CAPTURA DE RELAÇÕES SEMÂNTICAS COM A SIMILARIDADE DO COSSENO: UM EXEMPLO EM R

Afonso Xavier Canosa Rodrigues

Introdução

Introdução: o problema e marco teórico

R para o processamento de textos

Procedimento

Resultados

Conclusões

Introdução: trabalho relacionado

- Processamento de texto para trabalho de corpus
  https://github.com/afonsoxavier/semantics
- Trabalho com R com fins pedagógicos (cursos de Língua e Sociedade com grande tradição quantitativa e de análise de dados): LanguageandSociety(Spring 2015) | Afonso XavierCanosa - Academia.edu
- PLN e trabalho com R com fins experimentais (experimentos de linguística, tese de doutoramento sobre geoentidades). R como principal linguagem de programação https://citius.gal/research/publications/a-identificacao-e-referenciacao-de-entidades-geograficas-mencionadas-o-caso-da-peregrinacao-de-fernao-mendes-pinto

O problema

O problema a resolver consiste em determinar qual é o grau de proximidade entre entidades geográficas dados uns tipos geográficos e considerando unicamente as frequências de coocorrência.

I.e. Resolver uma relação semântica

As relações semânticas

Relações entre expressões do tipo:

Pertence_ao_tipo Ex. Pequim ∈ cidade

Pertencem_ao_mesmo_tipo Ex. {Pequim,Tóquio}

É_parte_de (meronímia) Ex. É_parte_de(Pequim,China)

Contém (holonímia) Ex. Contém(China, Pequim)

O marco teórico

Modelo distribucional na semântica, termos relacionados partilham contextos relacionados.

Na hipótese distribucional, palavras que ocorrem num contexto similar tendem a ser semanticamente similares (Mitchell & Lapata, 2010; Baroni, Bernardi & Zamparelli, 2014).

O modelo distribucional captura as propriedades semânticas de um termo através das coocorrências num corpus.

Materiais

Usamos um corpus médio não normalizado para realizar uma série de experimentos que provem o modelo distribucional.

R como linguagem de programação:

https://github.com/afonsoxavier/semantics/blob/master/cosine_similarity_results_article.R

Procedimento: a seleção dos termos

Selecionamos entidades geográficas mencionadas que respondem bem ao protótipo (bem caraterizadas) e pertencem a dous tipos geográficos diferentes.

EM	Freq	Tipo geográfico	Referência atual
Çamatra	14	Ilha	Sumatra, Indonésia
Iaoa	26	Ilha	Java, Indonésia
Martauão	35	Cidade	Martabão, Myanmar
Odiaa	19	Cidade	Aiutaia, Tailândia
Pequim	47	Cidade	Pequim, China
Tanixumaa	18	Ilha	Tanegaxima, Japão

Procedimento: a segmentação do corpus

Segmentamos o corpus nas unidades em que queremos medir as coocorrências. Neste caso utilizaremos a oração entendida como aquela unidade definida por um ponto como limite final.

Ex. de oração no corpus:

Os outros dous nauios que milagrosamente lhe escapamos, nos fizemos na volta do mar, & não podendo mais ferrar a terra por causa dos ventos Lestes que todo aquelle mès nos cursaraõ, nos foy forçado irmos demandar a costa da *laoa* bem contra nossa vontade.

Algumas considerações sobre as unidades de coocorrência neste corpus em particular

- Orações mais próximas ao nosso parágrafo atual.
- Maior longitude da oração vai permitir coocorrerem mais termos.
- A medida de coocorrência intui-se mais efetiva para medir a similaridade semântica do que a distância.
- Ex. concordâncias para Tanixumaa

Geração de matriz de coocorrências

No experimento queremos classificar as entidades como ilha ou cidade.

Geramos uma matriz para computar o número de vezes que os termos coocorrem com cada entidade.

	Çamatra	Iaoa	Martauão	Odiaa	Pequim	Tanixumaa
CIDADE	1	7	18	17	38	5
ILHA	13	11	1	1	1	11

A solução em R

O cálculo para cada uma das entidades é um processo repetitivo. Utilizamos R para todo o processo e para obter resultados para todas as entidades.

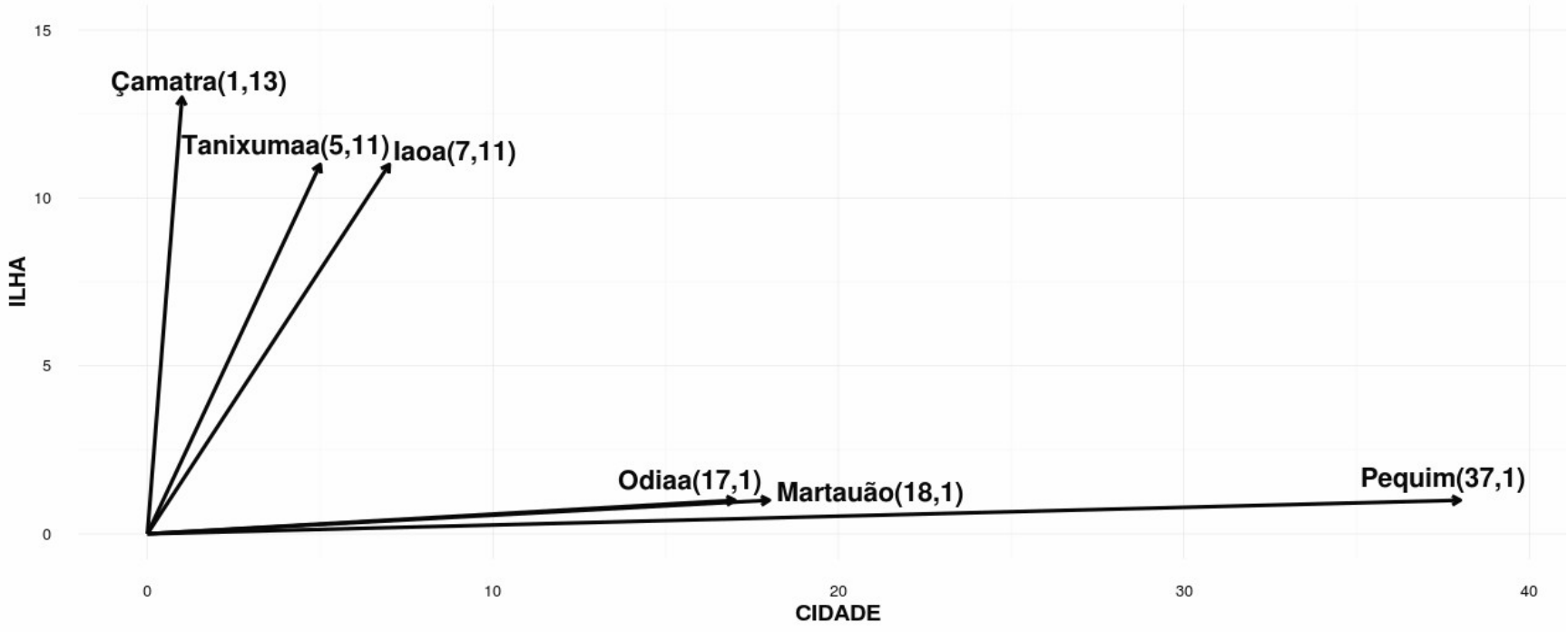
https://github.com/afonsoxavier/semantics/blob/master/cosine_similarity_results_article.R

```
# Create a matrix with documents as rows, terms as columns #


dtm <- DocumentTermMatrix(docs)
```

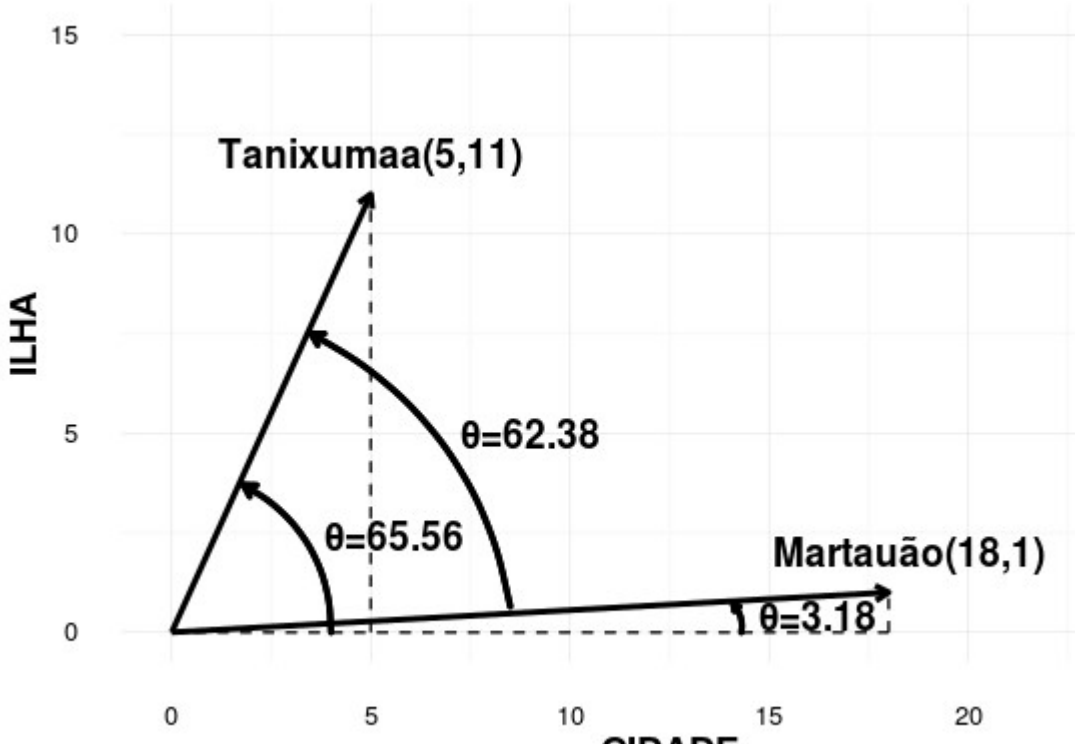
Representação num sistema de coordenadas

Uma vez elaborada a matriz, as coocorrências podem ser representadas como vetores



Representação num sistema de coordenadas

Deste modo cada entidade mencionada é definida como um vetor, as suas coordenadas os componentes da posição que podemos representar em forma geométrica com um valor de magnitude e direção.



A solução em sistema de coordenadas

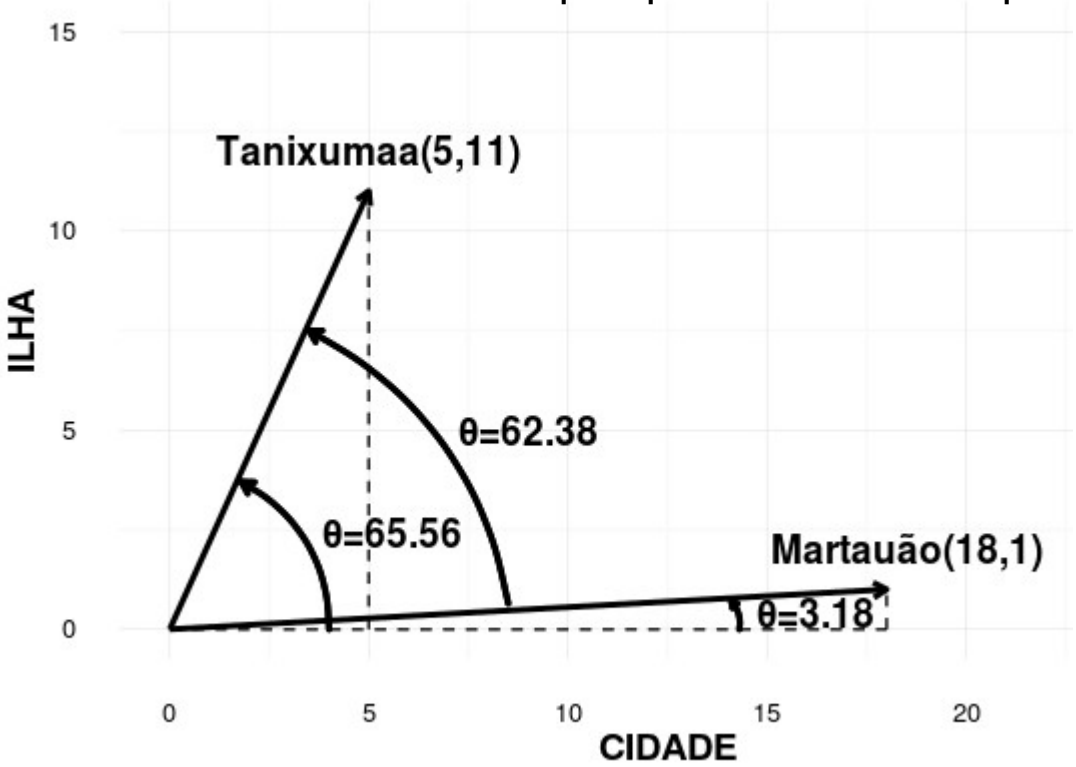
O cálculo para cada uma das entidades é um processo repetitivo. Utilizamos R para todo o processo e para obter resultados para todas as entidades.

```
# Calculate angles
pairs<-length(cidade1)
angles2<-c(1:length(ilha1))
for (i in 1:pairs){
  angles2[i]<-atan(ilha1[i]/cidade1[i])   # bring angle in radians
  print(deg(angles2[i]))
}
degangles<-deg(angles2) # show angles in degrees
```

Representação num sistema de coordenadas

Logo a distância semântica para os tipos CIDADE E ILHA das entidades geográficas mencionadas *Tanixumaa* e *Martauão* é:

θ (Tanixumaa) - θ (Martauão)| = |65.56° - 3.18°| = 62.38°



Representação num sistema de coordenadas

Independentemente da sua magnitude, a direção dos vetores fica entre os 0° e 90°. Obtemos assim uma medida para a similaridade ou distância semântica entre as entidades mencionadas a respeito dos tipos geográficos. Trazemos como exemplo o resultado obtido para a ilha de *Tanixumaa*:

cos(|θ (Tanixumaa) - θ (Çamatra)|) = cos (20.05°) = 0.94

cos(|θ (Tanixumaa) - θ (Iaoa)|)= cos(0.03°) = 0.99

cos(|θ (Tanixumaa) - θ (Martauão)|)= cos(62.38°) = 0.46

cos(|θ (Tanixumaa) - θ (Odiaa)|) = cos(62.19°) = 0.47

cos(|θ (Tanixumaa) - θ (Pequim)|) = cos(64.05°) = 0.44

A solução em R

Resolvemos a distância entre entidades

```
# Calculate  distance among entities
deganglesd<-diag(degangles)  # Identity matrix

for(i in 1:length(degangles)){
  for(g in 1:length(degangles)){
    deganglesd[i,g]<-abs(degangles[i]-degangles[g])
  }
}
cosanglesd<-cos(deganglesd)
```

Resultados

	Çamatra	Iaoa	Martauão	Odiaa	Pequim	Tanixumaa
Çamatra	1.0000000	**0.8823529**	0.1318850	0.1351132	0.1028992	**0.9394222**
Iaoa	**0.8823529**	1.0000000	0.5828468	0.5854906	0.5588836	**0.9902018**
Martauão	0.1318850	0.5828468	1.0000000	**0.9999947**	**0.9995740**	0.4636639
Odiaa	0.1351132	0.5854906	**0.9999947**	1.0000000	**0.9994737**	0.4665474
Pequim	0.1028992	0.5588836	**0.9995740**	**0.9994737**	1.0000000	0.4376085
Tanixumaa	**0.9394222**	**0.9902018**	0.4636639	0.4665474	0.4376085	1.0000000

Conclusões

Num corpus com desvios a respeito da norma que dificultam o seu processamento, se as entidades têm uma ocorrência estatisticamente relevante, a simples coocorrência de termos pode ser utilizada para a captura de relações semânticas sem necessidade de outro tipo de anotação ou enriquecimento de dados (anotação morfológico-sintática e semântica)

No caso estudado, em que a unidade de coocorrência é ampla (próxima ao parágrafo) a medida do cosseno captura melhor a similaridade do que a distância.

E possíveis aplicações

O cálculo da medida do cosseno foi aplicado no corpus para continuar os experimentos com aprendizado de máquina a partir de matrizes que medem a coocorrência entre todos os termos do corpus.

Duas versões do script (uma simples para o cálculo do cosseno e outra que cria também a matriz diretamente a partir de um corpus) estão disponíveis em:

https://github.com/afonsoxavier/semantics

Referências

Mitchell, J., & Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive science*, *34*(8), 1388-1429.

Baroni, M., Bernardi, R., & Zamparelli, R. (2014). Frege in space: A program of compositional distributional semantics. *LiLT (Linguistic Issues in Language Technology)*, *9*.