CareerBuilder Kaggle (2012)

Uma proposta de recomendação colaborativa

https://www.kaggle.com/c/job-recommendation

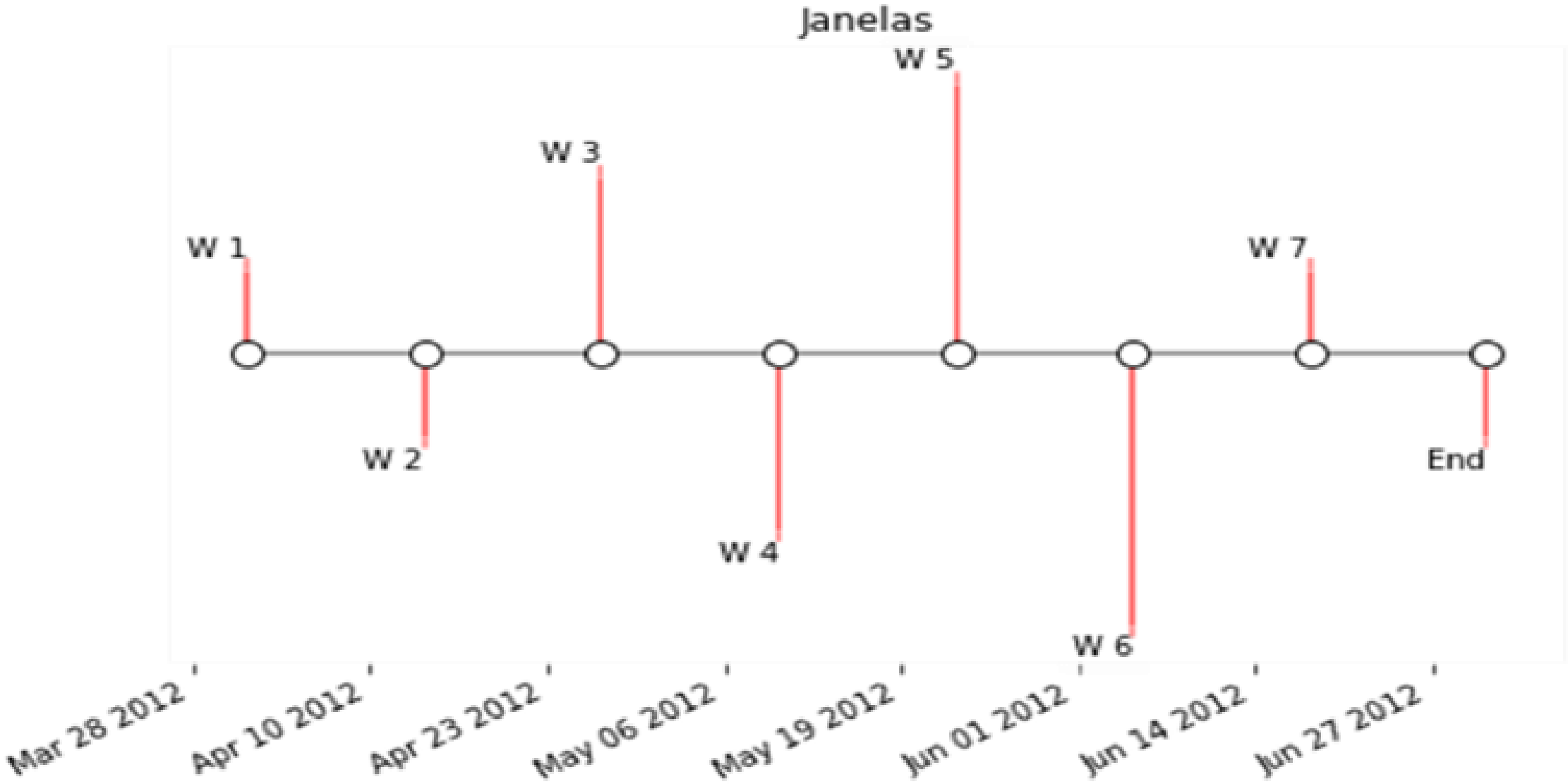
Jobs.tsv	**Dicionário de dados**		
	Tipo	***Descrição***	***Estruturado (S/N)***
Jobid	inteiro	ID vaga	S
WindowId	inteiro	ID janela	S
Title	texto	Título do cargo	N
Description	texto	Atribuições	N
Requirements	texto	Requisitos mínimos	N
City	texto	Localidade	S
State	texto	Estado	S
Country	texto	País	S
Zip5	inteiro	Código postal	S
StartDate	data	Início do anúncio	S
EndDate	data	Fim do anúncio	S

apps.tsv	**Dicionário de dados**		
	Tipo	***Descrição***	***Estrutura do (S/N)***
UserID	inteiro	ID candidato	S
WindowID	inteiro	ID janela	S
Split	texto	Treino/Teste	S
ApplicationDate	data	Data Candidatura	S
JobID	inteiro	ID Vaga	S

Users.tsv	**Dicionário de dados**		
	Tipo	***Descrição***	***Estrutura do (S/N)***
UserID	inteiro	ID candidato	S
WindowID	Inteiro	ID janela	S
Split	texto	Treino/Teste	S
City	texto	Cidade	S
State	texto	Estado	S
Country	texto	País	S
ZipCode	inteiro	Código postal	S
DegreeType	texto	Últ. formação	S
Major	texto	Área formação	N
GraduationDate	data	Data graduação	S
WorkHistoryCount	inteiro	Qtd. trabalhos	S
TotalYearsExperience	inteiro	Experiência	S
CurrentlyEmployed	bool	Situação atual	S
ManagedOthers	bool	Gerenciamento	S
ManagedHowMany	inteiro	Tamanho do time	S

Users_history.tsv	**Dicionário de dados**		
	Tipo	***Descrição***	***Estruturado (S/N)***
UserID	inteiro	ID candidato	S
WindowID	inteiro	ID janela	S
Split	texto	Treino/Teste	S
JobTitle	texto	Título vaga	N
Sequence	inteiro	Ordem decrescente	S

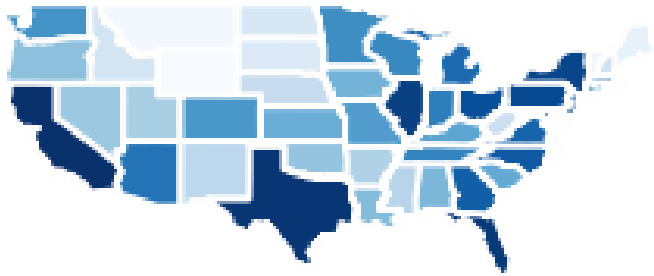
Window_dates.tsv	**Dicionário de dados**		
	Tipo	***Descrição***	***Estruturado (S/N)***
Window	inteiro	ID janela	S
Train Start	data	Início do treino	S
Train End/Test Start	data	Início teste	S
Test End	data	Fim teste	S

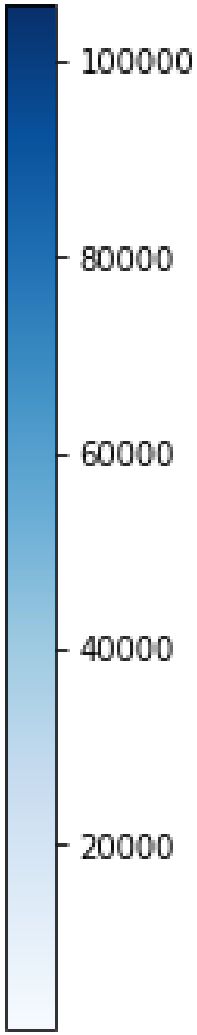

Janelas
W 1
W 2
W 3
W 4
W 5
W 6
W 7
End
Mar 28 2012
Apr 10 2012
Apr 23 2012
May 06 2012
May 19 2012
Jun 01 2012
Jun 14 2012
Jun 27 2012

Janelas	**Distribuição de candidatos**		
	Total	***Treino***	***Teste***
1	77.060 (19.77%)	71.641 (93%)	5.419 (7%)
2	58.228 (14.94%)	54.640 (93.8%)	3.588 (6.2%)
3	55.896 (14.34%)	52.126 (93.3%)	3.770 (6.7%)
4	53.449 (13.72%)	50.056 (93.7%)	3.393 (6.3%)
5	52.006 (13.34%)	48.914 (94.1%)	3.092 (5.9%)
6	43.334 (11.12%)	41.769 (96.4%)	1.565 (3.6%)
7	49.735 (12.76%)	47.724 (96%)	2.011 (4%)
Total	389.708	366.870 (média 94,30%)	22.838 (média 5.7%)

Jobs.tsv

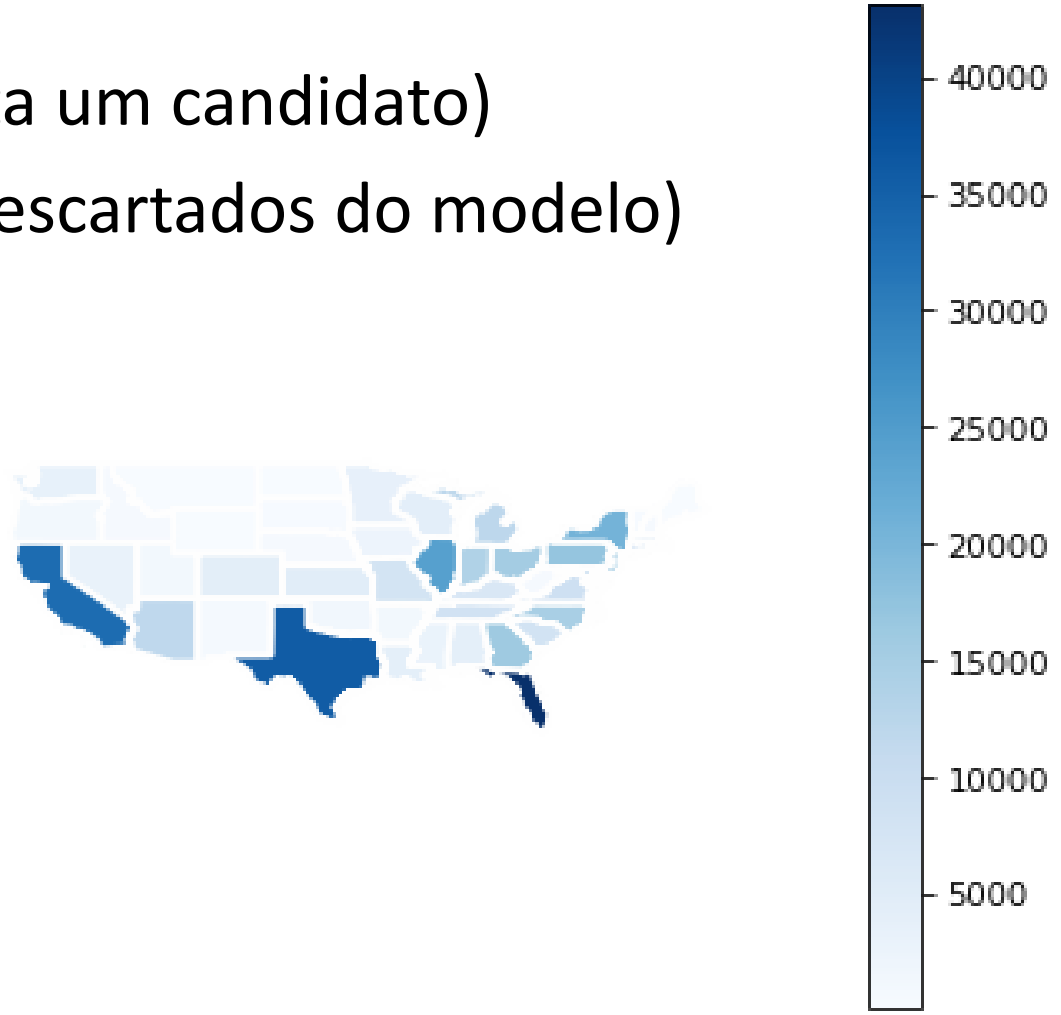
- 1.054.348 linhas (cada linha repreenta uma vaga)
- 0,13% vagas internacionais (descartadas do modelo)



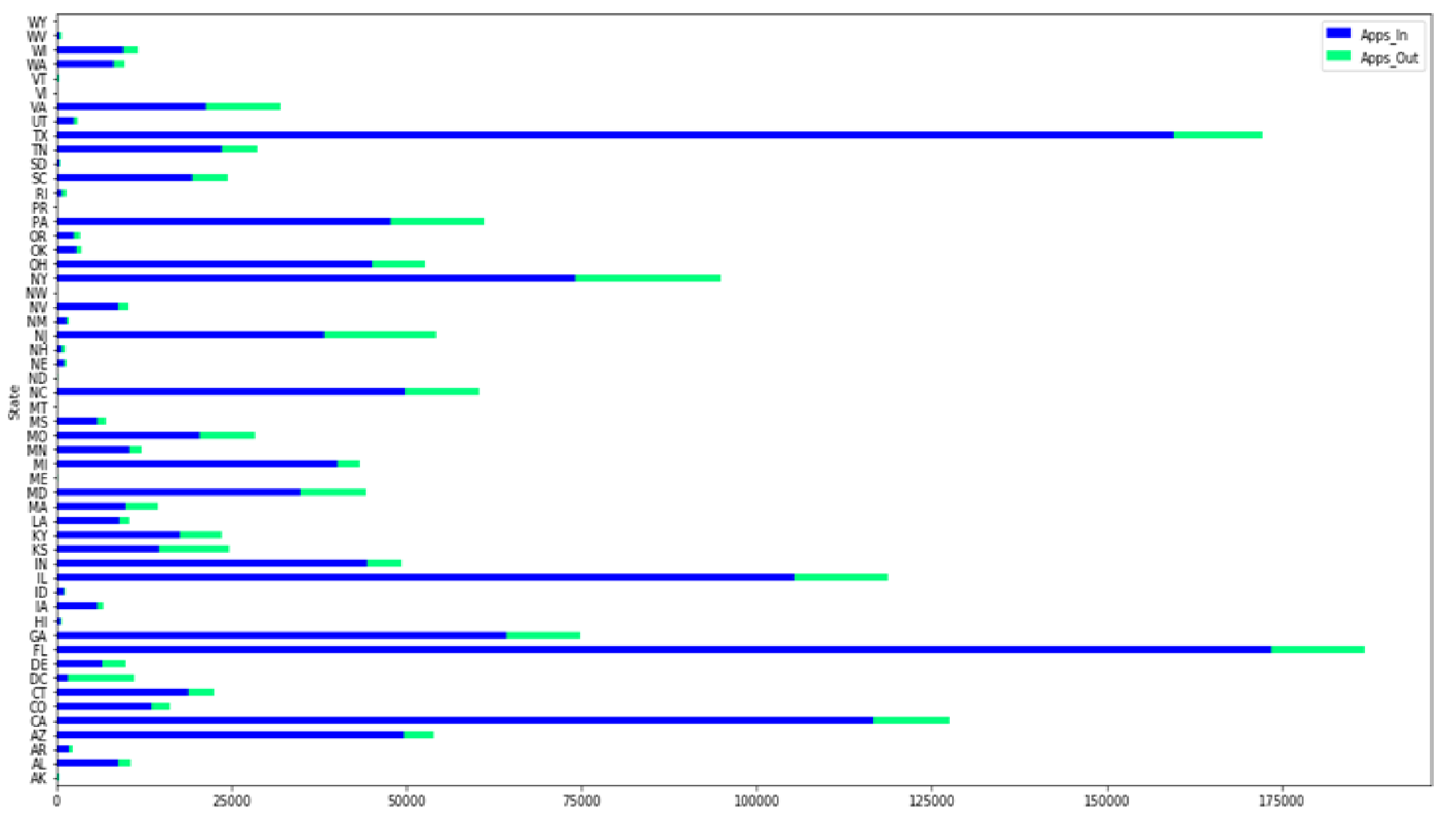


Users.tsv

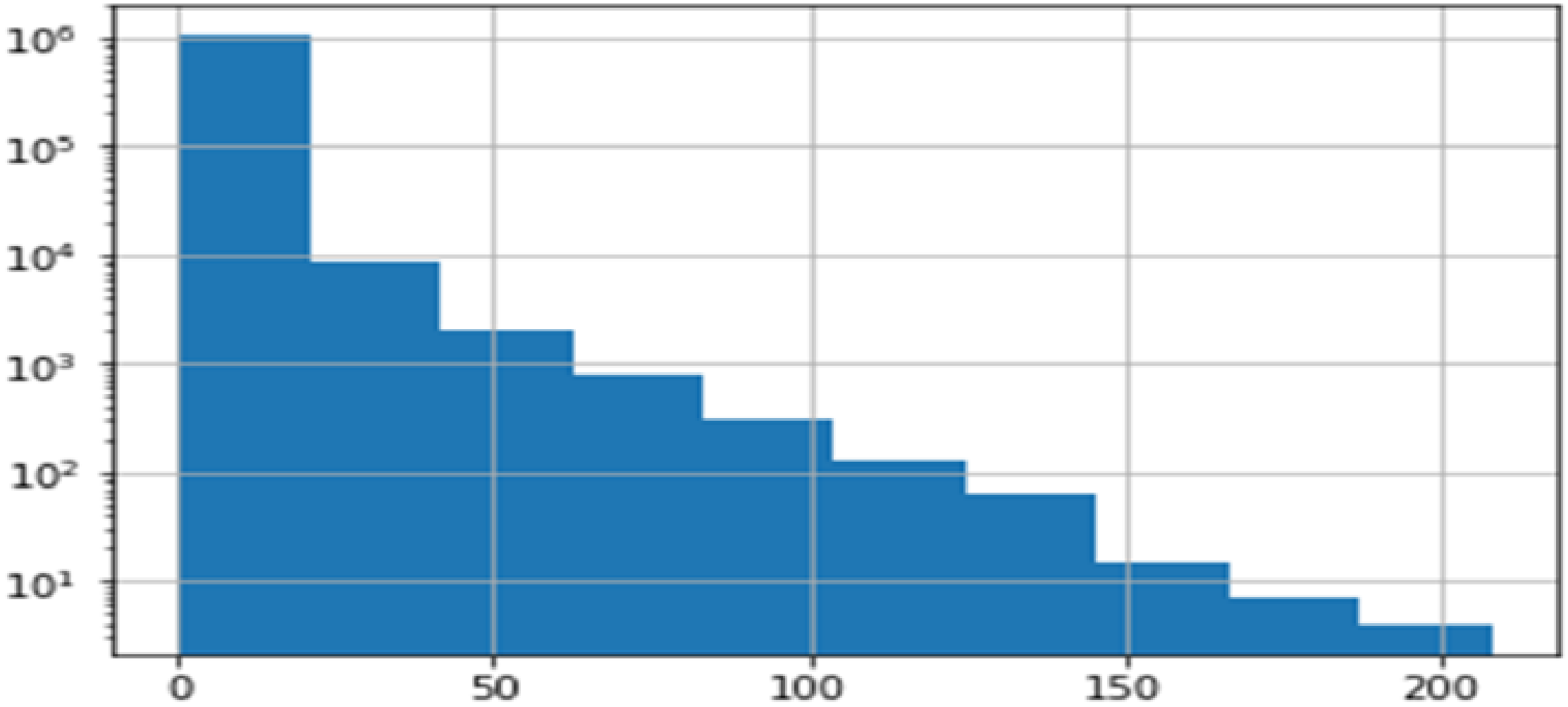
- 389.708 linhas (cada linha repreenta um candidato)
- 0,31% candidatos internacionais (descartados do modelo)



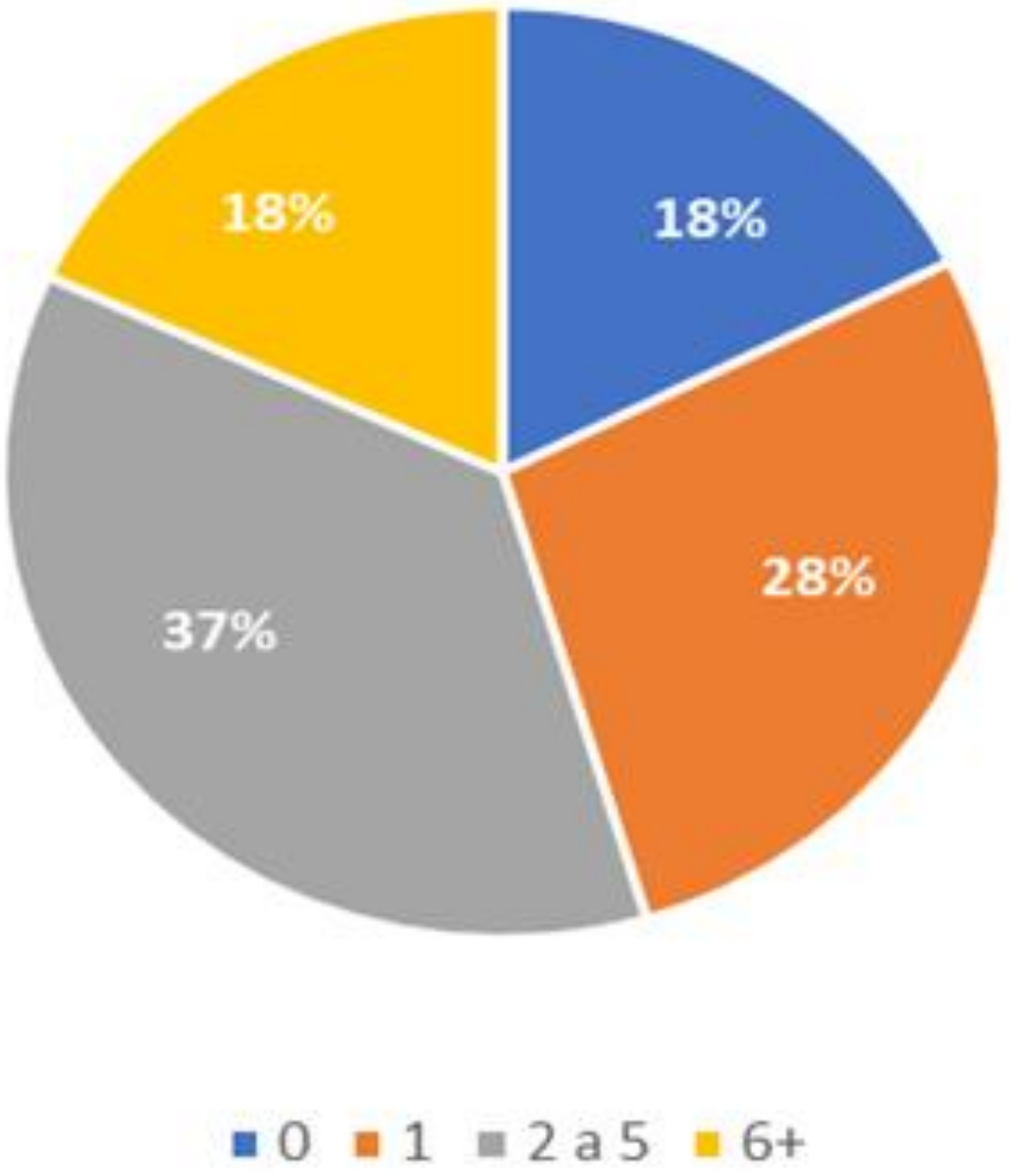
% vagas por candidato


Apps_In
Apps_Out
State
WY
WV
WI
WA
VT
VI
VA
UT
TX
TN
SD
SC
RI
PR
PA
OR
OK
OH
NY
NW
NV
NM
NJ
NH
NE
ND
NC
MT
MS
MO
MN
MI
ME
MD
MA
LA
KY
KS
IN
IL
ID
IA
HI
GA
FL
DE
DC
CT
CO
CA
AZ
AR
AL
AK
0
25000
50000
75000
100000
125000
150000
175000

candidaturas / vaga



#candidatura



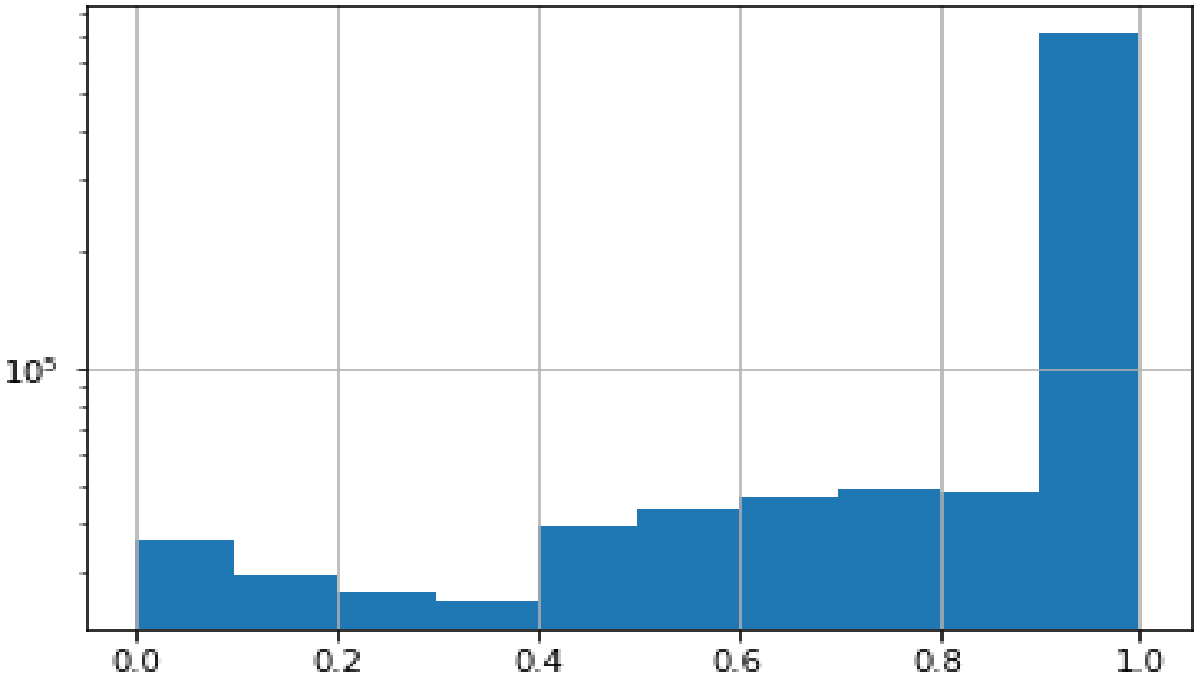
Usuários por Status Total		
Grupo	***Total***	***Média Candidaturas***
Empregados	199.888	3.92
Desempregados	189.820	4.31

Usuários por Status que realizaram candidaturas (Apps > 0)			
Grupo	***Total***	***% /Total***	***Média Candidaturas***
Empregados	163.102	81.60%	4.81
Desempregados	189.820	83.61%	5.17

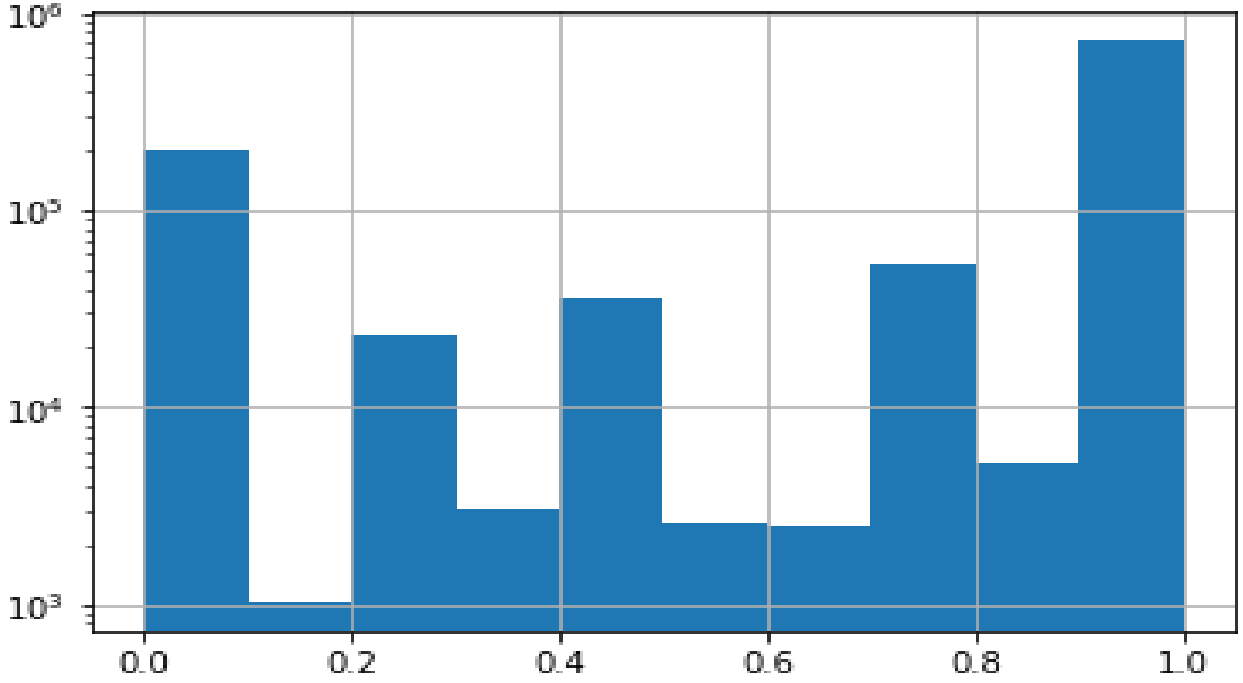
Usuários por Status que realizaram candidaturas (Apps = 0)			
Grupo	***Total***	***% /Total***	***Média Candidaturas***
Empregados	36.786	18.40%	0
Desempregados	31.687	16.69%	0

Probabilidade de visualização

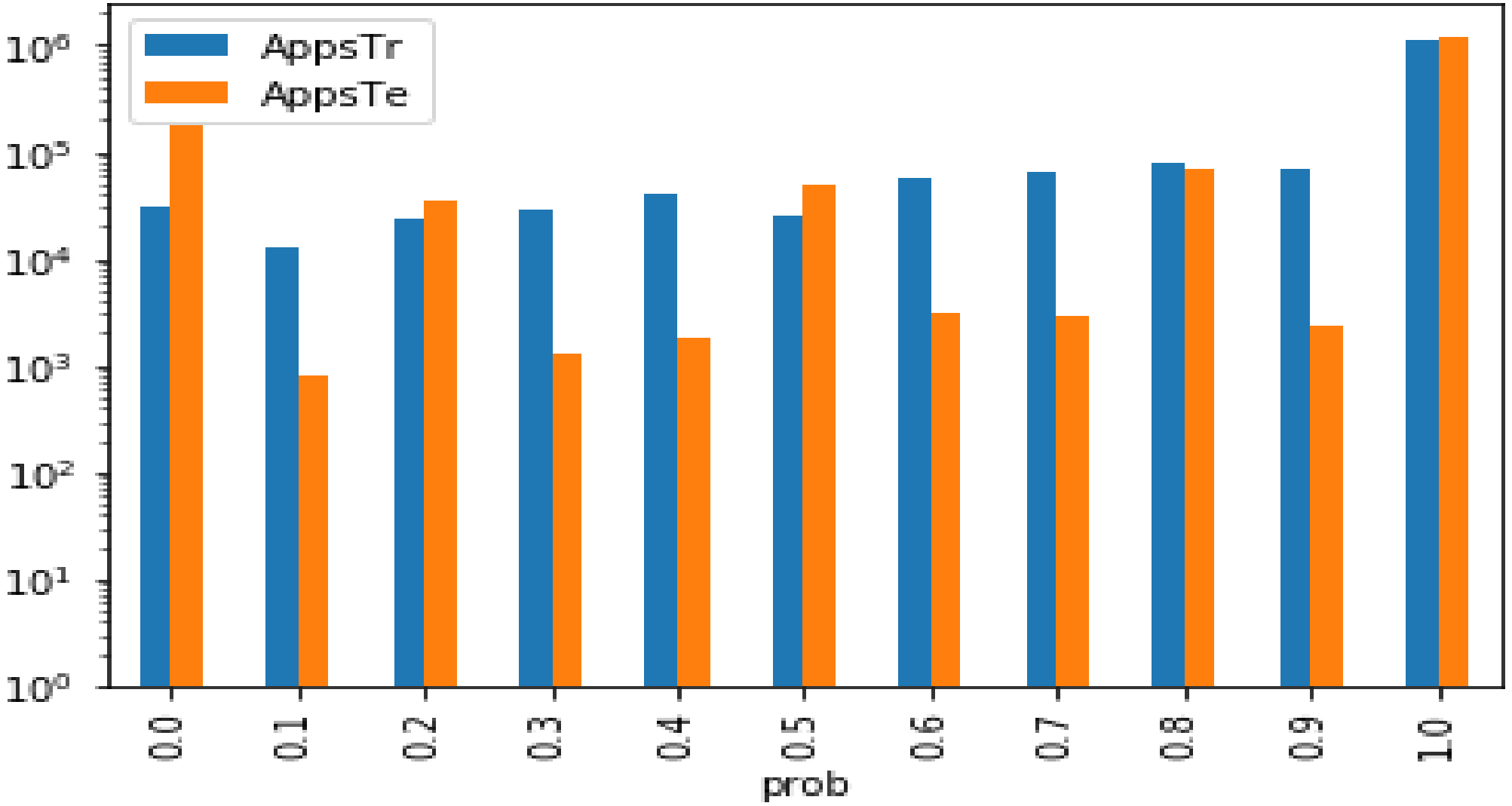
#treinamento



#teste



Candidaturas em função probabilidade visualização



Filtro colaborativo (distância de Jaccard)

$$S(U_a, U_b) = \frac{U_a \cap U_b}{U_a \cup U_b} \qquad (2)$$

Onde:

- $U_a \cap U_b$ = vagas que ambos se candidataram;
- $U_a \cup U_b$ = soma de todas as suas candidaturas

Complexidade do modelo

$$O(f(n)) = \frac{n \times (n-1)}{2} \qquad (3)$$

Candidaturas realizadas (Período Teste)		
Janela	Dimensão	# comparações
1	Total: 49.749 candidaturas realizadas por 3138 candidatos	4.921.953
2	Total: 28.311 candidaturas realizadas por 2052 candidatos	2.104.326
3	Total: 27.618 candidaturas realizadas por 2190 candidatos	2.396.955
4	Total: 27.772 candidaturas realizadas por 2046 candidatos	2.092.035
5	Total: 27.718 candidaturas realizadas por 1941 candidatos	1.882.770
6	Total: 11.822 candidaturas realizadas por 878 candidatos	385.003
7	Total: 12.607 candidaturas realizadas por 957 candidatos	457.446

Cálculo da Similaridade (Pseudocódigo)

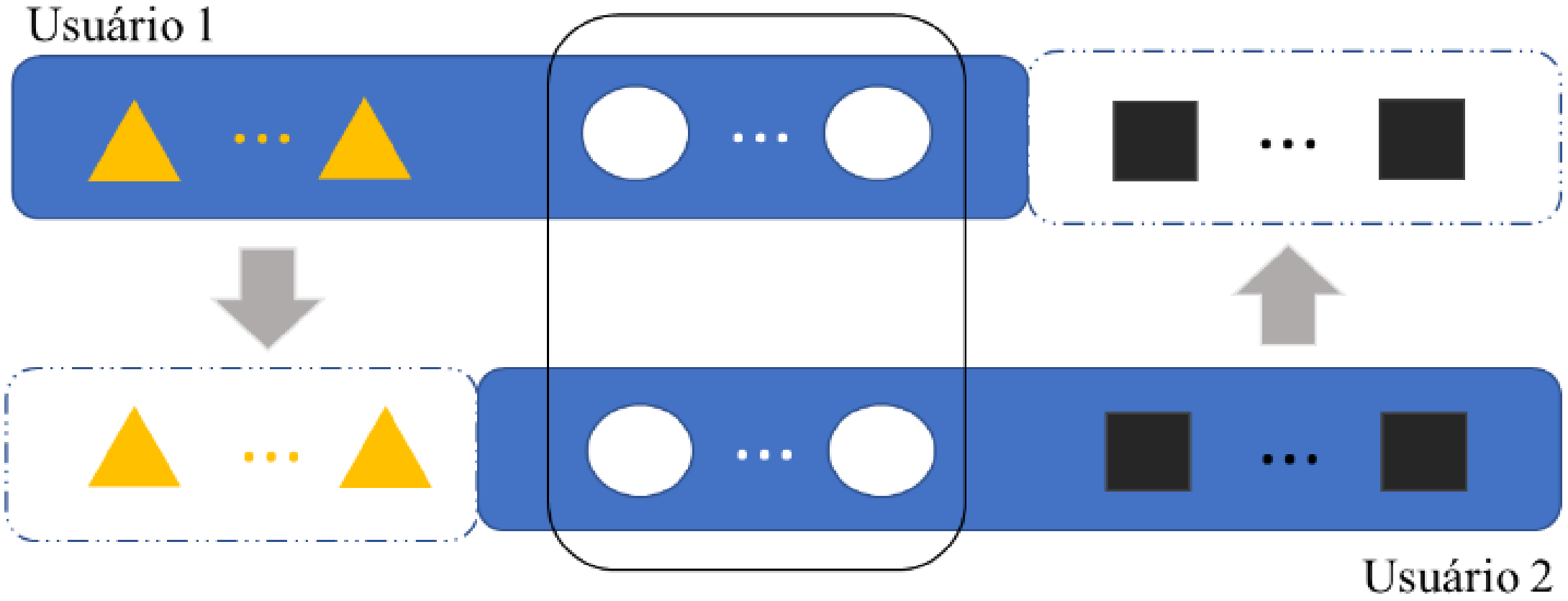
```
Arquivo Saída:
     1.   W – código da janela (de 1 a 7)
     2.   Usr1 – código do usuário 1
     3.   Usr2 – código do usuário 2
     4.   Qt1 – # candidaturas usuário 1
     5.   Qt2 – # candidaturas usuárias 2
     6.   QtJ – # candidaturas em Conjunto
     7.   S – % similaridade

1.   Faça de 1 a 7:

1.1. Apps[w]  = Candidaturas da janela [w] do período de Teste

1.2. Users[w] = Agrupar por usuários únicos

1.3. Em cada item de Users[w] (1.2): 0 a n -1

     1.3.1.  Usr1 = Localiza o usuário[u1] em Users[w]
     1.3.2.  apps_u1 = candidaturas de Usr1 em Apps[w]

     1.3.3.  Em cada item de Users[w] (1.2): u1+1 a n

       1.3.3.1.  Usr2 = Localiza o usuário[u2] em Users[w]
       1.3.3.2.  apps_u2 = candidaturas de Usr2 em Apps[w]
       1.3.3.3.  inner = Candidaturas em comum (Usr1 e Usr2)
       1.3.3.4.  union = Total Candidaturas (Usr1 e Usr2)
       1.3.3.5.  S = inner / union
       1.3.3.6.  Se S > 0, adiciona no arquivo de saída

1.4.  Salvar arquivo saída da janela_[w].tsv
```

Candidatos		
Qtd. Usuários Similares	Dimensão	% Total
1	1917	20%
2	1325	13.8%
3	906	9.5%
4	757	7.9%
5+	4681	48.8%
Total	9.586	Observação: representam 42% dos 22.838 usuários analisados

Similaridades		
Janela	***Quantidade***	***Média (candidatos similares / candidato)***
1	31.244	9.95
2	14.940	7.28
3	17.400	7.94
4	14.507	7.09
5	15.491	7.98
6	3.600	4.1
7	4262	4.45

Similaridades (> 5%)		
Janela	***Quantidade***	***Média (candidatos similares) / candidato***
1	9.722	3.09
2	5.590	2.72
3	6.939	3.1
4	5.514	2.69
5	5.232	2.69
6	1.406	1.6
7	1.667	1.74

Recomendações



Arquivo Saída:

1. W – Código da janela (de 1 a 7)
2. Usr_To – Usuário com a vaga recomendada
3. Usr_From – Usuário que partiu a recomendação
4. JobID – Vaga Recomendada

1. Faça de 1 a 7:
1.1. Apps[w] = Candidaturas da janela [w] do período de Teste
1.2. Top5 = Lista de usuários com as 5 maiores similaridades
1.3. Em cada item de Top5Ranking (1.2): 0 a n

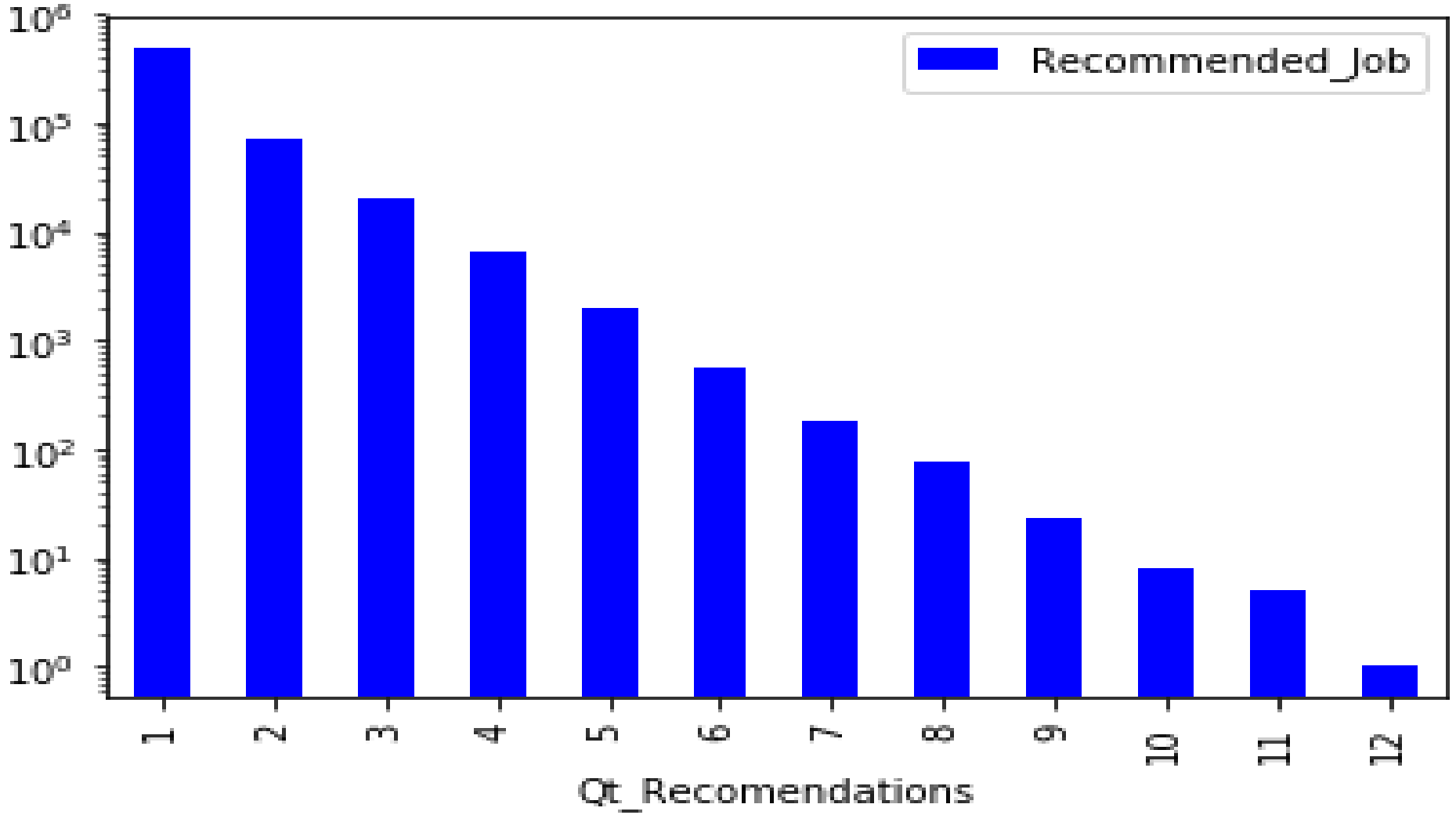
1.3.1. Usr1 = Coluna User1 em no Top5
1.3.2. Usr2 = Coluna User2 em no Top5
1.3.3. J1 = candidaturas de Usr1 em Apps[w]
1.3.4. J2 = candidaturas de Usr2 em Apps[w]
1.3.5. JTR1 = candidaturas de J2 (não em comum com J1)
1.3.6. JTR2 = candidaturas de J1 (não em comum com J2)
1.3.7. Adiciona JTR1 no arquivo saída: [J1,J2, JTR1]
1.3.8. Adiciona JTR2 no arquivo saída: [J2,J1, JTR2]

1.4. Salvar arquivo saída da recom_[w].tsv

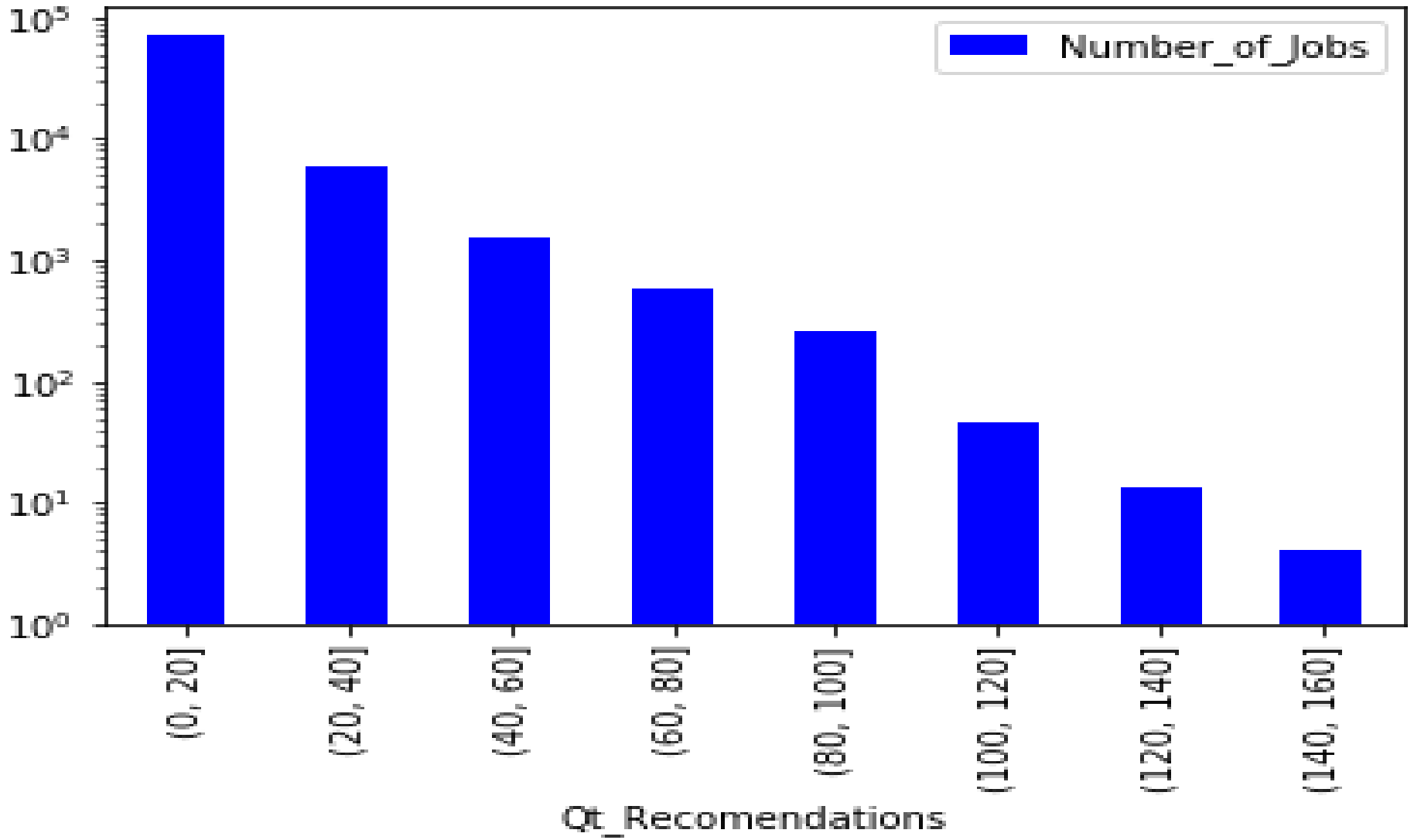
Resultado

747.863 recomendações

Vagas recomendadas # vezes para cada candidato



Vagas recomendadas para # candidatos



Limitações do modelo

- Não capturados pelo modelo:
  - Não efetuou qualquer candidatura (~ 18% dos usuários)
  - Candidatos que aplicaram somente a vagas que ninguém mais aplicou (sem similaridade)

- Dados esparsos (exemplo: Janela 2)
  - ~ 50.000 vagas
  - ~ 50.000 usuários
  - Candidaturas realizadas: ~ 200.000
  - Representa uma densidade menor do que 0,01%

Trabalhos Futuros

Recomendação baseada em conteúdo (I)

1. Comparação entre o título da vaga aberta com o título do histórico profissional dos candidatos

Jobs.tsv	**Dicionário de dados**		
	Tipo	***Descrição***	***Estruturado (S/N)***
Jobid	inteiro	ID vaga	S
WindowId	inteiro	ID janela	S
Title	texto	Título do cargo	N
Description	texto	Atribuições	N
Requirements	texto	Requisitos mínimos	N
City	texto	Localidade	S
State	texto	Estado	S
Country	texto	País	S
Zip5	inteiro	Código postal	S
StartDate	data	Início do anúncio	S
EndDate	data	Fim do anúncio	S

Users_history.tsv	**Dicionário de dados**		
	Tipo	***Descrição***	***Estruturado (S/N)***
UserID	inteiro	ID candidato	S
WindowID	inteiro	ID janela	S
Split	texto	Treino/Teste	S
JobTitle	texto	Título vaga	N
Sequence	inteiro	Ordem decrescente	S

Recomendação baseada em conteúdo (II)

Agrupamento das vagas, capturando informações do campo não estruturado na tentativa de identificar uma formação mínima requerida. Comparando com uma lista de níveis de formação. Assim, será possível tratar cada grupo destes de uma forma diferente.

Jobs.tsv	**Dicionário de dados**		
	Tipo	***Descrição***	***Estruturado (S/N)***
Jobid	inteiro	ID vaga	S
WindowId	inteiro	ID janela	S
Title	texto	Título do cargo	N
Description	texto	Atribuições	N
Requirements	texto	Requisitos mínimos	N
City	texto	Localidade	S
State	texto	Estado	S
Country	texto	País	S
Zip5	inteiro	Código postal	S
StartDate	data	Início do anúncio	S
EndDate	data	Fim do anúncio	S

Users.tsv -> DegreeType		
Título	***Descrição***	***Novo valor***
None	Candidatos sem escolaridade	0
High-school	Candidatos com baixa escolaridade, que não possuem curso superior	1
Associate's e Vocacional	Representam candidatos que possuem cursos técnicos ou graduações de curta duração	2
Bachelor's	Candidatos que possuem curso superior	3
Master's e Phd	Candidatos que possuem pós-graduação	4

Recomendação baseada em conteúdo (III)

Similaridade entre a vagas e recomendação das parecidas frente à alguma candidatura já realizada

Jobs.tsv	**Dicionário de dados**		
	Tipo	***Descrição***	***Estruturado (S/N)***
Jobid	inteiro	ID vaga	S
WindowId	inteiro	ID janela	S
Title	texto	Título do cargo	N
Description	texto	Atribuições	N
Requirements	texto	Requisitos mínimos	N
City	texto	Localidade	S
State	texto	Estado	S
Country	texto	País	S
Zip5	inteiro	Código postal	S
StartDate	data	Início do anúncio	S
EndDate	data	Fim do anúncio	S

Solução híbrida (em linhas gerais)

$$R = w_1C_1 + w_2C_2 \ldots + w_nC_n$$

Aplicando na prática

- Construção de um profile e captura de feedbacks implícitos
- Entender o momento profissional do candidato pela análise de suas atribuições anteriores
- Criar uma solução que permita a inclusão de novos motores de recomendação a um modelo existente