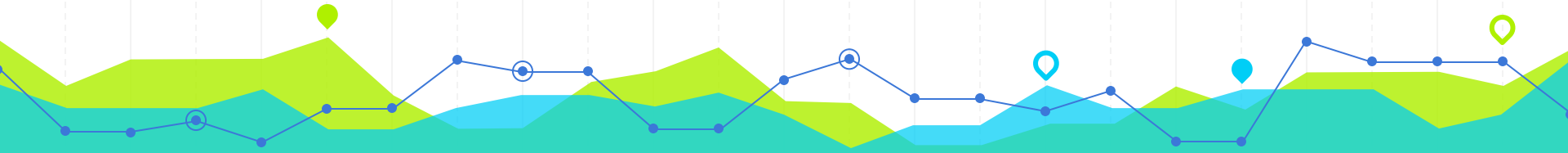
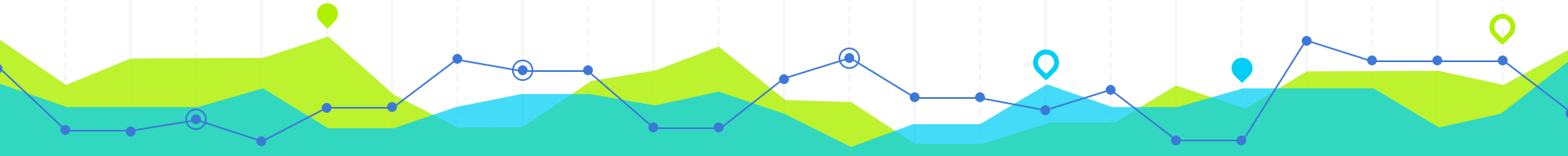


O preconceito da sociedade propagado pela Inteligência Artificial

Grupo:
Cinthya Lins (cml2)
Thiago Casa Nova (tcnl)

Inteligências Artificiais são realmente imparciais?

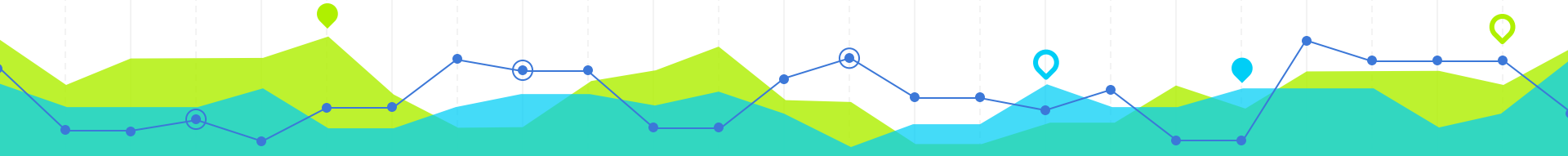


Como as Inteligências Artificiais priorizam as buscas que fazemos?

Importância

Mais pesquisados

Similaridade



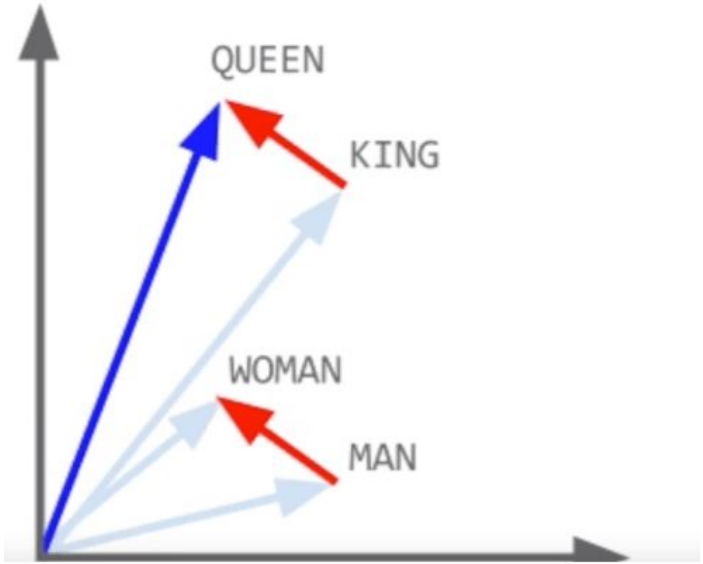
Teste de Associação Implícita

É uma forma de identificar maneiras de pensar, estereótipos e preconceitos.

Consiste em relacionar termos a diferentes conceitos da forma mais rápida possível.

Word embedding

Word embedding é um métodos que representa palavras e termos em vetores. Eles aprendem que vetores com semelhanças semânticas tendem a estar próximos um do outro. Ficou demonstrado também que a diferença entre vetores, podem tornar seus termos relacionados



$$\overrightarrow{\text{man}} - \overrightarrow{\text{woman}} \approx \overrightarrow{\text{king}} - \overrightarrow{\text{queen}}$$

Outras relações

Mulher - Homem

Filha - Filho

Enfermeira - Médico

Programador - Dona de Casa

Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embedding

- **Objetivo**: Desenvolver um algoritmo que reduza a conexão preconceituosa entre determinadas palavras sem perder conexões que são relevantes
  - Exemplo: recepcionista e mulher, rainha e mulher

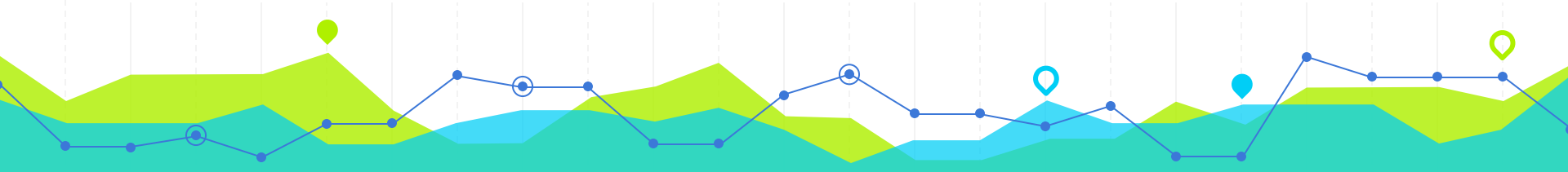
- **Método**:
  - Word2Vec
  - AMTurk
  - Questionários

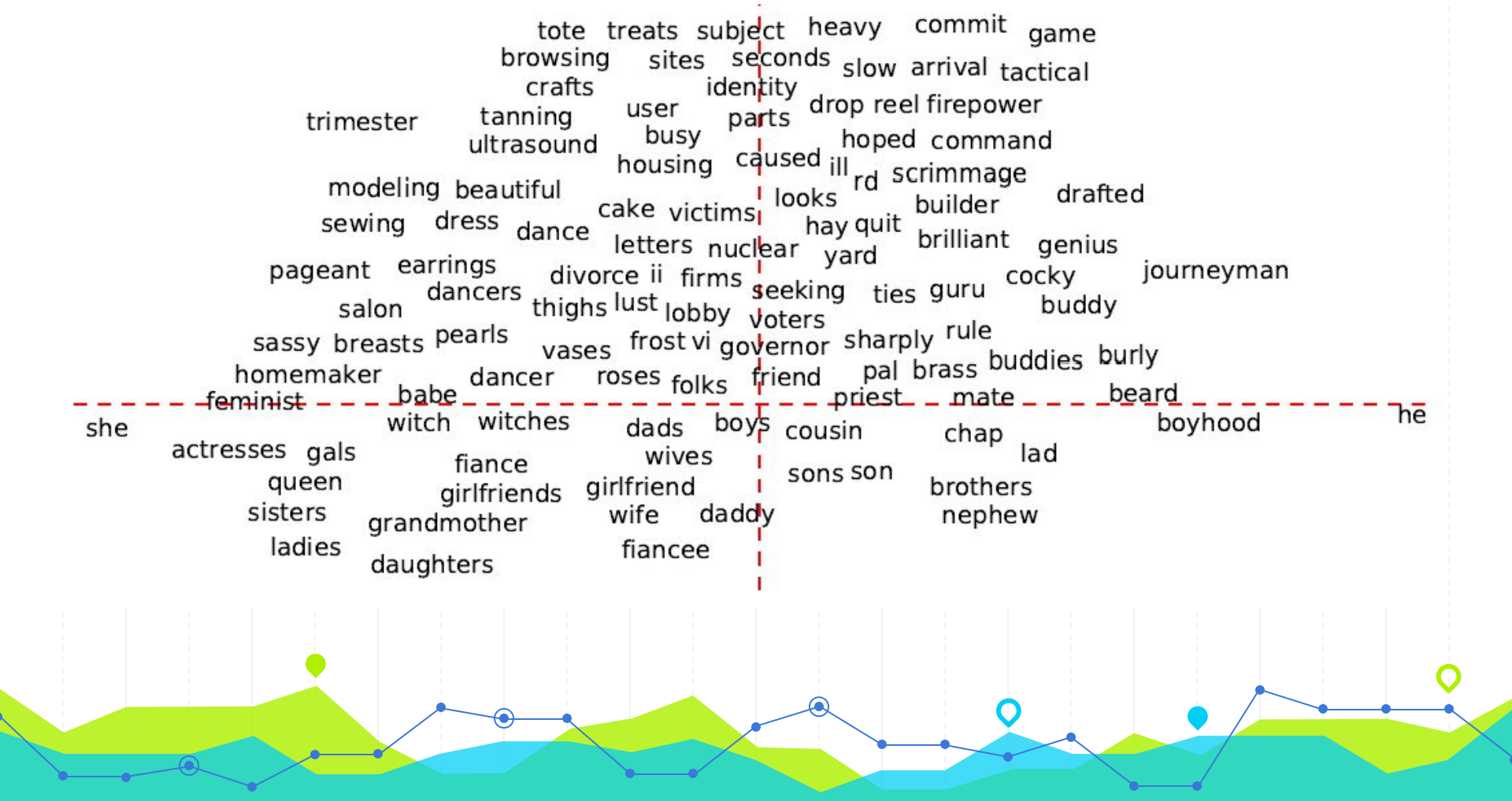
Extreme *she* occupations

1. homemaker
2. nurse
3. receptionist
4. librarian
5. socialite
6. hairdresser
7. nanny
8. bookkeeper
9. stylist
10. housekeeper
11. interior designer
12. guidance counselor

Extreme *he* occupations

1. maestro
2. skipper
3. protege
4. philosopher
5. captain
6. architect
7. financier
8. warrior
9. broadcaster
10. magician
11. figher pilot
12. boss




tote
treats
subject
heavy
commit
game
browsing
sites
seconds
slow
arrival
tactical
crafts
identity
drop
reel
firepower
trimester
tanning
user
parts
busy
hoped
command
ultrasound
housing
caused
ill
rd
scrimmage
modeling
beautiful
looks
builder
drafted
cake
victims
sewing
dress
dance
hay
quit
brilliant
genius
letters
nuclear
yard
pageant
earrings
divorce
ii
firms
cocky
journeyman
dancers
seeking
ties
guru
salon
thighs
lust
lobby
voters
buddy
sassy
breasts
pearls
frost
vi
governor
sharply
rule
vases
homemaker
dancer
roses
folks
friend
pal
brass
buddies
burly
feminist
babe
priest
mate
beard
she
witch
witches
dads
boys
cousin
chap
boyhood
he
actresses
gals
wives
lad
fiance
sons
son
queen
girlfriends
girlfriend
brothers
sisters
wife
daddy
nephew
grandmother
ladies
fiancee
daughters

- O algoritmo do estudo alterou a vetorização de palavras neutras removendo suas associações com gênero.
  - Enfermeiro(a) passa a ser relacionada igualmente a homem e mulher.
- Termos exclusivos de cada gênero não tiveram seu status alterados.

Semantics derived automatically from language corpora contain human-like biases

Reproduziu experimentos de outros estudos para analisar o viés negativo dos algoritmos.

Utilizou TAI e Word Embedding para quantificar os resultados.

Introdução - Flores e Insetos

- Estudo original - $1.35 / 10^{-8}$
- Reprodução - $1.50 / 10^{-7}$
  - Word Embedding e GloVe
  - Resultados

Viés Racial

- Estudo original - $1.17 / 10^{-6}$
- Reprodução - $1.41 / 10^{-8}$
  - Nomes
  - Balanceamento

Viés Racial - Entrevista

- Estudo original - 50%
  - 5000 Currículos iguais
- Reprodução
  - Corpus de documento mais atualizado

Homem/Mulher e Carreira/Família

- Estudo original $0.72 / 10^{-2}$
  - 38797 Realizaram o teste
  - Poucas palavras-chave
- Reprodução - $1.81 / 10^{-3}$

Conclusão

- Reflexo da sociedade
- Permitir acesso do público ao algoritmo
- Trabalho conjunto entre programadores e especialistas em desigualdades