Fundação Vanzolini

Dominando Big Data com o uso de Plataformas Gratuitas (nível intermediário)

Aula 4

LexisNexis® RISK SOLUTIONS

RELX Group

PRO

ESCOLA POLITÉCNICA DA UNIVERSIDADE DE SÃO PAULO

USP UNIVERSIDADE DE SÃO PAULO

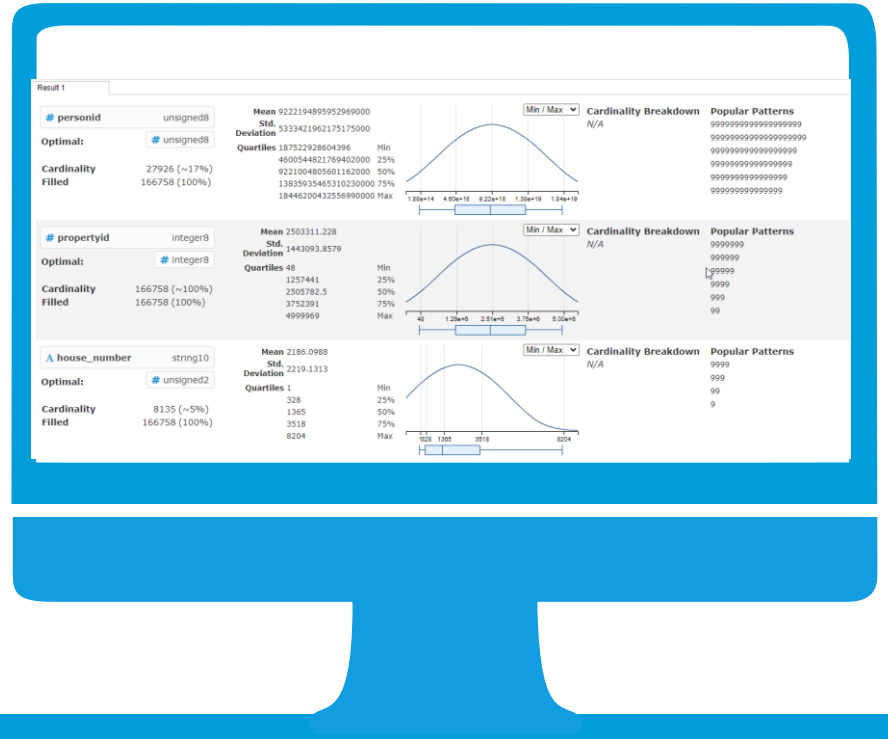
Bem-vindo! – Agenda da aula 4

- ✓Desafio Lending Club
- ✓DBScan
- ✓Intervalo
- ✓K-Means
- ✓NLP

Exercício prático:

Prepare o dataset do Lending Club

- Considere a aplicação de aprendizagem supervisionada
- Se baseie nos resultados do perfilamento de dados



Fundação Vanzolini

Aprendizagem não supervisionada

O que é Machine Learning?

- *"O estudo científico de algoritmos e modelos estatísticos que sistemas de computador usam para realizar uma tarefa específica sem usar instruções explícitas, baseando-se em padrões e inferência"*

- **Supervisionado -** quando apresentamos ao algoritmo dados de entrada e as respectivas saídas

- **Não supervisionado -** quando apresentamos somente os dados de entrada e o algoritmo descobre as saídas

- **Por reforço, profundo, etc -** o algoritmo utiliza tentativa e erro para encontrar uma solução para o problema, múltiplas camadas de aprendizado com dados complexos (imagens, vídeo, áudio), etc

Terminologia de ML

- Exemplo de aprendizado supervisionado:

  Dada uma amostra de registros:

```
Record1: Field1, Field2, Field3, … , FieldM
Record2: Field1, Field2, Field3, … , FieldM
…
RecordN: Field1, Field2, Field3, …, FieldM
```

Variáveis “Independentes”

  E um conjunto de valores a serem determinados,

```
Record1: TargetValue
Record2: TargetValue
…
RecordN: TargetValue
```

Variáveis “Dependentes”

  Aprenda a predizer valores para novas amostras.

Fundação Vanzolini

Exemplo prático de ML

- Dado o conjunto de dados sobre árvores em uma floresta:

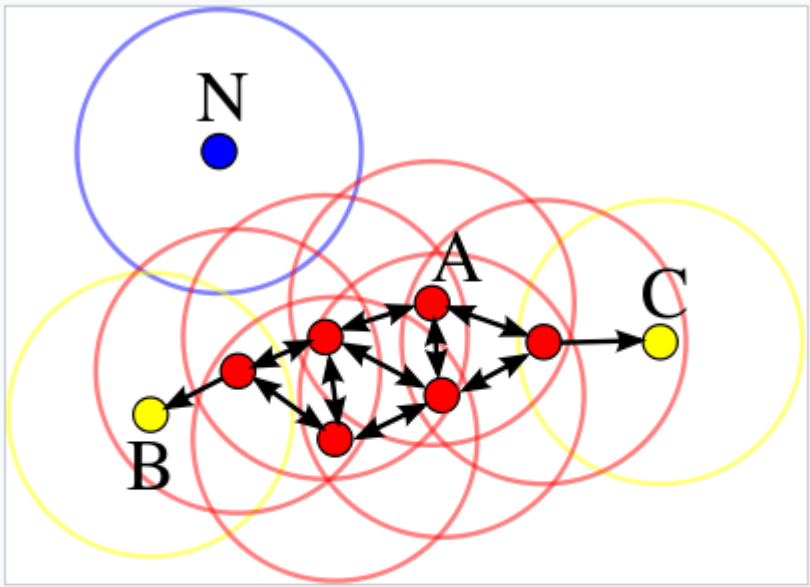
Altura	Diâmetro	Altitude	Pluviosidade	Idade
50	8	5000	12	80
56	9	4400	10	75
72	12	6500	18	60
47	10	5200	14	53

- Obtenha um modelo que determine a idade de uma árvore (variável dependente) a partir da sua altura, diâmetro, altura e pluviosidade do local (variáveis independentes).

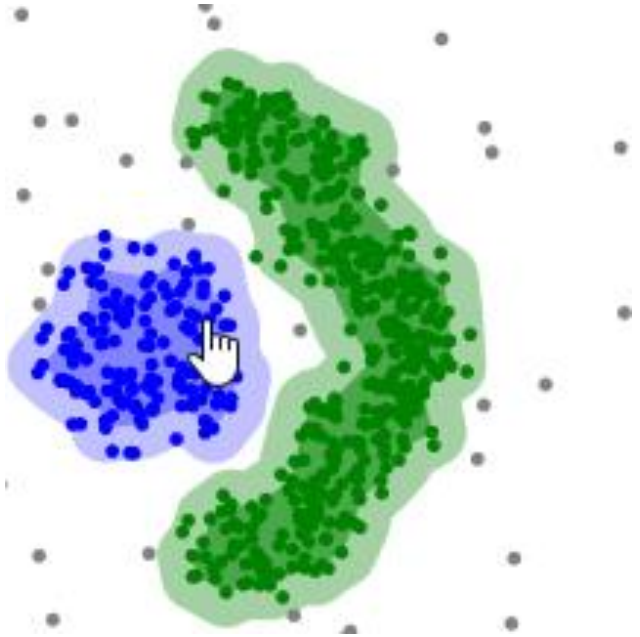
DBSCAN

Algoritmo de clusterização baseado na densidade de distribuição dos dados.

(https://hpccsystems.com/blog/DBSCAN)



In this diagram, minPts = 4. Point A and the other red points are core points, because the area surrounding these points in an ε radius contain at least 4 points (including the point itself). Because they are all reachable from one another, they form a single cluster. Points B and C are not core points, but are reachable from A (via other core points) and thus belong to the cluster as well. Point N is a noise point that is neither a core point nor directly-reachable.



Ref: https://en.wikipedia.org/wiki/DBSCAN

Tutorial de DBSCAN

Fluxo de aprendizagem de máquina

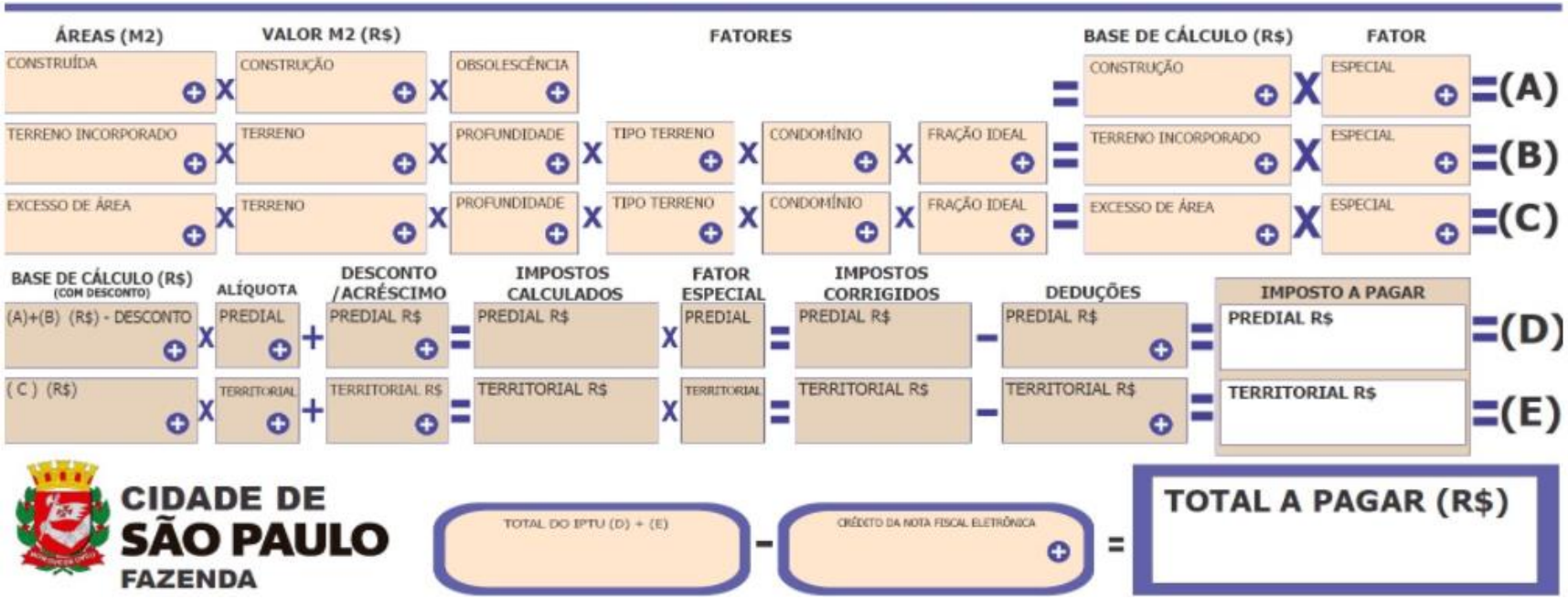
1. Definição do problema
2. Extração dos dados
3. Preparação dos dados
4. Conversão e segregação das variáveis
5. Treinamento e avaliação do modelo
6. Implantação do modelo

1. Definição do problema

“Dado um conjunto de atributos de uma propriedade (localização, metragem, ano de construção) é possível agrupá-los?"

http://geosampa.prefeitura.sp.gov.br/PaginasPublicas/_SBC.aspx

Property Tax formula: https://web1.sf.prefeitura.sp.gov.br/CartelaIPTU/



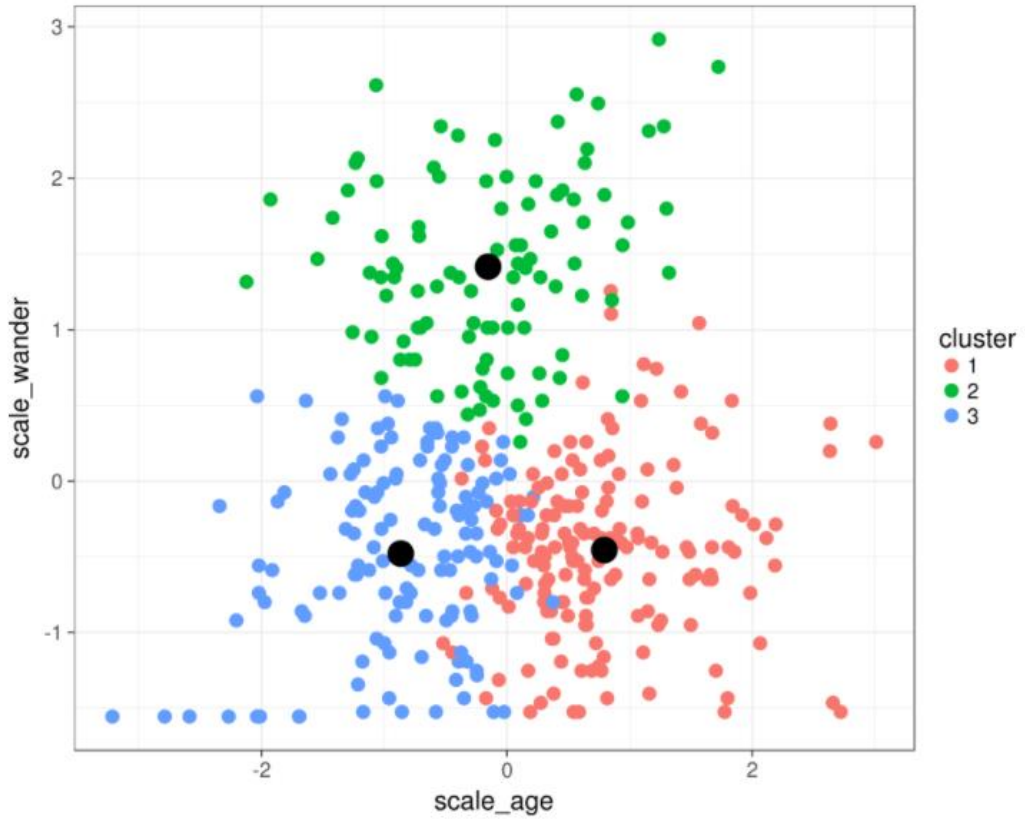
Fundação Vanzolini

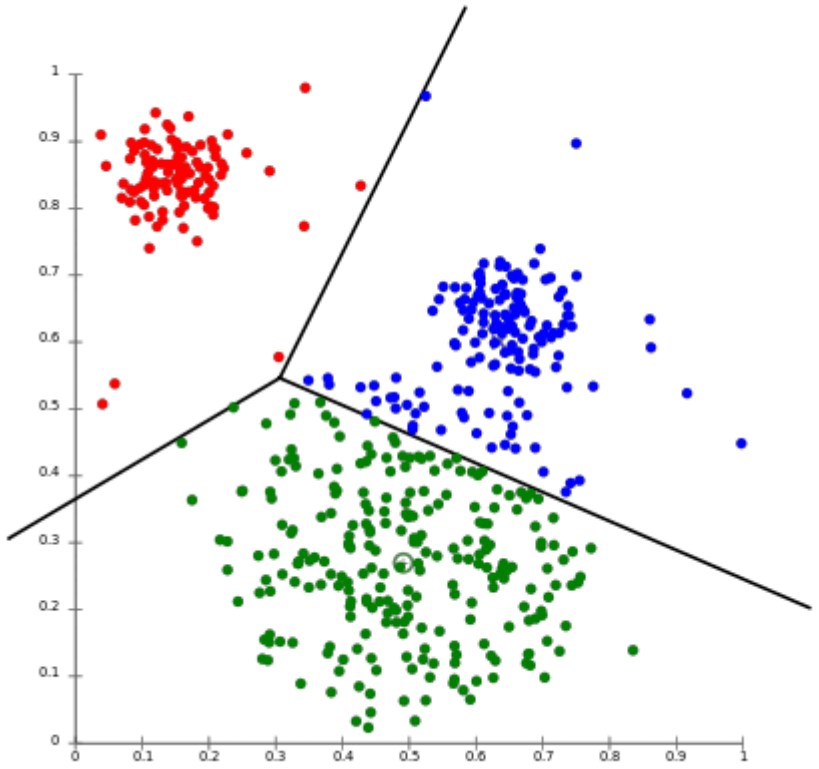
KMeans

K-Means

Algoritmo de clusterização para agrupamento de dados similares em um número pré-definido de grupos.

(https://hpccsystems.com/blog/kmeans)





Ref: https://rpubs.com/cyobero/k-means

Fundação Vanzolini

Tutorial de KMeans

Fluxo de aprendizagem de máquina

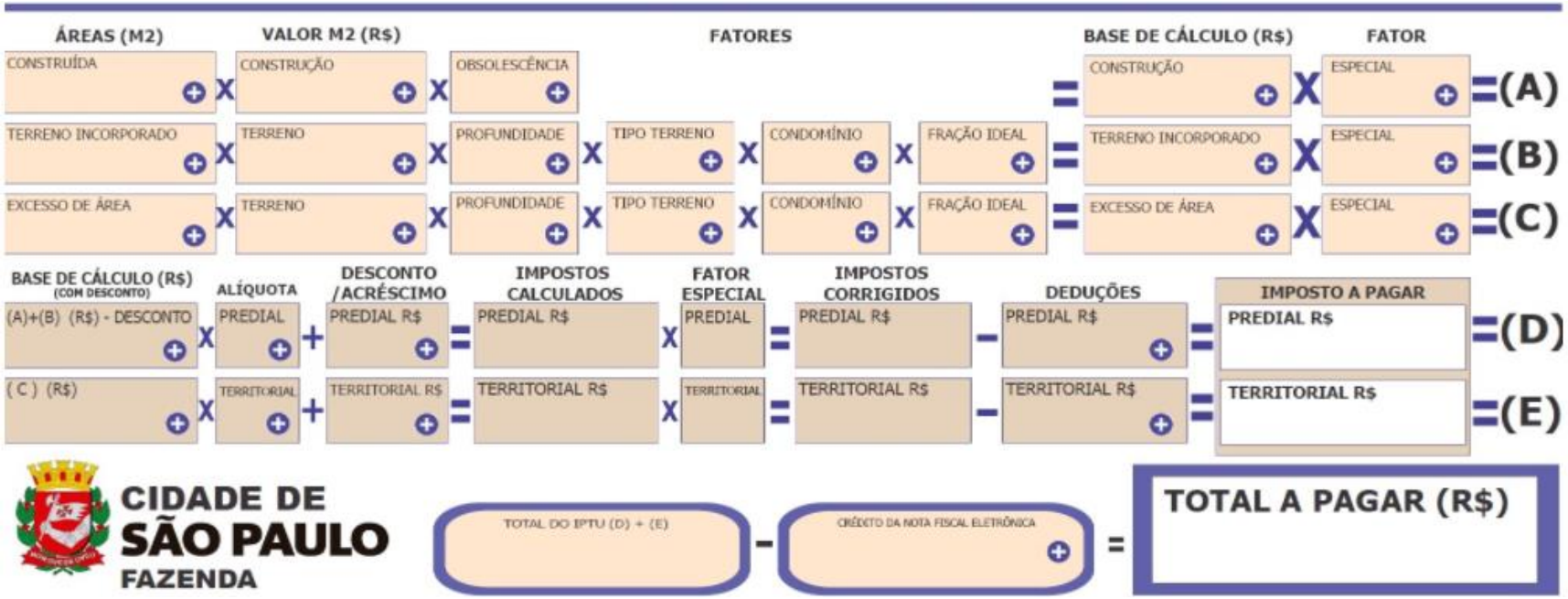
1. Definição do problema
2. Extração dos dados
3. Preparação dos dados
4. Conversão e segregação das variáveis
5. Treinamento e avaliação do modelo
6. Implantação do modelo

1. Definição do problema

“Dado um conjunto de atributos de uma propriedade (localização, metragem, ano de construção) é possível ordenar os seus outliers?"

http://geosampa.prefeitura.sp.gov.br/PaginasPublicas/_SBC.aspx

Property Tax formula: https://web1.sf.prefeitura.sp.gov.br/CartelaIPTU/



Fundação Vanzolini

Processamento de linguagem natural

Arquivos de texto

```
listing_id,id,date,reviewer_id,reviewer_name,comments
7202016,38917982,2015-07-19,28943674,Bianca,Cute and cozy place. Perfect location to everything!
7202016,39087409,2015-07-20,32440555,Frank,"Kelly has a great room in a very central location.
Beautiful building , architecture and a style that we really like.
We felt guite at home here and wish we had spent more time.
Went for a walk and found Seattle Center with a major food festival in progress. What a treat.
Visited the Space Needle and the Chihuly Glass exhibit. Then Pikes Place Market. WOW. Thanks for a great stay."
7202016,39820030,2015-07-26,37722850,Ian,"Very spacious apartment, and in a great neighborhood. This is the kind of apartment I wish I had!
Didn't really get to meet Kelly until I was on my out, but she was always readily available by phone.
I believe the only ""issue"" (if you want to call it that) was finding a place to park, but I sincerely doubt its easy to park anywhere in a residential area after 5 pm on a Friday"
7202016,40813543,2015-08-02,33671805,George,"Close to Seattle Center and all it has to offer - ballet, theater, museum, Space Needle, restaurants of all ilk just blocks away, and the Metropol
7202016,41986501,2015-08-10,34959538,Ming,"Kelly was a great host and very accommodating in a great neighborhood. She has some great coffee and while I wasn't around much during my stay the t
The apartment is in a great location and very close to the Seattle Center. The neighborhood itself has a lot of good food as well!"
7202016,43979139,2015-08-23,1154501,Barent,"Kelly was great, place was great, just what I was looking for-
clean, simple, well kept place.
```

```
#28 (2002) ->->->->->->2002
#7 Train: An Immigrant Journey, The (2000) ->->2000
$ (1971) ->->->->->->1971
$1,000 Reward (1913) ->->->->->1913
$1,000 Reward (1915) ->->->->->1915
$1,000 Reward (1923) ->->->->->1923
$1,000,000 Duck (1971) ->->->->->1971
$1,000,000 Reward, The (1920) ->->->->1920
$10,000 Under a Pillow (1921) ->->->->1921
$100,000 (1915) ->->->->->1915
$100,000 Pyramid, The (2001) (VG) ->->->2001
$1000 a Touchdown (1939) ->->->->1939
$20,000 Carat, The (1913) ->->->->1913
$21 a Day Once a Month (1941) ->->->->1941
$2500 Bride, The (1912) ->->->->1912
$30 (1999) ->->->->->1999
$30,000 (1920) ->->->->->1920
$300 y tickets (2002) ->->->->->2002
$40,000 (1996) ->->->->->1996
```

```
SOH NUL + NUL NUL NUL <area code="201" zone="Eastern Time Zone"/> STX NUL + NUL NUL NUL <area
NUL + NUL NUL NUL <area code="210" zone="Central Time Zone"/> VT NUL + NUL NUL NUL <area cod
NUL + NUL NUL NUL <area code="214" zone="Central Time Zone"/> SO NUL x NUL NUL NUL <area cod
SOH NUL NUL NUL <area code="835" description="PA Pennsylvania (Reading, Allentown, and
SOH ` NUL NUL NUL <area code="845" description="NY New York ( Poughkeepsie, Middletown,
```


Fundação Vanzolini

Regular Expressions (REGEX)

- Operadores para descrever **padrões e conjuntos** de cadeias de caracteres:
  - [A-Z]
    Qualquer caracter de “A” a “Z”
  - [a-z]
    Qualquer caracter de “a” a “z”
  - [A-Za-z]
    Qualquer caracter maiúsculo ou minúsculo
  - [A-Z][a-z]
    Qualquer caracter maiúsculo, seguido de um minúsculo. Ex.: “Oi”
  - [A-Z][a-z]+
    Qualquer caracter maiúsculo, seguido de um ou mais caracteres minúsculos. Ex.: “Oie”
  - [A-Z][a-z]?
    Qualquer caracter maiúsculo, seguido de nenhum ou um caracter minúsculo. Ex.: “Oi” e “A”
  - [A-Z][a-z]+ | [0-9]+
    Qualquer caracter maiúsculo, seguido de um ou mais caracteres minúsculos OU um ou mais dígitos. Ex.: “Oie” ou “1990”

Fundação Vanzolini

Regular Expressions (REGEX)

- ✓ Identificador de **caracteres especiais;**
  - ✓ \t = tab;
  - ✓ \n = quebra de linha;
  - ✓ \f = quebra de página.

- ✓ Descrever **conjuntos** de cadeias de caracteres;
  - ✓ El(e|a) representa as cadeias “Ele” e “Ela”;
  - ✓ A(u?)dição representa as cadeias “Adição” e “Audição”.

- ✓ Operadores para descrever e especificar **padrões**:
  - ✓ ? 0 **ou** 1 ocorrências da expressão precedente;
  - ✓ + 1 ou mais ocorrências da expressão precedente;
  - ✓ \ A expressão não deve ser considerada literalmente;
  - ✓ | Ocorrência da expressão precedente **ou** sucedente.

Visão geral do PLN em ECL

- ✓ Definições do tipo ***PATTERN, RULE* ou *TOKEN*:**
  - ✓ Usadas para detectar texto de interesse nos dados
- ✓ Funções de estrutura RECORD específicas (***MATCHTEXT***):
  - ✓ Utilizam as definições *PATTERN*, *RULE* ou *TOKEN* para obter e estruturar o texto de interesse
- ✓ A função ***PARSE***:
  - ✓ Implementa a operação de processamento e retorna o conjunto de registros

Exemplo de PLN

```
datafile := DATASET([{'And when Shechem the son of Hamor the Hivite, prince of Reuel'},
                     {'the son of Bashemath the wife of Esau.'}], {STRING10000 line});

PATTERN ws1                 := [' ','\t',','];
PATTERN ws                  := ws1 ws1?;
PATTERN article             := ['A','The','Thou','a','the','thou'];
TOKEN Name                  := PATTERN('[A-Z][a-zA-Z]+');
RULE Namet                  := name OPT(ws ['the','king of','prince of'] ws name);
PATTERN produced            := OPT(article ws) ['begat','father of','mother of'];
PATTERN produced_by         := OPT(article ws) ['son of','daughter of'];
PATTERN produces_with       := OPT(article ws) ['wife of'];
RULE relationtype           := ( produced | produced_by | produces_with );
RULE progeny                := namet ws relationtype ws namet;

results := {STRING60 Le                 :=  MATCHTEXT(Namet[1]);
            STRING60 Ri                 :=  MATCHTEXT(Namet[2]);
            STRING30 RelationPhrase     :=  MATCHTEXT(relationtype) };

outfile1 := PARSE(datafile,line,progeny,results,SCAN ALL);
outfile1;
```

le	ri	relationphrase
Shechem	Hamor the Hivite	the son of
Shechem	Hamor	the son of
Bashemath	Esau	the wife of

Fundação Vanzolini

Tipos de definição: PATTERN, TOKEN e RULE

PATTERN *patternid := parsepattern;*

TOKEN *tokenid := parsepattern;*

RULE *ruleid := parsepattern;*

- ✓ *patterned, tokenid, ruleid* – O nome do pattern, token ou ruleid.
- ✓ *parsepattern* – O padrão buscado, similar a uma expressão regular (regex).

O tipo **PATTERN** define uma expressão de parsing similar a uma expressão regular (regex).

O tipo **TOKEN** define uma expressão de parsing similar ao PATTERN, mas uma vez que a expressão seja encontrada, não busca combinações alternativas.

O tipo **RULE** define uma combinação de TOKENs e, da mesma forma que o PATTERN, busca combinações alternativas.

Fundação Vanzolini

Exemplo de PATTERN/TOKEN/RULE

```
ds := DATASET([{'quick brown fox'}],{STRING line});

PATTERN char := PATTERN('[A-Za-z]');
PATTERN ws := ' ';
PATTERN PatternWord := char+;
TOKEN   TokenWord := PatternWord;
RULE    RuleWords := TokenWord ws TokenWord OPT (ws TokenWord);
RULE    RuleWordsP := PatternWord ws PatternWord OPT(ws PatternWord);
RULE    RuleWordsM := TokenWord ws TokenWord;

PARSE(ds,line,PatternWord,{res := MATCHTEXT(PatternWord)});
PARSE(ds,line,TokenWord,  {res := MATCHTEXT(TokenWord)});
PARSE(ds,line,RuleWords,  {res := MATCHTEXT(RuleWords)});
PARSE(ds,line,RuleWordsP, {res := MATCHTEXT(RuleWordsP)});
PARSE(ds,line,RuleWordsM, {res := MATCHTEXT(RuleWordsM)});
PARSE(ds,line,RuleWordsM, {res := MATCHTEXT(TokenWord[1])});
PARSE(ds,line,RuleWordsP, {res := MATCHTEXT(RuleWordsP)}, WHOLE);
```


Fundação Vanzolini

Função de estruturas RECORD (PLN)

✓**MATCHED([*patternreference*])**

retorna TRUE/FALSE se o *patternreference* encontrou alguma equivalência.

✓**MATCHTEXT([*patternreference*])**

retorna o texto ASCII da *patternreference* encontrada ou vazio caso não encontre equivalência.

✓**MATCHROW([*patternreference*])**

retorna todo o registro do texto de equivalência do *patternreference*.

Fundação Vanzolini

Função PARSE

A função **PARSE** opera no dataset, usando o *pattern* e gerando o resultado no formato *result* especificado.

PARSE(*dataset,data,pattern,result,flags***)**

- ✓ *dataset* – conjunto de registros.
- ✓ *data* – O conteúdo a ser processado (geralmente um campo de um dataset).
- ✓ *pattern* – O pattern a ser utilizado.
- ✓ *result* – a estrutura RECORD de saída.
- ✓ *flags* – opções de parse.

Text Vectors

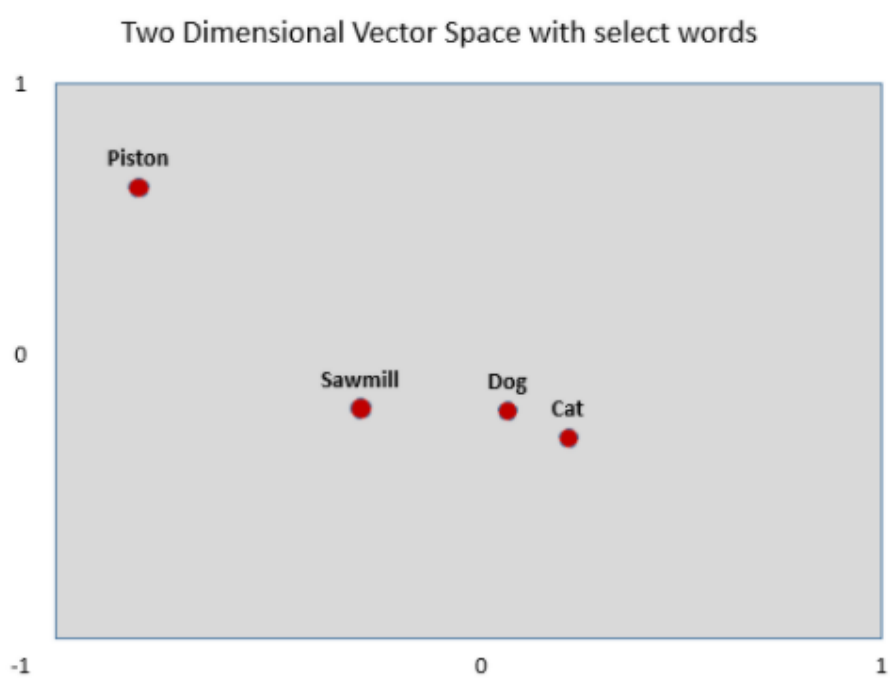


Figure 1 -- 2D Word Vector Space showing select words

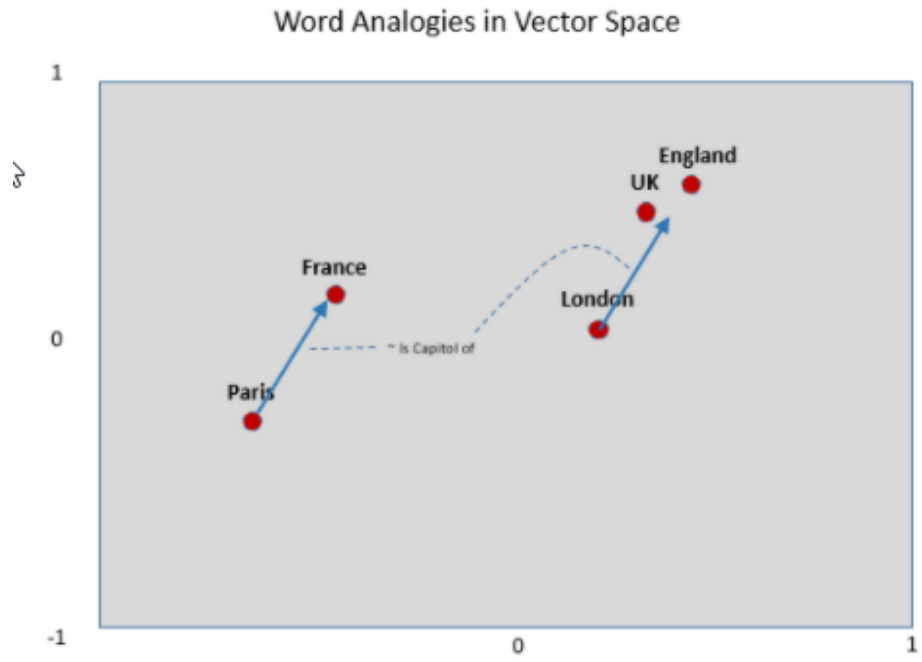


Figure 3 -- Word Analogies

Fundação Vanzolini

Sentenças de treinamento e saídas

##	sentid	text
1	1	Cute and cozy place. Perfect location to everything!
2	2	Kelly has a great room in a very central location. Beautiful building , architecture and a style that we really like.
3	3	Very spacious apartment, and in a great neighborhood. This is the kind of apartment I wish I had!Didn't really get to
4	4	Close to Seattle Center and all it has to offer - ballet, theater, museum, Space Needle, restaurants of all ilk just b
5	5	Kelly was a great host and very accommodating in a great neighborhood. She has some great coffee and while I wasn't ar
6	6	Kelly was great, place was great, just what I was looking for-clean, simple, well kept place.5 min walk to the Seattle
7	7	Kelly was great! Very nice and the neighborhood and place to stay was expected and comfortable. Overall great and woul
8	8	hola all bnb erz - Just left Seattle where I had a simply fantastic time for the weekend , no small part because of th
9	9	Kelly's place is conveniently located on a quiet street in Lower Queen Anne which is an easy walk or bus/cab ride to B

text	closest	similarity
	Item	Item
location is to quiet as place is to:	quiet,stay	1,0.999029278755188

text	closest	similarity
	Item	Item
neighbourhood	family,awesome,beautiful	0.9441138505935669,0.942000150680542,0.9396407008171082

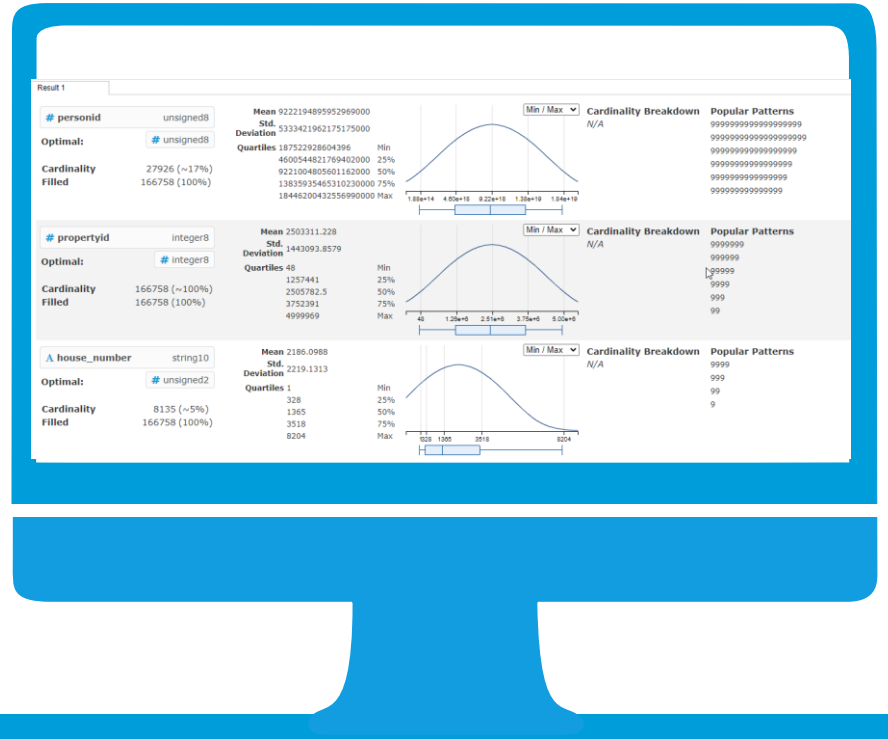
text	closest	similarity
	Item	Item
the apartment was spacious	Exactly as described, easy to get in and spacious.,Wonderful place! Clean, quiet, spacious, and very comfortable! Would definitely stay again!	0.9995267987251282,0.999
the neighbourhood was great	Nice quiet neighbourhood. Room was comfortable and clean.,Everything was accurate about the listing. Great location and neighbourhood.	0.9992449283599854,0.998

Desafio: Lending Club

Exercício prático:

Crie o data frame do dataset do Lending Club

- Considere a aplicação de aprendizagem supervisionada
- Se baseie nos resultados do perfilamento de dados



Fundação Vanzolini

Até a próxima aula!!!