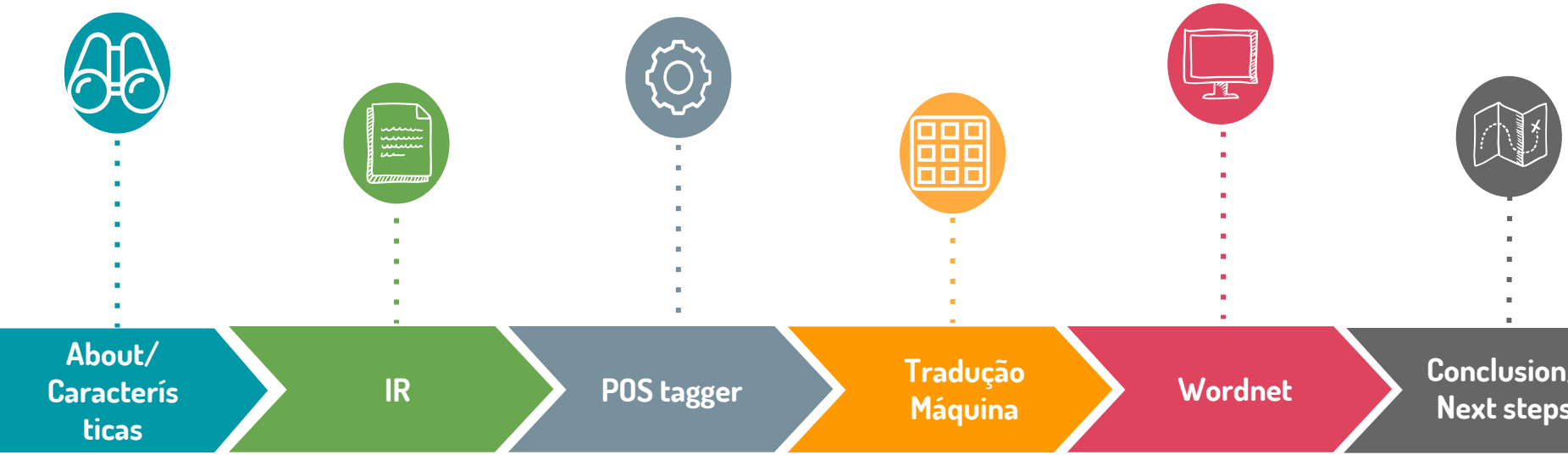
TEXT DATA

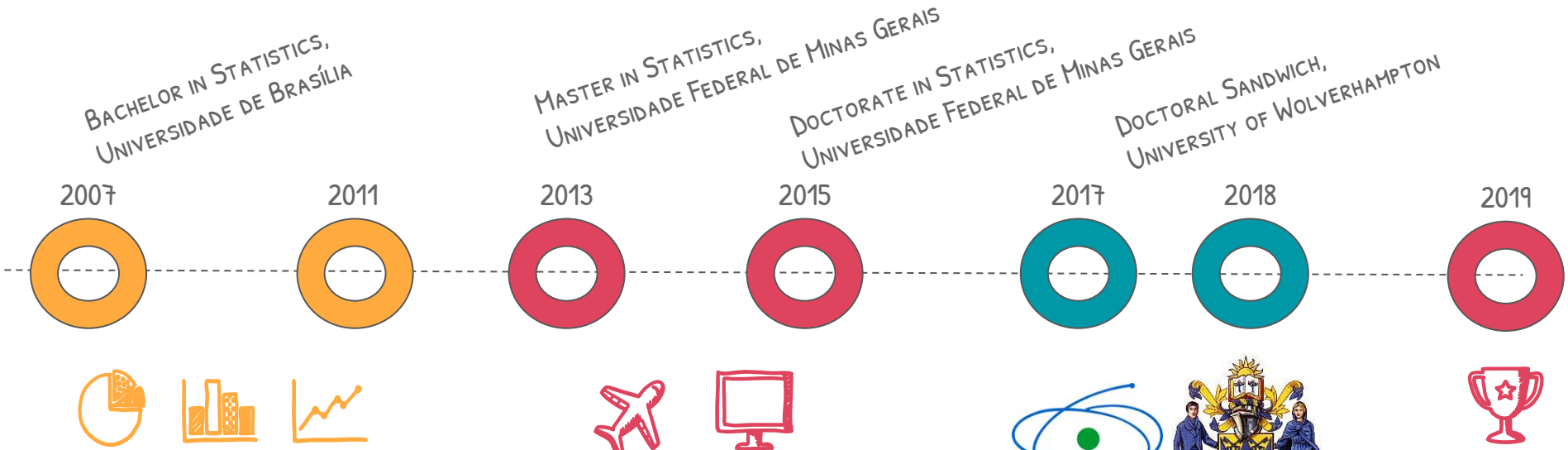
Larissa Sayuri Futino Castro dos Santos

14th August 2018

Agenda


About/ Caracterís ticas
IR
POS tagger
Tradução Máquina
Wordnet
Conclusion/ Next steps

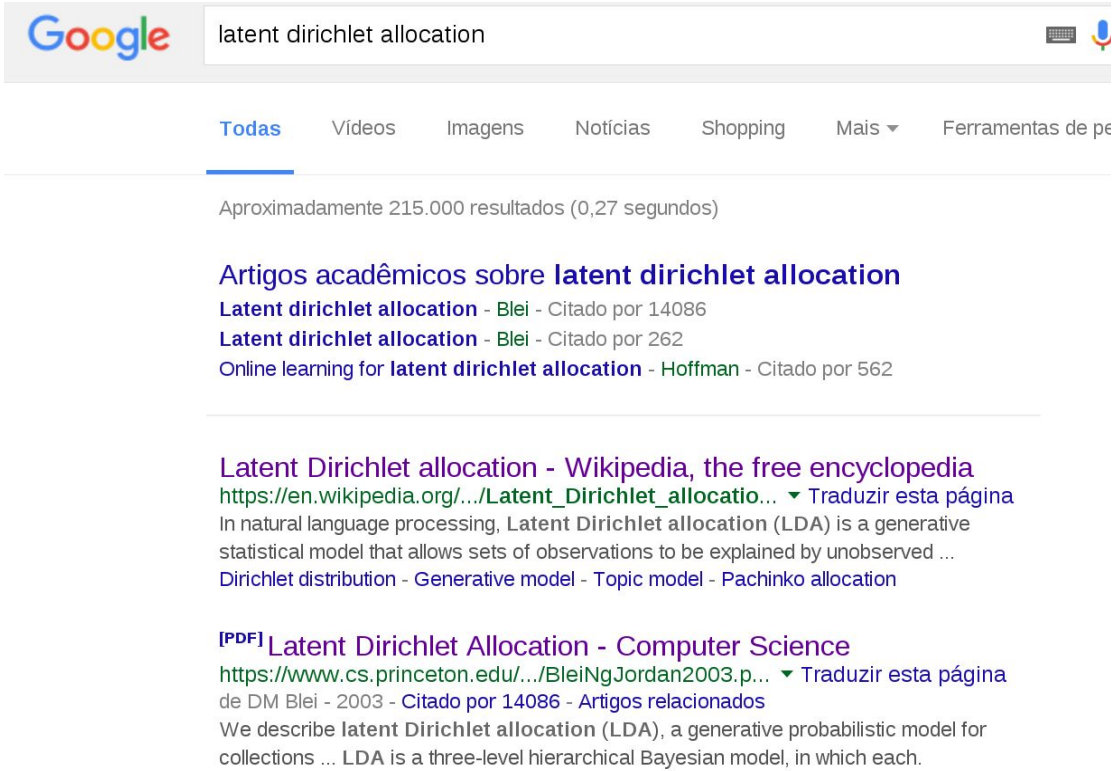
About


Bachelor in Statistics,
Universidade de Brasília
2007
2011
Master in Statistics,
Universidade Federal de Minas Gerais
2013
2015
Doctorate in Statistics,
Universidade Federal de Minas Gerais
2017
Doctoral Sandwich,
University of Wolverhampton
2018
2019
CAPES
INNOVATION
AND
OPPORTUNITY
UFMG
UNIVERSIDADE FEDERAL
DE MINAS GERAIS

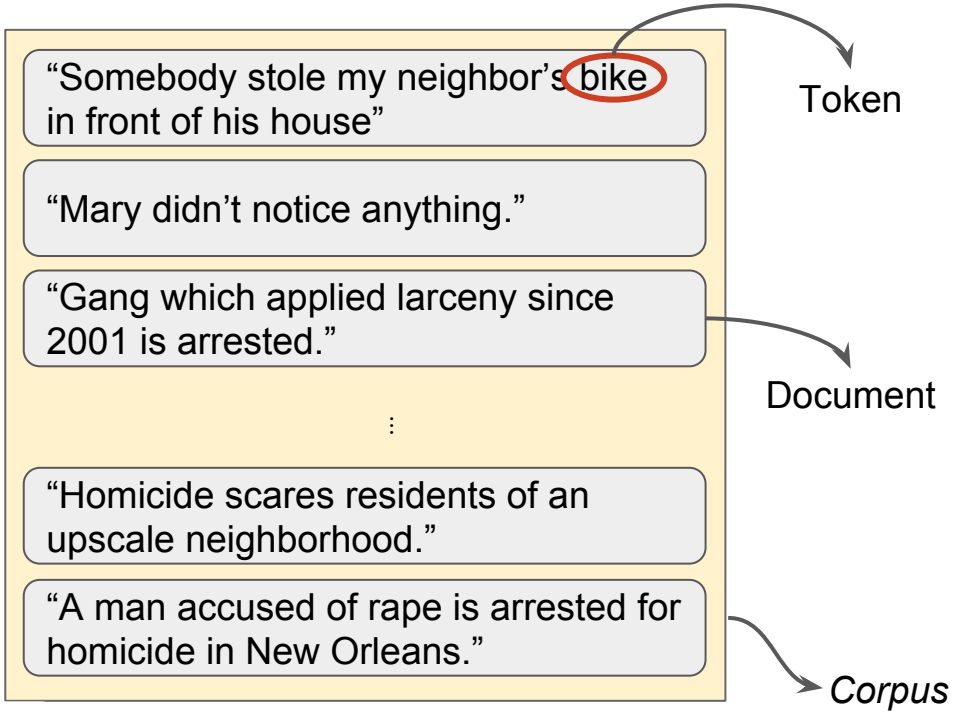
Characteristics

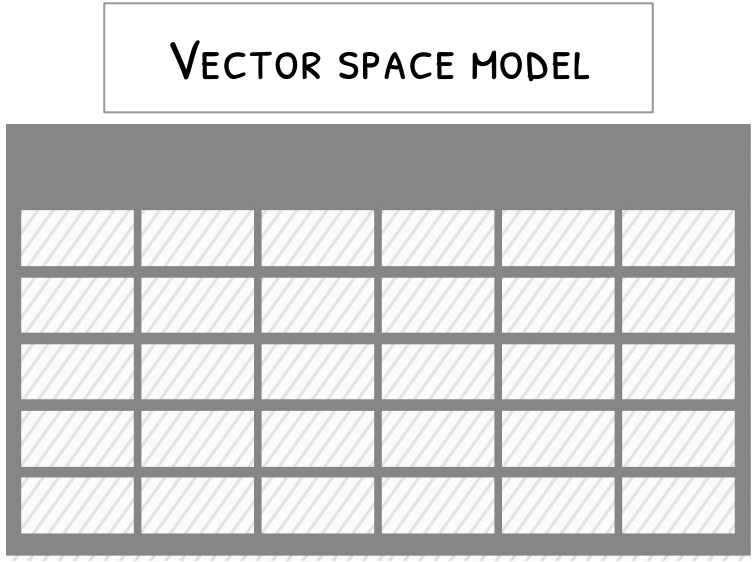
- Variabilidade de domínios
- Aspectos sócio econômicos
- Palavras são polissêmicas
- Formas abreviadas
- Gírias
- Linguagem informal
- Erros gramaticais
- Erros de ortografia
- Pontuação imprópria

Aplicação – Recuperação de Informação



Representação





TOKEN ~ FEATURE ~ VARIABLE

DOCUMENT ~ INSTANCE ~ELEMENTS

Modelo Bag of Words

CARSON, Calif., April 3 - Nissan Motor Corp said it is raising the suggested retail price for its cars and trucks.....

ESSE É UM MODELO MUIIITO SIMPLISTA... MAS ÚTIL!

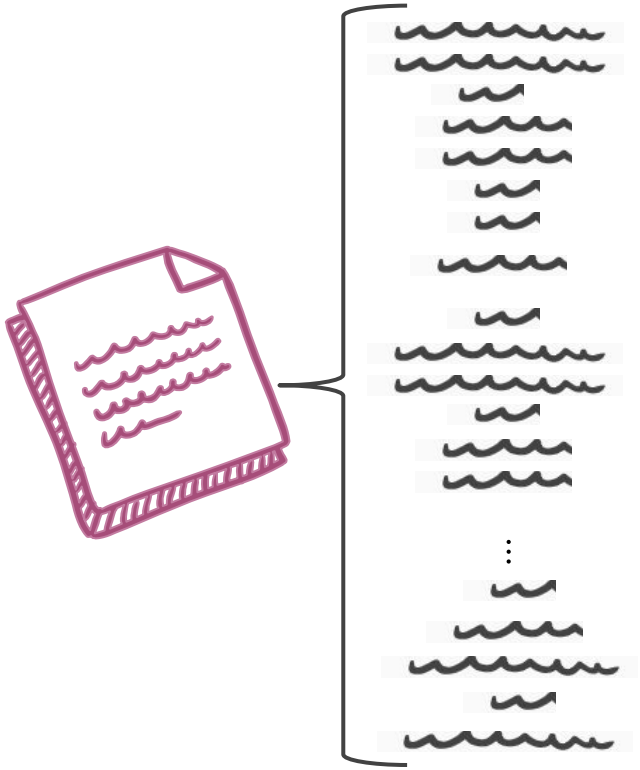
Cada palavra do texto, qualquer que seja a sua posição, é escolhida do dicionário com as mesmas probabilidades e independentemente das demais.

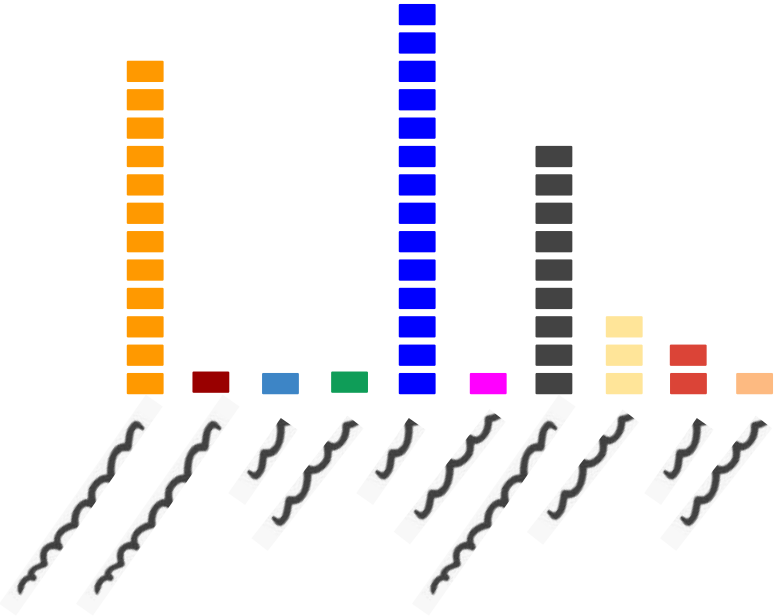
O texto resume-se ao vetor:

X = (N1, N2, N3, ..., NM),

Ni = nº de vezes que a i-ésima palavra do dicionário ocorreu no texto

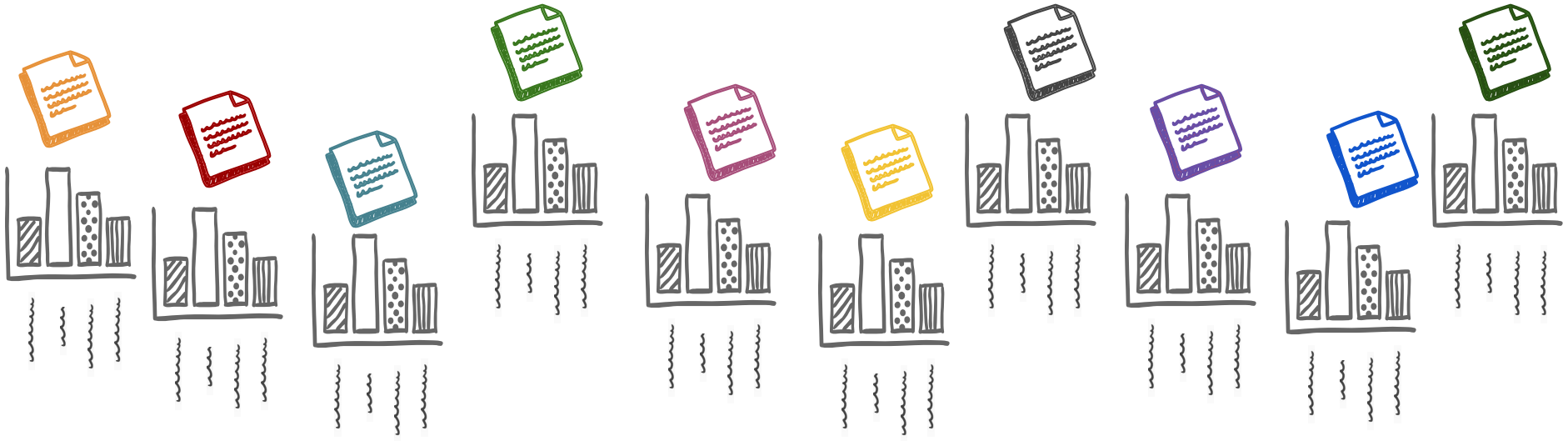
Modelo Bag of Words – Peso





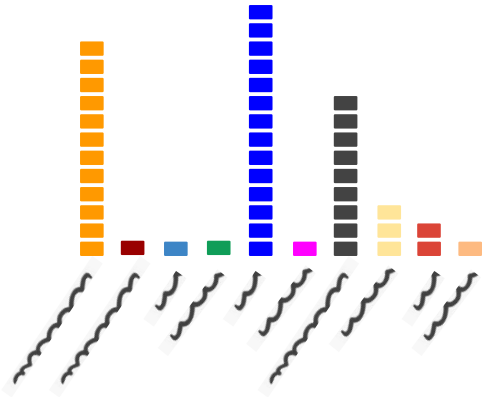
TF: Frequência do termo no documento

Modelo Bag of Words - Peso



IDF: Inverso frequência do termo no Corpus

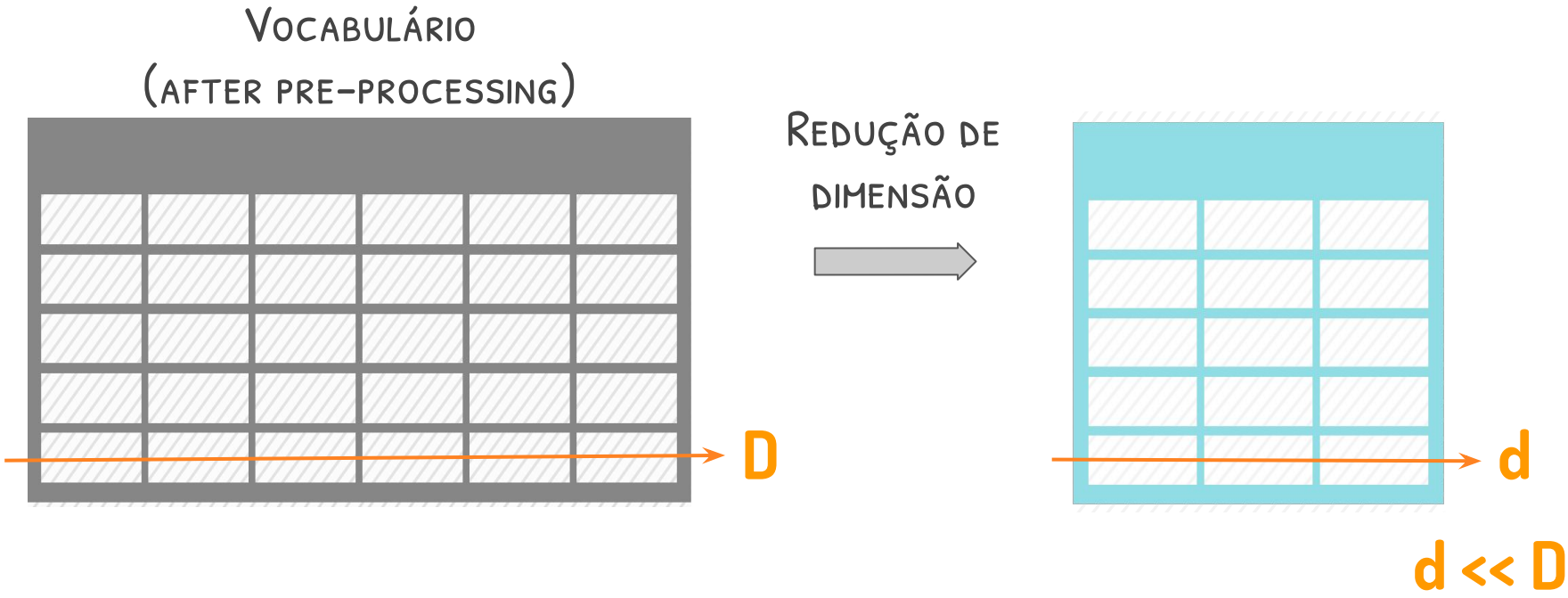
Modelo Bag of Words - Peso



TF: Frequência do termo no documento

IDF: Inverso frequência do termo no Corpus

Representação do texto



Vector Space Model

(Forte) pressuposto:

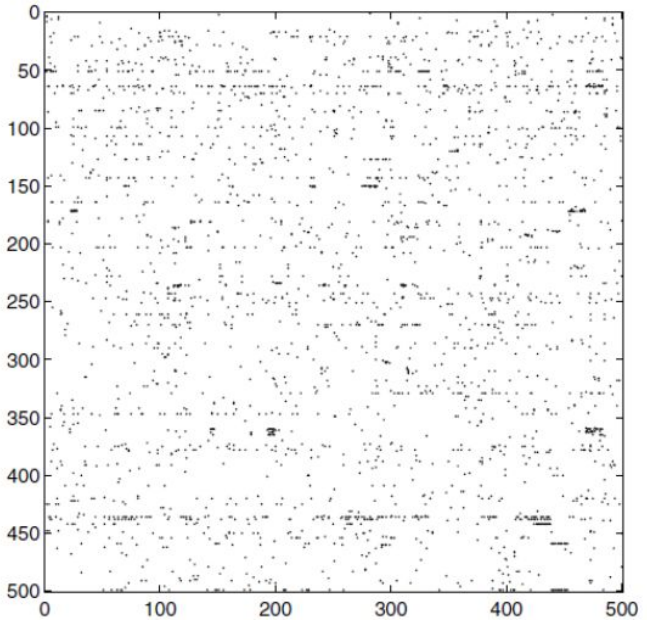
- A ordem dos termos não importa

Doc. representado pela presença dos seus termos:

- A lebre é mais veloz que a tartaruga
- A tartaruga é mais veloz que a lebre

Características:

- Grande dimensão
- Esparsidade



POS – tagger

POS Tagging:

"O processo de atribuir um rótulo de classe gramatical para cada termo em um Corpus."

Word	Lemma	Annotation	Code
ele	ele	Pron. Pes. 3aPes. Masc. Sing. Indef.	NCMS000
pára	parar	Verbo Princ. Indicat. Pres. 3aPes. Sing.	VMIP3S0
todos os dias	todos os dias	Adv. Geral	RG00000
para	para	Prepos. Simples	SP00000
ir	ir	Verbo Principal Infinit.	VMN0000
para	para	Prepos. Simples	SP00000
a	o	Determ. Artigo Fem. Sing. Indefin.	DAIFS00
faculdade	faculdade	Nome Comum Fem. Sing.	NCFS00
.	.	Pontuação	Fd

POS - tagger

Modelo estatístico:

"Um modelo para uma sequência de palavras"

Preposições
Conjunções
Artigos
Pronomes
Numerais

Substantivos
Verbos
Adjetivos
Adverbios

POS - tagger


ARTIGO
ARTIGO
THE
DOG
ATE
THE
BONE
SUBSTANTIVO
VERBO
ADJETIVO
ADVERBIO
SUBSTANTIVO
VERBO
ADJETIVO
ADVERBIO
SUBSTANTIVO
VERBO
ADJETIVO
ADVERBIO

POS - tagger

P(SubstantivolArtigo) = 0.64

P(VerbolArtigo) = 0.07

P(AdjetivolArtigo) = 0.29

P(AdvérbiolArtigo) = 0

POS - tagger

DOG	ATE
Substantivo	Substantivo
Verbo	Verbo
Adjetivo	Adjetivo
Adverbio	Adverbio

P(SubstantivoISubstantivo) = 0.08

P(VerboISubstantivo) = 0.17

P(AdjetivoISubstantivo) = 0

P(AdvérbioISubstantivo) = 0

P(SubstantivoIAdjetivo) = 0.72

P(VerboIAdjetivo) = 0

P(AdjetivoIAdjetivo) = 0.03

P(AdvérbioIAdjetivo) = 0

POS - tagger

SUBSTANTIVO
VERBO
ADJETIVO
ADVERBIO

P(ArtigoISubstantivo) = 0.08

P(ArtigoIVerbo) = 0.17

P(ArtigoIAdjetivo) = 0

P(ArtigoIAdvérbio) = 0

POS - tagger

THE BONE

Substantivo

Verbo

Adjetivo

Adverbio

P(SubstantivolArtigo) = 0.08

P(VerbolArtigo) = 0.17

P(AdjetivolArtigo) = 0

P(AdvérbiolArtigo) = 0

The dog ate the bone

P(Artigo)

P(Substantivo|Artigo)
P(Verbo|Artigo)
P(Adjetivo|Artigo)
P(Adverbio|Artigo)

P(Substantivo|Substantivo)
P(Verbo|Substantivo)
P(Adjetivo|Substantivo)
P(Adverbio|Substantivo)

P(Substantivo|Verbo)
P(Verbo|Verbo)
P(Adjetivo|Verbo)
P(Adverbio|Verbo)

P(Substantivo|Adjetivo)
P(Verbo|Adjetivo)
P(Adjetivo|Adjetivo)
P(Adverbio|Adjetivo)

P(Substantivo|Adverbio)
P(Verbo|Adverbio)
P(Adjetivo|Adverbio)
P(Adverbio|Adverbio)

P(Artigo|Substantivo)
P(Artigo|Verbo)
P(Artigo|Adjetivo)
P(Artigo|Adverbio)

P(Substantivo|Artigo)
P(Verbo|Artigo)
P(Adjetivo|Artigo)
P(Adverbio|Artigo)

THE DOG ATE THE BONE

P(ARTIGO)

P(SUBSTANTIVO|ARTIGO)
P(VERBO|ARTIGO)
P(ADJETIVO|ARTIGO)
P(ADVERBIO|ARTIGO)

P(SUBSTANTIVO|SUBSTANTIVO)
P(VERBO|SUBSTANTIVO)
P(ADJETIVO|SUBSTANTIVO)
P(ADVERBIO|SUBSTANTIVO)

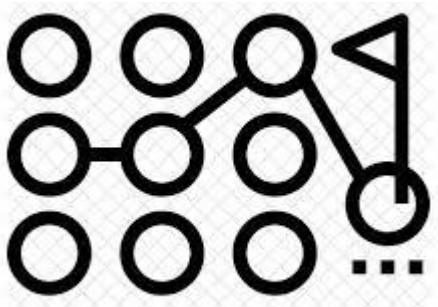
P(SUBSTANTIVO|VERBO)
P(VERBO|VERBO)
P(ADJETIVO|VERBO)
P(ADVERBIO|VERBO)

P(SUBSTANTIVO|ADJETIVO)
P(VERBO|ADJETIVO)
P(ADJETIVO|ADJETIVO)
P(ADVERBIO|ADJETIVO)

P(SUBSTANTIVO|ADVERBIO)
P(VERBO|ADVERBIO)
P(ADJETIVO|ADVERBIO)
P(ADVERBIO|ADVERBIO)

P(ARTIGO|SUBSTANTIVO)
P(ARTIGO|VERBO)
P(ARTIGO|ADJETIVO)
P(ARTIGO|ADVERBIO)

P(SUBSTANTIVO|ARTIGO)
P(VERBO|ARTIGO)
P(ADJETIVO|ARTIGO)
P(ADVERBIO|ARTIGO)



A SEQUÊNCIA DE ANOTAÇÕES MAIS PLAUSÍVEL É A DE MAIOR PROBABILIDADE

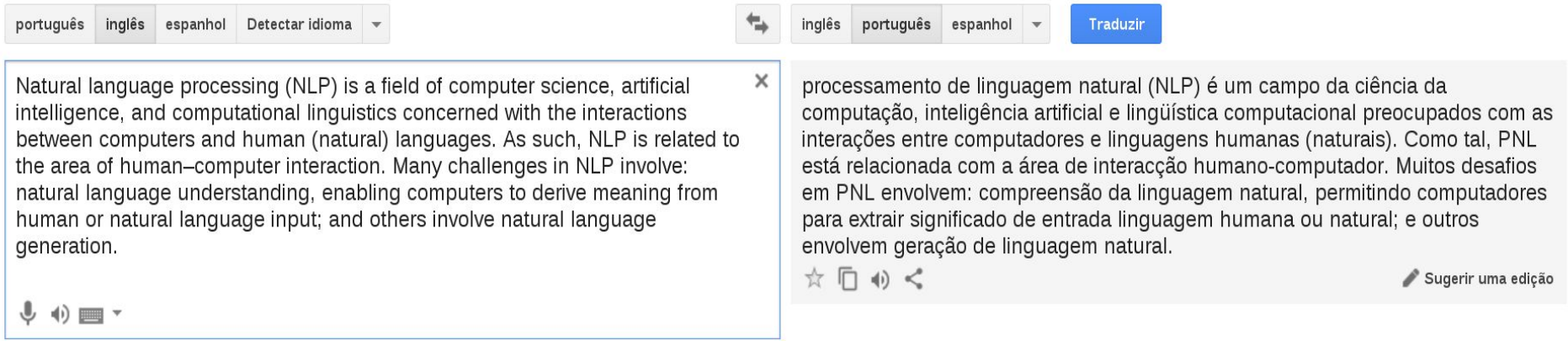
POS - tagger

Probabilidades estimadas como Frequências relativas

Exige o emprego de um Corpus representativo da língua

A ordem de dependência da cadeia (n-gram) é um parâmetro a ser considerado.
Grandes dimensões são reconhecidamente pouco representativas da língua

Aplicação – Tradução de Máquina



Aplicação - Tradução de Máquina

FONTE:

Completo entendimento

ALVO:

Conhecimento sofisticado, poético e criativo

Aplicação - Tradução de Máquina

"Diferenças sintáticas estão relacionadas com diferenças semânticas em como as línguas mapeiam conceitos em palavras." (Jurafsky)

English	*brother*	Japanese	*otooto* (younger)
		Japanese	*oniisan* (older)
		Mandarin	*gege* (older)
		Mandarin	*didi* (older)
English	*wall*	German	*Wand* (inside)
		German	*Mauer* (outside)
English	*know*	French	*connaître* (be acquainted with)
		French	*savoir* (know a proposition)
English	*they*	French	*ils* (masculine)
		French	*elles* (feminine)
German	*berg*	English	*hill*
		English	*mountain*
Mandarin	*tā*	English	*he*, *she*, or *it*

Figure 21.1 Differences in specificity.

"Algumas línguas dividem um dado conceito em mais detalhes do que outras." (Jurafsky)

Tradução de Máquina

SVO:

Alemão
Francês
Inglês

SOV:

Hindu
Japonês

VSO:

Irlandês
Árabe clássico
Hebraico bíblico

Ideia geral:

Alterar a estrutura da língua Fonte para que esteja em conformidade com as regras da língua Alvo.

Tradução de Máquina

Ideia geral:

Alterar a estrutura da língua Fonte para que esteja em conformidade com as regras da língua Alvo.

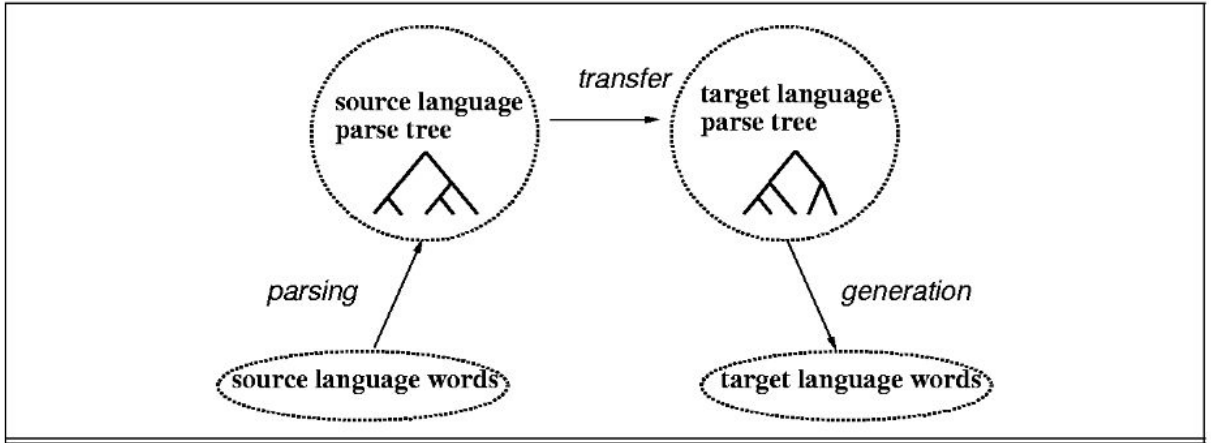


Figure 21.3 The transfer architecture for Machine Translation.

Tradução de Máquina – Transformações Sintáticas

Figure 21.4 A simple transformation that reorders adjectives and nouns

Conjunto distinto de regras de referência por par de língua

Tradução de Máquina - Interlingua

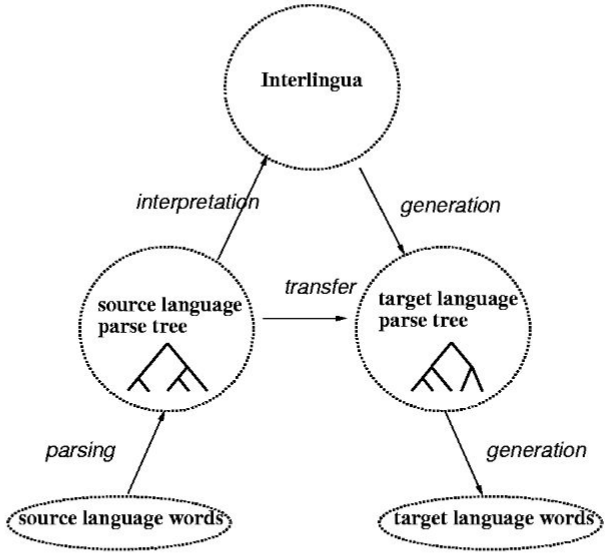


Figure 21.7 Diagram Suggesting the Relation Between the Transfer and Interlingua Models, generally credited to Vauqois.

Necessidade de uma Ontologia: inventário de conceitos e relações

Tradução de Máquina - Tradução Direta

	English to French:
1.	$NP \rightarrow Adjective_1\ Noun_2$ $\Rightarrow$ $NP \rightarrow Noun_2\ Adjective_1$
	Japanese to English:
2.	$Existential\text{-}There\text{-}Sentence \rightarrow There_1\ Verb_2\ NP_3\ Postnominal_4$ $\Rightarrow$ $Sentence \rightarrow (NP \rightarrow NP_3\ Relative\text{-}Clause_4)\ Verb_2$
3.	$NP \rightarrow NP_1\ Relative\ Clause_2$ $\Rightarrow$ $NP \rightarrow Relative\text{-}Clause_2\ NP_1$

Figure 21.5 An informal description of some transformations.

Tendência a serem conservadores

Tradução de Máquina – Tradução Direta

Input:	watashihatsukuenouenopenwojonniageta.
After stage 1:	watashi ha tsukue no ue no pen wo jon ni ageru PAST.
After stage 2:	I ha desk no ue no pen wo John ni give PAST.
After stage 3:	I ha pen on desk wo John to give PAST.
After stage 4:	I give PAST pen on desk John to.
After stage 5:	I give PAST the pen on the desk to John.
After stage 6:	I gave the pen on the desk to John.

Figure 21.9 An Example of Processing in a Direct System

Tradução de Máquina - Tradução Estatística

Traduzir: Adona roi

The lord is my shepherd
The lord will look after me
The lord is for me like somebody who looks after animals with cotton-like hair

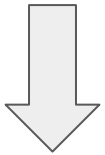
Conclusão: Sempre há uma metáfora, construção, palavra ou tempo verbal sem um paralelo na língua alvo

Transfer, Interlingua, Direct Models:

Qual representação empregar

Quais passos seguir para traduzir

Focar no resultado e não no processo



Tradução Estatística

Tradução de Máquina

O que define uma boa tradução?

Fidelidade à língua Fonte

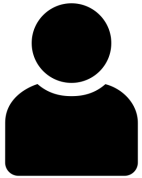
Naturalidade/Fluência para a língua Alvo

Tradução de Máquina

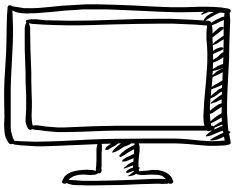
O que define uma boa tradução?

Fidelidade à língua Fonte

Naturalidade/Fluência para a língua Alvo

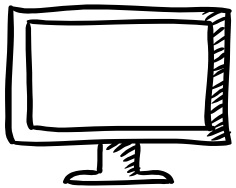


Encontram o tolerável para os dois critérios



?

Tradução de Máquina



Função que represente a importância da fidelidade e da fluência da tradução

$$melhorTraducao(\hat{T}) = argmax_T\, fluencia(T) * fidelidade(T, S)$$

$$melhorTraducao(\hat{T}) = argmax_T\, P(T) * P(T, S)$$

Tradução de Máquina

P(T) ~ Fluência:

Pode ser estimada com modelos do tipo n-gram

Qual tradução é mais fluente?

1. The car was fast
2. The fast car was
3. Fast the was car

Tradução de Máquina – P(SIT) ~ Fidelidade

2-gram (reliable Corpus):

P(car|the) = a
P(was|car) = b
P(fast|was) = c
P(fast|the) = d
P(car|fast) = e
P(the|fast) = f
P(was|the) = g
P(car|was) = h

1. The car was fast 
2. The fast car was
3. Fast the was car

Tradução de Máquina – P(T) ~ Fluência

Das auto was schnell

1. The car was fast
2. The car was red
3. The bicycle was blue

Tradução de Máquina – P(T) ~ Fluência

Bilingual Probabilistic Dictionary

Alemão	Inglês	Prob
Das	the	1
auto	car	0.9
auto	bicycle	0.1
war	was	1
schnell	fast	0.9
schnell	red	0.05
schnell	blue	0.05

Alinhamento a nível de palavra

1. The(1) car(.9) was(1) fast(.9)
2. The(1) car(.9) was(1) red(.05)
3. The(1) bicycle(.1) was(1) blue(.05)

Semantic tagger

Semantic Tags:

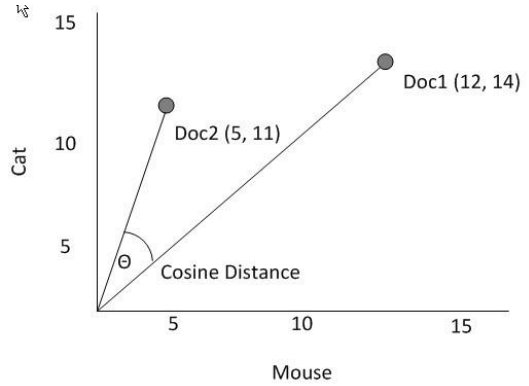
"Sets of words related at some level of generality with the same mental concept."

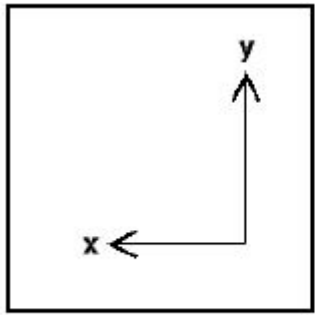
A general and abstract terms	**B** the body and the individual	**C** arts and crafts	**E** emotion
F food and farming	**G** government and public	**H** architecture, housing and the home	**I** money and commerce in industry
K entertainment, sports and games	**L** life and living things	**M** movement, location, travel and transport	**N** numbers and measurement
O substances, materials, objects and equipment	**P** education	**Q** language and communication	**S** social actions, states and processes
T Time	**W** world and environment	**X** psychological actions, states and processes	**Y** science and technology
Z names and grammar			

Semantic tagger

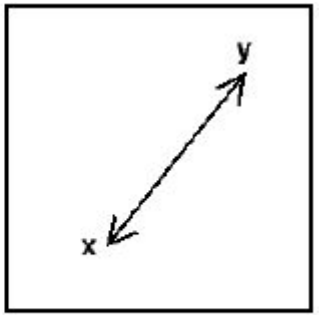
Qual a similaridade entre:

Menina e Menino?
Abelha e Colméia?
Bola e futebol?



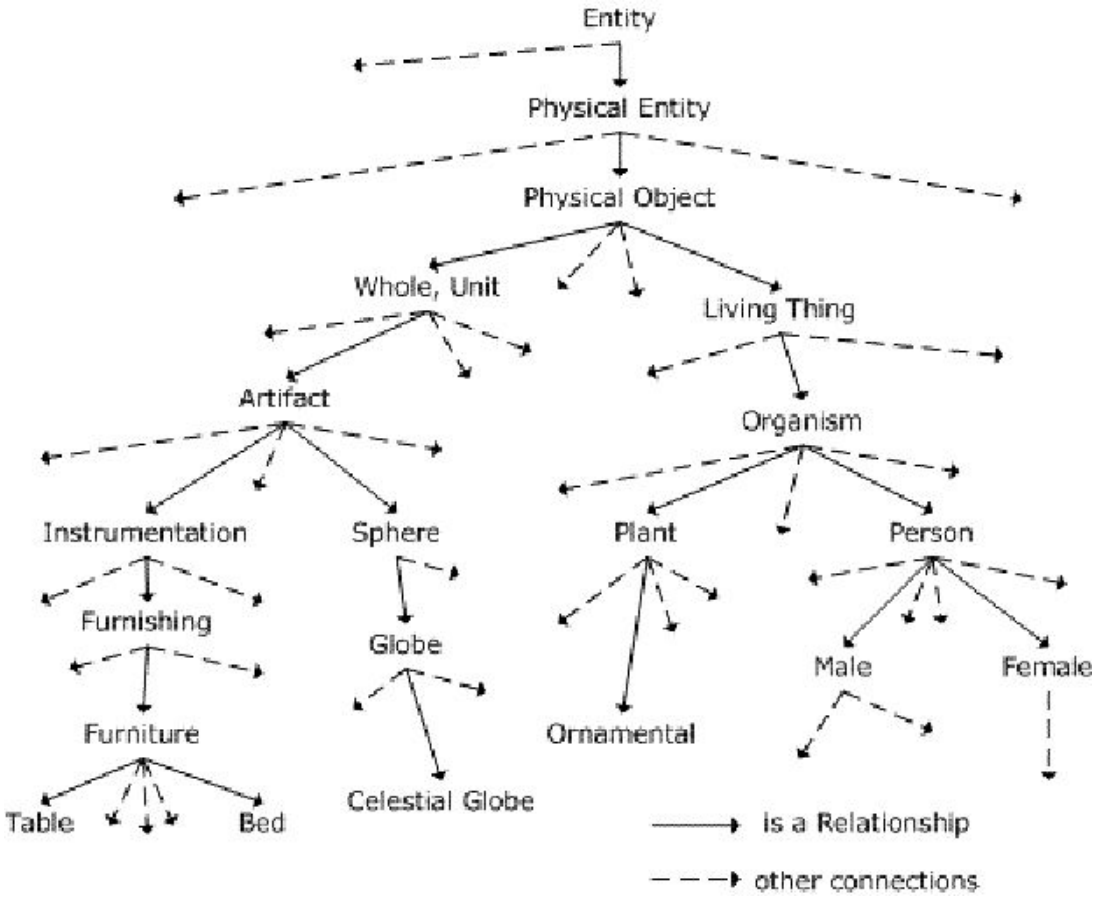


Manhattan

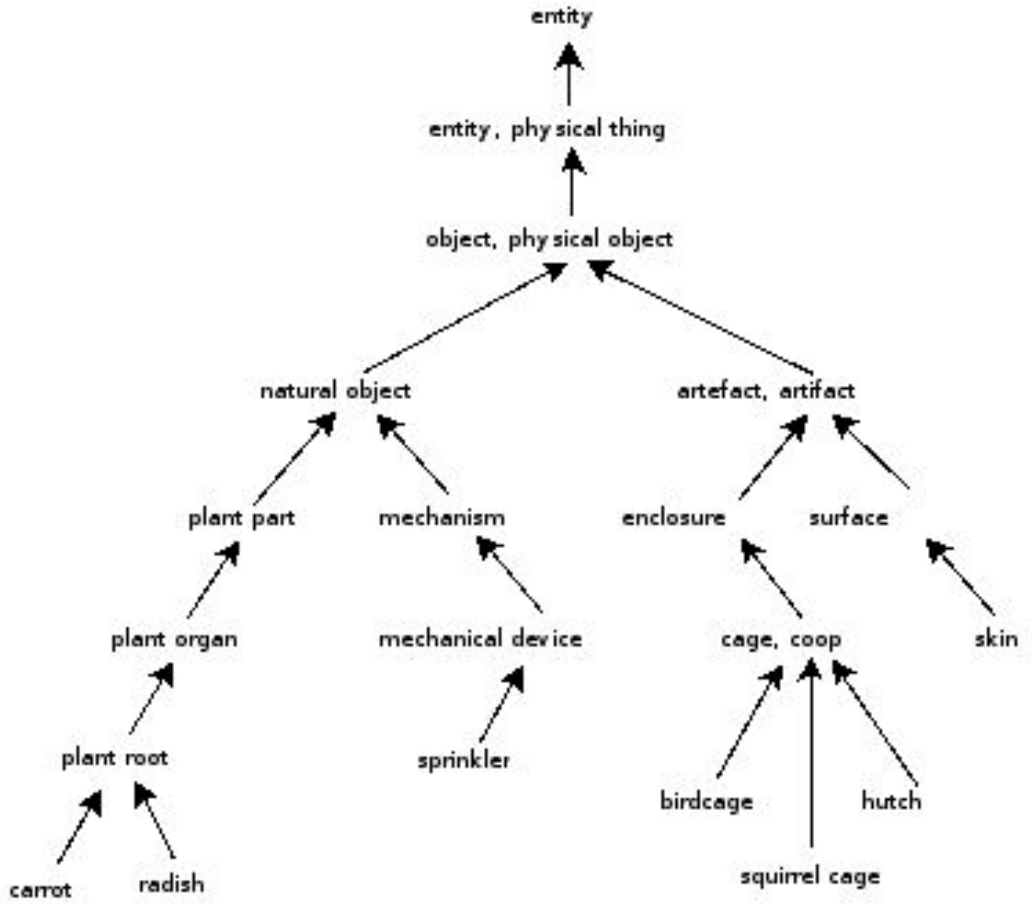


Euclidean

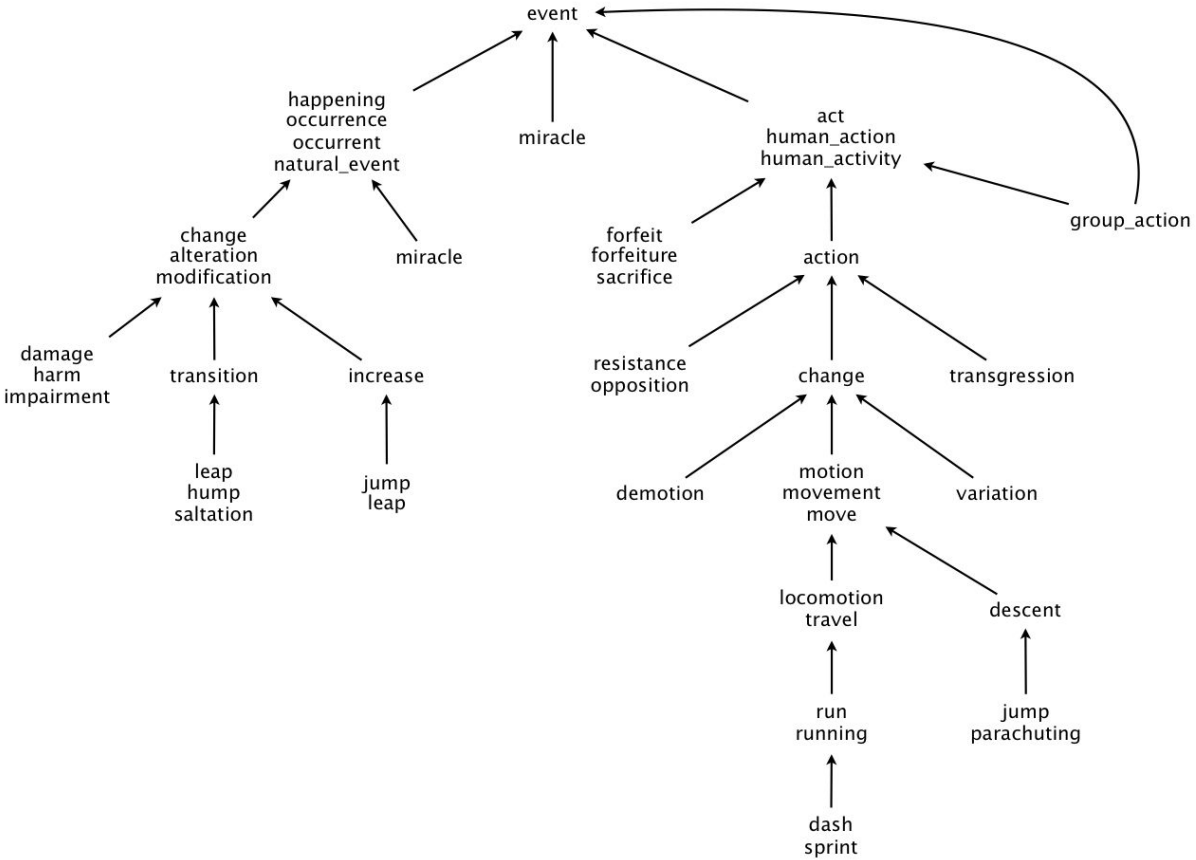
Wordnet/Semantics



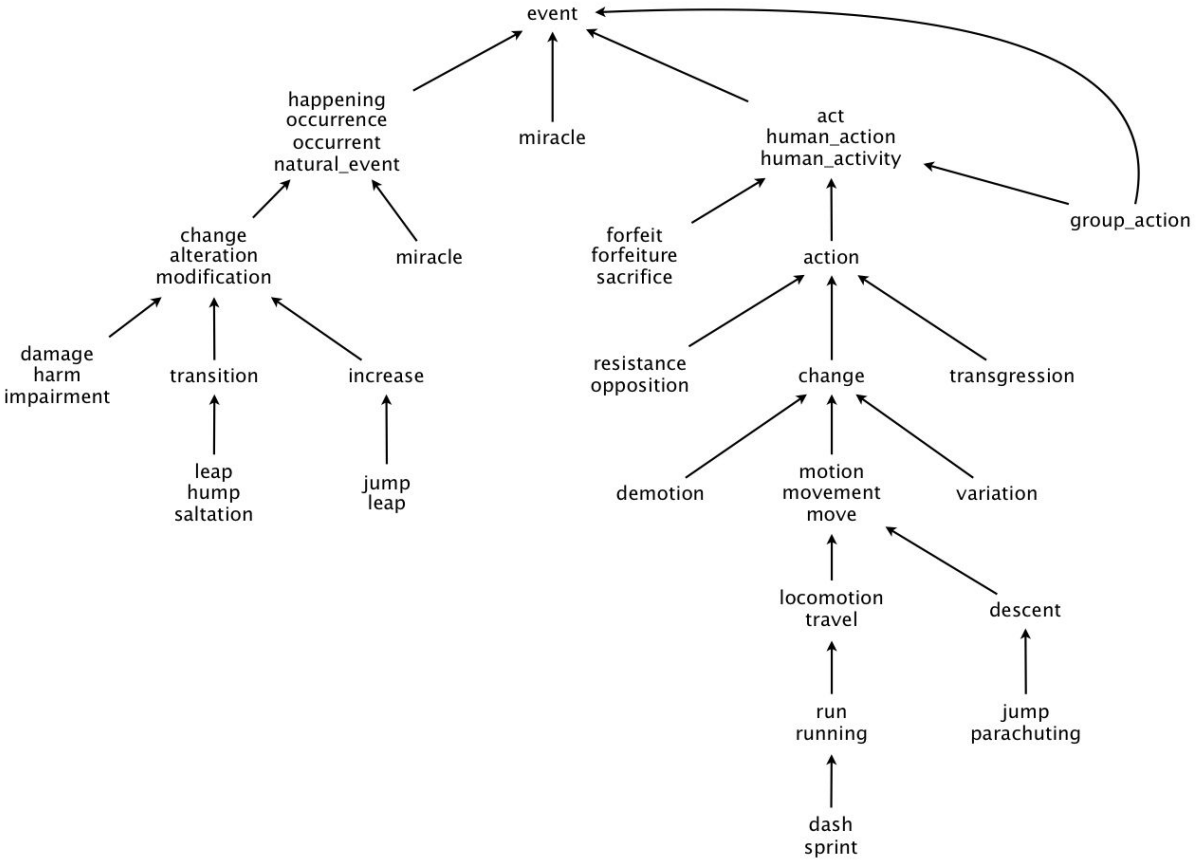
Wordnet/Semantics



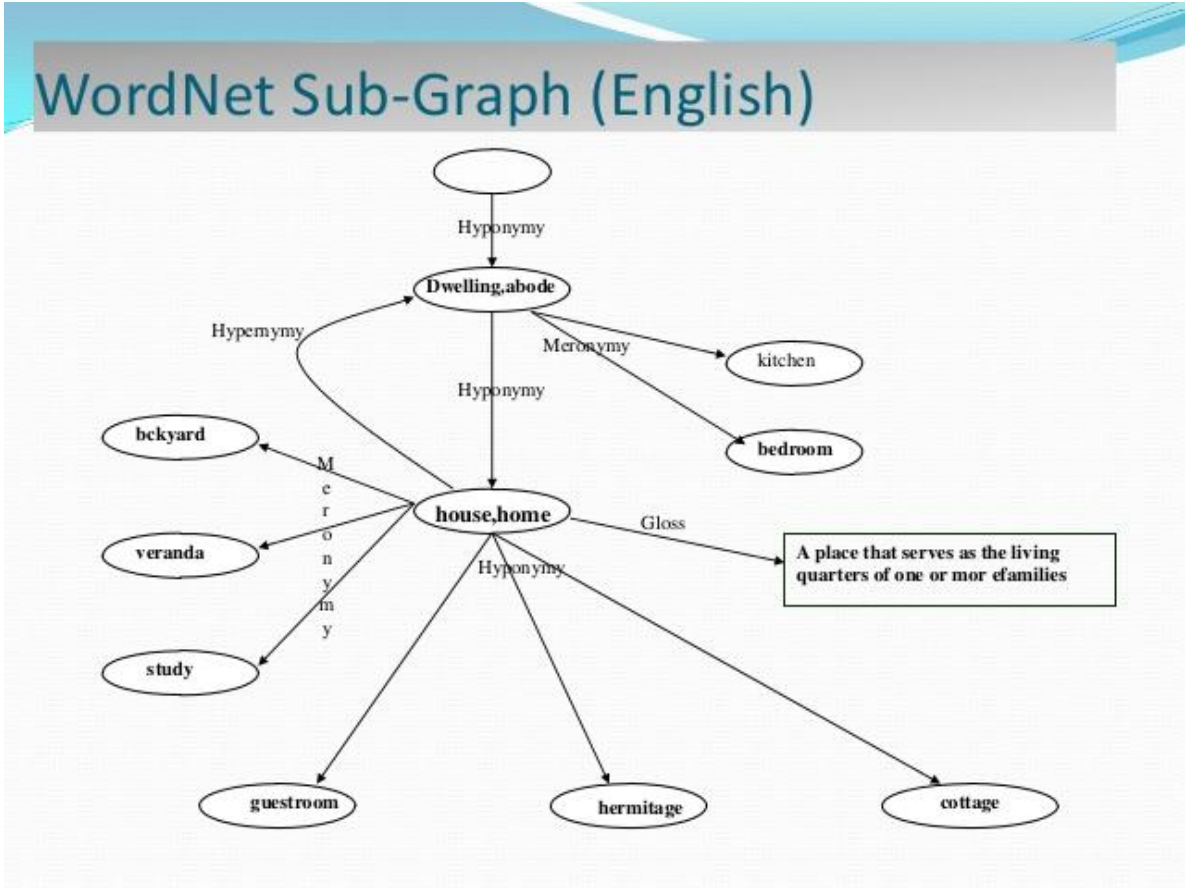
Wordnet/Semantics



Wordnet/Semantics



Wordnet/Semantics



Wordnet/Semantics


```
In [32]: # we can calculate the similarity by looking at the shortest path
         right.path_similarity(minke)

Out[32]: 0.25

In [33]: right.path_similarity(orca)

Out[33]: 0.16666666666666666

In [34]: # Leacock-Chodorow Similarity takes into consideration the distance and
         # the depth of the taxonomy
         right.lch_similarity(minke)

Out[34]: 2.2512917986064953

In [35]: right.lch_similarity(orca)

Out[35]: 1.845826690498331
```

Wordnet/Semantics

- Dicionário digital (muito completo)
- Feito por lexicógrafos
- Disponível para um número limitado de línguas 

Conclusions and Future Work

Uma ferramenta de NLP é input para outra ferramenta de NLP. Os erros só aumentam.

Ferramentas de NLP crescentemente incorporam redes Neurais

Ferramentas de NLP são desenvolvidas com especificidade para domínios.

Ferramentas de NLP dependem (demais) da base de dados. Muita relevância para quem constrói os Corpus.

Ferramentas de NLP são desenvolvidas prioritariamente no mercado, não na academia.