)^2 + (2)^2 + (2)^2 + (4)^2 + (4)^2 + (4)^2} = \sqrt{16+16+16+4+4+16+16+16} = \sqrt{104} = 10.19$

Problema: enfatizar as palavras que realmente importam

- Palavras comum: ocorrem com frequência no texto
  - It, that, a, and, in, to, with, for
  - Dominam a medida de similaridade
- Palavras raras: que não ocorrem com tanta frequência mas de fato descrevem os dados
  - Migrating, birds, flight, calls

Solução: atribuir pesos

- Cada palavra possui um peso e, uma parcela é descontada de acordo com a frequência que a palavra aparece em diversos textos
- O que define uma palavra importante?
  - Comum localmente: aparece diversas vezes em um documento
  - Aparece raramente em diversos documentos (globalmente rara)
  - Palavra importante: tradeoff entre os dois

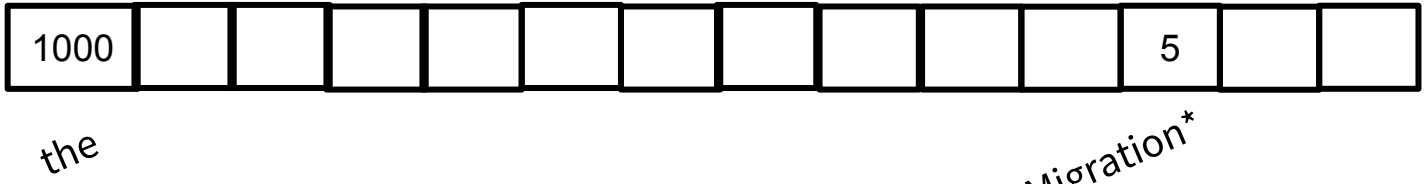
Representação de um documento (TF-IDF)

- TF: term-frequency
- IDF: inverse document frequency

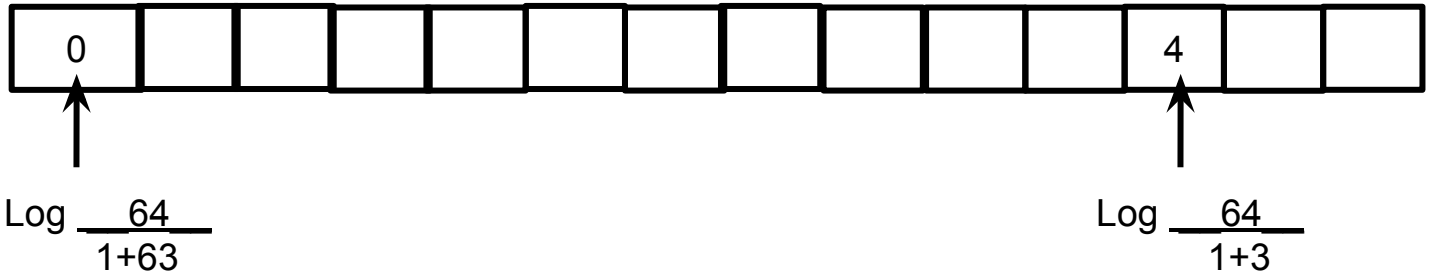
$$\log \frac{\#\ \text{docs}}{1 + \#\ \text{docs using word}}$$

Representação de um documento (TF-IDF)

- TF: term-frequency



- IDF: inverse document frequency



Agrupando (clustering) documentos

”

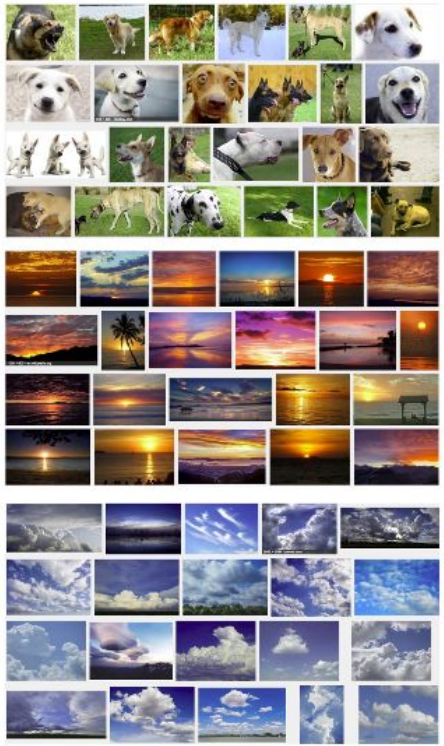
Problema de classificação com múltiplas classes

ESPORTES

NOTÍCIAS

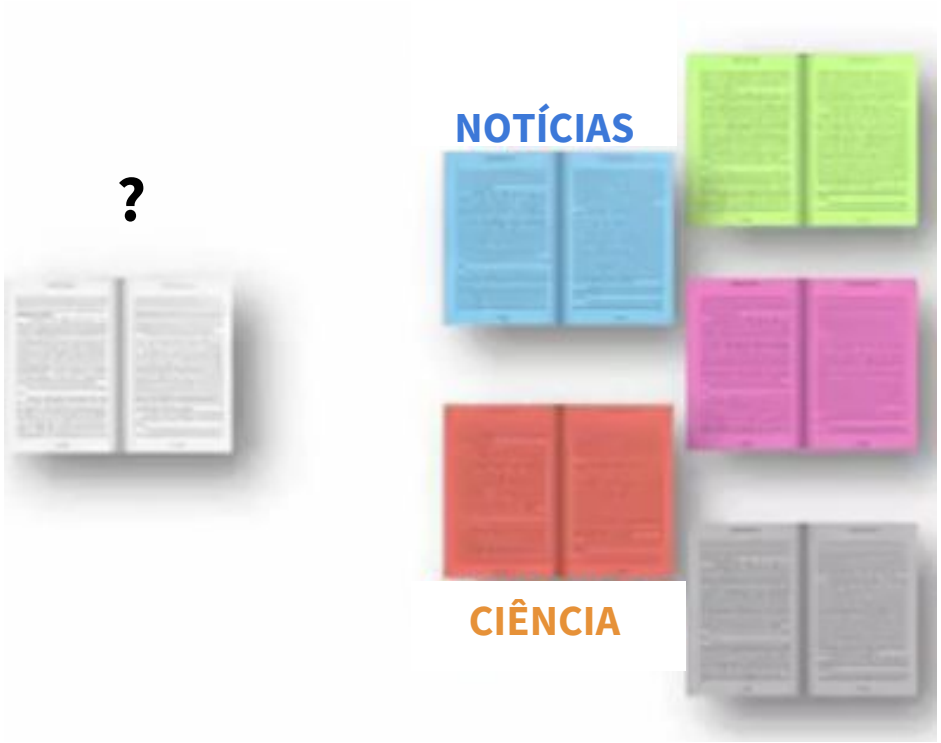
ENTRETENIMENTO

CIÊNCIA



Problema de classificação com múltiplas classes

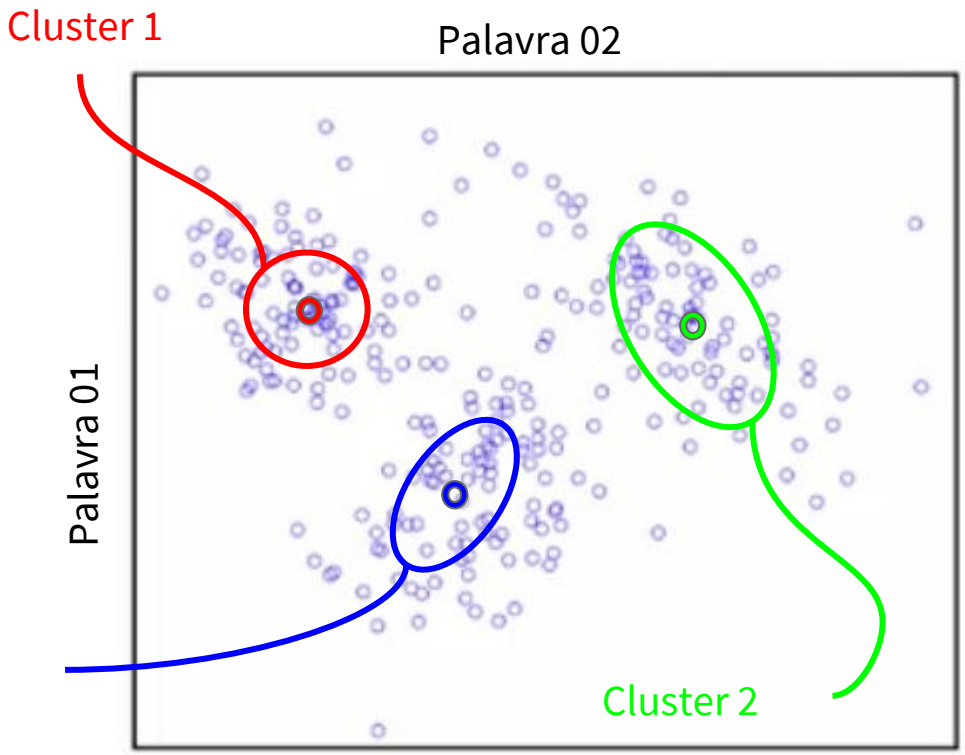
APRENDIZAGEM SUPERVISIONADA



Clustering

APRENDIZAGEM não SUPERVISIONADA

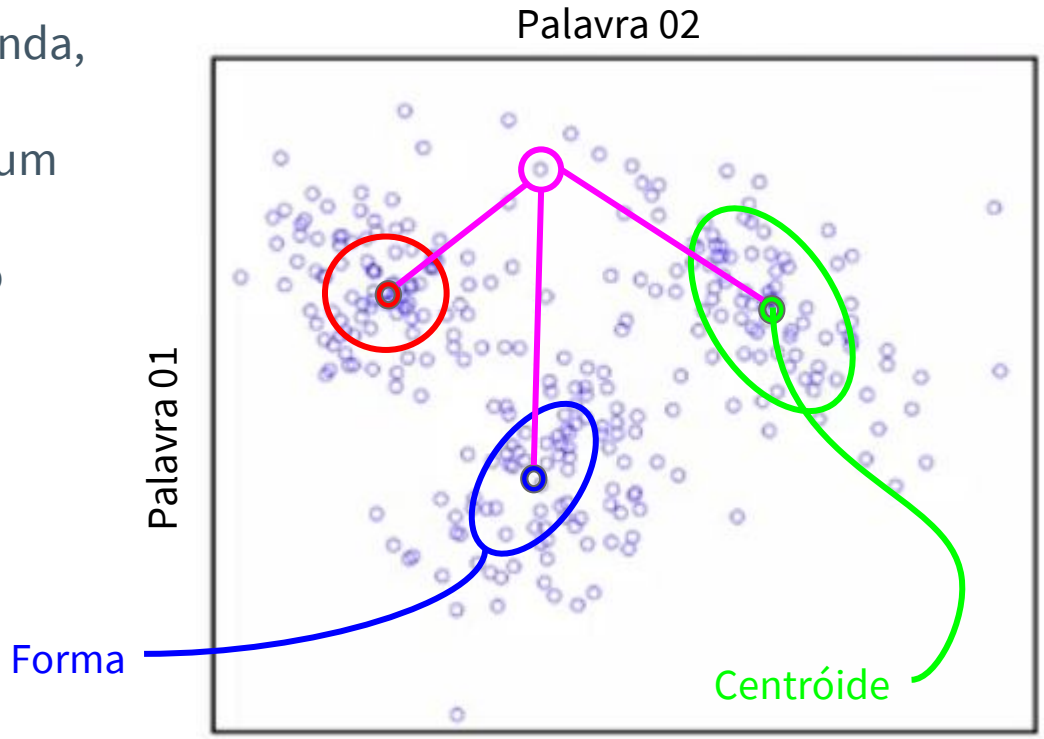
- Nenhum rótulo é provido
- Entrada: vetores de palavras representando os documentos
- Saída: rótulos dos clusters



Como é definido um cluster

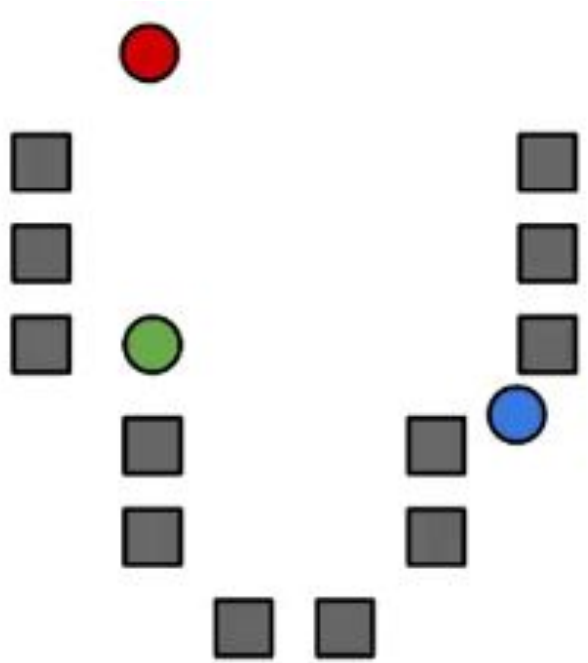
APRENDIZAGEM não SUPERVISIONADA

- Centróide uma forma (redonda, elíptica)
- Como associar um artigo à um cluster?
  - Menor distância para o centro do cluster



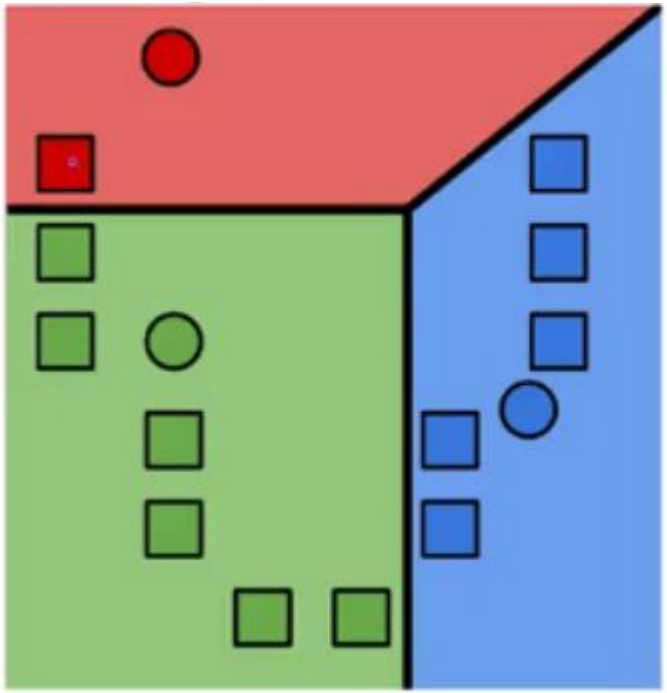
Algoritmo K-médias

1. Inicializar os centróides dos cluster (randomicamente)



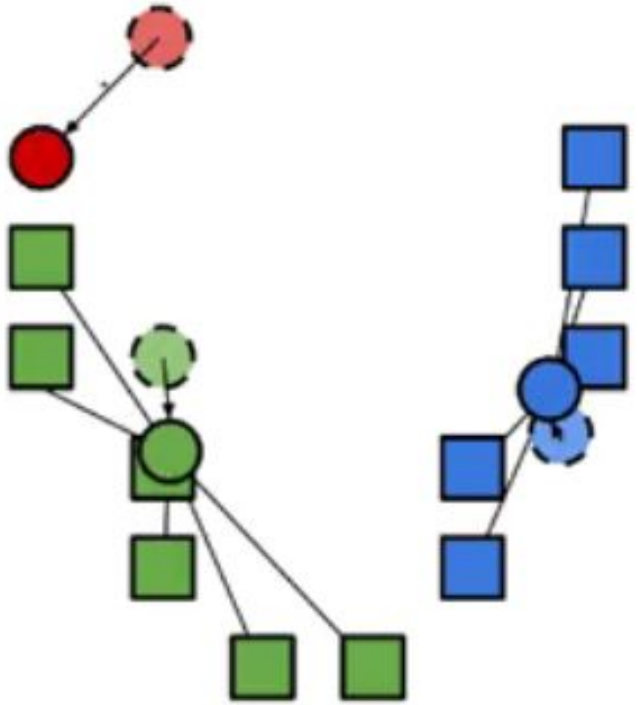
Algoritmo K-médias

1. Inicializar os centróides dos cluster (randomicamente)
2. Associar cada observação ao centróide do cluster mais próximo (célula de Voronoi)



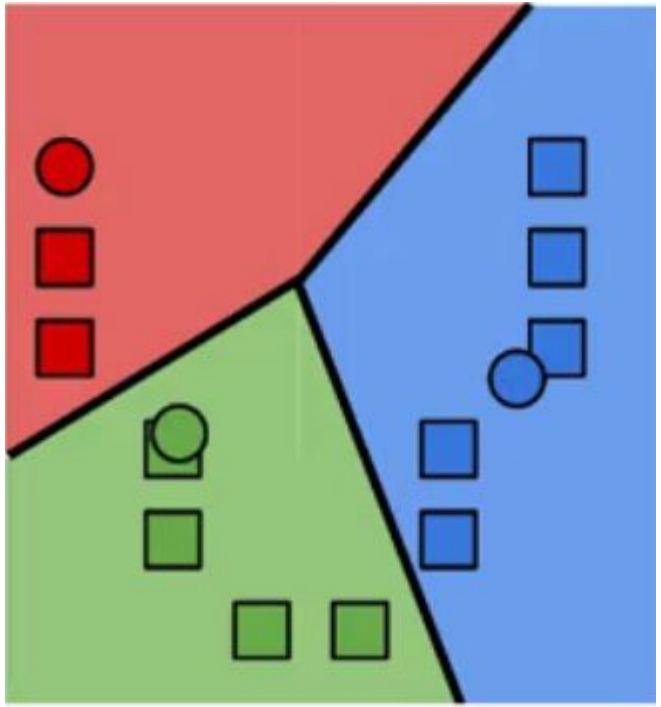
Algoritmo K-médias

1. Inicializar os centróides dos cluster (randomicamente)
2. Associar cada observação ao centróide do cluster mais próximo (célula de Voronoi)
3. A cada nova iteração, novas amostras chegam e a posição dos centróides é recalculada através da média



Algoritmo K-médias

1. Inicializar os centróides dos cluster (randomicamente)
2. Associar cada observação ao centróide do cluster mais próximo (célula de Voronoi)
3. A cada nova iteração, novas amostras chegam e a posição dos centróides é recalculada através da média
4. Repetir os passos 1 e 2 até convergir



Prática

”

Passo 1: Importar graphlab e carregar os dados

```
import graphlab
pessoas = graphlab.SFrame('people_wiki.gl/')
```

Passo 1: páginas de pessoas na Wikipedia

```
In [4]: len(pessoas)
Out[4]: 59071
```

```
In [5]: pessoas
```

Out[5]:

URI	name	text
<http://dbpedia.org/resource/Digby_Morrell> ...	Digby Morrell	digby morrell born 10 october 1979 is a former ...
<http://dbpedia.org/resource/Alfred_J._Lewy> ...	Alfred J. Lewy	alfred j lewy aka sandy lewy graduated from ...
<http://dbpedia.org/resource/Harpdog_Brown> ...	Harpdog Brown	harpdog brown is a singer and harmonica player who ...
<http://dbpedia.org/resource/Franz_Rottensteiner> ...	Franz Rottensteiner	franz rottensteiner born in waidmannsfeld lower ...
<http://dbpedia.org/resource/G-Enka> ...	G-Enka	henry krvits born 30 december 1974 in tallinn ...
<http://dbpedia.org/resource/Sam_Henderson> ...	Sam Henderson	sam henderson born october 18 1969 is an ...
<http://dbpedia.org/resource/Aaron_LaCrate> ...	Aaron LaCrate	aaron lacrate is an american music producer ...

Passo 1: páginas de pessoas na Wikipedia

```
bowie = pessoas[pessoas['name'] == 'David Bowie']
```

```
bowie
```

URI	name	text
<http://dbpedia.org/resource/David_Bowie> ...	David Bowie	david bowie born david robert jones 8 january ...

[? rows x 3 columns]

Note: Only the head of the SFrame is printed. This SFrame is lazily evaluated.

You can use sf.materialize() to force materialization.

Passo 1: páginas de pessoas na Wikipedia

```
bowie['text']
```

```
dtype: str
Rows: ?
['david bowie born david robert jones 8 january 1947 is an english singer songwriter multiins
trumentalist record producer arranger and actor he is also a painter and collector of fine ar
t bowie boi has been a major figure in the world of popular music for over four decades and i
s renowned as an innovator particularly for his work in the 1970s he is known for his distinc
tive voice as well as the intellectual depth and eclecticism of his work aside from his music
al abilities he is recognised for his androgynous beauty which was an iconic element to his i
mage particularly in the 1970s and 1980sbowie first caught the eye and ear of the public in j
uly 1969 when his song space oddity reached the top five of the uk singles chart after a thre
eyear period of experimentation he reemerged in 1972 during the glam rock era with the flambo
yant androgynous alter ego ziggy stardust spearheaded by the hit single starman and the album
the rise and fall of ziggy stardust and the spiders from mars bowies impact at that time as d
escribed by biographer david buckley challenged the core belief of the rock music of its day
and created perhaps the biggest cult in popular culture the relatively shortlived ziggy perso
na proved merely one facet of a career marked by continual reinvention musical innovation and
striking visual presentationin 1975 bowie achieved his first major american crossover success
with the numberone single fame and the hit album young americans which the singer characteris
ed as plastic soul the sound constituted a radical shift in style that initially alienated ma
ny of his uk devotees he then confounded the expectations of both his record label and his am
erican audiences by recording the minimalist album low 1977the first of three collaborations
with brian eno over the next two years low heroes and lodger the socalled berlin trilogy albu
ms all reached the uk top five and received lasting critical praise after uneven commercial s
uccess in the late 1970s bowie had uk number ones with the 1980 single ashes to ashes its par
```

Passo 2: contar as palavras do "text"

```
bowie['word_count'] = graphlab.text_analytics.count_words(bowie['text']))
```

Passo 2: saída

```
bowie['word_count'] = graphlab.text_analytics.count_words(bowie['text'])
```

```
print bowie['word_count']
```

```
[{'all': 3, 'abilities': 1, 'particularly': 2, 'producer': 1, 'from': 2, 'dance': 1, 'over':
2, 'both': 1, 'contemporary': 1, 'years': 1, 'four': 1, 'known': 1, 'including': 1, 'spiders'
: 1, 'world': 1, '1977the': 1, 'eno': 1, 'fine': 1, 'its': 2, 'one': 1, 'constituted': 1, 'st
yle': 1, 'uneven': 1, 'gold': 2, 'also': 1, 'recognised': 1, 'had': 1, '29': 1, 'silver': 1,
'late': 1, 'to': 3, 'lets': 1, 'critical': 1, 'under': 1, '8': 1, 'of': 19, 'has': 7, 'decade
s': 1, 'alter': 1, 'styles': 1, 'then': 2, 'queen': 1, 'march': 1, 'song': 1, 'songwriter': 1
, 'innovation': 1, 'period': 1, 'next': 2, 'ones': 1, 'five': 3, 'intellectual': 1, 'fall': 1
, 'not': 2, 'during': 1, 'unique': 1, 'continued': 1, 'him': 1, 'minimalist': 1, 'success': 2
, 'impact': 1, 'january': 1, 'yielded': 1, 'list': 2, 'berlin': 1, 'public': 1, 'ashes': 2, '
rock': 2, 'popular': 3, '2000s': 1, '2013david': 1, 'expectations': 1, 'lasting': 1, 'says':
1, 'achieved': 1, '1947': 1, 'soul': 2, 'culture': 1, 'bowies': 2, 'stardust': 2, 'uk': 5, 'a
': 7, 'reached': 3, 'estimated': 1, 'collector': 1, 'best': 1, 'innovator': 1, 'album': 6, 't
hroughout': 2, 'beauty': 1, 'for': 4, 'placed': 1, 'distinctive': 1, 'ear': 1, 'since': 2, 'l
abel': 1, 'singles': 2, 'experiment': 1, 'socalled': 1, 'permeated': 1, '23rd': 1, 'britons':
1, 'new': 1, 'multiinstrumentalist': 1, 'monsters': 1, 'core': 1, 'comparable': 1, 'belief':
1, 'parent': 1, 'initially': 1, 'sold': 1, 'million': 1, 'commercial': 2, 'iconic': 1, 'press
ure': 1, 'ranked': 1, 'studio': 1, 'shortlived': 1, 'element': 1, 'singers': 1, 'eye': 1, '10
0': 2, 'super': 1, 'by': 4, 'plastic': 1, 'scary': 1, 'stone': 1, 'certifications': 2, 'creep
s': 1, 'english': 1, 'career': 2, 'many': 1, 'bowie': 7, 'eclecticism': 1, 'david': 3, 'facet
': 1, 'characterised': 1, 'tour': 1, '1970s': 3, 'buckley': 2, 'top': 2, 'first': 3, 'major':
2, 'industrial': 1, 'striking': 1, 'radical': 1, 'merely': 1, 'number': 2, 'marked': 1, 'two'
: 1, '39th': 1, 'cult': 1, 'americans': 1, 'artists': 1, 'reinvention': 1, '1990s': 1, 'poll'
: 1, 'experimentation': 1, 'platinum': 2, 'art': 1, 'described': 1, '1983': 1, '1980': 1, '19
```

Passo 3: transformar contagem em tabela

```
bowie_tabela_numero_palavras = bowie[['word_count']].stack('word_count',
new_column_name = ['word','count'])
```

Passo 3: saída

```
bowie_tabela_numero_palavras
```

word	count
on	2
39th	1
him	1
stone	1
silver	1
eight	1
singers	1
certifications	2
awarded	1
million	1

[291 rows x 2 columns]

Passo 4: ordenar as palavras por frequência

```
bowie_tabela_numero_palavras.sort('count',ascending=False)
```

Passo 4: saída da ordenação

```
bowie_tabela_numero_palavras.sort('count',ascending=False)
```

word	count
the	38
and	25
of	19
in	15
his	13
he	9
has	7
with	7
bowie	7
a	7

[291 rows x 2 columns]

Passo 5: cálcular o tfidf e adicionar resultado na tabela pessoas

```
tfidf = graphlab.text_analytics.tf_idf(pessoas['word_count'])
pessoas['tfidf']=tfidf
```

Passo 5: saída do cálculo e add na tabela

tfidf
{'selection': 3.836578553093086, ...
{'precise': 6.44320060695519, ...
{'just': 2.7007299687108643, ...
{'all': 1.6431112434912472, ...
{'they': 1.8993401178193898, ...
{'currently': 1.637088969126014, ...
{'exclusive': 10.455187230695827, ...
{'taxi': 6.0520214560945025, ...

Passo 6: explorando o TF-IDF para o Bowie

```
bowie = pessoas[pessoas['name'] == 'David Bowie']
```

```
bowie[['tfidf']].stack('tfidf',new_column_name=['word','tfidf']).sort('tfidf',ascending=False)
```

Passo 6: saída do TF-IDF para o Bowie

word	tfidf
bowie	42.2146783214
ziggy	23.972289347
androgynous	16.0841128201
bowies	15.883945903
certifications	15.2383991185
album	14.7075950119
stardust	14.545846645
uk	13.9530499896
buckley	12.8652369952
ashes	12.7827510782

Passo 7: o David Bowie é mais próximo ao Taylor Swift ou Arnold Schwarzenegger?

```
swift = pessoas[pessoas['name'] == 'Taylor Swift']
```

```
arnold = pessoas[pessoas['name'] == 'Arnold Schwarzenegger']
```

```
graphlab.distances.cosine(bowie['tfidf'][0], swift['tfidf'][0])
```

```
0.90731925097562 84
```

```
graphlab.distances.cosine(bowie['tfidf'][0], arnold['tfidf'][0])
```

```
0.9818825183588984
```

Passo 8: Que são as personalidades mais próximas ao David Bowie?

```
knn_model = graphlab.nearest_neighbors.create(pessoas,features=['tfidf'],label='name')
```

```
Starting brute force nearest neighbors model training.
```

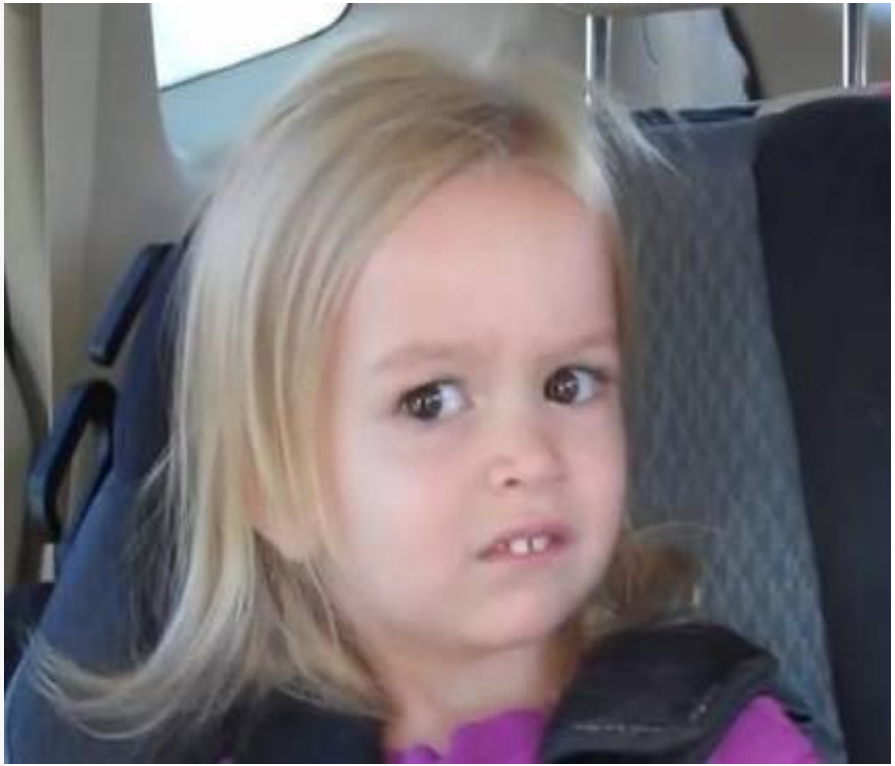
```
knn_model.query(bowie)
```

Passo 8: rank dos 5

query_label	reference_label	distance	rank
0	David Bowie	0.0	1
0	Phil Collins	0.80701754386	2
0	George Michael	0.819628647215	3
0	Alicia Keys	0.81975308642	4
0	Carrie Underwood	0.823255813953	5

Atividade prática individual (valendo 2.5 pts)

- Calcule a distância entre o Bruno Mars e o Barak Obama. Utilize pelo menos 2 distâncias diferentes da distância Cosine, utilizada na prática;
- Defina quem está mais próximo da Brad Pitt, o Ryan Gosling ou a Taylor Swift?
- Quais são as 5 personalidades mais próximas do Justin Bieber?



Dúvidas?Sugestões? Inquietações? Aconselhamentos?

- Desabafe em: **deborah.vm@gmail.com**