Carros, barcos, mulheres e outras "coisas semelhantes"

(ou "Porque é que nem tudo é o que parece!")

Luís Sarmento
FEUP & Linguateca
http://www.fe.up.pt/~las

Auto-Apresentação

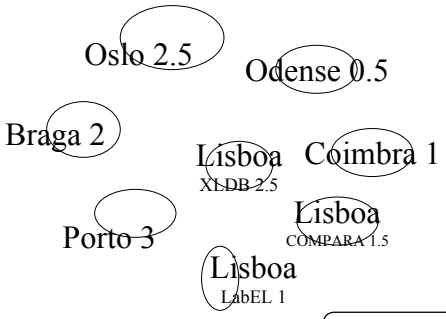
- Aluno de doutoramento Engenharia Informática
  - Faculdade de Engenharia da Universidade do Porto
  - Análise Semântica Robusta do Português
- Orientadores:
  - Prof. Eugénio Oliveira (LIACC: www.fe.up.pt/~eol)
  - Dra. Diana Santos (Linguateca: www.linguateca.pt)
- Últimos 4 anos tenho colaborado com a Linguateca
  - Pólo do Porto (Prof. Belinda Maia)
  - Projectos: Corpógrafo, REPENTINO, SIEMÊS…
- Mais informação sobre mim: www.fe.up.pt/~las/

Linguateca, um projecto para o português

- Centro de recursos distribuído para o processamento computacional da língua portuguesa
- Projecto financiado pela FCT através do POSI (2000-2006)
- Primeiro pólo no SINTEF ICT, Oslo, começou em 2000 (actividade no SINTEF começou em 1998 com o projecto **Processamento Computacional do Português**)

Modelo IRA

- **I**nformação
- **R**ecursos
- **A**valiação



Linguateca num relance, www.linguateca.pt

- > 1000 links Mais de 1.500.000 visitas ao site
- AC/DC, CETEMPúblico, COMPARA … Recursos valiosos para o processamento do português
- *Morfolimpíadas, parte portuguesa do CLEF, HAREM* Avaliação conjunta para o português

- Recursos públicos
- Incentivar a investigação e colaboração
- Medida e comparação formal

- Uma língua, muitas culturas
- Cooperação usando a Web
- Não à adaptação directa das aplicações para o inglês

Contacto: Diana Santos@sintef.no

LIACC -- NIAD&R

- LIACC unidade de investigação criada em 1998 na Universidade Porto
- Composta por 3 Núcleos:
  - NCC- Ciências da Computação - Fac. de Ciências
  - NIAAD- Int, Artificial e Análise de Dados - Fac. Economia
  - **NIAD&R- Robótica, Int. Artificial Distribuida -Fac. Engenharia**
- Comité Organizador :
  - NCC: Prof. Miguel Filgueiras, Dr. Luís Damas
  - NIAAD: Prof. Pavel Brazdil
  - NIAD&R: Prof. Eugénio Oliveira

NIAD&R

- Tópicos de Investigação:
  - Sistemas Multi-Agente
  - Tecnologias de Negócio Electrónico
  - Aprendizagem Automática
  - Robótica
  - Ontologias
  - *Análise Semântica do Português*
- Contacto:
  - Prof Eugénio de Oliveira eco@fe.up.pt

Uma palavra de compreensão…

- Esta não se trata de uma palestra machista!
- Trata-se de uma tentativa de demonstrar as potencialidades dos mecanismos de analogia em processamento automático de linguagem
- Envolve tentar identificar em texto:
  - elementos "semelhantes"
  - elementos "diferentes"
- Algo que por vezes gera resultados "estranhos"…
- Lembrem-se: "Nem tudo é o que parece!!"

Aliás…

- Toda a gente sabe que as mulheres têm uma ***capacidade superior*** de detectar diferenças.
- Por exemplo, uma cena do quotidiano de um casal, enquanto a esposa tenta ajudar o marido a escolher uma camisa para o fato novo:
  - **Maria:** Não percebes? São diferentes! Uma camisa é **cor-de-rosa**, a outra é **salmão**!
  - **José:** Não percebo!… Para mim são iguais… **Salmão** é um **peixe**, não é uma **cor**!
- É precisamente sobre este "**problema masculino**" que iremos falar ao longo desta apresentação…

Motivação (1)

- Sistemas de Processamento de Linguagem Natural:
  - Sistemas de Extracção de Informação
    - Resposta Automática a Pergunta
  - Sistemas de Sumarização Automática
  - Sistemas de Tradução Automática
  - …
- Envolvem capacidades de Análise Semântica:
  - Identificação e Extracção de Terminologia
  - Identificação e Classificação de Entidades Mencionadas
  - Descoberta de Relações Semânticas
  - Análise de Papeis Semânticos
  - Desambiguação de Sentidos
  - …

Motivação (2)

- Sistemas de Análise Semântica:
  - Complexos de construir
  - Envolvem normalmente grandes conjuntos de regras:
    - Tornam-se complexos de manter!
  - Normalmente são especializados
    - num determinado **domínio** (ex: texto médico)
    - num determinado **género de texto** (ex: artigos científicos)
  - Por vezes é difícil adaptá-los para outras situações
    - Nem sempre lidam bem com situações "novas"

Motivação (3)

- Humanos têm a capacidade de
  - estabelecer semelhança entre elementos do mundo
  - traçar analogias entre vários objectos
- Estas capacidades ajudam a que sejamos capazes de compreender mais facilmente certa informação presente em texto
- Permite ultrapassar a ausência ou desconhecimento sobre alguns elementos visto em texto
  - XXX deve beber-se fresca
  - Eu gosto de XXX enquanto estou a ver futebol
  - Se beber muita XXX ganha barriga

Motivação (4)

- Esta capacidade seria de grande utilidade na análise automática de texto porque:
  - muitas vezes os sistemas automáticos são confrontados com situações completamente novas
  - é impossível codificar regras para tratar de todos os casos explicitamente
  - é impossível codificar informação sobre todos os elementos do mundo na bases de conhecimento dos sistemas.
- A capacidades de identificar semelhanças e estabelecer analogias em texto seria por isso muito útil na Análise Semântica

Exemplo Concreto (1)

- Reconhecimento de Entidades Mencionadas
  - Tarefa de Análise Semântica
  - Objectivo: classificar Nomes em Contexto
- Por exemplo:
  - A aldeia foi inundada pelo **Lima**
- Mas "Lima":
  - Sobrenome, Capital do Perú, Rio em Portugal, etc…
    - A conferência foi organizada pelo **Lima**
    - A conferência foi organizada em **Lima**
- Como conseguir Analisar correctamente:
  - Será que deveremos ter uma regra para TODAS as possibilidades

Exemplo Concreto (2)

- Como resolver esta ambiguidade?
  - "A aldeia foi inundada pelo **Lima"**
- Pesquisando situações "Semelhantes":
  - Usando o Google: {foi inundada pelo X}, X =
    - "rio","lago", "mar", etc…
    - "Rio Jaboatão", "Rio Beberibe", etc.
- SEM usar qualquer regra especializada **parece ser possível** obter informação importante acerca de objectos "Semelhantes"…
- Ajuda a resolver o problema de ambiguidade

Algumas Questões em Pesquisa

- Como se pode **estabelecer** e **medir** a semelhança entre:
  - Elementos? - ex: “banana” ~ “morango”
  - Contextos? - ex: “beber uma” ~ “tomar uma”
- Como **identificar** os contextos em que determinados elementos podem ser “semelhantes” (ou não)?
  - Cor-de-rosa ~ salmão (cores)
  - Bacalhau ~ salmão (peixes / alimentos)
  - Cor-de-rosa ~ Bacalhau?
- Como incluir estas capacidades em sistemas de Análise Semântica?
  - Como calcular semelhança?
  - Que recurso extra são necessários?

Uma coisa de cada vez…

- Vamos focar numa questão apenas:
  - Dado um elemento, ou um conjunto de **elementos semelhantes**, tentar encontrar em texto outros **elementos semelhantes**
- Ex:
  - [azul, salmão] -> [verde, preto, carmim, …]
  - [bacalhau, salmão] -> [sardinha, truta, peixe-espada…]
- Exemplo inspirador:
  - GoogleSets - labs.google.com/sets/

Motivações para começar por este caso

- Essencialmente por **questões práticas**:
  - auxílio à construção de recursos léxico-semânticos para a língua Portuguesa do tipo WordNet:
    - Partindo de uma "pequena" WordNet expandir a rede por pesquisa de semelhantes…
  - construção de *topic maps* para navegação em colecções
  - expansão que expressões de pesquisa em motores
- E também, porque *parecia* mais fácil…
  - Mas relembrando: "Nem tudo é o que parece…"

Formalizando um pouco mais

- Dado um conjunto de elementos da mesma **classe conceptual**, pretende-se encontrar mais elementos dessa **classe conceptual**.
- Os elementos da classe estabelecem uma relação **especialização** relativamente ao conceito da classe:
  - Classe "cor" -> [amarelo, verde, branco, etc]
    - "amarelo", "verde", "branco" são **Hipónimos** de "cor"
    - "cor" é **Hiperónimo** de "amarelo", "verde", "branco"
    - "amarelo", "verde", "branco" são **co-hipónimos** entre si (de "Cor")
- O nosso objectivo: **pesquisa de Co-Hipónimos!**

Duas Estratégias Alternativas para resolver este problema

- Métodos Linguisticamente Informados
- Baseiam-se numa prévia Análise Morfológica
- Tentam explorar regularidades a nível sintáctico

- Métodos "Data-driven" que recorrem a convergência estatística
- Recorrem a quantidades massivas de texto
- Tentam encontrar padrões sobre as ocorrências lexicais

Duas Estratégias Alternativas para resolver este problema (2)

- Métodos Linguisticamente Informados
- Permitem aproximações de mais alto nível mas necessitam de Analisadores e estes nem sempre estão disponíveis…
- Apesar de haver analisadores de grande precisão, na prática podem apresentar alguma falta de robustez em lidar com texto menos "bem-comportado"

- Tecnicas "Data-driven"
- Permitem aproximações muito simples do ponto de vista algorítmico, mas necessitam de uma quantidade de texto enorme até se obter resultados interessantes…
- Permitem obter resultados razoáveis quase sempre, embora quase nunca resultados excelentes.
- São dificeis de "controlar"

Estratégia Escolhida: “Data-driven”:

- Usando o WPT03
  - colecção web disponível publicamente
  - cerca de 1.5 Milhões de Documentos Web
  - Cerca de 1000 Milhões de palavras
- Difícil de pesquisar directamente:
  - Codificação em formato MySQL: BACO
  - Criação de várias tabelas de auxiliares pesquisa
  - Já iremos ver em mais detalhe…

Voltando ao problema…

- Pesquisa de “Elementos Semelhantes”
  - Mas o que são “Elementos Semelhantes”?
  - Como se encontram “Elementos Semelhantes”?
- Definição operacional:
  - Elementos Semelhantes são elementos que ocorrem em “situações semelhantes” (contextos semelhantes) ao longo do texto

Contextos Semelhantes

- "abacaxi" e "pitanga" são semelhantes porque ocorrem em contextos semelhantes:
  - Eu gosto de suco de abacaxi
    Eu gosto de suco de pitanga
  - O abacaxi está maduro
    A pitanga está madura
- Mas "alumínio" e "pitanga" não são semelhantes
  - Comprei uma porta em pitanga (?)
  - O alumínio está maduro (?)

Um método para encontrar elementos semelhantes…

- Ao fim de 24 slides, chegou o momento…
  - Obrigado pela paciência!
- Dado um conjunto de elementos semelhantes
  1. Procurar os contextos em que este elementos ocorrem
  2. Procurar elementos que ocorrem nos MESMOS contextos
- Estes elementos deverão ser semelhantes aos iniciais
  - Serão tanto mais semelhantes quantos os contextos que os novos elementos possuírem em comum com os iniciais
  - Mesmo que "alumínio" e "pitanga" tenham contextos em comum, "abacaxi" e "pitanga" devem ter muitos mais…

Mas como definir o contexto?

- Há muitas possibilidades:
  - Frases / sentenças
  - Palavras que antecedem? "p1,p2…pn banana"
    - Quantas palavras?
  - Palavras que sucedem? "banana p1,p2 …pn"
    - Quantas palavras?
  - Palavras que co-ocorrem?
    - Banana co-ocorre com "madura" "fruta" "iogurte" "comer"…
    - Algumas destas palavras apenas (verbos, nomes…)?
- ???

Reformulando…

- Qual o **contexto mínimo** que permite identificar elementos semelhantes?
  - É importante identificar o "mínimo" porque permite utilizar aproximações mais simples…
- Contexto pode ser apenas 1 palavra antes:
  - "**comer** banana"
  - "**descascar** banana"
- Embora pareça muito pouco:
  - "**comer** rápido"
  - "**descascar** tudo"

Experimentando…

- Vamos assumir que o contexto pode ser definido pelas **2 palavras anteriores** ao elemento:
  - {P1 P2} ELEMENTO
    - {iogurte de} morango
    - {cadeira em} acrílico
    - {whisky com} gelo
    - {descascar uma} banana
- A pesquisa de contextos de um determinado elemento torna-se assim bastante mais simples

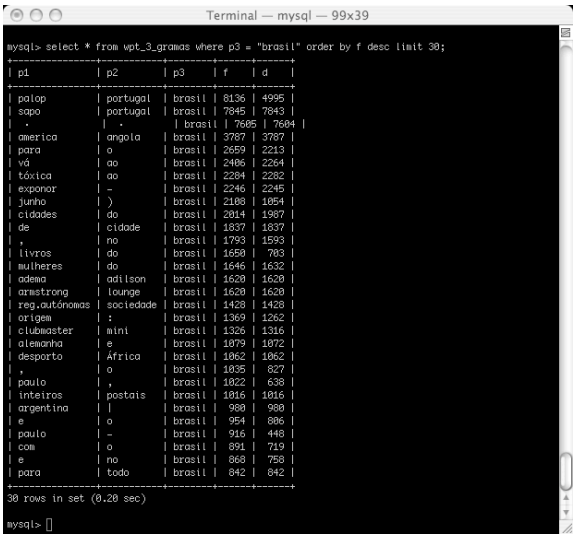
Onde pesquisar os contextos

- WPT03 ~ 6GB ~ 1.5 M Docs ~ 1000 palavras
- Foi transformado numa BD chamada BACO
- BACO possui várias tabelas:
  - Frases: 35M
  - n-gramas (sequências 1, 2, **3**, 4 palavras): *273*M
  - co-ocorrências: 780M
- Dados + Índices BACO ~ 80 GB
  - Muita redundância de dados tentando obter representações que permitam pesquisas rápidas (?)

Usando o BACO

- Tabela de 3-Gramas:
  - Contagem da sequências de 3 palavras previamente compilada: cerca de 90 M de tuplos
- Permite pesquisas "rápidas" para obter:
  - contextos de um elemento
  - os elementos que ocorrem num dado contexto
- Será a informação base para o nosso processo de obtenção de semelhanças…

P1 P2 "Brasil" no BACO



Temos então tudo para começar...

- Para os elementos do conjunto S = $\{s_1, s_2, .. s_n\}$
  - Procurar no BACO contextos de 2 palavras: c_1, c_2, s_i
- Para contextos "representativos" c_{i1}, c_{i2}, X_i:
  - Procurar elementos X_i que ocorram nesses contextos
- Medida de Semelhança:
  - número de "contextos representativos" em comum entre o candidato e os elementos do conjunto inicial
- Filtram-se os candidatos como preposições, artigos, marcas de pontuação e certos verbos muito frequentes

Contextos "representativos"

- Contextos "representativos" co-ocorrem com:
  - Um determinado número mínimo de elementos iniciais
    - Parâmetro **L**
    - Objectivo: garantir a especificidade dos contextos
  - Mas não co-ocorrem com um número demasiado elevado de palavras (250).
    - Por exemplo:
      - "tipo de X" --> Há mais de 7000 possibilidades X
      - "por exemplo X" --> Há mais de 1500 possibilidades para X
    - Objectivo: ignorar contextos demasiado genéricos e sem grande valor semântico…

Resultados (1)

#	Conjunto inicial	L	#C	Resultado
1	amarelo, vermelho, azul	3	183	verde (97), branco (81), preto (79), castanho (53), cinzento (48), negro (41), laranja (39), cor (37), rosa (36), cinza (28), dourado (24), cores (23), escuro (17), roxo (16), branca (16), cor-de-rosa (14), vermelha (14), ouro (14), está (14), creme (13), prateado (13), violeta (13), bem (13), natural (12), sépia (12), encarnado (12), púrpura (12), tipo (12), amarela (12), claro (11), ocre (11), lilás (11), preta (11)
2	rosa salmão	2	24	sousa (4), oliveira (4), verde (4), amarelo (4), email (3), azul (3), lima (3), andrade (3), preto (3), data (3), campos (3), sá (3),…
3	bacalhau, salmão	2	74	peixe (38), atum (27), arroz (26), carne (25), camarão (22), frango (21), queijo (20), robalo (19), polvo (19), feijão (17), pescada (16), vitela (16), marisco (15), tamboril (15), pato (14), cabrito (14), garoupa (14), lulas (14), pão (14), linguado (14), legumes (14), cherne (13), …
4	rosa, laranja	2	126	verde (37), azul (35), vermelho (30), amarelo (30), preto (24), cinzento (21), cinza (20), castanho (20), branco (20), violeta (18), branca (13), lilás (12), púrpura (12), vermelha (12), cor (11), castanha (11), amarela (11), creme (10), roxo (10)…

Resultados (2)

#	Conjunto inicial	L	#C	Resultado
5	Curitiba, florianopolis recife	2	39	lisboa (19), leiria (14), part (13), instituto (13), portugal (13), estados (13), exterior (12), pt (12), medico (12), pesquisa (11), nutricionista (11), gerente (11), origem (11), oferta (11), procura (11), trabalho (11), informal (11), feira (11), procurar (11), pleno (11), são (10), quero (10), primeiro (9), net (9), entrevista (8), porto (7), sp (4), emprego (4), londres (4), campinas (4), brasília (4), rio (4), …
6	mármore granito	2	166	pedra (83), madeira (58), betão (29), forma (28), vidro (27), ferro (25), calcário (25), portugal (23), alumínio (23), aço (21), cimento (20), metal (20), xisto (20), cantaria (19), material (17), ouro (17), papel (16), …
7	whiskey, rum, vodka	2	6	cerveja (4), vinho (4), sumo (4), tequila (4), aguardente (4), azeite (3), coca-cola (3), brandy (3), leite (3), refrigerante (3), vodka (3), **vidro** (3), pinga (3), 33 (3), licor (3), vinagre (3), cachaça (3)…
8	carro, barco	2	**459**	sistema (70), novo (68), grupo (67), mundo (67), jogo (61), país (58), dia (56), programa (55), trabalho (55), automóvel (55), avião (55), veículo (54), corpo (53), homem (49), processo (49), campo (48), lugar (47), conjunto (46), serviço (45), autocarro (45), filme (44), tempo (44), meio (43), livro (42), ponto (42), comboio (42)

Comentários a esta aproximação (1)

- O topo da lista é quase sempre constituído pelos candidatos que se esperavam (+)
  - Excepção para Conj. 6: palavras mais "raras"
    - Falta matéria textual para a convergência
  - Excepção para Conj. 8 foram gerados demasiados contextos…
- Alguns candidatos estão incompletos (-)
  - "rio", "são" (Conj. 5)
  - limitação intrínseca ao facto de estarmos a procurar em contextos de 3 palavras…

Comentários a esta aproximação (2)

- Sensibilidade a certas ambiguidades (-)
  - "garrafa de X": X = "vodka" ou X = "vidro"
  - Será grave?
  - Como detectar estes casos e filtrá-los?
- Algoritmo elegante mas pouco eficiente (-)
  - Tempo de execução: 1m até 25m iBook G4
  - Impede a sua execução em tempo-real, como auxílio a sistemas de análise semântica "on-line"

Estamos contentes com este método?

- Tem alguns pontos interessantes:
  - Elegante, e é fácil de compreender
  - Pode ser usado com auxílio à construção de Ontologias
- Mas:
  - demasiado lento
    - Por enquanto permite apenas a utilização off-line
  - critérios de selecção de contextos relevantes e de elementos semelhantes é talvez demasiado simples
    - Necessidade de recorrer a Medidas de Associação Estatística mais apropriadas…

Método 2: Uma alternativa simples

- Observação:
  - Elementos semelhantes ocorrem no contexto de **coordenações**, isto é são **enumeráveis / listáveis**
    - "gelado de morango, laranja e limão"
    - "prova de matemática e física ou química"
- A obtenção de elementos semelhantes pode ser feita procurando elementos que se coordenam com os elementos do conjunto inicial
  - Basta procurar padrões com possíveis coordenações:
    - , X e Y…
    - , X ou Y…

Método 2: Preparação dos dados

- Pré-selecção de tuplos para ***alguns*** padrões (12) usando a tabela de 4-gramas do BACO
- Geração de tabela auxiliar par(x,y)
- Tabelas resultante mais pequena o que permite pesquisas mais rápidas

Padrão	Tuplos recolhidos
, X e Y	179415
, X ou Y	25.203
, X , Y	399.013
X , Y ,	428.746
X , Y e	202.619
X, Y ou	28.941
X e o Y	112.746
X e a Y	153.477
X ou o Y	6.824
X ou a Y	13.083
X , o Y	207.068
X , a Y	271.152
Total	**2.028.287**

Coordenações com “Brasil”

...ysql — 39x39

```
mysql> select * from wpt_pares where p1
 = "brasil" order by f desc limit 30;
+--------+-------------+------+------+
| p1     | p2          | f    | d    |
+--------+-------------+------+------+
| brasil | cabo        |  189 |  184 |
| brasil | varig       |  170 |  159 |
| brasil | gol         |  159 |  159 |
| brasil | tam         |  159 |  159 |
| brasil | transbrasil |  159 |  159 |
| brasil | portugal    |  155 |  155 |
| brasil | uruguai     |  128 |  128 |
| brasil | canadá      |  118 |  115 |
| brasil | ásia        |  114 |  114 |
| brasil | espanha     |  112 |  110 |
| brasil | canadá      |  107 |  107 |
| brasil | angola      |  103 |  102 |
| brasil | espanha     |  101 |   96 |
| brasil | argentina   |   94 |   86 |
| brasil | chile       |   89 |   84 |
| brasil | chile       |   83 |   78 |
| brasil | 101         |   78 |   41 |
| brasil | portugal    |   70 |   68 |
| brasil | uruguai     |   66 |   66 |
| brasil | argentina   |   60 |   58 |
| brasil | que         |   60 |   55 |
| brasil | lisboa      |   58 |   39 |
| brasil | portugal    |   58 |   42 |
| brasil | angola      |   58 |   57 |
| brasil | em          |   58 |   52 |
| brasil | macau       |   58 |   38 |
| brasil | frança      |   56 |   52 |
| brasil | e           |   54 |   53 |
| brasil | frança      |   53 |   52 |
| brasil | para        |   52 |   52 |
+--------+-------------+------+------+
30 rows in set (0.08 sec)
```

...ysql — 39x39

```
mysql> select * from wpt_pares where p2
 = "brasil" order by f desc limit 30;
+-----------+--------+------+------+
| p1        | p2     | f    | d    |
+-----------+--------+------+------+
| comercial | brasil |  647 |  159 |
| paulo     | brasil |  444 |  229 |
| portugal  | brasil |  317 |  239 |
| focinhos  | brasil |  244 |  244 |
| angola    | brasil |  242 |  229 |
| argentina | brasil |  216 |  211 |
| portugal  | brasil |  168 |  157 |
| portugal  | brasil |  152 |  126 |
| )         | brasil |  130 |  130 |
| bolívar   | brasil |  124 |  124 |
| espanha   | brasil |  122 |   90 |
| janeiro   | brasil |  118 |  106 |
| sp        | brasil |  115 |   79 |
| azeituna  | brasil |  105 |  104 |
| austrália | brasil |   83 |   61 |
| rs        | brasil |   83 |   72 |
| bélgica   | brasil |   81 |   74 |
| bélgica   | brasil |   78 |   74 |
| ms        | brasil |   76 |    7 |
| espanha   | brasil |   69 |   65 |
| argentina | brasil |   66 |   64 |
| frança    | brasil |   66 |   65 |
| espanha   | brasil |   65 |   65 |
| bahia     | brasil |   64 |   55 |
| 2002      | brasil |   64 |   64 |
| austráiio | brasil |   61 |   41 |
| canadá    | brasil |   58 |   51 |
| frança    | brasil |   56 |   56 |
| angola    | brasil |   56 |   56 |
| espanha   | brasil |   50 |   50 |
+-----------+--------+------+------+
30 rows in set (22.83 sec)
```

Método 2: Algoritmo (1)

- Para os elementos do conjunto inicial S = $\{s_1, s_2, ..s_n\}$, procurar em par(X,Y):
  - X para os quais Y = s_i
  - Y para os quais X = s_i
  - … contabilizando número de contextos distintos / pares em os elementos ocorrem coordenados
- Medida de Semelhança:
  - número de contexto distintos de coordenação em que o candidato ocorre:
    - 12 por cada elemento do conjunto S, há 12 possibilidades de coordenação

Método 2: Algoritmo (2)

- Possibilidade de um parâmetro L de qualidade:
  - L número mínimo de elementos de S com os quais o candidato ocorre coordenado
- Eliminar elementos "ruidosos" da lista de candidatos:
  - preposições, artigos, marcas de pontuação e verbos frequentes
- Resultados…

Método 2: Resultados (2)

#	Conjunto inicial	L	Resultado
1	amarelo, vermelho, azul	2	verde (57), branco (52), preto (51), laranja (48), rosa (43), castanho (36), cinzento (34), cinza (28), roxo (27), violeta (24), negro (24), magenta (20), encarnado (19), dourado (19), lilás (18), ocre (16), creme(16), bege (15), cor (15), púrpura (14), vermelha (14), amarela (14), prateado (13), prata (13), ouro (13), bordeaux (12), salmão(12), branca(11), cor-de-rosa(10), cores(10), mar(10), preta(10), escuro (10)…
2	rosa salmão	2	amarelo (22), azul (21), verde (18), laranja (17), branco (15), vermelho (12), castanho (8), cinza (7), creme (7), lilás (7), bege (7), cinzento (4), bordeaux (4), adagio (4), dourado (4), cor (3), ocre (2), exemplo (2), peixe (2)
3	bacalhau, salmão	2	atum (17), sardinha (16), lampreia (13), cavala (13), sardinhas (11), truta (10), batatas (10), sável (10), pescada (10), linguado (8), presunto (8), arroz (8), queijo (8), pão (7), verde (6), tomate (6), peixe (6), gambas (6), charcutaria (5), cebola (5), camarão (5), sarrabulho (5), arenque (5), espadarte (5), tamboril (5), brócolos (4), mel (4), garoupa (4), lulas (4)…
4	rosa, laranja	2	amarelo (33), azul (30), verde (29), vermelho (28), branco (21), limão (16), roxo (16), castanho (16), preto (14), violeta (14), amarela (13), **ananás** (12), lilás (12), **maçã** (12), cinza (11), pêssego (11), preta (10), canela (10), creme (9), jasmim (9), damasco (9), vermelha (9), bege (8), **banana** (8), branca (8),…

Método 2: Resultados (2)

#	Conjunto inicial	L	Resultado
5	Curitiba, florianopolis recife	2	salvador (18), fortaleza (15), brasília (13), brazil (10), brasil (9), campinas (7), paulo (6), porto (6), florianópolis (6), alegre (5), bahia (5), santos (4), 2000 (4), rio (4), capital (4), janeiro (4), belo (4), 1996 (3), 2001 (3), cuiabá (3), ed. (3), manaus (3)
6	mármore granito	2	madeira (17), calcário (17), pedra (12), ferro (10), quartzo (8), cimento (8), barro (7), vidro (7), tijoleira (7), cerâmica (5), travertino (5), mosaico (5), ardósia (5), rochas (5), areia (5), tijolo (5), papel (4), cantaria (4), lioz (4), pedreiras (4), betão (3), aço (3), crómio (3), quartzito (3), região (3), alabastro (2), alvenaria (2), chapa (2), marfim (2), ornamentais (2), decoração (2), portuguesa (2), agricultura (2), rocha (2), chão (2), reboco (2), produção (2), alumínio (2)
7	whiskey, rum, vodka	2	brandy (10), gin (9), whisky (8), açúcar (7), tabaco (6), uísque (5), cerveja (4), sol (4), aguardente (4), conhaque (4),..
8	carro, barco	2	comboio (19), avião (18), autocarro (12), motor (11), mota (9), mar (9), automóvel (8), moto (7), porto (7), pé (7), navio (6), trabalho (6), viagens (6), metro (6), bicicleta (6), praia (5), camião (5), táxi (4), metropolitano (4), primeira (4), maior (4), ponte (4), viagem (4), transporte (4), água (4), vento (4), cavalo (4), melhor (4), velocidade (4)…

Comentários a esta aproximação (1)

- O topo da lista é mais uma vez povoado por elementos "verdadeiros" (+)
  - excepção para Conj. 8: palavras "raras"
  - Em certos casos aparenta ser superior ao método anterior (2,3,5,…)
- Muito rápido (+):
  - Após criação de tabela auxiliar (~1h) pesquisas 5-10s
- Imune à ambiguidade da "garrafa de vidro" (+)
  - Dificilmente "whisky" e "vidro" ocorrem coordenados

Comentários a esta aproximação (2)

- Método ruidoso:
  - obtém elementos relacionados MAS por vezes a relação é apenas contextual (sintagmática).
    - "whiskey" e "tabaco"
- Alguns casos continuam difíceis já que os elementos podem ser individualmente muito ambíguos (rosa, laranja)
  - Esses elementos podem aparecer coordenados com muitas outras palavras em contextos totalmente diferentes:
  - "Rosa" e "laranja" são ambos "cores" ou "vegetais"…

Breve Discussão

- Ambos os métodos:
  - possuem diferentes problemas:
    - possibilidade de validação cruzada;
  - apresentam problemas quando os dados são esparsos:
    - problema típico das técnicas "data-driven";
  - usam medidas de mérito muito simples baseadas no número de co-ocorrências distintas;
  - podem melhorar a sua precisão com filtros relativamente simples, quer estatísticos quer linguísticos, ou até por alteração dos parâmetros de pesquisa (especialmente o L)
  - podem desde já auxiliar a construção semi-automática de recursos léxico-semânticos.

Questões incómodas…

- O que são de facto elementos "semelhantes"?
- Semelhantes a (Granito, mármore)?
  - em construção civil? "aço", "madeira"…
  - em escultura? "gesso", "calcário"…
  - em geologia? "basalto", "feldspato"…
- Forte dependência do domínio / contexto:
  - Como identificar o contexto para poder pesquisar em conformidade?

Mais questões incómodas

- Que elementos semelhantes é que pretendemos?
- Semelhantes a ("Porto Alegre", "Florianópolis)?
  - Cidades no Sul do Brasil? "Curitiba", "Navegantes"…
  - Cidades do Brasil? "Belém", "Manaus", "Teresina"…
  - Cidades da América do Sul? "Lima", "Bogotá", "Santiago"…
  - Cidades do Hemisfério Sul? "Perth", "Joanesburgo"…
  - Cidades? "Aveiro", "Tóquio", "Dakar"
- Forte dependência do nível de especialização
- Como detectar diferentes níveis de especialização?
  - E pesquisar elementos "semelhantes" e conformidade?

A questão mais incómoda!

- Mas porquê o título da apresentação??
  - **Carros, barcos, mulheres e outras "coisas semelhantes"!**
- L é o parâmetro de filtragem de resultados
  - Fazendo L = 1 os resultandos "dispersam-se"…
  - Na verdade a primeira versão do algoritmo não tinha parâmetro de qualidade L, logo L era sempre = 1

#	Conjunto inicial	L	Resultado
1	carro, barco	1	comboio (19), avião (18), casa (16), autocarro (12), motor (11), mota (9), mar (9), automóvel (8), moto (7), porto (7), **mulher** (7), pé (7), navio (6), trabalho (6), viagens (6), piloto (6), metro (6), bicicleta (6), equipa (5), praia (5), camião (5), motorista (5), televisão (5), telemóvel (5), táxi (4), metropolitano (4), roupa (4), primeira (4), maior (4), escola (4), mergulho (4), ponte (4), caravela (4)…

A questão mais incómoda! (2)

- Aliás o título poderia até ter sido:
  - **Rum, whiskey, mulheres e ….**
    outras "coisas perigosas"?!

#	Conjunto inicial	L	Resultado
1	whiskey, rum, vodka	1	brandy (10), gin (9), whisky (8), açúcar (7), tabaco (6), uísque (5), cerveja (4), sol (4), aguardente (4), conhaque (4), água (4), limão (3), sal (3), porto (3), sumos (3), champanhe (3), pinga (3), **mulheres** (3), licor (3), sumo (3), tequila (3), gelo (3), laranja (3), charutos (3), artesanato (2), hortelã (2), bagaço (2), vinho (2), calvados (2), anseios (2), mel (2), peles (2), creme (2), bourbon (2), martini (2), velho (2), rhum (2), gemas (2), aspirina (2), menta (2), espumantes (2), absinto (2), cointreau (2)…

Comentários Finais

1. É possível extrair alguma semântica em português com métodos simples e "ingénuos"
   - Há muita margem para evolução em cada um dos métodos
   - Podem ser aplicadas técnicas de validação cruzada tentando anular os problemas típicos de cada método
2. É possível a aplicação prática dos métodos apresentados com optimizações simples
   - O primeiro método é mais lento mas poderá ser utilizado em aplicações "off-line"
   - O segundo método pode ser usado em aplicações on-line

Comentários Finais

- Há muitas questões teóricas (filosóficas?) ainda por resolver:
  - Como lidar com os diferentes contextos?
  - Como lidar com diferentes níveis de especialização?
- Ainda não começamos a tratar das semelhanças entre contextos:
  - Problema parece intrincado com o da semelhança de elementos: "o ovo e a galinha"?

Obrigado!

DEMONSTRAÇÃO
E / OU
PERGUNTAS….

las@fe.up.pt www.fe.up.pt/~las