FIAP

MBA+

Artificial Intelligence & Machine Learning

Processamento de Linguagem Natural

Agenda

- **Recuperação de Informação**
- **Bayesian Sets**
- **Trabalho Final**

Recuperação de Informação

Recuperação de Informação

FIAP

- Anos 90: sistemas de recuperação de informação não eram confiáveis. Por isso, pessoas preferiam obter informações de outras pessoas.

- Exemplo: preferiam um agente de viagens para programar suas viagens do que confiar no que viam na internet.

Recuperação de Informação

FIAP

- Entretanto, ao longo dos anos, a evolução desses sistemas foi tanta que a busca pela web se tornou padrão e a fonte de informação preferida na maioria das vezes.

- É importante salientar que o campo da recuperação de informação não começou com a web, mas sim com publicações científicas e registros de bibliotecas e logo se espalhando para campos como jornalismo, direito e medicina

Recuperação de Informação

- Ainda assim, com o advento da World Wide Web, a escala de publicação foi elevada para a casa da dezena de milhares de criadores de conteúdo.

- Porém, essa explosão de publicações seria discutível se a informação não pudesse ser encontrada, anotada e analisada a fim de que o usuário possa achar rapidamente uma informação que é ao mesmo tempo relevante e compreensiva para suas necessidades

Recuperação de Informação

FIAP

- O significado de recuperação de informação pode ser muito amplo. O simples fato de tirar o cartão de crédito da carteira para consultar seu número pode ser entendido como recuperação de informação.

- Contudo, como campo de estudo acadêmico, podemos definir recuperação de informação como:
  - “encontrar material (geralmente documentos) de natureza não estruturada (geralmente texto) que satisfaça uma necessidade de informação a partir de grande coleções de dados”

Recuperação de Informação

FIAP

- Apesar de ser a principal definição, ela não abrange todo campo de estudo. Podemos falar também de classificação e agrupamento de textos, objetos de aulas posteriores.

- Recuperação de informação pode ser definido em termos de escala também:
  - Web search: sistema que provê busca de bilhões de documentos em bilhões de computadores
  - Pessoal: classificação pessoal de email
  - Enterprise search: busca interna de documentos, patentes, artigos, etc.

Recuperação de Informação - Exemplo

FIAP

- Source: coleção de livros de Shakspeare

- Busca: peças que contenham a palavra *Brutus* AND *Caesar* AND NOT *Calpurnia*

- Como fazer?
  - A partir do começo, ler todo o texto, e encontrar o que foi buscado
  - Isso pode ser feito com GREP (Regex)

Recuperação de Informação - Exemplo

FIAP

- Mas isso é escalável?

- Eu preciso de 3 pré-requisitos para uma busca de propósitos múltiplos:
  - Processar grandes coleções de documentos de forma rápida
  - Permitir operações de matching mais flexíveis
  - Permitir busca ranqueada

Recuperação de Informação - Exemplo

FIAP

- Uma maneira de evitar uma procura linear de texto para cada query é indexar os documentos de antemão.

- Além disso, podemos providenciar uma maneira binária de marcar, para cada livro, quais personagens estão presentes ou não.
  - Isso é conhecido como TDM (*term-document matrix)*

Recuperação de Informação - Exemplo

FIAP

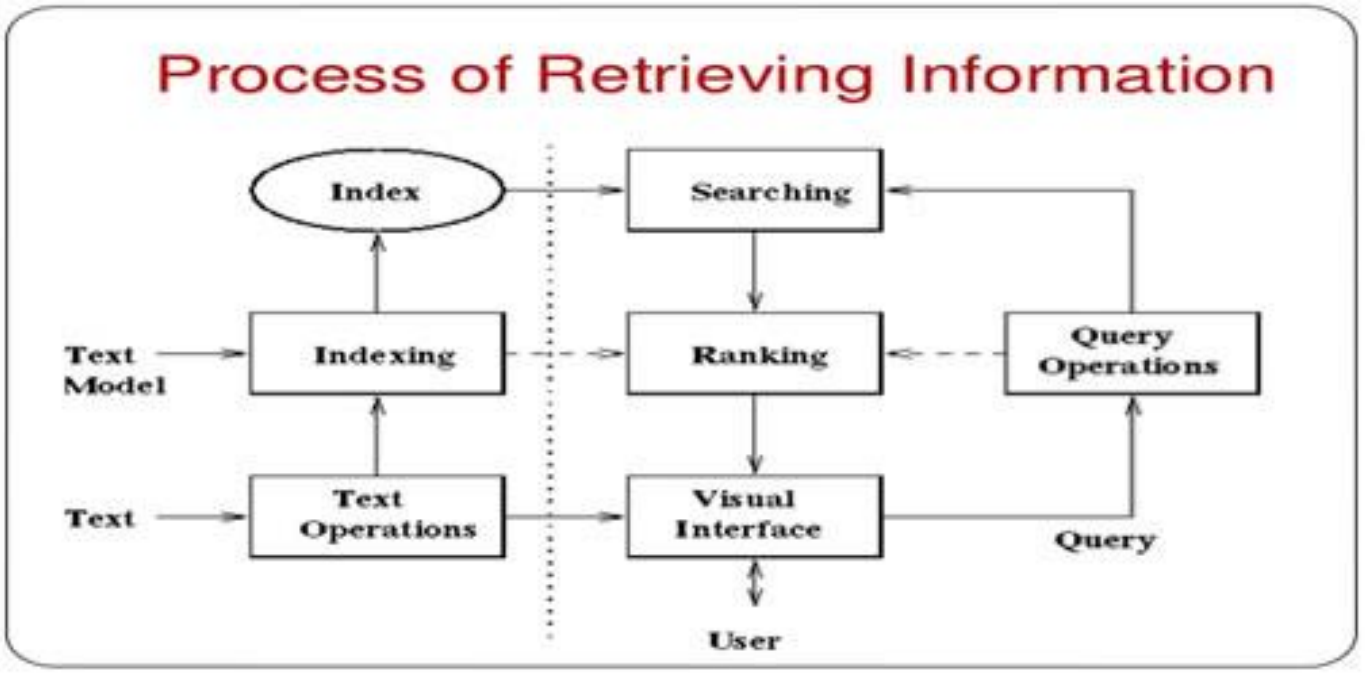
- Veja a TDM do exemplo:

	Antony and Cleopatra	Julius Caesar	The Tempest	Hamlet	Othello	Macbeth	...
Antony	1	1	0	0	0	1	
Brutus	1	1	0	1	0	0	
Caesar	1	1	0	1	1	1	
Calpurnia	0	1	0	0	0	0	
Cleopatra	1	0	0	0	0	0	
mercy	1	0	1	1	1	1	
worser	1	0	1	1	1	0	
...							

- Assim, para responder a pergunta, pegamos os vetores de *Brutus, Caesar e Calpurnia* (o complemento)
  - 110100 AND 110111 AND 101111 = 100100

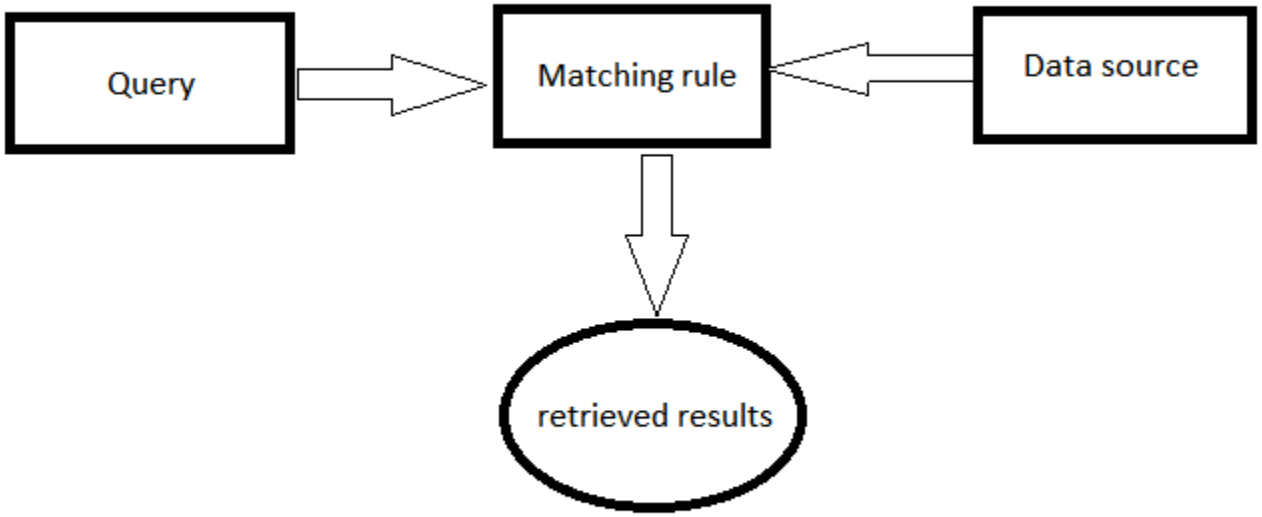
Recuperação de Informação

- Assim, definimos um modelo de arquitetura para um sistema de recuperação da informação, conforme diagrama abaixo:



Recuperação de Informação

- Por limitação de escopo (técnica e de tempo também), focaremos numa arquitetura mais simples:



basic model of an information retrieval system

Recuperação de Informação

FIAP

- A partir dessa arquitetura, vamos abstrair o conceito de recuperação de informação utilizando como exemplo uma técnica conhecida como Bayesian Sets

Bayesian Sets

Conjuntos Bayesianos

FIAP

- O que Jesus e Darwin tem em comum?
  - Além de estarem associados com duas diferentes visões da origem do homem, ambos também possuem faculdades na universidade de Cambridge em suas homenagens.

- Mas como encontrar esse tipo de relação a partir de um conjunto de dados?

Conjuntos Bayesianos

FIAP

- Conjuntos Bayesianos, indicados para resolver esse tipo de problema, usam um conceito baseado em modelo de um cluster e classifica itens usando uma pontuação que avalia a probabilidade marginal de cada item pertencer a um cluster contendo os itens da consulta.

- O foco são conjuntos de dados esparsos binários cujo score pode ser obtido por uma multiplicação de matriz, tornando o algoritmo possível de ser aplicado a grandes conjuntos de dados.

Fundamentos

FIAP

- Considere um universo de itens $\mathcal{D}$ que, dependendo da aplicação, tal conjunto seja composto de páginas web, filmes, pessoas, palavras, frases, imagens ou qualquer outro objeto sobre o qual desejamos formar consultas.

- Considere $\mathcal{D}_c \subset \mathcal{D}$ um conjunto de queries fornecidas pelo usuário cujos elementos são exemplos de algum conceito/classe/cluster dos dados

Fundamentos

- O algoritmo, então, deve providenciar uma conclusão ao subconjunto $\mathcal{D}_c$, isto é, um conjunto $\mathcal{D}'_c \subset \mathcal{D}$ que inclua todos os elementos de $\mathcal{D}_c$ e outros elementos de $\mathcal{D}$ que também estão nesse conceito/classe/cluster.

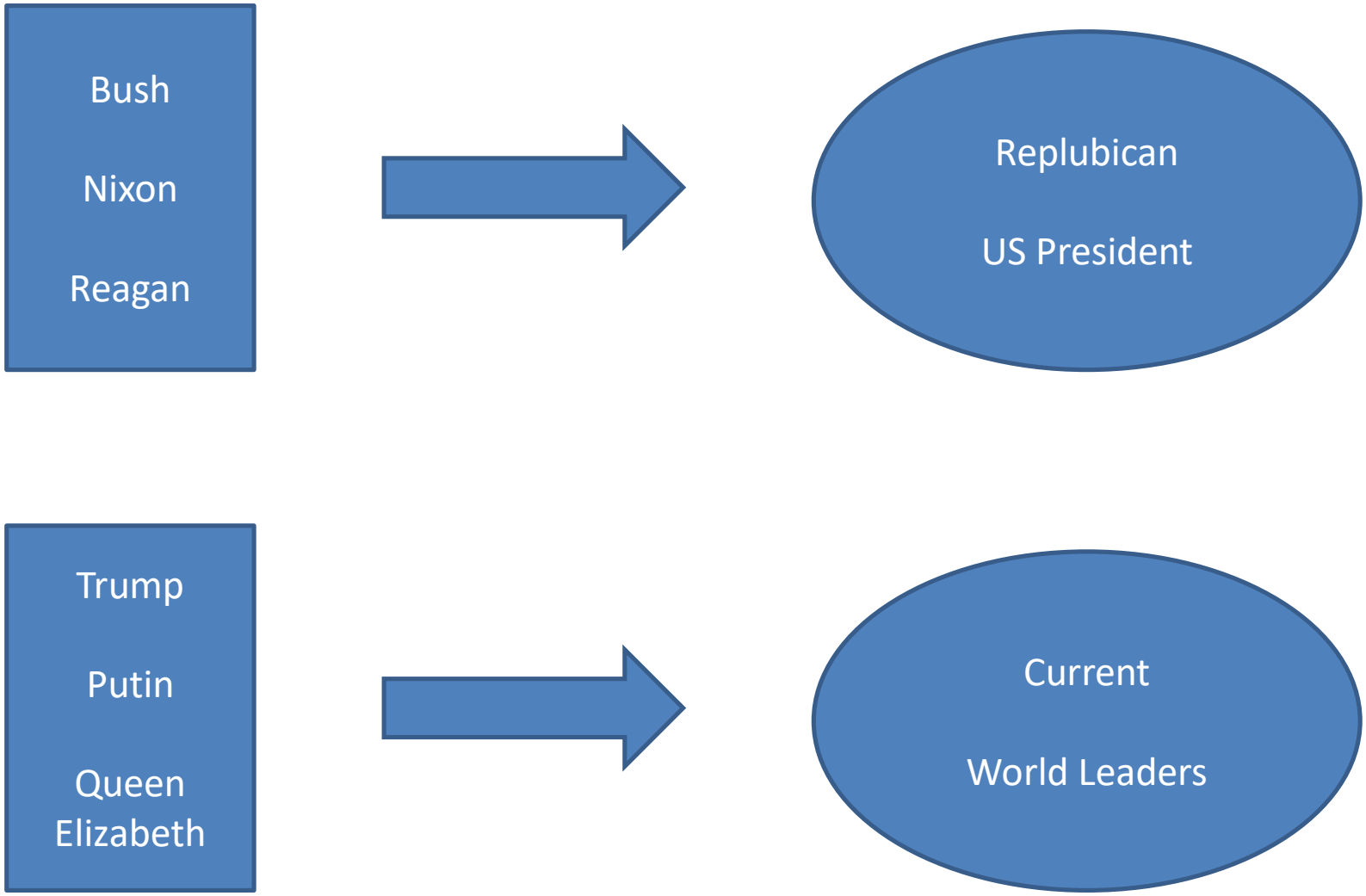
Fundamentos

FIAP

- Aqui, então, estamos lidando com um problema semi-supervisionado, já que, enquanto a maioria dos algoritmos de clustering são completamente não-supervisionados, as consultas aqui fornecidas proveem dicas supervisionadas (ou restrições) de pertencimento a um cluster particular.

- De certa forma, isso é um problema de *Feature Selection*, já que, de antemão, eu indico quais características são relevantes para a formação do cluster

Fundamentos - Exemplo



Fundamentos

- De outro ponto de vista, o objetivo desse algoritmo é resolver algum tipo de problema de recuperação de informação.

- Como em todo problema desse tipo, o resultado deve ser relevante para a consulta e faz sentido limitar o resultado para os melhores itens ordenados por relevância.

- Assumiremos essa abordagem daqui em diante.

Conjuntos Bayesianos

FIAP

- Seja:
  - $\mathcal{D}$ um conjunto de dados de itens;
  - $x \in \mathcal{D}$ um item desse conjunto;
  - $\mathcal{D}_c \subset \mathcal{D}$ um conjunto de consultas provido pelo usuário

- Nosso objetivo é ordenar (ranquear) os elementos de $\mathcal{D}$ por quão bem eles se ajustam (são semelhantes) ao subconjunto $\mathcal{D}_c$

- Intuição:
  - Se o conjunto $\mathcal{D}$ é composto por todos os filmes e o conjunto de consultas $\mathcal{D}_c$ consiste em dois filmes animados da Disney, esperamos que outros filmes animados da Disney sejam altamente ranqueados

Conjuntos Bayesianos

FIAP

- Utilizamos um modelo probabilístico para mensurar quão bem os itens se ajustam a $\mathcal{D}_c$.

- Tendo observado $\mathcal{D}_c$ como pertencente a algum conceito, queremos saber quão provável é que x também pertença a $\mathcal{D}_c$:
  - $p(x|\mathcal{D}_c)$

- Entretanto, isso incorre no problema de sensibilidade:
  - a probabilidade de uma imagem diminui com o número de pixels

Conjuntos Bayesianos

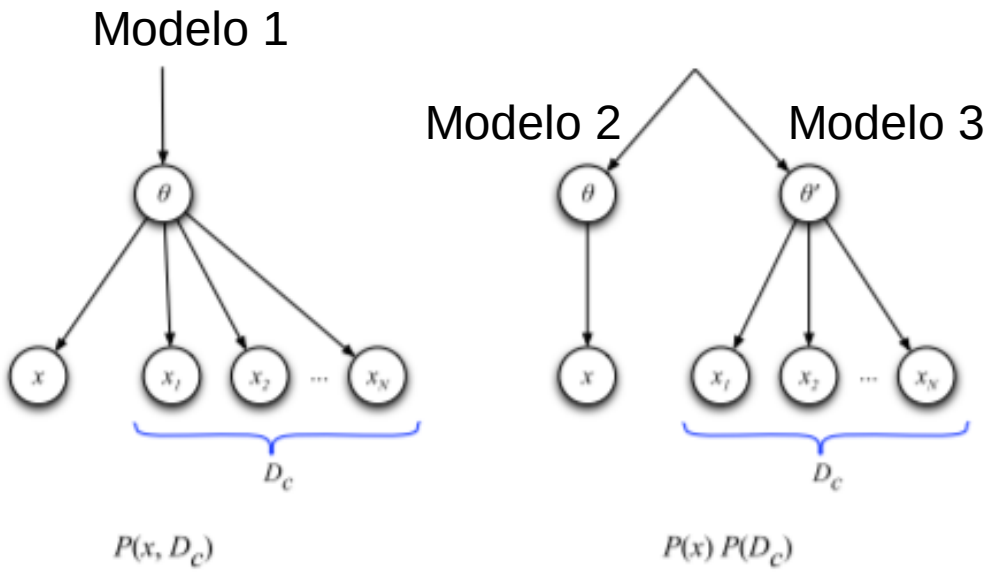
- Para remover esse efeito, computamos a proporção:
  - $score(x) = \frac{p(x|\mathcal{D}_c)}{p(x)}$ (1)

- Utilizando a Regra de Bayes, podemos reescrever (1) da seguinte maneira:
  - $score(x) = \frac{p(x,\mathcal{D}_c)}{p(x)p(\mathcal{D}_c)}$ (2)

Conjuntos Bayesianos

- Isto pode ser interpretado como a proporção da probabilidade conjunta de observar x e $\mathcal{D}_c$ e a probabilidade de independentemente observar x e $\mathcal{D}_c$.

- Intuitivamente, a figura abaixo mostra o significado da proporção expressa na equação (2):



Conjuntos Bayesianos

- Da discussão anterior, ainda não está claro como $p(x|\mathcal{D}_c)$ e $p(x)$ podem ser calculados.

- Uma maneira natural de definir um cluster é assumir que seus pontos vieram independentemente e igualmente distribuídos a partir de algum simples modelo estatístico parametrizável.

- Se todos os pontos em $\mathcal{D}_c$ pertencem a um mesmo cluster, então, do ponto de vista dessa definição, eles foram gerados a partir da mesma configuração de parâmetros. Entretanto, essa configuração é desconhecida, então precisamos calcular a média dos possíveis valores de parâmetros ponderados por alguma densidade anterior nos valores dos parâmetros, $p(\theta)$.

Conjuntos Bayesianos

- Assim, usando o teorema de Bayes, podemos estimar os parâmetros:
  - $p(\theta|\mathcal{D}_c) = \frac{p(\mathcal{D}_c|\theta)p(\theta)}{p(\mathcal{D}_c)}$
    - Na prática, usaremos parâmetros empíricos adotados pelos próprios autores, mas que satisfazem a relevância da recuperação da informação.

Conjuntos Bayesianos

FIAP

- Assim, estabelecemos o algoritmo dos conjuntos bayesianos:

Bayesian Sets Algorithm

background: a set of items $\mathcal{D}$, a probabilistic model $p(\mathbf{x}|\theta)$ where $\mathbf{x} \in \mathcal{D}$, a prior on the model parameters $p(\theta)$

input: a query $\mathcal{D}_c = \{\mathbf{x}_i\} \subset \mathcal{D}$

for all $\mathbf{x} \in \mathcal{D}$ **do**

compute $\quad \text{score}(\mathbf{x}) = \dfrac{p(\mathbf{x}|\mathcal{D}_c)}{p(\mathbf{x})}$

end for

output: return elements of $\mathcal{D}$ sorted by decreasing score

Detalhes da implementação

FIAP

- Apesar de todo fundamento matemático da distribuição de Bernoulli, distribuição Beta e função Gama, o score pode ser obtido através de multiplicação de matriz. Vamos entender como chegar até lá:
  - Ler um arquivo csv, fazer o processamento do texto usando tudo que foi visto até agora
  - Criar uma lista de strings que será sua consulta e fazer o mesmo processamento nela
  - Criar um DTM (transposto de TDM) de $\mathcal{D}$
  - Criar um DTM da lista ajustado (fit) ao DTM de $\mathcal{D}$

Detalhes da Implementação

- Assim, o score é obtido da seguinte forma:
  - $s = nc + Xq$
  - X é o DTM de $\mathcal{D}$
  - $q = \log(\tilde{\alpha}) - \log(\alpha) - \log(\tilde{\beta}) + \log(\beta)$
  - $nc = \sum \log(\alpha + \beta) - \log(\alpha + \beta + N) + \log(\tilde{\beta}) - \log(\beta)$
  - $\alpha = c \times m$
  - $\beta = c \times (1 - m)$
  - $\tilde{\alpha} = \alpha + \sum x_{ij}$
  - $\tilde{\beta} = \beta + N - \sum x_{ij}$
  - $c = 2$
  - $m = mean\ vector(X)$
  - $x_{ij} = vetor\ de\ características\ do\ DTM\ ajustado$

FIAP

MBA+