Tarefa 1: Descrever um problema

Pedro Paulo Vezzá Campos - 7538743

14 de agosto de 2013

1 Definição do problema

Sou o atual mantenedor do sistema MatrUSP (`http//bcc.ime.usp.br/matrusp`), um combinador de matrícula para alunos de graduação da USP. Para o seu funcionamento é necessário extrair diversas informações sobre disciplinas oferecidas a alunos de graduação, principalmente do JupiterWeb mas também de outros lugares relevantes aos alunos como por exemplo do CEPE ou cursos de idiomas da FFLCH, CAVC Idiomas, Poliglota, etc. Atualmente um script acessa esses sistemas e em seguida, faz o *parse* das informações disponíveis, estruturando os dados disponíveis em um banco de dados simples.

O sistema não possui vínculo oficial com a Universidade ou outras entidades e portanto não possui nenhum acesso privilegiado aos bancos de dados. Assim, é necessário extrair a informação das páginas web de cada um dos serviços. Com isso surgem diversos problemas como inconsistências das informações disponíveis e a fragilidade do programa frente a mudanças radicais no design das páginas.

2 Caracterização dos dados (como representar?)

Imagino que a representação de cada uma das páginas seria próxima ao formato *bag of words* onde cada *word* seria um tipo de dado procurado: String para título da matéria e nome do professor; Data para período de início ou término da matéria; Hora para marcar o início e término de uma aula etc.

3 O que/como seria uma solução (computacional) boa do seu problema?

É muito importante que o sistema não gere falsos positivos, exibir matérias inexistentes ou com dados errados é muito prejudicial para a experiência do usuário, portanto, o sistema deve ter uma alta confiança na resposta dada.

O tempo de processamento não é um problema. O *parsing* é feito em lotes, uma vez por dia. Qualquer tempo de execução de no máximo algumas horas é razoável.

4 Quais informações de contexto podem ser úteis para o reconhecimento?

- Disciplinas de universidades tem duração de 4, 6 ou 12 meses.
- Informações como um dia da semana próximo de um horário é um forte indicativo que o conjunto represente um dia de aula.
- O nome de uma matéria normalmente está próximo a um código que a identifica.

5 Você vê desafios? Quais?

O principal desafio é a grande variabilidade de disposição das informações buscadas em uma página. Cada site tem seu *layout* e criar um algoritmo realmente robusto e que seja capaz de estruturar os dados disponíveis é um trabalho de pesquisa atual tanto na academia quanto na indústria.

6 Ousaria desenhar os passos para chegar à solução desejada? Quais seriam esses passos?

Em princípio esse problema encaixa-se na área de processamento de linguagem natural , mais especificamente no tópico de extração de informação.

- Criar um *lexer* capaz de dividir o documento nos átomos relevantes ao problema descritos acima.
- Realizar um aprendizado supervisionado para permitir que o sistema infira relações entre os átomos que permita que ele deduza qual informação está sendo processada.

  Em [1] são citadas algumas técnicas recomendadas para a tarefa:

  - Expressões regulares escritas manualmente
  - Usando classificadores tais como o Naïve de Bayes ou modelos de entropia máxima
  - Usando modelos sequenciais tais como cadeias de Markov ocultas ou campos aleatórios condicionais

7 Referências

Referências

[1] WIKIPEDIA, "Information extraction", Disponível em `http://en.wikipedia.org/w/index.php?title=Information_extraction&oldid=550994368`. Acesso em 14 de agosto de 2013.

8 Anexos

8.1 Uma página do JupiterWeb

```
Universidade de Sao Paulo
BRASIL

  * Graduac,ao
  * [1]Ajuda
  * [2]Guia USP Acessivel
  * [3]Matricula interativa
  * [4]Informac,oes Academicas
  * [5]Calendario USP
  * [6]Disciplinas
  * [7]Turmas



  * Acesso Restrito
  * [8]Entrar
  * [9]Esqueci a Senha
  * [10]Primeiro Acesso



Disciplinas oferecidas



[11][print_edit.gif]  Preparar para impressao


[Logo_usp2.gif]
               Jupiter - Sistema de Graduac,ao


          Instituto de Matematica e Estatistica

                 Ciencia da Computac,ao

Disciplina: MAC0300 - Metodos Numericos da Algebra Linear


[12]Clique para consultar os Requisitos desta Disciplina MAC0300

Lista de Turmas oferecidas


Codigo da Turma: 2013245
         Inicio: Aug 1 2013
```

```
           Fim: Dec 10 2013
  Tipo da Turma: Teorica
   Observac,oes:

Horario            Prof(a).
qua     10:00 11:40 Leonidas de Oliveira Brandao
sex     08:00 09:40 Leonidas de Oliveira Brandao


                                  Vagas Inscritos Pendentes Matriculados
Obrigatoria                         80        49         1           43
       IME - Ciencia da Computac,ao 80        44         1           43
Optativa Livre                       0         1         0            0
Alunos Especiais                     2         0         -            0
Extracurricular                      0         1         0            0
    ------------------------------------------------------------------------------------

[13]Clique para consultar as Informac,oes da Disciplina MAC0300
  ------------------------------------------------------------------------------------

                                          [14]Creditos | [15]Fale conosco
              (c) 1999 - 2013 - Departamento de Informatica da Codage/USP
```

References

Visible links
1. https://uspdigital.usp.br/jupiterweb/jupAjuda.jsp?codmnu=2209
2. https://uspdigital.usp.br/jupiterweb/grdGuiaUSPAcessivel.jsp?codmnu=2210
3. https://uspdigital.usp.br/jupiterweb/grdMatriculainterativa.jsp?codmnu=2211
4. https://uspdigital.usp.br/jupiterweb/grdInformacoesAcademicas.jsp?codmnu=2212
5. https://uspdigital.usp.br/jupiterweb/jupCalendario.jsp?codmnu=2213
6. https://uspdigital.usp.br/jupiterweb/jupDisciplinaBusca?tipo=D&codmnu=2214
7. https://uspdigital.usp.br/jupiterweb/jupDisciplinaBusca?tipo=T&codmnu=2215
8. https://uspdigital.usp.br/jupiterweb/webLogin.jsp
9. https://uspdigital.usp.br/jupiterweb/esqueciSenha
10. https://uspdigital.usp.br/jupiterweb/primeiroAcesso
11. javascript:OpenWindowToPrint()
12. https://uspdigital.usp.br/jupiterweb/listarCursosRequisitos?coddis=MAC0300
13. https://uspdigital.usp.br/jupiterweb/obterDisciplina?sgldis=MAC0300
14. https://uspdigital.usp.br/jupiterweb/creditos.jsp
15. https://uspdigital.usp.br/jupiterweb/jupColegiadoEmailLista

Hidden links:
16. http://www.usp.br/

8.2 Uma página de um curso do CEPE

[1][logo_usp.jpg]
[2]Inicio

Information

Menu

- [3]CEPEUSP
  - [4]Historico
  - [5]Infraestrutura
  - [6]Normas
  - [7]Como frequentar
- [8]Cursos
  - [9]Comunidade USP
  - [10]Diferenciados
  - [11]Infanto-juvenil
  - [12]Publico Adulto
  - [13]Terceira Idade
- [14]Eventos
- [15]Nucleos
  - [16]LAtiS
  - [17]NUMES
  - [18]NUPSEA
  - [19]NURI
  - [20]PET

Modalidade - Tenis

Comunidade USP - Curso

- [21]Cursos
  - [22]Comunidade USP
    - [23]Alongamento
    - [24]Badminton
    - [25]Basquetebol
    - [26]Boxe Educativo
    - [27]Caminhada / Alongamento
    - [28]Canoagem
    - [29]Capoeira
    - [30]Corrida
    - [31]Deep Running
    - [32]Exerc. Individual para iniciantes
    - [33]Exerc. Localizados
    - [34]Fitness
    - [35]Futebol
    - [36]Futebol mais de 40 anos
    - [37]Ginastica Olimpica
    - [38]Hidroginastica
    - [39]Judo
    - [40]Karate
    - [41]Mat Pilates
    - [42]Musculac,ao

- [43]Natac,ao
- [44]Orientac,ao Nutricional
- [45]Preparac,ao Fisica
- [46]Programa Emagrecimento
- [47]Remo
- [48]Soft Tenis
- [49]Tenis
- [50]Treinamento Funcional
- [51]Voleibol
- [52]Yoga

[53]Inicio > [54]Cursos > [55]Comunidade USP

O que e

O curso de iniciac,ao destina-se a pessoas interessadas em adquirir experiencia na modalidade. O aluno aprende os movimentos basicos do jogo: batida de direita ("forehand"), batida de esquerda ("backhand") e saque ("service"). Aprende tambem a movimentar-se na quadra, a utilizar o paredao, inicio da execuc,ao do voleio ("volley"). Serao dadas noc,oes de regras, arbitragem e estrategia de jogo.

A turma de aperfeic,oamento tem por objetivo aprimorar as habilidades adquiridas na iniciac,ao (batida de direita, batida de esquerda, paralelas e cruzadas e saque). Aprende-se o voleio, o "lob", o "smash" e a movimentac,ao na quadra, levando ‘a participac,ao em jogos internos individuais e em duplas.

Horarios

Sexo	Dias	Horarios	Nivel	Professor	Idade	Local	Vagas	Distrib.Vagas
Masculino e Feminino	3-a/5-a	07h00-08h00	II	Thales	18 +	Q.Tenis 4 a 9	20	6/8/2013
Masculino e Feminino	3-a/5-a	08h00-09h00	II	Thales	18 +	Q.Tenis 4 a 9	20	6/8/2013
Masculino e Feminino	3-a/5-a	11h00-12h00	II	Thales	18 +	Q.Tenis 4 a 9	20	6/8/2013
Masculino e Feminino	3-a/5-a	12h00-13h00	I	Thales	18 +	Q.Tenis 4 a 9	20	6/8/2013
Masculino e Feminino	2-a/4-a	16h30-18h00	I	Eduardo	18 +	Q.Tenis 6 e 7	12	5/8/2013
Masculino e Feminino	2-a/4-a	18h00-19h30	I	Eduardo	18 +	Q.Tenis 6 e 7	12	5/8/2013
Masculino e Feminino	3-a/5-a	16h30-18h00	I	Eduardo	18 +	Q.Tenis 6 e 7	12	6/8/2013
Masculino e Feminino	3-a/5-a	18h00-19h30	I	Eduardo	18 +	Q.Tenis 6 e 7	12	6/8/2013

[56]Inicio | [57]Horario | [58]Localizac,ao | [59]Contato

Prac,a 02, Prof. Rubiao Meira, 61 - Cidade Universitaria, Sao Paulo, SP - CEP 05508-110

References

1. http://www.usp.br/
2. http://www.cepe.usp.br/site/
3. http://www.cepe.usp.br/site/?q=cepeusp
4. http://www.cepe.usp.br/site/?q=cepeusp/historico
5. http://www.cepe.usp.br/site/?q=cepeusp/infraestrutura

6. http://www.cepe.usp.br/site/?q=cepeusp/normas
7. http://www.cepe.usp.br/site/?q=cepeusp
8. http://www.cepe.usp.br/site/?q=cursos
9. http://www.cepe.usp.br/site/?q=cursos/comunidade-usp/
10. http://www.cepe.usp.br/site/?q=cursos/diferenciados
11. http://www.cepe.usp.br/site/?q=cursos/infanto-juvenil
12. http://www.cepe.usp.br/site/?q=cursos/publico-adulto
13. http://www.cepe.usp.br/site/?q=cursos/terceira-idade
14. http://www.cepe.usp.br/site/?q=eventos
15. http://www.cepe.usp.br/site/?q=nucleos
16. http://www.cepe.usp.br/site/?q=nucleos/latis
17. http://www.cepe.usp.br/site/?q=nucleos/numes
18. http://www.cepe.usp.br/site/?q=nucleos/nupsea
19. http://www.cepe.usp.br/site/?q=nucleos/nuri
20. http://www.cepe.usp.br/site/?q=nucleos/pet
21. http://www.cepe.usp.br/site/?q=cursos
22. http://www.cepe.usp.br/site/?q=cursos/comunidade-usp
23. http://www.cepe.usp.br/site/?q=cursos/comunidade-usp/alongamento
24. http://www.cepe.usp.br/site/?q=cursos/comunidade-usp/badminton
25. http://www.cepe.usp.br/site/?q=cursos/comunidade-usp/basquetebol
26. http://www.cepe.usp.br/site/?q=cursos/comunidade-usp/boxe-educativo
27. http://www.cepe.usp.br/site/?q=cursos/comunidade-usp/caminhada-/-alongamento
28. http://www.cepe.usp.br/site/?q=cursos/comunidade-usp/canoagem
29. http://www.cepe.usp.br/site/?q=cursos/comunidade-usp/capoeira
30. http://www.cepe.usp.br/site/?q=cursos/comunidade-usp/corrida
31. http://www.cepe.usp.br/site/?q=cursos/comunidade-usp/deep-running
32. http://www.cepe.usp.br/site/?q=cursos/comunidade-usp/exerc-individual-para-iniciantes
33. http://www.cepe.usp.br/site/?q=cursos/comunidade-usp/exerc-localizados
34. http://www.cepe.usp.br/site/?q=cursos/comunidade-usp/fitness
35. http://www.cepe.usp.br/site/?q=cursos/comunidade-usp/futebol
36. http://www.cepe.usp.br/site/?q=cursos/comunidade-usp/futebol-mais-de-40-anos
37. http://www.cepe.usp.br/site/?q=cursos/comunidade-usp/ginastica-olimpica
38. http://www.cepe.usp.br/site/?q=cursos/comunidade-usp/hidroginastica
39. http://www.cepe.usp.br/site/?q=cursos/comunidade-usp/judo
40. http://www.cepe.usp.br/site/?q=cursos/comunidade-usp/karate
41. http://www.cepe.usp.br/site/?q=cursos/comunidade-usp/mat-pilates
42. http://www.cepe.usp.br/site/?q=cursos/comunidade-usp/musculacao
43. http://www.cepe.usp.br/site/?q=cursos/comunidade-usp/natacao
44. http://www.cepe.usp.br/site/?q=cursos/comunidade-usp/orientacao-nutricional
45. http://www.cepe.usp.br/site/?q=cursos/comunidade-usp/preparacao-fisica
46. http://www.cepe.usp.br/site/?q=cursos/comunidade-usp/programa-emagrecimento
47. http://www.cepe.usp.br/site/?q=cursos/comunidade-usp/remo
48. http://www.cepe.usp.br/site/?q=cursos/comunidade-usp/soft-tenis
49. http://www.cepe.usp.br/site/?q=cursos/comunidade-usp/tenis
50. http://www.cepe.usp.br/site/?q=cursos/comunidade-usp/treinamento-funcional
51. http://www.cepe.usp.br/site/?q=cursos/comunidade-usp/voleibol
52. http://www.cepe.usp.br/site/?q=cursos/comunidade-usp/yoga
53. http://www.cepe.usp.br/site/
54. http://www.cepe.usp.br/site/?q=cursos
55. http://www.cepe.usp.br/site/?q=cursos/comunidade-usp

56. http://www.cepe.usp.br/site/?q=inicio
57. http://www.cepe.usp.br/site/?q=horario
58. http://www.cepe.usp.br/site/?q=localizacao
59. http://www.cepe.usp.br/site/?q=contato