Introdução à Recuperação de Informações

https://github.com/fccoelho/curso-IRI

IRI 2: Vocabulário de termos e lista de "postings"

Flávio Codeço Coelho

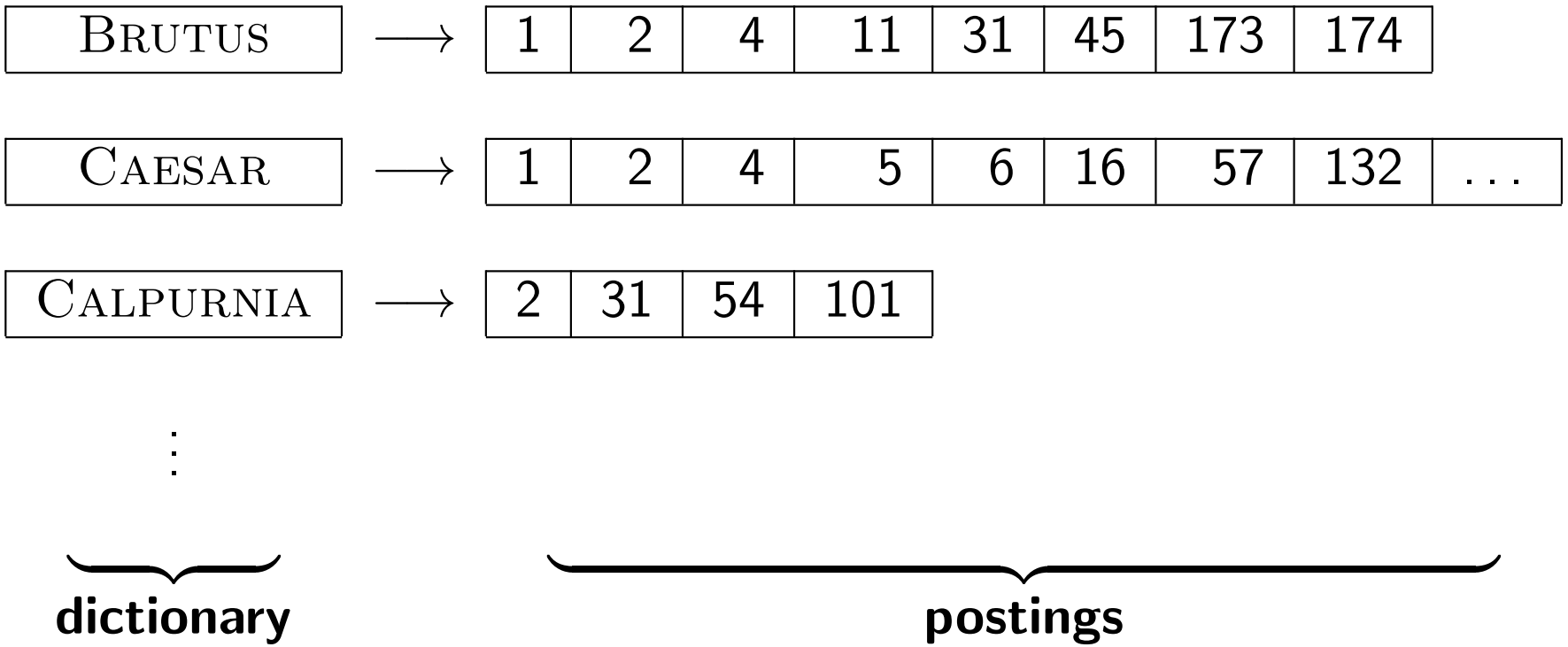
Escola de Matemática Aplicada, Fundação Getúlio Vargas

Sumário da Aula

1. Recapitulação
2. Documentos
3. Termos
   - Genericos + outras línguas
   - English
4. Consultas de frases

Índice invertido

Para cada termo t, armazenamos uma lista de todos os documentos que contém t.



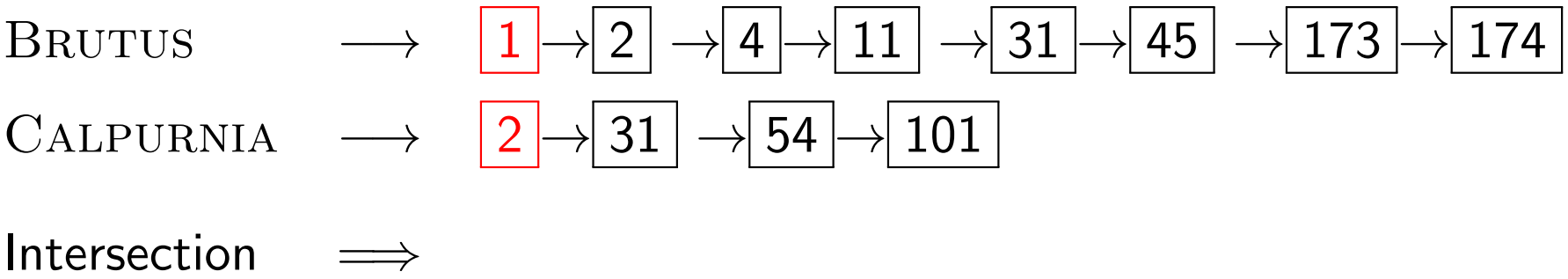
Interseção de duas listas de "postings"

BRUTUS $\longrightarrow$ 1 → 2 → 4 → 11 → 31 → 45 → 173 → 174

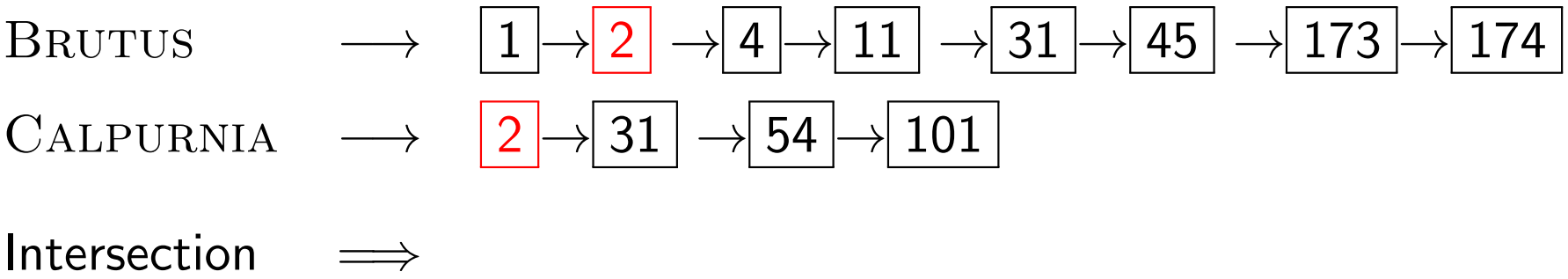
CALPURNIA $\longrightarrow$ 2 → 31 → 54 → 101

Intersection $\Longrightarrow$

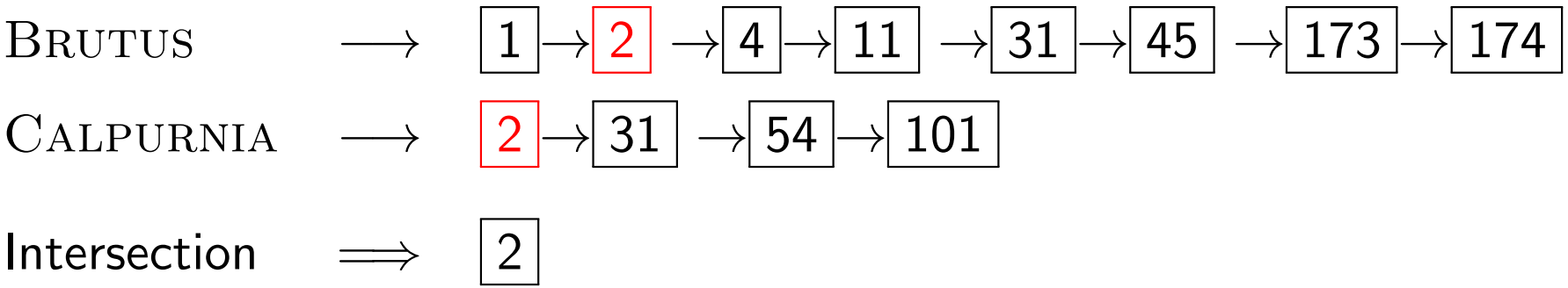
Interseção de duas listas de "postings"



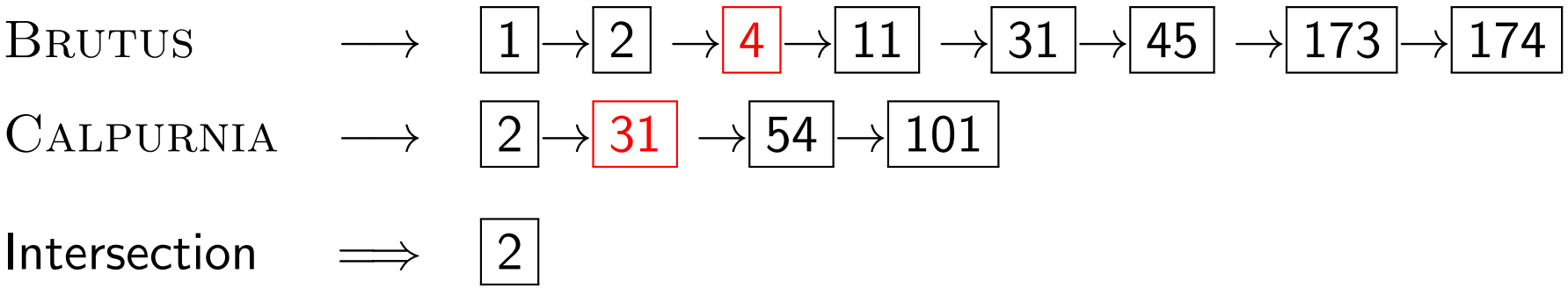
Interseção de duas listas de "postings"



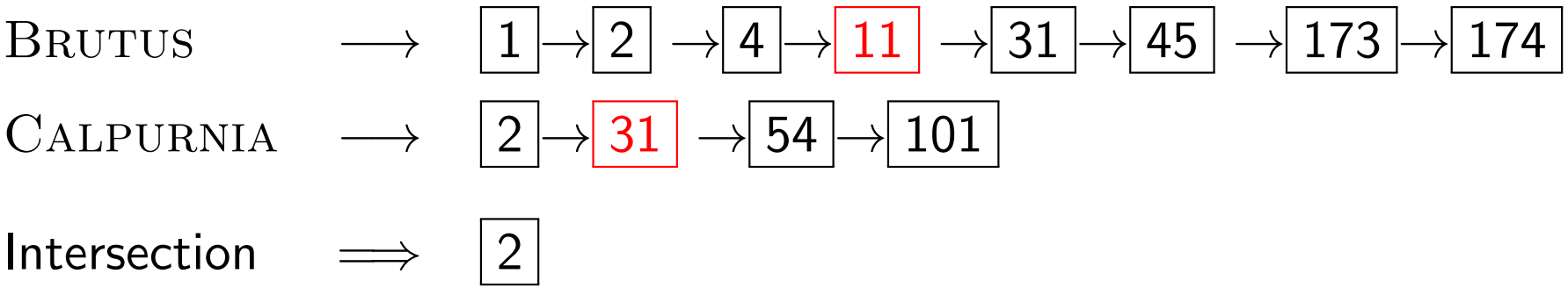
Interseção de duas listas de "postings"



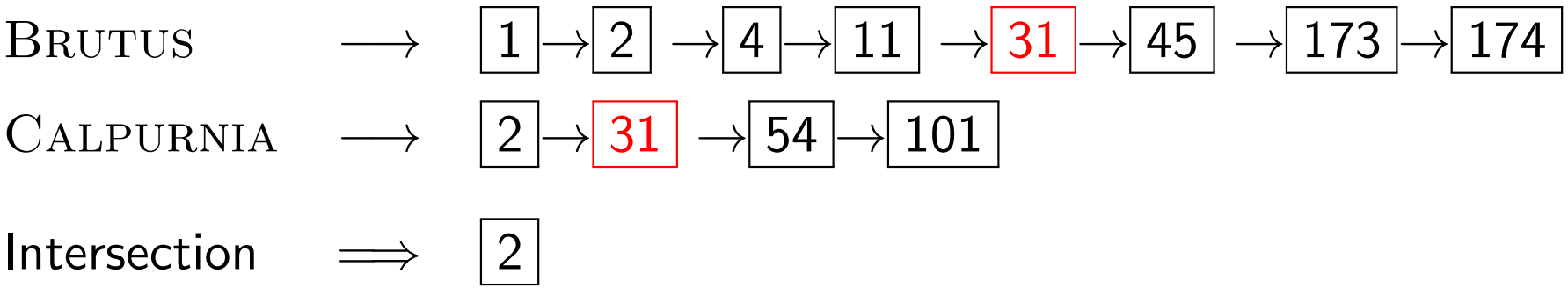
Interseção de duas listas de "postings"



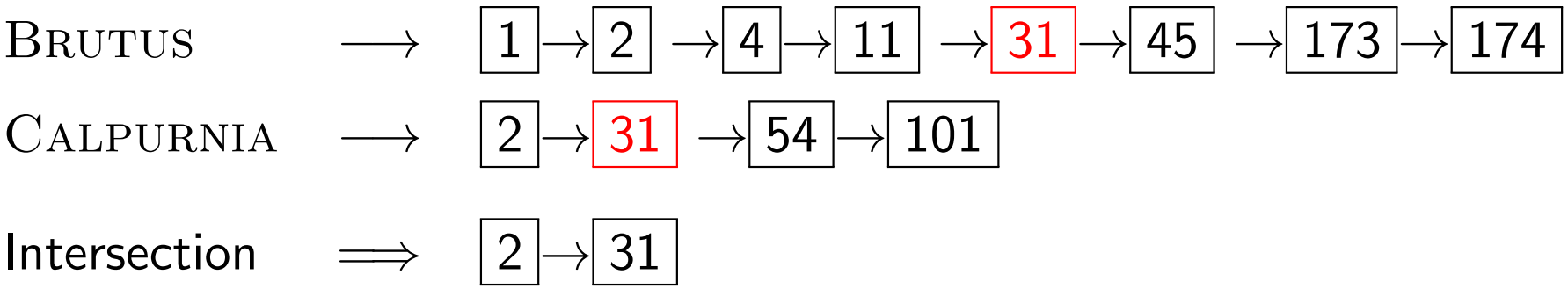
Interseção de duas listas de "postings"



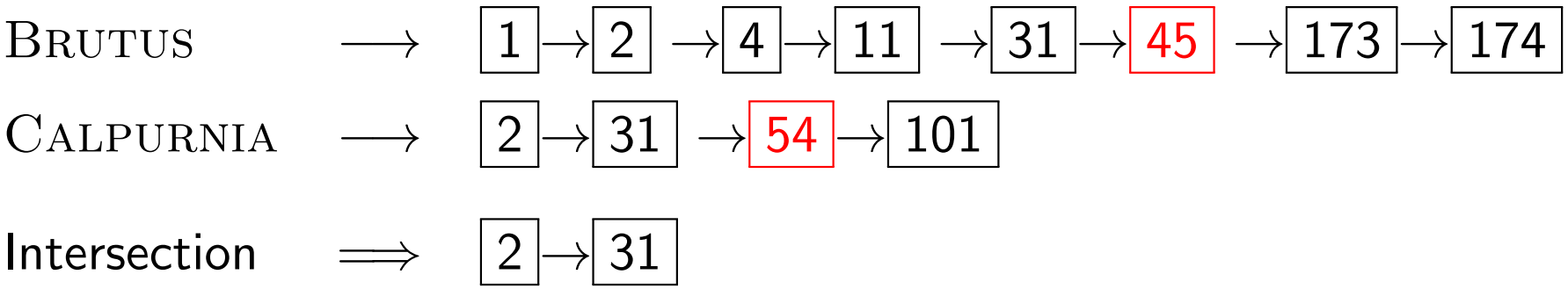
Interseção de duas listas de "postings"



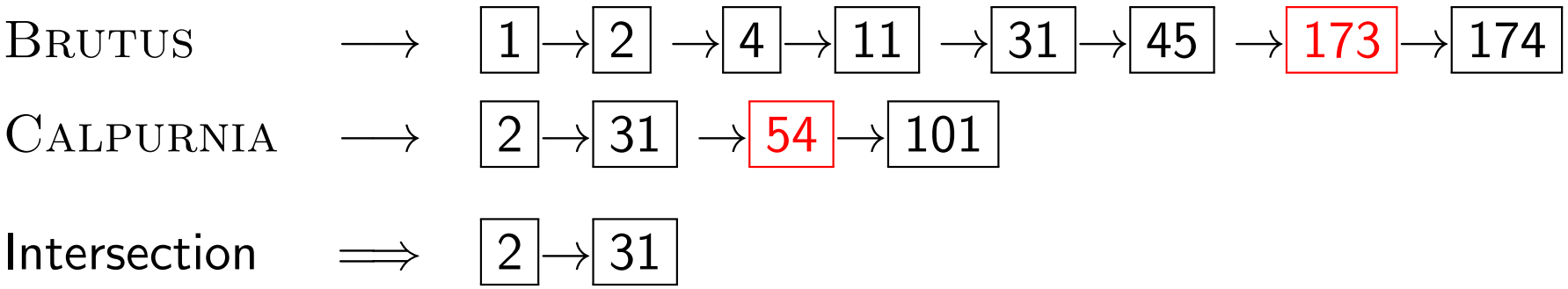
Interseção de duas listas de "postings"



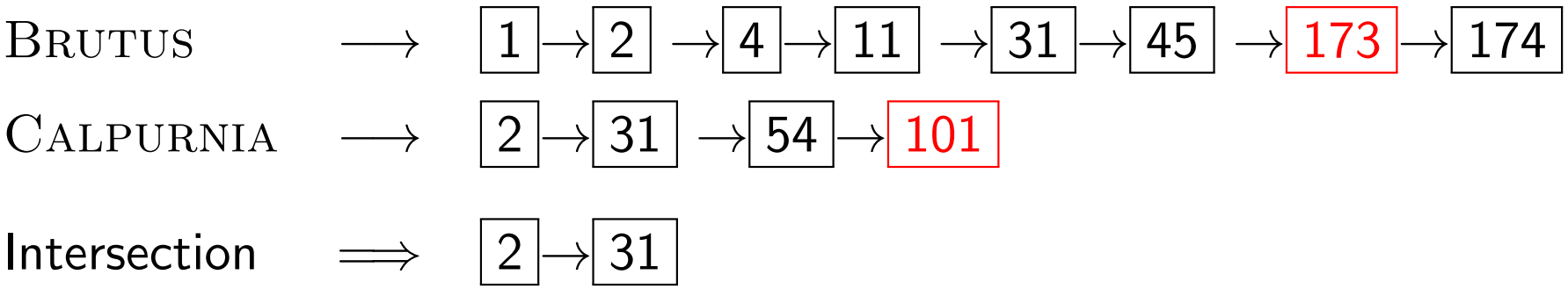
Interseção de duas listas de "postings"



Interseção de duas listas de "postings"

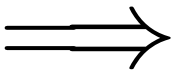


Interseção de duas listas de "postings"



Construindo o índice invertido: ordenando os "postings"

term	docID
I	1
did	1
enact	1
julius	1
caesar	1
I	1
was	1
killed	1
i'	1
the	1
capitol	1
brutus	1
killed	1
me	1
so	2
let	2
it	2
be	2
with	2
caesar	2
the	2
noble	2
brutus	2
hath	2
told	2
you	2
caesar	2
was	2
ambitious	2



term	docID
ambitious	2
be	2
brutus	1
brutus	2
capitol	1
caesar	1
caesar	2
caesar	2
did	1
enact	1
hath	1
I	1
I	1
i'	1
it	2
julius	1
killed	1
killed	1
let	2
me	1
noble	2
so	2
the	1
the	2
told	2
you	2
was	1
was	2
with	2

O Google usa busca Booleana?

- No Google, a interpretação defaul de uma consulta [w_1 w_2 ... w_n] é w_1 AND w_2 AND ... AND w_n
- Casos em que você recebe resultados que não contém uma das w_i:
  - âncora de texto
  - página contém variantes da w_i (morfologia, correção ortográfica , sinônimo)
  - consultas longas (n é grande)
  - expressões booleanas geram poucos resultados.
- Booleana simples vs. Ordenação do conjunto de resultados
  - A busca booleana simples não gera ordenação dos documentos.
  - O Google (e a maioria dos buscadores booleanos bem feitos) ordena o conjunto de resultados – com os melhores resultados (de acordo com um estimador de relevância) no topo.

Documentos

- Última aula: Sistema de recuperação Booleana simples

Documentos

- Última aula: Sistema de recuperação Booleana simples
- Nossos pressupostos eram:

Documentos

- Última aula: Sistema de recuperação Booleana simples
- Nossos pressupostos eram:
  - Nós sabemos o que é um documento.

Documentos

- Última aula: Sistema de recuperação Booleana simples
- Nossos pressupostos eram:
  - Nós sabemos o que é um documento.
  - Documentos são "legíveis por máquina".

Documentos

- Última aula: Sistema de recuperação Booleana simples
- Nossos pressupostos eram:
  - Nós sabemos o que é um documento.
  - Documentos são "legíveis por máquina".
- Na prática isto pode ser bem complicado.

Parseando um Documento

- Precisamos lidar com formato e lingua de cada documento.

Parseando um Documento

- Precisamos lidar com formato e lingua de cada documento.
- Em que formato ele está? pdf, word, excel, html etc.

Parseando um Documento

- Precisamos lidar com formato e lingua de cada documento.
- Em que formato ele está? pdf, word, excel, html etc.
- Em que língua ele está?

Parseando um Documento

- Precisamos lidar com formato e lingua de cada documento.
- Em que formato ele está? pdf, word, excel, html etc.
- Em que língua ele está?
- Em que codificação está?

Parseando um Documento

- Precisamos lidar com formato e lingua de cada documento.
- Em que formato ele está? pdf, word, excel, html etc.
- Em que língua ele está?
- Em que codificação está?
- Cada uma destas perguntas implica um problema de classificação.

Parseando um Documento

- Precisamos lidar com formato e lingua de cada documento.
- Em que formato ele está? pdf, word, excel, html etc.
- Em que língua ele está?
- Em que codificação está?
- Cada uma destas perguntas implica um problema de classificação.
- Alternativa: usar heurísticas.

Complicações de formato e Língua

- Um mesmo índice normalmente contém termos de diversas línguas.

Complicações de formato e Língua

- Um mesmo índice normalmente contém termos de diversas línguas.
  - Algumas vezes um documento ou seus componentes contém múltiplas línguas e formatos.

Complicações de formato e Língua

- Um mesmo índice normalmente contém termos de diversas línguas.
  - Algumas vezes um documento ou seus componentes contém múltiplas línguas e formatos.
  - Email em francês com PDF em espanhol anexado

Complicações de formato e Língua

- Um mesmo índice normalmente contém termos de diversas línguas.
  - Algumas vezes um documento ou seus componentes contém múltiplas línguas e formatos.
  - Email em francês com PDF em espanhol anexado
- Qual a unidade do documento para fins de indexação?

Complicações de formato e Língua

- Um mesmo índice normalmente contém termos de diversas línguas.
  - Algumas vezes um documento ou seus componentes contém múltiplas línguas e formatos.
  - Email em francês com PDF em espanhol anexado
- Qual a unidade do documento para fins de indexação?
- Um arquivo?

Complicações de formato e Língua

- Um mesmo índice normalmente contém termos de diversas línguas.
  - Algumas vezes um documento ou seus componentes contém múltiplas línguas e formatos.
  - Email em francês com PDF em espanhol anexado
- Qual a unidade do documento para fins de indexação?
- Um arquivo?
- Um email?

Complicações de formato e Língua

- Um mesmo índice normalmente contém termos de diversas línguas.
  - Algumas vezes um documento ou seus componentes contém múltiplas línguas e formatos.
  - Email em francês com PDF em espanhol anexado
- Qual a unidade do documento para fins de indexação?
- Um arquivo?
- Um email?
- Um email com 5 anexos?

Complicações de formato e Língua

- Um mesmo índice normalmente contém termos de diversas línguas.
  - Algumas vezes um documento ou seus componentes contém múltiplas línguas e formatos.
  - Email em francês com PDF em espanhol anexado
- Qual a unidade do documento para fins de indexação?
- Um arquivo?
- Um email?
- Um email com 5 anexos?
- Um grupo de arquivos(ppt ou latex in HTML)?

Complicações de formato e Língua

- Um mesmo índice normalmente contém termos de diversas línguas.
  - Algumas vezes um documento ou seus componentes contém múltiplas línguas e formatos.
  - Email em francês com PDF em espanhol anexado
- Qual a unidade do documento para fins de indexação?
- Um arquivo?
- Um email?
- Um email com 5 anexos?
- Um grupo de arquivos(ppt ou latex in HTML)?
- Em suma: Responder à pergunta "O que é um documento?" não é trivial e requer algumas decisões.

Definições

- Palavra – Uma string delimitada de caracteres como aparece no texto.

Definições

- Palavra – Uma string delimitada de caracteres como aparece no texto.
- Termo – Uma palavra "normalizada" (capitalização, morfologia, ortografia, etc); Uma classe equivalente de palavras.

Definições

- Palavra – Uma string delimitada de caracteres como aparece no texto.
- Termo – Uma palavra "normalizada" (capitalização, morfologia, ortografia, etc); Uma classe equivalente de palavras.
- Token – Uma instancia de uma palavra ou termo ocorrendo em um documento.

Definições

- Palavra – Uma string delimitada de caracteres como aparece no texto.
- Termo – Uma palavra "normalizada" (capitalização, morfologia, ortografia, etc); Uma classe equivalente de palavras.
- Token – Uma instancia de uma palavra ou termo ocorrendo em um documento.
- Tipo – O mesmo que termo na maioria dos casos: Uma classe de equivalência de tokens.

Normalização

- Precisamos "normalizar" os termos no texto indexado, assim como o termos da consulta da mesma maneira.

Normalização

- Precisamos "normalizar" os termos no texto indexado, assim como o termos da consulta da mesma maneira.
- Exemplo: Queremos associar *U.S.A.* com *USA*

Normalização

- Precisamos "normalizar" os termos no texto indexado, assim como o termos da consulta da mesma maneira.
- Exemplo: Queremos associar *U.S.A.* com *USA*
- A maneira mais comum é criar classes de equivalência de termos.

Normalização

- Precisamos "normalizar" os termos no texto indexado, assim como o termos da consulta da mesma maneira.
- Exemplo: Queremos associar *U.S.A.* com *USA*
- A maneira mais comum é criar classes de equivalência de termos.
- Alternativamente: fazer expansões assimétricas

Normalização

- Precisamos "normalizar" os termos no texto indexado, assim como o termos da consulta da mesma maneira.
- Exemplo: Queremos associar *U.S.A.* com *USA*
- A maneira mais comum é criar classes de equivalência de termos.
- Alternativamente: fazer expansões assimétricas
  - window $\rightarrow$ window, windows

Normalização

- Precisamos "normalizar" os termos no texto indexado, assim como o termos da consulta da mesma maneira.
- Exemplo: Queremos associar *U.S.A.* com *USA*
- A maneira mais comum é criar classes de equivalência de termos.
- Alternativamente: fazer expansões assimétricas
  - window $\rightarrow$ window, windows
  - windows $\rightarrow$ Windows, windows

Normalização

- Precisamos "normalizar" os termos no texto indexado, assim como o termos da consulta da mesma maneira.
- Exemplo: Queremos associar *U.S.A.* com *USA*
- A maneira mais comum é criar classes de equivalência de termos.
- Alternativamente: fazer expansões assimétricas
  - window $\rightarrow$ window, windows
  - windows $\rightarrow$ Windows, windows
  - Windows (sem expansão)

Normalização

- Precisamos "normalizar" os termos no texto indexado, assim como o termos da consulta da mesma maneira.
- Exemplo: Queremos associar *U.S.A.* com *USA*
- A maneira mais comum é criar classes de equivalência de termos.
- Alternativamente: fazer expansões assimétricas
  - window $\rightarrow$ window, windows
  - windows $\rightarrow$ Windows, windows
  - Windows (sem expansão)
- Mais poderoso, porém menos eficiente

Normalização

- Precisamos "normalizar" os termos no texto indexado, assim como o termos da consulta da mesma maneira.
- Exemplo: Queremos associar *U.S.A.* com *USA*
- A maneira mais comum é criar classes de equivalência de termos.
- Alternativamente: fazer expansões assimétricas
  - window $\rightarrow$ window, windows
  - windows $\rightarrow$ Windows, windows
  - Windows (sem expansão)
- Mais poderoso, porém menos eficiente
- Porque não colocamos *window*, *Window*, *windows*, e *Windows* na mesma classe de equivalência?

Normalização: outras línguas

- Normalização e detecção de linguagem interagem.

Normalização: outras línguas

- Normalização e detecção de linguagem interagem.
- *PETER WILL NICHT MIT.* $\rightarrow$ MIT = mit

Normalização: outras línguas

- Normalização e detecção de linguagem interagem.
- *PETER WILL NICHT MIT.* $\rightarrow$ MIT = mit
- *He got his PhD from MIT.* $\rightarrow$ MIT $\neq$ mit

Lembre-se: Construção de índices invertidos

- Entrada:

Friends, Romans, countrymen. | So let it be with Caesar | ...

Lembre-se: Construção de índices invertidos

- Entrada:
  Friends, Romans, countrymen. | So let it be with Caesar | ...
- Saída:
  friend | roman | countryman | so | ...

Lembre-se: Construção de índices invertidos

- Entrada:
  Friends, Romans, countrymen. | So let it be with Caesar | ...
- Saída:
  friend | roman | countryman | so | ...
- Cada token é candidato a uma entrada de posting.

Lembre-se: Construção de índices invertidos

- Entrada:
  Friends, Romans, countrymen. | So let it be with Caesar | ...
- Saída:
  friend | roman | countryman | so | ...
- Cada token é candidato a uma entrada de posting.
- Quais são os tokens válidos?

Exercícios

In June, the dog likes to chase the cat in the barn. – Quantos tokens? Quantos tipos de Palavras?

Porque a tokenização é difícil – mesmo em inglês. Tokenize: *Mr. O'Neill thinks that the boys' stories about Chile's capital aren't amusing.*

Problemas de tokenização: Uma ou duas palavras? (ou várias)

- Hewlett-Packard

Problemas de tokenização: Uma ou duas palavras? (ou várias)

- Hewlett-Packard
- State-of-the-art

Problemas de tokenização: Uma ou duas palavras? (ou várias)

- Hewlett-Packard
- State-of-the-art
- co-education

Problemas de tokenização: Uma ou duas palavras? (ou várias)

- Hewlett-Packard
- State-of-the-art
- co-education
- the hold-him-back-and-drag-him-away maneuver

Problemas de tokenização: Uma ou duas palavras? (ou várias)

- Hewlett-Packard
- State-of-the-art
- co-education
- the hold-him-back-and-drag-him-away maneuver
- data base

Problemas de tokenização: Uma ou duas palavras? (ou várias)

- Hewlett-Packard
- State-of-the-art
- co-education
- the hold-him-back-and-drag-him-away maneuver
- data base
- San Francisco

Problemas de tokenização: Uma ou duas palavras? (ou várias)

- Hewlett-Packard
- State-of-the-art
- co-education
- the hold-him-back-and-drag-him-away maneuver
- data base
- San Francisco
- Los Angeles-based company

Problemas de tokenização: Uma ou duas palavras? (ou várias)

- Hewlett-Packard
- State-of-the-art
- co-education
- the hold-him-back-and-drag-him-away maneuver
- data base
- San Francisco
- Los Angeles-based company
- cheap San Francisco-Los Angeles fares

Problemas de tokenização: Uma ou duas palavras? (ou várias)

- Hewlett-Packard
- State-of-the-art
- co-education
- the hold-him-back-and-drag-him-away maneuver
- data base
- San Francisco
- Los Angeles-based company
- cheap San Francisco-Los Angeles fares
- York University vs. New York University

Números

- 3/20/91

Números

- 3/20/91
- 20/3/91

Números

- 3/20/91
- 20/3/91
- Mar 20, 1991

Números

- 3/20/91
- 20/3/91
- Mar 20, 1991
- B-52

Números

- 3/20/91
- 20/3/91
- Mar 20, 1991
- B-52
- 100.2.86.144

Números

- 3/20/91
- 20/3/91
- Mar 20, 1991
- B-52
- 100.2.86.144
- (800) 234-2333

Números

- 3/20/91
- 20/3/91
- Mar 20, 1991
- B-52
- 100.2.86.144
- (800) 234-2333
- 800.234.2333

Números

- 3/20/91
- 20/3/91
- Mar 20, 1991
- B-52
- 100.2.86.144
- (800) 234-2333
- 800.234.2333
- Sistemas mais antigos de RI costumavam não indexar números. . .

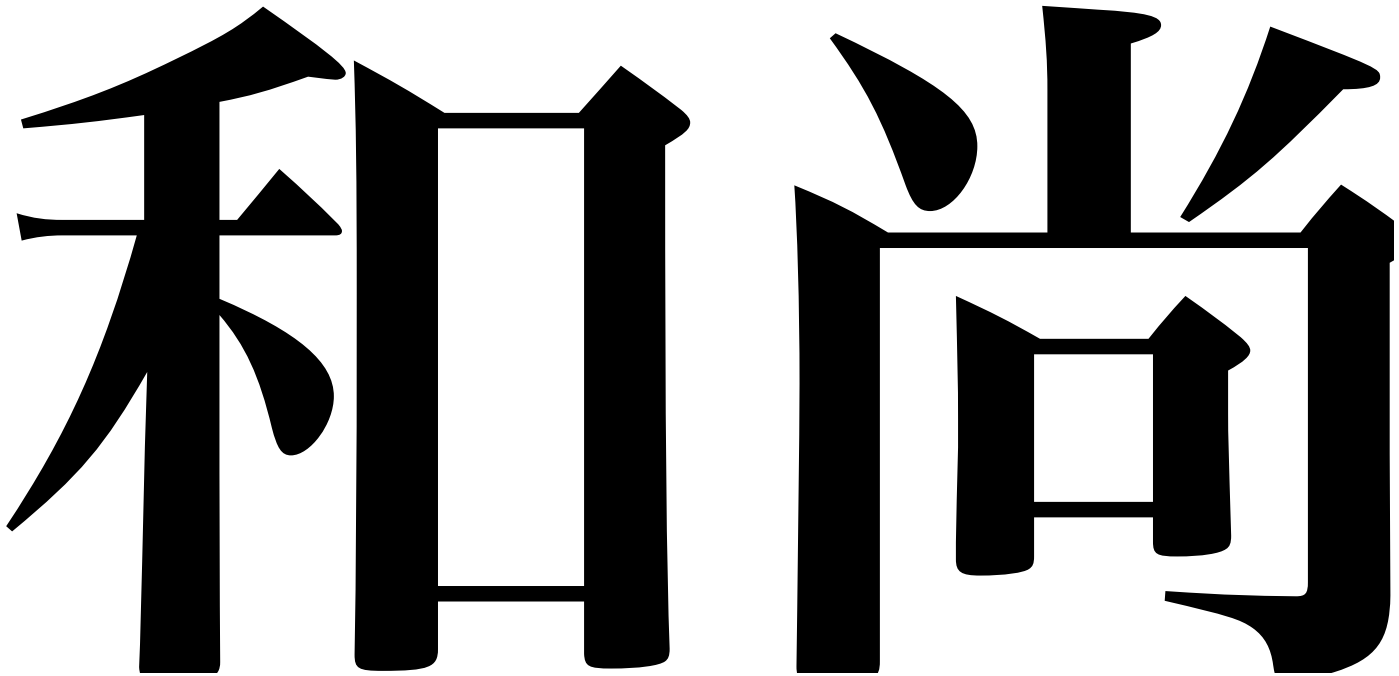
Números

- 3/20/91
- 20/3/91
- Mar 20, 1991
- B-52
- 100.2.86.144
- (800) 234-2333
- 800.234.2333
- Sistemas mais antigos de RI costumavam não indexar números. . .
- . . . mas as vezes isto pode ser útil.

Chinês: Ná há espaços

莎拉波娃现在居住在美国东南部的佛罗里达。今年４月９日，莎拉波娃在美国第一大城市纽约度过了１８岁生日。生日派对上，莎拉波娃露出了甜美的微笑。

Segmentação ambígua em chinês



Os dois caracteres podem ser tratados como uma palavra significando ‘monge’ ou como uma sequência de duas palavras significando “e” e “parado”.

Outros casos de ausência de espaços

- Palavras compostas em Holandês, Alemão, Sueco in Dutch, German, Swedish

Outros casos de ausência de espaços

- Palavras compostas em Holandês, Alemão, Sueco in Dutch, German, Swedish
- Computerlinguistik $\rightarrow$ Computer + Linguistik

Outros casos de ausência de espaços

- Palavras compostas em Holandês, Alemão, Sueco in Dutch, German, Swedish
- Computerlinguistik $\rightarrow$ Computer + Linguistik
- Lebensversicherungsgesellschaftsangestellter

Outros casos de ausência de espaços

- Palavras compostas em Holandês, Alemão, Sueco in Dutch, German, Swedish
- Computerlinguistik $\rightarrow$ Computer + Linguistik
- Lebensversicherungsgesellschaftsangestellter
- $\rightarrow$ leben + versicherung + gesellschaft + angestellter

Outros casos de ausência de espaços

- Palavras compostas em Holandês, Alemão, Sueco in Dutch, German, Swedish
- Computerlinguistik $\rightarrow$ Computer + Linguistik
- Lebensversicherungsgesellschaftsangestellter
- $\rightarrow$ leben + versicherung + gesellschaft + angestellter
- Inuit: tusaatsiarunnanngittualuujunga (Não consigo escutar direito.)

Outros casos de ausência de espaços

- Palavras compostas em Holandês, Alemão, Sueco in Dutch, German, Swedish
- Computerlinguistik $\rightarrow$ Computer + Linguistik
- Lebensversicherungsgesellschaftsangestellter
- $\rightarrow$ leben + versicherung + gesellschaft + angestellter
- Inuit: tusaatsiarunnanngittualuujunga (Não consigo escutar direito.)
- Muitas outras línguas apresentam dificuldades de segmentação: Finlandês, Urdu, ...

Japonês

ノーベル平和賞を受賞したワンガリ・マータイさんが名誉会長を務めるＭＯＴＴＡＩＮＡＩキャンペーンの一環として、毎日新聞社とマガジンハウスは「私の、もったいない」を募集します。皆様が日ごろ「もったいない」と感じて実践していることや、それにまつわるエピソードを８００字以内の文章にまとめ、簡単な写真、イラスト、図などを添えて１０月２０日までにお送りください。大賞受賞者には、５０万円相当の旅行券とエコ製品２点の副賞が贈られます。

Japonês

ノーベル平和賞を受賞したワンガリ・マータイさんが名誉会長を務めるＭＯＴＴＡＩＮＡＩキャンペーンの一環として、毎日新聞社とマガジンハウスは「私の、もったいない」を募集します。皆様が日ごろ「もったいない」と感じて実践していることや、それにまつわるエピソードを８００字以内の文章にまとめ、簡単な写真、イラスト、図などを添えて１０月２０日までにお送りください。大賞受賞者には、５０万円相当の旅行券とエコ製品２点の副賞が贈られます。

4 “alfabetos” diferentes: Caracteres chineses, hiragana silabário para finalizações inflecionais e palavras funcionais, katakana silabário para transcrições de palavras estrangeiras. Sem espaços (como no chinês).

Japonês

ノーベル平和賞を受賞したワンガリ・マータイさんが名誉会長を務めるＭＯＴＴＡＩＮＡＩキャンペーンの一環として、毎日新聞社とマガジンハウスは「私の、もったいない」を募集します。皆様が日ごろ「もったいない」と感じて実践していることや、それにまつわるエピソードを８００字以内の文章にまとめ、簡単な写真、イラスト、図などを添えて１０月２０日までにお送りください。大賞受賞者には、５０万円相当の旅行券とエコ製品２点の副賞が贈られます。

4 “alfabetos” diferentes: Caracteres chineses, hiragana silabário para finalizações inflecionais e palavras funcionais, katakana silabário para transcrições de palavras estrangeiras. Sem espaços (como no chinês).
Um usuário pode fazer uma consulta inteiramente em hiragana!

Árabe

كِتَابٌ ⇐ ك ِ ت ا ب ٌ
un b ā t i k
/kitābun/ *'a book'*

Árabe: Bidirecionalidade

استقلت الجزائر في سنة 1962 بعد 132 عاما من الاحتلال الفرنسي.

← START ← → ← →

'Algeria achieved its independence in 1962 after 132 years of French occupation.'

Árabe: Bidirecionalidade

استقلت الجزائر في سنة 1962 بعد 132 عاما من الاحتلال الفرنسي.

← → ← → ← START

'Algeria achieved its independence in 1962 after 132 years of French occupation.'

Bidirecionalidade não é um problema se o texto estiver codificado em Unicode.

Acentos e diacríticos

- Acentos: Árvore vs. Arvore (Simples omissão de acento)

Acentos e diacríticos

- Acentos: Árvore vs. Arvore (Simples omissão de acento)
- Tremas: Universität vs. Universitaet (substituição por sequencia especial de caracteres: "ae")

Acentos e diacríticos

- Acentos: Árvore vs. Arvore (Simples omissão de acento)
- Tremas: Universität vs. Universitaet (substituição por sequencia especial de caracteres: "ae")
- Critério mais importante: Qual a forma mais comumente adotada por usuários em suas consultas?

Acentos e diacríticos

- Acentos: Árvore vs. Arvore (Simples omissão de acento)
- Tremas: Universität vs. Universitaet (substituição por sequencia especial de caracteres: "ae")
- Critério mais importante: Qual a forma mais comumente adotada por usuários em suas consultas?
- Mesmo em línguas que possuem acentos por padrão, freuentemente os usuários não os digitam.

Capitalização

- Reduzir todas as letras para minúsculas

Capitalização

- Reduzir todas as letras para minúsculas
- Possíveis exceções: Palavras capitalizadas no meio da frase

Capitalização

- Reduzir todas as letras para minúsculas
- Possíveis exceções: Palavras capitalizadas no meio da frase
- MIT vs. mit

Capitalização

- Reduzir todas as letras para minúsculas
- Possíveis exceções: Palavras capitalizadas no meio da frase
- MIT vs. mit
- Fed vs. fed

Capitalização

- Reduzir todas as letras para minúsculas
- Possíveis exceções: Palavras capitalizadas no meio da frase
- MIT vs. mit
- Fed vs. fed
- Frequentemente é melhor converter tudo para minúsculas pois os usuários normalmente fazem suas consultas sem utilizar a capitalização correta.

“Stop words”

- “stop words” = Palavras extremamente comuns que podem parecer ser de pouco valor na seleção de documentos correspondentes à necessidade do usuário.

“Stop words”

- “stop words” = Palavras extremamente comuns que podem parecer ser de pouco valor na seleção de documentos correspondentes à necessidade do usuário.
- Exemplos: artigos, preposições, etc.

"Stop words"

- "stop words" = Palavras extremamente comuns que podem parecer ser de pouco valor na seleção de documentos correspondentes à necessidade do usuário.
- Exemplos: artigos, preposições, etc.
- Eliminação de "Stop word" costumava ser padrão em sistemas de RI mais antigos.

"Stop words"

- "stop words" = Palavras extremamente comuns que podem parecer ser de pouco valor na seleção de documentos correspondentes à necessidade do usuário.
- Exemplos: artigos, preposições, etc.
- Eliminação de "Stop word" costumava ser padrão em sistemas de RI mais antigos.
- Mas estas palavras são necessárias para consultas frasais, p.ex. "Reino da Dinamarca"

"Stop words"

- "stop words" = Palavras extremamente comuns que podem parecer ser de pouco valor na seleção de documentos correspondentes à necessidade do usuário.
- Exemplos: artigos, preposições, etc.
- Eliminação de "Stop word" costumava ser padrão em sistemas de RI mais antigos.
- Mas estas palavras são necessárias para consultas frasais, p.ex. "Reino da Dinamarca"
- A maioria dos engines de busca na web, indexa "stop words".

Mais classes de equivalência

- Soundex: equivalencia fonética, Muller = Mueller)

Mais classes de equivalência

- Soundex: equivalencia fonética, Muller = Mueller)
- Thesauri: equivalencia semântica, carro = automóvel)

Lematização

- Reduz variações e inflexões a uma forma básica

Lematização

- Reduz variações e inflexões a uma forma básica
- Exemplo: *sou*, *são*, *é* $\rightarrow$ *ser*

Lematização

- Reduz variações e inflexões a uma forma básica
- Exemplo: *sou*, *são*, *é* $\rightarrow$ *ser*
- Exemplo: *carro*, *carros* $\rightarrow$ *carros*

Lematização

- Reduz variações e inflexões a uma forma básica
- Exemplo: *sou, são, é* $\rightarrow$ *ser*
- Exemplo: *carro, carros* $\rightarrow$ *carros*
- Exemple: *Os carros do rapaz são de diferentes cores* $\rightarrow$ *o carro do rapaz ser de diferente cor*

Lematização

- Reduz variações e inflexões a uma forma básica
- Exemplo: *sou*, *são*, *é* $\rightarrow$ *ser*
- Exemplo: *carro*, *carros* $\rightarrow$ *carros*
- Exemple: *Os carros do rapaz são de diferentes cores* $\rightarrow$ *o carro do rapaz ser de diferente cor*
- Lematização implica fazer a redução "correta" às forma mais fundamental segundo o dicionário o lemma).

Lematização

- Reduz variações e inflexões a uma forma básica
- Exemplo: *sou*, *são*, *é* $\rightarrow$ *ser*
- Exemplo: *carro*, *carros* $\rightarrow$ *carros*
- Exemple: *Os carros do rapaz são de diferentes cores* $\rightarrow$ *o carro do rapaz ser de diferente cor*
- Lematização implica fazer a redução "correta" às forma mais fundamental segundo o dicionário o lemma).
- Morfologia inflexional (*cortando* $\rightarrow$ *cortar*) vs. morfologia derivacional (*destruição* $\rightarrow$ *destruir*)

Truncagem ("Stemming")

- Definição de truncagem: Heurística bruta que corta fora as terminações das palavras na esperança de alcançar o mesmo objetivo da lematização sem a necessidade de conhecimento linguístico.

Truncagem ("Stemming")

- Definição de truncagem: Heurística bruta que corta fora as terminações das palavras na esperança de alcançar o mesmo objetivo da lematização sem a necessidade de conhecimento linguístico.
- Depende da língua

Truncagem ("Stemming")

- Definição de truncagem: Heurística bruta que corta fora as terminações das palavras na esperança de alcançar o mesmo objetivo da lematização sem a necessidade de conhecimento linguístico.
- Depende da língua
- Frequentemente inflexional e derivacional

Truncagem ("Stemming")

- Definição de truncagem: Heurística bruta que corta fora as terminações das palavras na esperança de alcançar o mesmo objetivo da lematização sem a necessidade de conhecimento linguístico.
- Depende da língua
- Frequentemente inflexional e derivacional
- Exemplo de truncagem derivacional: *automatizar*, *automático*, *automatização* se reduzem a *automat*

Algoritmo de Porter

- Algoritmo mais comum para truncagem do inglês

Algoritmo de Porter

- Algoritmo mais comum para truncagem do inglês
- Resultados sugerem que é tão bom quanto outras opções de truncagem

Algoritmo de Porter

- Algoritmo mais comum para truncagem do inglês
- Resultados sugerem que é tão bom quanto outras opções de truncagem
- Convenções + 5 fases de reduções

Algoritmo de Porter

- Algoritmo mais comum para truncagem do inglês
- Resultados sugerem que é tão bom quanto outras opções de truncagem
- Convenções + 5 fases de reduções
- Fases são aplicadas sequencialmente

Algoritmo de Porter

- Algoritmo mais comum para truncagem do inglês
- Resultados sugerem que é tão bom quanto outras opções de truncagem
- Convenções + 5 fases de reduções
- Fases são aplicadas sequencialmente
- Cada fase consiste de um conjunto de comandos.

Algoritmo de Porter

- Algoritmo mais comum para truncagem do inglês
- Resultados sugerem que é tão bom quanto outras opções de truncagem
- Convenções + 5 fases de reduções
- Fases são aplicadas sequencialmente
- Cada fase consiste de um conjunto de comandos.
  - Exemplo de comando: Deleta o *ement* final se o que sobrar for mais longo do que um caractere

Algoritmo de Porter

- Algoritmo mais comum para truncagem do inglês
- Resultados sugerem que é tão bom quanto outras opções de truncagem
- Convenções + 5 fases de reduções
- Fases são aplicadas sequencialmente
- Cada fase consiste de um conjunto de comandos.
  - Exemplo de comando: Deleta o *ement* final se o que sobrar for mais longo do que um caractere
  - replacement $\rightarrow$ replac

Algoritmo de Porter

- Algoritmo mais comum para truncagem do inglês
- Resultados sugerem que é tão bom quanto outras opções de truncagem
- Convenções + 5 fases de reduções
- Fases são aplicadas sequencialmente
- Cada fase consiste de um conjunto de comandos.
  - Exemplo de comando: Deleta o *ement* final se o que sobrar for mais longo do que um caractere
  - replacement $\rightarrow$ replac
  - cement $\rightarrow$ cement

Algoritmo de Porter

- Algoritmo mais comum para truncagem do inglês
- Resultados sugerem que é tão bom quanto outras opções de truncagem
- Convenções + 5 fases de reduções
- Fases são aplicadas sequencialmente
- Cada fase consiste de um conjunto de comandos.
  - Exemplo de comando: Deleta o *ement* final se o que sobrar for mais longo do que um caractere
  - replacement $\rightarrow$ replac
  - cement $\rightarrow$ cement
- Exemplo de convenção: Das regras em comando composto, selecione aquela que se aplica ao sufixo mais longo.

Trucador Porter: Algumas regras

Regra			**Exemplo**		
SSES	$\rightarrow$	SS	caresses	$\rightarrow$	caress
IES	$\rightarrow$	I	ponies	$\rightarrow$	poni
SS	$\rightarrow$	SS	caress	$\rightarrow$	caress
S	$\rightarrow$		cats	$\rightarrow$	cat

Três Truncadores: Uma comparação

Texto exemplo: Such an analysis can reveal features that are not easily visible from the variations in the individual genes and can lead to a picture of expression that is more biologically transparent and accessible to interpretation

Truncador Porter: such an analysi can reveal featur that ar not easili visibl from the variat in the individu gene and can lead to a pictur of express that is more biolog transpar and access to interpret

Truncador Lovins: such an analys can reve featur that ar not eas vis from th vari in th individu gen and can lead to a pictur of expres that is mor biolog transpar and acces to interpres

Truncador Paice: such an analys can rev feat that are not easy vis from the vary in the individ gen and can lead to a pict of express that is mor biolog transp and access to interpret

Truncagem melhora efetividade?

- Em geral, a truncagem aumenta a efetividade para algumas consultas e diminui para outras.

Truncagem melhora efetividade?

- Em geral, a truncagem aumenta a efetividade para algumas consultas e diminui para outras.
- Consultas em que a truncagem pode ajudar: [tartan sweaters], [sightseeing tour san francisco]

Truncagem melhora efetividade?

- Em geral, a truncagem aumenta a efetividade para algumas consultas e diminui para outras.
- Consultas em que a truncagem pode ajudar: [tartan sweaters], [sightseeing tour san francisco]
- (classes de equivalência: {sweater,sweaters}, {tour,tours})

Truncagem melhora efetividade?

- Em geral, a truncagem aumenta a efetividade para algumas consultas e diminui para outras.
- Consultas em que a truncagem pode ajudar: [tartan sweaters], [sightseeing tour san francisco]
- (classes de equivalência: {sweater,sweaters}, {tour,tours})
- Classes de equivalência para *oper* do truncador Porter contém todas estas formas: *operate operating operates operation operative operatives operational*.

Truncagem melhora efetividade?

- Em geral, a truncagem aumenta a efetividade para algumas consultas e diminui para outras.
- Consultas em que a truncagem pode ajudar: [tartan sweaters], [sightseeing tour san francisco]
- (classes de equivalência: {sweater,sweaters}, {tour,tours})
- Classes de equivalência para *oper* do truncador Porter contém todas estas formas: *operate operating operates operation operative operatives operational*.
- Consultas em que a truncagem prejudica: [operational AND research], [operating AND system], [operative AND dentistry]

Exercise: O que o Google faz?

- "Stop words"
- Normalização
- Tokenização
- minusculização
- Truncagem
- Alfabetos não latinos
- Tremas
- Palavras compostas
- Números

Consultas de frases

- Queremos responder a uma consulta como [stanford university] – como uma frase.

Consultas de frases

- Queremos responder a uma consulta como [stanford university] – como uma frase.
- Logo *The inventor Stanford Ovshinsky never went to university* não deve ser uma resposta.

Consultas de frases

- Queremos responder a uma consulta como [stanford university] – como uma frase.
- Logo *The inventor Stanford Ovshinsky never went to university* não deve ser uma resposta.
- O conceito de consulta de frase é facilmente compreensível para usuários.

Consultas de frases

- Queremos responder a uma consulta como [stanford university] – como uma frase.
- Logo *The inventor Stanford Ovshinsky never went to university* não deve ser uma resposta.
- O conceito de consulta de frase é facilmente compreensível para usuários.
- Cerca de 10% das consultas na web são consultas de frases.

Consultas de frases

- Queremos responder a uma consulta como [stanford university] – como uma frase.
- Logo *The inventor Stanford Ovshinsky never went to university* não deve ser uma resposta.
- O conceito de consulta de frase é facilmente compreensível para usuários.
- Cerca de 10% das consultas na web são consultas de frases.
- COnseguências para o índice invertido: Não é mais suficiente guardar docIDs em listas de postings.

Consultas de frases

- Queremos responder a uma consulta como [stanford university] – como uma frase.
- Logo *The inventor Stanford Ovshinsky never went to university* não deve ser uma resposta.
- O conceito de consulta de frase é facilmente compreensível para usuários.
- Cerca de 10% das consultas na web são consultas de frases.
- COnseguências para o índice invertido: Não é mais suficiente guardar docIDs em listas de postings.
- Duas maneiras de estender os índice invertido:

Consultas de frases

- Queremos responder a uma consulta como [stanford university] – como uma frase.
- Logo *The inventor Stanford Ovshinsky never went to university* não deve ser uma resposta.
- O conceito de consulta de frase é facilmente compreensível para usuários.
- Cerca de 10% das consultas na web são consultas de frases.
- COnseguências para o índice invertido: Não é mais suficiente guardar docIDs em listas de postings.
- Duas maneiras de estender os índice invertido:
  - índice de bigramas

Consultas de frases

- Queremos responder a uma consulta como [stanford university] – como uma frase.
- Logo *The inventor Stanford Ovshinsky never went to university* não deve ser uma resposta.
- O conceito de consulta de frase é facilmente compreensível para usuários.
- Cerca de 10% das consultas na web são consultas de frases.
- COnseguências para o índice invertido: Não é mais suficiente guardar docIDs em listas de postings.
- Duas maneiras de estender os índice invertido:
  - índice de bigramas
  - Índice posicional

Índices de bigramas

- Indexa cada par consecutivo de termos no texto como uma frase.

Índices de bigramas

- Indexa cada par consecutivo de termos no texto como uma frase.
- Por exemplo, *Friends, Romans, Countrymen* vai gerar dois bigramas: *"friends romans"* e *"romans countrymen"*

Índices de bigramas

- Indexa cada par consecutivo de termos no texto como uma frase.
- Por exemplo, *Friends, Romans, Countrymen* vai gerar dois bigramas: *"friends romans"* e *"romans countrymen"*
- Cada um destes bigramas é agora um termo do vocabulário.

Índices de bigramas

- Indexa cada par consecutivo de termos no texto como uma frase.
- Por exemplo, *Friends, Romans, Countrymen* vai gerar dois bigramas: *"friends romans"* e *"romans countrymen"*
- Cada um destes bigramas é agora um termo do vocabulário.
- Agora consultas de frases de duas palavras podem ser facilmente respondidas.

Consultas de frases mais longas

- Uma frase longa como *"stanford university palo alto"* pode ser representada como a consulta Booleana "STANFORD UNIVERSITY" AND "UNIVERSITY PALO" AND "PALO ALTO"

Consultas de frases mais longas

- Uma frase longa como *"stanford university palo alto"* pode ser representada como a consulta Booleana "STANFORD UNIVERSITY" AND "UNIVERSITY PALO" AND "PALO ALTO"
- É necessário pós-filtrar o conjunto de respostas para identificar o subconjunto que contém a frase de quatro palavras.

Problemas com índices de bigramas

- Porque estes índices são raramente usados?

Problemas com índices de bigramas

- Porque estes índices são raramente usados?
- Falsos Positivos, como indicado acima

Problemas com índices de bigramas

- Porque estes índices são raramente usados?
- Falsos Positivos, como indicado acima
- Índice explode devido ao tamanho exagerado do vocabulário

Índices posicionais

Índices posicionais

- Índices posicionais são mais eficientes que indices de bigramas.

Índices posicionais

- Índices posicionais são mais eficientes que indices de bigramas.
- Listas de Postings em um índice não posicional: Cada posting é apenas um docID

Índices posicionais

- Índices posicionais são mais eficientes que indices de bigramas.
- Listas de Postings em um índice não posicional: Cada posting é apenas um docID
- Listas de Postings em um índice posicional: Cada posting é um docID e uma lista de posições

Índices positionais: Exemplo

Índices positionais: Exemplo

Query: "to_1 be_2 or_3 not_4 to_5 be_6"

Índices positionais: Exemplo

Query: *"to*$_1$ *be*$_2$ *or*$_3$ *not*$_4$ *to*$_5$ *be*$_6$*"*

TO, 993427:

⟨ 1: ⟨7, 18, 33, 72, 86, 231⟩;
2: ⟨1, 17, 74, 222, 255⟩;
4: ⟨8, 16, 190, 429, 433⟩;
5: ⟨363, 367⟩;
7: ⟨13, 23, 191⟩; . . . ⟩

BE, 178239:

⟨ 1: ⟨17, 25⟩;
4: ⟨17, 191, 291, 430, 434⟩;
5: ⟨14, 19, 101⟩; . . . ⟩

Índices positionais: Exemplo

Query: *"to*$_1$ *be*$_2$ *or*$_3$ *not*$_4$ *to*$_5$ *be*$_6$*"*

TO, 993427:

⟨ 1: ⟨7, 18, 33, 72, 86, 231⟩;
2: ⟨1, 17, 74, 222, 255⟩;
4: ⟨8, 16, 190, 429, 433⟩;
5: ⟨363, 367⟩;
7: ⟨13, 23, 191⟩; . . . ⟩

BE, 178239:

⟨ 1: ⟨17, 25⟩;
4: ⟨17, 191, 291, 430, 434⟩;
5: ⟨14, 19, 101⟩; . . . ⟩

Índices positionais: Exemplo

Query: "to_1 be_2 or_3 not_4 to_5 be_6"

TO, 993427:

⟨ 1: ⟨7, 18, 33, 72, 86, 231⟩;
2: ⟨1, 17, 74, 222, 255⟩;
4: ⟨8, 16, 190, 429, 433⟩;
5: ⟨363, 367⟩;
7: ⟨13, 23, 191⟩; . . . ⟩

BE, 178239:

⟨ 1: ⟨17, 25⟩;
4: ⟨17, 191, 291, 430, 434⟩;
5: ⟨14, 19, 101⟩; . . . ⟩

Índices positionais: Exemplo

Query: "to_1 be_2 or_3 not_4 to_5 be_6"

TO, 993427:

⟨ 1: ⟨7, 18, 33, 72, 86, 231⟩;
2: ⟨1, 17, 74, 222, 255⟩;
4: ⟨8, 16, 190, 429, 433⟩;
5: ⟨363, 367⟩;
7: ⟨13, 23, 191⟩; . . . ⟩

BE, 178239:

⟨ 1: ⟨17, 25⟩;
4: ⟨17, 191, 291, 430, 434⟩;
5: ⟨14, 19, 101⟩; . . . ⟩

Índices positionais: Exemplo

Query: "to_1 be_2 or_3 not_4 to_5 be_6"

TO, 993427:

⟨ 1: ⟨7, 18, 33, 72, 86, 231⟩;
2: ⟨1, 17, 74, 222, 255⟩;
4: ⟨8, 16, 190, 429, 433⟩;
5: ⟨363, 367⟩;
7: ⟨13, 23, 191⟩; . . . ⟩

BE, 178239:

⟨ 1: ⟨17, 25⟩;
4: ⟨17, 191, 291, 430, 434⟩;
5: ⟨14, 19, 101⟩; . . . ⟩

Índices positionais: Exemplo

Query: *"to*$_1$ *be*$_2$ *or*$_3$ *not*$_4$ *to*$_5$ *be*$_6$*"*

TO, 993427:

⟨ 1: ⟨7, 18, 33, 72, 86, 231⟩;
2: ⟨1, 17, 74, 222, 255⟩;
4: ⟨8, 16, 190, 429, 433⟩;
5: ⟨363, 367⟩;
7: ⟨13, 23, 191⟩; . . . ⟩

BE, 178239:

⟨ 1: ⟨17, 25⟩;
4: ⟨17, 191, 291, 430, 434⟩;
5: ⟨14, 19, 101⟩; . . . ⟩

Índices positionais: Exemplo

Query: *"to*$_1$ *be*$_2$ *or*$_3$ *not*$_4$ *to*$_5$ *be*$_6$*"*

TO, 993427:

⟨ 1: ⟨7, 18, 33, 72, 86, 231⟩;
2: ⟨1, 17, 74, 222, 255⟩;
4: ⟨8, 16, 190, 429, 433⟩;
5: ⟨363, 367⟩;
7: ⟨13, 23, 191⟩; . . . ⟩

BE, 178239:

⟨ 1: ⟨17, 25⟩;
4: ⟨17, 191, 291, 430, 434⟩;
5: ⟨14, 19, 101⟩; . . . ⟩

Índices positionais: Exemplo

Query: *"to*$_1$ *be*$_2$ *or*$_3$ *not*$_4$ *to*$_5$ *be*$_6$*"*

TO, 993427:

⟨ 1: ⟨7, 18, 33, 72, 86, 231⟩;
2: ⟨1, 17, 74, 222, 255⟩;
4: ⟨8, 16, 190, 429, 433⟩;
5: ⟨363, 367⟩;
7: ⟨13, 23, 191⟩; . . . ⟩

BE, 178239:

⟨ 1: ⟨17, 25⟩;
4: ⟨17, 191, 291, 430, 434⟩;
5: ⟨14, 19, 101⟩; . . . ⟩

Índices positionais: Exemplo

Query: *"to*$_1$ *be*$_2$ *or*$_3$ *not*$_4$ *to*$_5$ *be*$_6$*"*

TO, 993427:

⟨ 1: ⟨7, 18, 33, 72, 86, 231⟩;
2: ⟨1, 17, 74, 222, 255⟩;
4: ⟨8, 16, 190, 429, 433⟩;
5: ⟨363, 367⟩;
7: ⟨13, 23, 191⟩; . . . ⟩

BE, 178239:

⟨ 1: ⟨17, 25⟩;
4: ⟨17, 191, 291, 430, 434⟩;
5: ⟨14, 19, 101⟩; . . . ⟩

Índices positionais: Exemplo

Query: *"to*$_1$ *be*$_2$ *or*$_3$ *not*$_4$ *to*$_5$ *be*$_6$*"*

TO, 993427:

⟨ 1: ⟨7, 18, 33, 72, 86, 231⟩;
2: ⟨1, 17, 74, 222, 255⟩;
4: ⟨8, 16, 190, 429, 433⟩;
5: ⟨363, 367⟩;
7: ⟨13, 23, 191⟩; . . . ⟩

BE, 178239:

⟨ 1: ⟨17, 25⟩;
4: ⟨17, 191, 291, 430, 434⟩;
5: ⟨14, 19, 101⟩; . . . ⟩

Índices positionais: Exemplo

Query: *"to*$_1$ *be*$_2$ *or*$_3$ *not*$_4$ *to*$_5$ *be*$_6$*"*

TO, 993427:

⟨ 1: ⟨7, 18, 33, 72, 86, 231⟩;
2: ⟨1, 17, 74, 222, 255⟩;
4: ⟨8, 16, 190, 429, 433⟩;
5: ⟨363, 367⟩;
7: ⟨13, 23, 191⟩; . . . ⟩

BE, 178239:

⟨ 1: ⟨17, 25⟩;
4: ⟨17, 191, 291, 430, 434⟩;
5: ⟨14, 19, 101⟩; . . . ⟩

Índices positionais: Exemplo

Query: *"to*$_1$ *be*$_2$ *or*$_3$ *not*$_4$ *to*$_5$ *be*$_6$*"*

TO, 993427:

⟨ 1: ⟨7, 18, 33, 72, 86, 231⟩;
2: ⟨1, 17, 74, 222, 255⟩;
4: ⟨8, 16, 190, 429, 433⟩;
5: ⟨363, 367⟩;
7: ⟨13, 23, 191⟩; . . . ⟩

BE, 178239:

⟨ 1: ⟨17, 25⟩;
4: ⟨17, 191, 291, 430, 434⟩;
5: ⟨14, 19, 101⟩; . . . ⟩

Índices positionais: Exemplo

Query: "to_1 be_2 or_3 not_4 to_5 be_6"

TO, 993427:

⟨ 1: ⟨7, 18, 33, 72, 86, 231⟩;
2: ⟨1, 17, 74, 222, 255⟩;
4: ⟨8, 16, 190, 429, 433⟩;
5: ⟨363, 367⟩;
7: ⟨13, 23, 191⟩; . . . ⟩

BE, 178239:

⟨ 1: ⟨17, 25⟩;
4: ⟨17, 191, 291, 430, 434⟩;
5: ⟨14, 19, 101⟩; . . . ⟩

Índices positionais: Exemplo

Query: *"to*$_1$ *be*$_2$ *or*$_3$ *not*$_4$ *to*$_5$ *be*$_6$*"*

TO, 993427:

⟨ 1: ⟨7, 18, 33, 72, 86, 231⟩;
2: ⟨1, 17, 74, 222, 255⟩;
4: ⟨8, 16, 190, 429, 433⟩;
5: ⟨363, 367⟩;
7: ⟨13, 23, 191⟩; . . . ⟩

BE, 178239:

⟨ 1: ⟨17, 25⟩;
4: ⟨17, 191, 291, 430, 434⟩;
5: ⟨14, 19, 101⟩; . . . ⟩

Índices positionais: Exemplo

Query: *"to*$_1$ *be*$_2$ *or*$_3$ *not*$_4$ *to*$_5$ *be*$_6$*"*

TO, 993427:

⟨ 1: ⟨7, 18, 33, 72, 86, 231⟩;
2: ⟨1, 17, 74, 222, 255⟩;
4: ⟨8, 16, 190, 429, 433⟩;
5: ⟨363, 367⟩;
7: ⟨13, 23, 191⟩; . . . ⟩

BE, 178239:

⟨ 1: ⟨17, 25⟩;
4: ⟨17, 191, 291, 430, 434⟩;
5: ⟨14, 19, 101⟩; . . . ⟩

Índices positionais: Exemplo

Query: *"to*$_1$ *be*$_2$ *or*$_3$ *not*$_4$ *to*$_5$ *be*$_6$*"*

TO, 993427:

⟨ 1: ⟨7, 18, 33, 72, 86, 231⟩;
2: ⟨1, 17, 74, 222, 255⟩;
4: ⟨8, 16, 190, 429, 433⟩;
5: ⟨363, 367⟩;
7: ⟨13, 23, 191⟩; . . . ⟩

BE, 178239:

⟨ 1: ⟨17, 25⟩;
4: ⟨17, 191, 291, 430, 434⟩;
5: ⟨14, 19, 101⟩; . . . ⟩

Índices positionais: Exemplo

Query: "to_1 be_2 or_3 not_4 to_5 be_6"

TO, 993427:

⟨ 1: ⟨7, 18, 33, 72, 86, 231⟩;
2: ⟨1, 17, 74, 222, 255⟩;
4: ⟨8, 16, 190, 429, 433⟩;
5: ⟨363, 367⟩;
7: ⟨13, 23, 191⟩; . . . ⟩

BE, 178239:

⟨ 1: ⟨17, 25⟩;
4: ⟨17, 191, 291, 430, 434⟩;
5: ⟨14, 19, 101⟩; . . . ⟩

Documento 4 é a resposta!

Busca por proximidade

- Acabamos de ver como usar um índice posicional para busca de frases.

Busca por proximidade

- Acabamos de ver como usar um índice posicional para busca de frases.
- Também podemos usá-los para busca por proximidade.

Busca por proximidade

- Acabamos de ver como usar um índice posicional para busca de frases.
- Também podemos usá-los para busca por proximidade.
- Por exemplo: employment /4 place

Busca por proximidade

- Acabamos de ver como usar um índice posicional para busca de frases.
- Também podemos usá-los para busca por proximidade.
- Por exemplo: employment /4 place
- Encontrar todos os documentos que contêm EMPLOYMENT e PLACE a até 4 palavras words de distância entre si.

Busca por proximidade

- Acabamos de ver como usar um índice posicional para busca de frases.
- Também podemos usá-los para busca por proximidade.
- Por exemplo: employment /4 place
- Encontrar todos os documentos que contêm EMPLOYMENT e PLACE a até 4 palavras words de distância entre si.
- *Employment agencies that place healthcare workers are seeing growth* é uma resposta.

Busca por proximidade

- Acabamos de ver como usar um índice posicional para busca de frases.
- Também podemos usá-los para busca por proximidade.
- Por exemplo: employment /4 place
- Encontrar todos os documentos que contêm EMPLOYMENT e PLACE a até 4 palavras words de distância entre si.
- *Employment agencies that place healthcare workers are seeing growth* é uma resposta.
- *Employment agencies that have learned to adapt now place healthcare workers* não é uma resposta.

Busca por proximidade

- Use o índice posicional

Busca por proximidade

- Use o índice posicional
- Algoritmo mais simples: Olhe para o Produto vetorial das posições de (i) EMPLOYMENT no documento e (ii) PLACE no documento

Busca por proximidade

- Use o índice posicional
- Algoritmo mais simples: Olhe para o Produto vetorial das posições de (i) EMPLOYMENT no documento e (ii) PLACE no documento
- Muito ineficiente para palavras frequentes, especialmente "stop words"

Busca por proximidade

- Use o índice posicional
- Algoritmo mais simples: Olhe para o Produto vetorial das posições de (i) EMPLOYMENT no documento e (ii) PLACE no documento
- Muito ineficiente para palavras frequentes, especialmente "stop words"
- Note que queremos retornar as posições encontradas e não apenas uma lista de documentos.

Busca por proximidade

- Use o índice posicional
- Algoritmo mais simples: Olhe para o Produto vetorial das posições de (i) EMPLOYMENT no documento e (ii) PLACE no documento
- Muito ineficiente para palavras frequentes, especialmente "stop words"
- Note que queremos retornar as posições encontradas e não apenas uma lista de documentos.
- Isto é importante para sumarizações dinâmicas etc.

Interseção de "proximidade"

```
POSITIONALINTERSECT(p1, p2, k)
 1  answer ← ⟨ ⟩
 2  while p1 ≠ NIL and p2 ≠ NIL
 3  do if docID(p1) = docID(p2)
 4        then l ← ⟨ ⟩
 5             pp1 ← positions(p1)
 6             pp2 ← positions(p2)
 7             while pp1 ≠ NIL
 8             do while pp2 ≠ NIL
 9                do if |pos(pp1) − pos(pp2)| ≤ k
10                      then ADD(l, pos(pp2))
11                      else if pos(pp2) > pos(pp1)
12                              then break
13                   pp2 ← next(pp2)
14                while l ≠ ⟨ ⟩ and |l[0] − pos(pp1)| > k
15                do DELETE(l[0])
16                for each ps ∈ l
17                do ADD(answer, ⟨docID(p1), pos(pp1), ps⟩)
18                pp1 ← next(pp1)
19             p1 ← next(p1)
20             p2 ← next(p2)
21        else if docID(p1) < docID(p2)
22                then p1 ← next(p1)
23                else p2 ← next(p2)
24  return answer
```

Esquema Combinado

- Índices de bigramas e posicionais podem ser combinados proveitosamente.

Esquema Combinado

- Índices de bigramas e posicionais podem ser combinados proveitosamente.
- Muitos bigramas são extremamente frequentes: Michael Jackson, Britney Spears, etc.

Esquema Combinado

- Índices de bigramas e posicionais podem ser combinados proveitosamente.
- Muitos bigramas são extremamente frequentes: Michael Jackson, Britney Spears, etc.
- Para estes bigramas, o aumento de velocidade proporcionado pela indexação de bigramas, é substancial.

Esquema Combinado

- Índices de bigramas e posicionais podem ser combinados proveitosamente.
- Muitos bigramas são extremamente frequentes: Michael Jackson, Britney Spears, etc.
- Para estes bigramas, o aumento de velocidade proporcionado pela indexação de bigramas, é substancial.
- Esquema de combinação: Incluir bigramas frequentescomo vocabulário no índice. Todas as outras frases são recuperadas por interseção posicional.

Consultas "Posicionais" no Google

- Para buscas na Web, consultas posicionais são muito mais caras do que consultas Booleanas regulares.

Consultas "Posicionais" no Google

- Para buscas na Web, consultas posicionais são muito mais caras do que consultas Booleanas regulares.
- Vejamos o exemplo de consultas de frases.

Consultas "Posicionais" no Google

- Para buscas na Web, consultas posicionais são muito mais caras do que consultas Booleanas regulares.
- Vejamos o exemplo de consultas de frases.
- Porque são mais custosas?

Consultas "Posicionais" no Google

- Para buscas na Web, consultas posicionais são muito mais caras do que consultas Booleanas regulares.
- Vejamos o exemplo de consultas de frases.
- Porque são mais custosas?
- Você consegue demonstrar no Google que as consultas de frases são mais custosas que consultas Booleanas?